# Characterization of Emotional Speech for Gender Classification

[1]Anil Kumar Patra, [2]Jyoti Mohanty, [3]Mihir N. Mohanty, [4]Hemanta Kumar Palo
[1]Dept. of ECE, Kalam Institute of Technology, Berhampur, Odisha, India
[2,3,4]Dept. of ECE, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

## Abstract

Researchers have argued that, recognition systems trained with gender specific speech signals outperform than corresponding systems not taking gender information into account. In this context, detection of speech signals colored with emotion remains a challenge as the emotional contents in speech signal is little understood. Recognition of gender information in speech signal is crucial in speaker annotation, speaker diarization, speaker clustering, social robots, interactive TV programming, multi-media retrieval, bio-metrics and host of other such application. One such attempt is taken here to characterize and classify speech emotions based on gender specific information using frequency informative features. Fundamental frequency, formants, magnitude and amplitude of spectrum are extracted from angry, fear and happy speech emotions for characterizing the emotions to distinguish different gender. The information is used as features for classification of speech emotions using K-means classifier. The results are promising and have provided an improved accuracy above 5% for gender specific emotion recognition as compared to corresponding gender independent speech emotion recognition.

## Keywords

Emotional Speech; Gender Classification; Spectrum; Fundamental Frequency; Formants Ion.

## I. Introduction

Application domain such as speaker diarization, speaker indexing, annotation, speaker clustering, voice synthesis, social robots, biometrics, retrieval of multimedia databases, interactive TV, designing effective programming environment etc. needs an effective human computer interface that can recognize speaker gender. It remains a complex issue when the speech signal is colored with emotions, since the emotions are ill-defined due to overlapping and confusing nature. Expressive emotions vary widely between genders hence it poses difficulty in gender based emotion classification. The commonly used pitch feature for gender detection and emotion identification is time variant and based on the emotional state of the speaker [1]. Further, gender information is phoneme independent and time invariant in nature that makes the area challenging [2].

Literatures in this direction advocates inclusion of gender information for improvement of human speech emotion recognition. Comparison on two emotional databases (Danish and Berlin) for classification of gender using a pool of 1379 features has been attempted in [3]. The authors applied branch and bound feature selection approach to reduce the features for identification of human gender with Support vector machine classifier. The gender dependent models outperformed the gender independent model as claimed by the authors. An improvement in accuracy has been observed in recognition of human speech emotion when gender information was included in its recognition [4]. Recognition of gender specific speech emotion has been attempted using utterances of SmartKom and BDES database with highest accuracy of 91.65% [5]. Ververidis et al., have considered gender information in classifying five classes of speech emotions from Danish Emotional Speech database using prosodic statistics [6]. Using Bayes classifier with Gaussian probabilistic density functions, the authors reported an improved accuracy of about 7% for male speech and 3% for female speech as compared to corresponding gender independent speech emotion recognition.

Extracting suitable features of emotional speech signal that provides gender information is an issue needs further attention. The problem is aggravated as emotional database that comprising both gender are not readily available. Among speech features, pitch can provide distinctive physiologically trait of a speaker gender. It is the fundamental frequency at which the vocal cord vibrates during conversation. Due to longer and thicker male vocal folds, the pitch is lower for male than female and children [7]. The authors have proposed gender recognition for smartphone application based on pitch based features. Further, the vocal tract resonates at different frequencies during emotional outbursts. Arguably, the frequency information contained in the signal will vary between genders and between emotions. The resonant frequencies known as formant frequencies can be of considerable importance for gender specific emotion recognition. This has been proved in our result using first four formant frequencies of male and female emotional utterances collected from different sources. Along with this the fundamental frequency has been used to distinguish the gender specific emotions such as angry, fear and happy. To characterize speech emotions based on gender two spectrums based features as spectrum magnitude and spectrum amplitude have been extracted from the male and female utterances. The maximum values of the proposed spectrum based features have been used to distinguish the speech emotions based on gender that have not been attempted earlier.

Rest of the sections is organized as follows: The methodology of feature extraction has been described in section II. Section III provides a detailed description of the classification scheme employed. The results obtained and characterization of speech emotions based on gender using different features is shown in section IV. The classification accuracy obtained with different gender is compared with that of gender independent emotion recognition results. Section V concludes the work.

## II. Feature Extraction Techniques

Feature extraction from the database and classification of the emotions based on the features are two important module of an emotion recognition system. These sections are elaborated in this section with our proposed feature extraction method characterizing human speech emotions based on gender.

Fig. 1 provides the proposed method of gender based speech emotion recognition. It consists of the feature extraction module that provides the necessary gender specific speech emotion features to the chosen classifier. Classification of both genders are done both individually and also in combination of total utterances to compared the results.

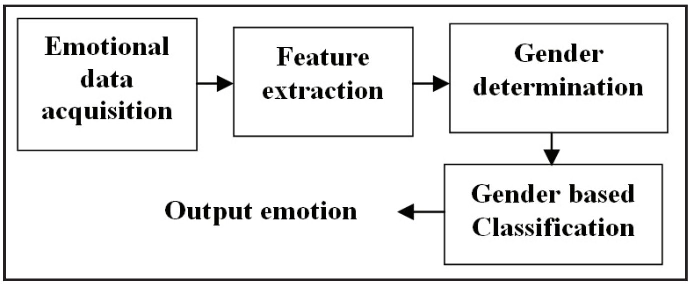

Fig. 1: Proposed Method of Gender Based Speech Emotion Recognition

### A. Fundamental Frequency

Pitch or fundamental frequency is one of the relevant prosodic features that can characterize speech, emotions adequately [8-9]. It is both gender and emotion specific. It varies with age and among speakers. Pitch is estimated here using autocorrelation method in this work to obtain the frequency content of the human gender during emotional encounter [10]. The extraction method is phase insensitive and is simple due to use of single multiplier and an accumulator. Ease of hardware implementation, direct and straightforward method of estimation from the signal waveform instead of using complex spectral method of extraction makes the method popular. For estimation of fundamental frequency, the ACFs (autocorrelation coefficients) are computed as given by

$$\mathrm{A}\ (\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau) \qquad (1)$$

Where, τ denotes the time lag. ACF can attain the highest value for the condition given byx(n)=x(n+τ). For a time period T of x(n) and any integer I the ACF will attain the peaks at τ=IT. The highest value among ACFs is given by A(0) with lower peaks at increasing in value of τ. From the position of the peaks, the fundamental frequencies can be computed at τ=T.

### B. Formant Frequencies

Formant frequencies represent the resonant frequencies of the vocal tract. These are unique to gender, speaker type, emotions and age. Due to varied shape and dimension of vocal tract of different gender, the formant frequencies differ. France et. al., 2001 observed that, the formants differ with stressed speech than neutral one due to variation in efforts in articulating voiced sounds [11]. Hence, the feature is relevant in the context of speech emotion recognition. Although, during conversation human vocal tract attains resonance a number of times, usually, the first five or six resonant frequencies are sufficient to study the analysis band of the speech signal. It is observed that, the first formant frequency generally occurs between 300 to 800 Hz, second between 700-2200 Hz, third between 1800 to 2800 Hz, fourth below 3500 Hz and fifth around 4500 Hz [8, 12]. In this work, the first four formants are considered for the proposed gender based emotional speech classification. The formants are estimated here using linear prediction method. For an L-order all-pole model representing the vocal tract, the angles of the poles $\hat{\theta}(z)$ considered to be far away from the z-plane represents the formant and is given by [8].

$$\hat{\theta}(z) = \frac{1}{1-\sum_{i=1}^{L}\hat{p}(i)z^{-i}} \qquad (2)$$

### C. Spectrum Features

The vibration of vocal tracts and vocal folds often gives the frequency information of a speaker affective state. Hence, frequency domain analysis of the signal is more informative than time domain analysis particularly for speech and emotion recognition [13]. The proposed spectrum based features uses the magnitude of the spectrum to distinguish gender specific information of emotional utterances. The magnitude of the spectrum is estimated here using the absolute value of the fast Fourier transform (FFT) of the signal. Considering a signal x(n), its FFT as X(K), the relationship can be given by

$$X^{'}(K) = abs\left(X(K)\right) = abs(\sum_{n=0}^{n-1} x(n)e^{-j2\pi/n}) \qquad (3)$$

Convolution of the spectrum magnitude with unit impulse function is done to obtain the spectrum amplitude of the signal.

## III. K-means Classification

To recognize gender based speech emotions K-means classifier is used in this work. This is a hard clustering algorithm which sections the m-dimension extracted features sets of each gender representing an emotional class into K-clusters. The K-clusters belongs to the designated gender each having a cluster center. A minimization of the objective function is achieved to obtain optimum convergence using squared error function [14]. Denoting the cluster center as $Z_l$ having feature point $x_q^{(l)}$ and distance measure norm between the feature points and cluster center as‖·‖, the objective function can be defined by

$$O = \sum_{l=1}^{K}\sum_{q=1}^{m}\left\|s_q^{(l)} - Z_l\right\|^2 \qquad (4)$$

The K-means algorithm has been implemented in the following steps

1. **Centroid computation:** Select the K-points centroids.
2. **Iteration:** Form K-cluster by repeating the procedure and allotting corresponding feature points of the desired gender containing the emotional class to the nearest centroid.
3. **Re-estimate:** The cluster centers are re-estimated iteratively unless no change in cluster center is manifested.

## IV. Results and Discussion

The objective of this paper is to extract gender specific information from emotional speech utterances and characterize these using suitable speech features. The emotional utterances involving both genders are collected from B-Tech students of our university that contained a large pool of students of both genders. Twenty utterances of angry, fear and happy emotional states are finally selected from approximately three hundred samples collected randomly and tested for adequate emotional expressions by three linguistic professors of our institute. The data are recorded in a good quality Samsung mobile set in MP3 format and converted to .wav format using format factory software. The signals are digitized in the process of conversion and a sampling rate of 16 KHz is maintained in the process.

The emotions are separated based on gender for the characterization with our chosen feature extraction techniques. As frequency informative features are more reliable particularly for speech emotion recognition, hence few efficient frequency domain approaches have been adapted for extracting suitable gender information from these signals. Fundamental frequency, first four formants as F1, F2, F3 and F4, spectrum magnitude and spectrum amplitude information are extracted initially. Average spectrum magnitude of each utterance of corresponding gender is extracted representing the concerned emotional class. The plot of spectrum magnitude obtained in this way of both gender is shown in fig. 2 through fig. 4.

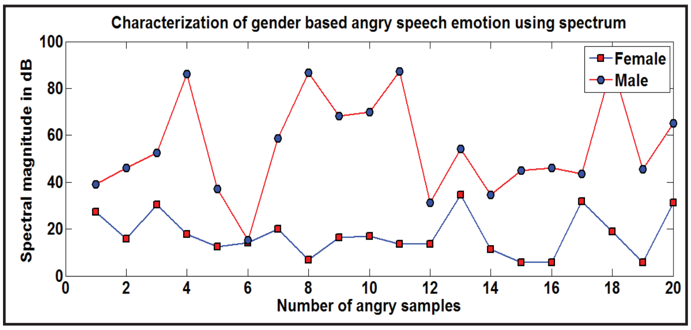

Fig. 2: Characterization of Gender Based Angry Speech Emotion Using Spectrum Magnitude in dB

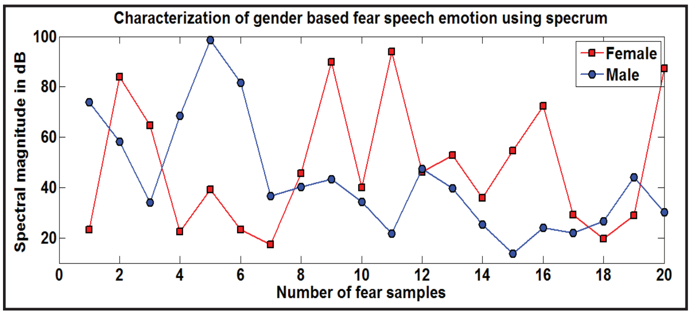

Fig. 3: Characterization of Gender Based Fear Speech Emotion Using Spectrum Magnitu de in dB

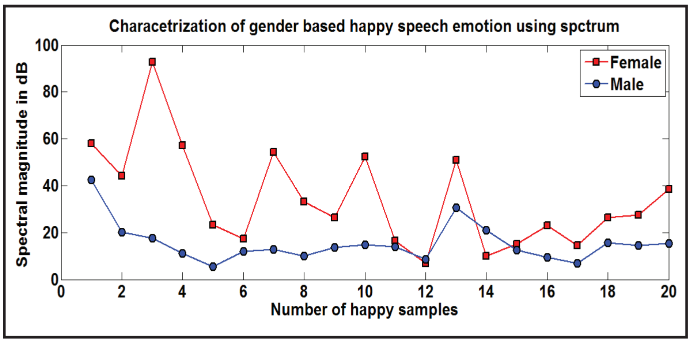

Fig. 4: Characterization of Gender Based Happy Speech Emotion Using Spectrum Magnitude in dB

It is observed that, male have higher average spectrum magnitude as compared to that of female for all the chosen emotional state as shown in these Figures. The range of magnitude varied from 200 to 900 for male as against 100 to 400 in case of female for angry state as found from fig. 2. Similarly, the range of magnitude is 100 to 500 for male and 5 to 300 for female using fear state (Figure 3). It is from 50 to 230 for male and from 5 to 180 for female using happy state (Fig. 4). It can be concluded that, angry and fear states have higher spectrum magnitude as compared to the happy state.

Use of frequency of the signal in describing human gender in speech signal is shown in fig. 5 through fig. 7. As claimed in literature, the fundamental frequency of female is higher than that of male. It varied from 200 Hz to 1000 Hz and from 30 to 300 for female and male speakers respectively for angry state as shown in fig. 5.

The corresponding range for female fear and happy states are from 500 Hz to 1000 Hz (Fig. 6) and from 200 Hz to 650 Hz for happy state (Fig. 7). Male utterances have shown lower fundamental frequencies as compared to female as observed from these Figures.

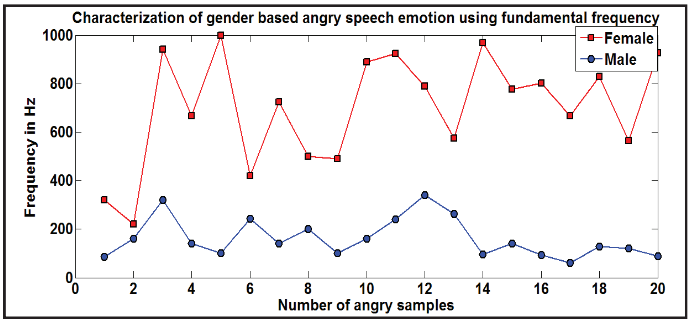

Fig. 5: Characterization of Gender Based Angry Speech Emotion Using Signal Frequency

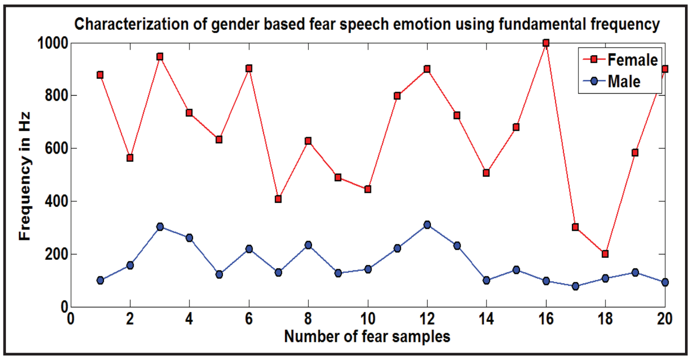

Fig. 6: Characterization of Gender Based Fear Speech Emotion Using Signal Frequency

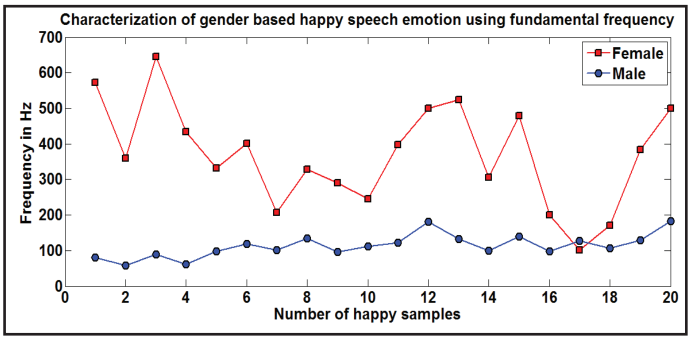

Fig. 7: Characterization of Gender Based Happy Speech Emotion Using Signal Frequency

The variation of spectrum amplitude of male and female for angry, fear and happy state of a single utterance is shown in Figure 8 through Figure 10 respectively. The spectrum amplitude has been higher for female compared to that of male in these figure for the chosen emotional states. This may be due to higher energy of female voices as compared to that of male voices. However, a more intense study in this regard using power and energy based algorithm is required to completely authenticate this claim.

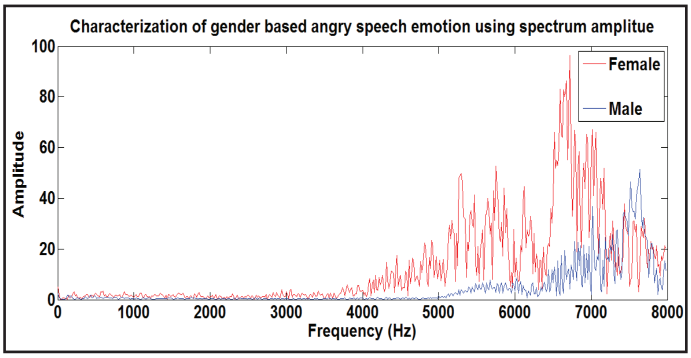

Fig. 8: Variation of Spectrum Amplitude for Gender Based Angry Speech Emotion

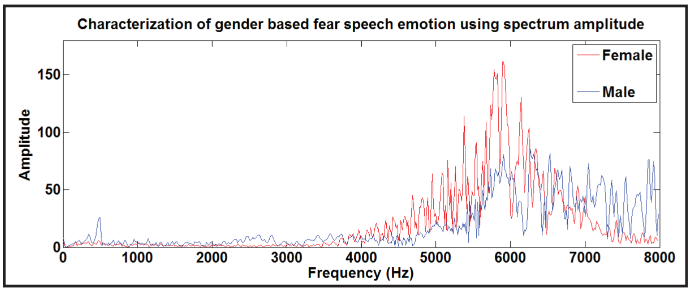

Fig. 9: Variation of Spectrum Amplitude for Gender Based Fear Speech Emotion

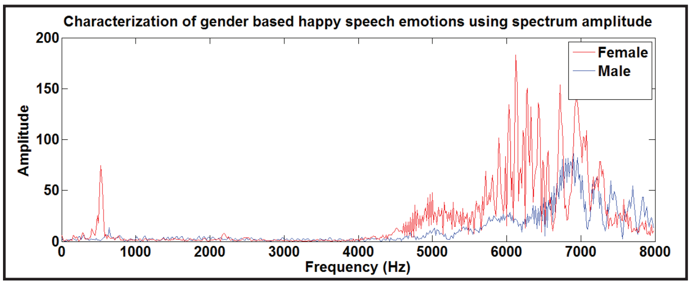

Fig. 10: Variation of Spectrum Amplitude for Gender Based Happy Speech Emotion

Four formant frequencies as F1, F2, F3 and F4 of each utterance for both genders are extracted. The average of F1 of all formants are computed and tabulated for both gender. Similarly, average of F2, F3 and F4 of both gender are also computed and tabled for comparison. Table 1, provides the desired gender based comparison of all the chosen state using formant frequencies. It can be argued that, the female have higher of these values as compared to male speakers.

Table 1: Comparison of Gender based Speech emotion Using formant Frequencies F1=First Formant, F2 =Second Formant, F3 =Third Formant, F4 =Fourth Formant

| Emotions | Average F1 | | Average F2 | | Average F3 | | Average F4 | |
|---|---|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male | Female | Male |
| Angry | 981.8 | 778.9 | 1084.1 | 953.8 | 2090.7 | 1660.0 | 2487.0 | 2099.3 |
| Fear | 923.6 | 687.2 | 992.6 | 867.3 | 1875.9 | 1632.4 | 2234.7 | 1921.1 |
| Happy | 774.7 | 560.4 | 905.2 | 703.9 | 1464.7 | 1152.8 | 1839.3 | 1679.5 |

Table 2: Gender based K-means Classification of different Emotional State

| Classification accuracy | | Signal Frequency | Spectrum magnitude | Spectrum amplitude | Formants |
|---|---|---|---|---|---|
| Male | Angry-fear | 60.8% | 63.9% | 64.4% | 60.3% |
| | Happy-Fear | 62.2% | 64.7% | 64.9% | 61.8% |
| | Angry-Happy | 64.1% | 65.3% | 65.6% | 63.7% |
| Female | Angry-fear | 56.6% | 57.2% | 58.7% | 56.0% |
| | Happy-Fear | 59.3% | 58.3% | 59.1% | 58.8% |
| | Angry-Happy | 59.3% | 59.7% | 60.4% | 58.6% |
| Both gender | Angry-fear | 55.1% | 58.6% | 59.8% | 55.1% |
| | Happy-Fear | 55.5% | 57.7% | 57.9% | 55.5% |
| | Angry-Happy | 56.2% | 56.8% | 57.7% | 56.2% |

The recognition accuracy using K-means classifier using gender information of the emotions is shown in Table 2. Highest recognition accuracy has been observed with spectrum amplitude features followed by spectrum magnitude. Formant frequencies and fundamental frequency provided meager difference in accuracy. The male recognition accuracy for speech emotion has been more than the female accuracy. Combination of both gender in recognition of human speech emotions provided the lowest accuracy among all.

## V. Conclusion

Gender identification is a crucial issue in speech recognition due to varied application as explained earlier. The challenge remains wide open when the speech signal is colored with emotions. The recognition accuracy improves involving gender information with emotion recognition as reported in literature surveyed in this work. Hence an attempt is taken to characterize gender information in emotional speech signal using frequency informative features. Other effective feature that can describe gender information in speech emotion is an area need to be explored further. Further the classification based K-means has been attempted with good amount of accuracy. Comparison of other classifiers in similar condition with efficient features will give new feature directions.

## References

[1] Castellano, P.; Slomka S.; Barger, P.,"Gender gates for telephone-based automatic speaker recognition", Digital Signal Processing, Vol. 7, No. 2, 1997, pp. 65–79, 1997.

[2] Wu, K.; Childers, D. G.,"Gender recognition from speech. Part I: Coarse analysis", Journal of Acoustic Society of America, Vol. 90, No. 4, 1991, pp. 1828–1840, 1991.

[3] Kotti, M.; Kotropoulos, C.,"Gender classification in two emotional speech databases", Proceedings of 19th International Conference on Pattern Recognition, 2008, Dec, Tampa, p 1-4, 2008.

[4] Xiao, Z.; Dellandr´ea, E.; Dou, W.; Chen, L.,"Hierarchical classification of emotional speech", Technical Report RR-LIRIS-2007-06, LIRIS UMR 5205 CNRS, 2007.

[5] Vogt, T.; Andr`e E.,"Improving automatic emotion recognition from speech via gender differentiation", Proceedings of Language Resources and Evaluation Conference, 2006, May, Genoa, Italy, pp. 1123-1126, 2006.

[6] Ververidis, D.; Kotropoulos, C.,"Automatic speech classification to five emotional states based on gender information", Proceedings of 12th European Signal Processing Conference, 2004 Sept, Austria, pp. 341-344, 2004.

[7] Bisio, I.; Delfino, A.; Lavagetto, F.; Marchese, M.; Sciarrone, A.,"Gender-driven motion recognition through speech signals for ambient intelligence applications," IEEE Transactions on emerging topics in computing", Vol. 1, No. 2, pp. 244-257, 2013.

[8] Ververidis, D.; Kotropoulos, C.,"Emotional speech recognition: Resources, features and methods," Speech Communication, Elsevier, Vol. 48, No. 9, 2006, pp. 1162-1181, 2006.

[9] Palo, H. K.; Mohanty, M. N.,"Efficient Feature Extraction for Fear State Analysis from Human Voice", Indian Journal of Science and Technology, Vol. 9, No. 38, 2016, pp. 1-11, 2016.

[10] Rabiner, L. R.,"On the Use of Autocorrelation Analysis for Pitch Detection", IEEE transactions on acoustics,

speech, and signal processing, Vol. ASSP-25, No. 1, pp. 24–33, (1977).

[11] France, D. J.; Shiavi, R. G.; Silverman, S.; Silverman, M.; Wilkes, M. (2010),"Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Transaction on Biomedical Engineering, Vol. 7, pp. 829-837, 2000.

[12] Palo, H. K.; Mohanty, J.; Mohanty, M.; Chandra, M., "Recognition of Anger, Irritation and Disgust Emotional States based on Similarity Measures", Indian Journal of Science and Technology, Vol. 9, No. 38, pp. 1-9, 2016.

[13] Palo, H. K.; Mohanty, M. N.,"Classification of emotions of angry and disgust", Smart Computing Review, Vol. 5, No. 3, 2015, pp. 151-158, 2015.

[14] Trabelsi, I.; Benayed, D.; Ellouze, N.,"Comparison between GMM SVM sequence kernel and GMM: Application to speech emotion recognition", Journal of Engineering Science and Technology, 2016.

Anil Kumar Patra has completed his 'A.M.I.E.' from Institute of Engineers, India, masters for Berhampur university, Orissa, India and Ph.D. from Monard university, Uttarpradesh,India. Currently he is working as the principal at Kalam Institute of Technology, Berhampur, Orissa, India. His field of interest is digital signal processing.

Jyoti Mohanty has completed his M-Tech from Technical Education and Research, Siksha 'O' nusandhan University, Bhubaneswar, Odisha, India.

Currently he is working as an Assistant professor in ITER, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India in the Department of ECE and is a member of IEEE..

Mihir Narayan Mohantyhas published more than 100 papers in International/ National journals and Conferences along with approximately 20 years of teaching experience.

He is an active member of many professional societies like IEEE, IET, IETE, EMC & EMI Engineers India, IE (I), ISCA, ACEEE, IAEng etc. He has received his M.Tech. degree in Communication System Engineering from the Sambalpur University, Sambalpur, Odisha and Ph.D. in Applied Signal Processing from BijuPattanaik Technical University, Odisha. He is currently working as an Associate Professor and was Head in the Department of Electronics and Instrumentation Engineering, Institute of Technical Education and Research, Siksha O‟ Anusandhan University, Bhubaneswar, Odisha. His area of research interests includes Applied Signal and image Processing, Digital Signal/Image Processing, Biomedical Signal Processing, Microwave Communication Engineering and Bioinformatics.

Hemanta Kumar Palo has completed his Master of Engineering in Electronicsand Communication engineering from "Birla Institute of Technology", Mesra, Ranchi in 2011. He is having 20 years' industrial experience and is currently serving as an Assistant Professor in the department of Electronics and Communication Engineering in the Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India. He is the life member of IEI, India and is the member of IEEE. His area of research includes speech and emotional analysis. He has published more than 25 papers in national and international journals and conferences.