International Journal of Engineering Research and General Science Volume 3, Issue 2, March-April, 2015
ISSN 2091-2730

Gender based Affection Recognition of Speech Signals using Spectral & Prosodic Feature Extraction

Mrs. Jibi Raj, Mr. Sujith Kumar

Abstract— Speech is one of the most fundamental and natural means of communication between human beings. Human beings use emotions extensively for expressing their intentions through speech. Affection Recognition through speech signals is a current research topic in the field of human-computer interaction with wide range of applications. The proposed system recognizes the emotional state of a person based on gender. The system compose of two subsystems: Gender Recognition and Emotion Recognition.Here, a speech emotion recognition system using both the spectral & prosodic features is proposed. Since both the spectral & prosodic features contain emotion information, the combination of these features improves the performance of the system. The gender recognized speech and the features extracted are given to the emotion recognition subsystem, where the emotions are recognized based on two classifiers (i.e., two support vector machines): the one trained on the basis of signals recorded by male speakers and the other one trained by that of female speakers.

Keywords— Affection, Energy, Formants, Mel Frequency Cepstral Coefficients, Pitch, Principal Component Analysis, Support Vector Machines.

INTRODUCTION

Speech is the most natural form of human communication. Emotion is an individual mental state that arises spontaneously rather than through conscious effort. Human beings use emotions extensively for conveying their intentions through speech. In some situations, it is even more important than the logical information contained in the speech. Emotional states are correlated with particular physiological states, which in turn make predictable effects on speech features, especially on pitch, timing, and voice quality.

When a person is in a state of anger, joy or fear, the sympathetic nervous system gets aroused. Then the heart rate and blood pressure increases, the mouth becomes dry and there will be occasional muscle tremors. The speech is then fast, loud and with strong high frequency energy. When someone is sad or bored, the parasympathetic nervous system gets arouse. Then the heart rate and blood pressure decreases and salivation increases, which results in slow, low-pitched speech with a weak high frequency energy.

Speech emotion recognition is particularly useful for applications which require natural man-machine interaction such as web and computer tutorial applications where the response of those systems to the user depends on the detected emotion. It is applicable in cases such as telemedicine, call centers and E-learning, where the key step is to identify the speaker and his emotions and take appropriate action based on the emotions.

The gender based speech emotion recognition system has the gender recognized emotional speech as an input and the classified emotion as an output. The first step is to extract the main features of the input speech that will differentiate between the different emotions and the gender is recognized based on the features. The most popular features used in speech emotion recognition are prosody and spectral. However the performance of the system degrades substantially, when these acoustic features are employed individually, i.e either prosody or spectral. Then the feature selection, removal and standardization algorithms are applied to get the optimum feature vectors. The vector is then given to the classifier in training and testing scheme. The final output is the classified emotion according to the input speech.

METHODOLOGY

Emotion from the speech is represented by the large number of parameters that is contained in the speech. Due to change in these parameters, there will be corresponding change in emotions. This method presents a gender-driven emotion recognition system whose

aim, starting from speech signals, is to determine the gender of speakers and then, on the basis of this information, to classify the emotion characterizing the speech signals. The proposed system composed of two functional blocks: Gender Recognition (GR) and Emotion Recognition (ER). The former can be implemented by a Pitch Estimation method, and the latter by two Support Vector Machine (SVM) classifiers, which exploits the GR subsystem output. The system recognizes the emotions such as anger, happy, boredom and sadness.

The basic block diagram for the gender based speech emotion recognition system is as shown in the Figure 1. Proposed system is based on prosodic and spectral features of speech. It consists of the emotional speech as input. Prosodic and spectral features were extracted from the speech signal. Gender recognition, based on pitch extraction, provides information about the gender of the speaker. Reduced feature sets, obtained by feature selection, performed through Principal Component Analysis were provided to the Support Vector Machine for training of the classifiers. The classifier classifies the test sample into one of the emotions and gives output. The recognized emotions are then grouped as positive and negative emotions by means of a classifier such as Support Vector Machines to make the system more reliable for specific applications like call centers, E-learning etc.

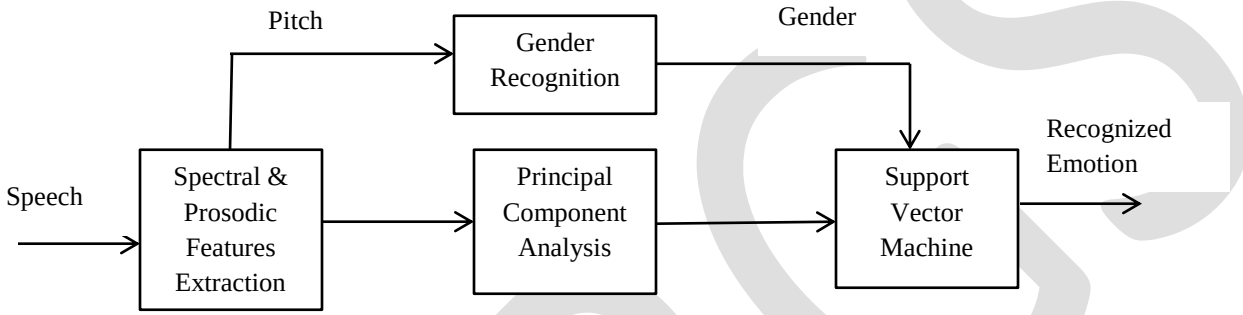


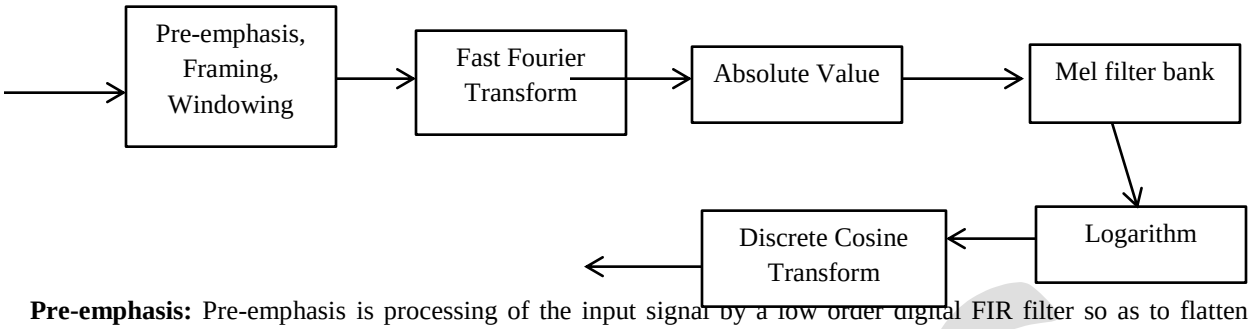
Fig.1. Block Diagram of Gender based Speech Emotion Recognition

Spectral & Prosodic Feature Extraction

The first step in speech emotion recognition system is to select a significant feature which carries large emotional information about the speech signal. Several researches have shown that effective parameters to distinguish a particular emotional states with potentially high efficiency are spectral features such as Mel Frequency Cepstrum Coefficients (MFCC) and prosodic features such as pitch, speech energy and formants. Prosodic features are related to the excitation of the vocal tract and speaking style of a person. Variation in vocal tract shapes/sizes and dynamic changes in articulator movements mainly cause the change in speaking rate. Spectral features characterize signal properties in the frequency domain, thus providing useful additions to prosodic features. For the purpose of feature extraction, spectral analysis algorithm such as Mel-frequency Cepstral Coefficients, MFCCs can be used.

- **Mel Frequency Cepstral Coefficients**

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency t measured in Hz, a subjective pitch is measured on a scale called the Mel Scale. The Mel frequency scale is linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. A compact representation would be provided by a set of mel frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a Mel frequency scale. The calculation of the MFCC includes the following steps.



Pre-emphasis: Pre-emphasis is processing of the input signal by a low order digital FIR filter so as to flatten spectrally the input signal in favor of vocal tract parameters. It makes the signal less susceptible to later finite precision effects.

$$s_p(n) = s(n) - as(n-1) \qquad (1)$$

where a is a pre-emphasis coefficient lying usually in an interval of (0.9,1), s(n) is the original signal and $s_p(n)$ is a pre-emphasized signal.

Framing: The input speech signal s(n) has always a finite length N_{total} but is usually not processed whole due to its quasi-stationary nature. Since speech signal is not stationary, the signal is divided into short segments called frames, within which speech signal can be considered as stationary. The signal is framed into pieces of length $N \ll N_{total}$ samples. Overlapping is performed because after framing of the signal, windowing is applied which causes the loss of information at the beginning and end of each frame. Overlapping reincorporates this information back into our extracted features.

Windowing: Windowing is done for minimizing the disruptions at the starting and at the end of each frame. The concept is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. Hamming window is widely used.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1 \qquad (2)$$

Where N is the total number of samples.

Fast Fourier Transform (FFT): FFT is performed to convert each frame of N samples from time domain into frequency domain. FFT gets log magnitude spectrum to determine MFCC.

Mel Filter Bank and Frequency Warping*:* The frequency range in FFT spectrum is very wide and voice signal does not follow the linear scale. Set of triangular filters are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the center frequency and decrease linearly to zero at center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. The frequencies f in Hz are converted to Mel scale using the following conversion formula.

$$F(mel) = 2595 * \log_{10}\left[1 + \frac{f}{700}\right] \qquad (3)$$

Logarithm*:* The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition.

Discrete Cosine Transform*:* It is used to orthogonalise the filter energy vectors. Because of this orthogonalisation step, the information of the filter energy vector is compacted into the first number of components and shortens the vector into number of components.

- **Speech Energy**

The energy is the basic and most important feature in speech signal. The energy of the speech signal provides a convenient representation that reflects the amplitude variation. In order to obtain the energy feature, we use short-term energy function to extract the value of energy in each speech frame. The energy of each frame is calculated by

$$E = \sum_{n=0}^{N-1} |s_i(n)|^2 \qquad (4)$$

where $s_i(n)$ denotes the i^{th} frame of the speech signal $s(n)$.

❖ **Formants**

Formants are a distinguishing or meaningful frequency components of human speech. They are the resonance frequencies of the vocal tract. Formant frequencies are the dominant frequency components of human speech, so the slight variation in their properties may cause a major difference. The formants are physically defined as poles in a system function expressing the characteristics of a vocal tract. A simple method to estimate formant frequency relies on linear prediction analysis. They are obtained by finding the roots of the prediction polynomial obtained by Linear Prediction Coefficient (LPC) analysis. LPC determines the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense.Being$Z_i = r_i e^{\pm j\theta_i}$, the i^{th} complex root pair of the prediction polynomial, the frequency, called γ_i, of the i^{th} formant related to the i^{th} complex root pair of the LPC polynomial, can be estimated by applying the following formula

$$\gamma_i = \frac{F_s}{2\pi}\theta_i \qquad (5)$$

❖ **Pitch**

Pitch period is defined as the time interval between two consecutive voiced excitation cycles i.e. the distance in time from one peak to the next peak. It is the fundamental frequency of the excitation source. The pitch signal is produced due to the vibration of the vocal folds, tension of the vocal folds and sub glottal air pressure, so the pitch is different for each emotion. Pitch of the speech signal can be estimated using the autocorrelation method. The autocorrelation gives a measure of the correlation of a speech signal with a delayed copy of itself. For a discrete time speech signal s(n), the autocorrelation function is given by

$$R(\tau) = \sum_{n=0}^{N-1} s(n)\, s(n+\tau) \qquad \tau \,\varepsilon\, [0,1,\ldots N-1] \qquad (6)$$

$R(\tau)$ is the autocorrelation of lag τ. The pitch period is defined as

$$\tau_{pitch} = arg_{\tau} max R(\tau) \qquad (7)$$

The frequency of pitch is computed as

$$\rho_{pitch} = \frac{F_s}{\tau_{pitch}} \qquad (8)$$

where F_s is the sampling frequency of the speech signal.

Pitch based Gender Recognition

Gender identification is an important step in speaker and speech recognition systems. Due to physiological differences such as vocal fold thickness or vocal tract length, and differences in speaking style, there are gender based differences in human speech. For speech signal based gender identification, the most commonly used feature is pitch period. The main reason for using the pitch period comes from the fact that the pitch values of male speakers are on an average lower than pitch values of female speakers because male vocal folds are longer and thicker compared to female ones. For every speaker, a set of pitch period estimations are obtained from his/her speech signal. Pitch is the most distinctive feature between male and female. Pitch depends on the relative highness or lowness of a tone as perceived by the human ear. The commonly used method to estimate the pitch is based on detecting the highest value of the autocorrelation. The correlation between two waveforms is the measure of their similarity. The waveforms are compared at different time intervals, and their similarity is calculated at each interval. The result of a correlation is a measure of similarity as a function of time shift between the beginnings of the two waveforms. The autocorrelation function is the correlation of a waveform with itself. One would expect exact similarity at a time shift of zero, with increasing dissimilarity as the time shift increases. In the case of voiced speech, the main peak in short-time autocorrelation function normally occurs at a lag equal to the pitch-period.

Principal Component Analysis

The performance of a pattern recognition system highly depends on the discriminative ability of the features. Selecting the most relevant subset from the original feature set, we can increase the performance of the classifier and on the other hand decrease the computational complexity. Principle Component Analysis (PCA) is a mathematical method that uses transformations to identify

patterns in data. The components are arranged in a particular pattern with the component having highest variance occuring at the topmost level, followed by other components with high variance but totally uncorrelated with the previous components. The algorithm for feature reduction for N-dimensional vectors $\{s(t)\}\ t = (1,2,\dots,N)$using PCA is as follows:

Step 1: By making use of transformation of coordinate translation, we set average vector s_m as the origin of new coordinate system in the form $s * t = s_t - s_m$

$$s_m = \frac{1}{N}\sum_{t=1}^{N} s_t \qquad (9)$$

Step 2: Find out overall covariance matrix R

$$R = \frac{1}{N}\sum_{t=1}^{N} s_t {s_t}^T \qquad (10)$$

Step 3: Find out eigenvalues $(\lambda_1, \lambda_2, \dots, \lambda_N)$ and related eigenvectors $(q_1, q_2, \dots, q_N)$.

Step 4: The sort order for each eigenvalue is descending order. We get a transformation matrix $A = (q_1, q_2, \dots, q_M)$ $(M < N)$.

Step 5: Transform N-dimensional original vector to M dimensional new vector in the form $y_t = A^T s_t$.

SVM Classification

SVMs are supervised learning methods that transforms input data from the initial dimensionality onto a higher dimension by using a kernel function to find an optimal separating hyperplane. SVM achieves optimum classification in the new feature space, where a clear distinction among features obtained by the optimum placement of a separation hyperplane under the precondition of linear separability. The hyperplane is obtained using Sequential Minimal Optimization (SMO) algorithm with no data points allowed to violate the Karush-Kuhn-Tucker (KKT) conditions.

Two kernel based Support Vector Machines are employed in this work. The first SVM is used if a male speaker is recognized by the Gender Recognition block. The other SVM in the case of a female speaker. The two SVMs are trained by using the speech signals of the emotional database of male and female speaker's. The two SVMs are trained by the traditional Quadratic Programming (QP) problem. i.e., the following problem has been solved for each gender g:

$$min_{\lambda_g}\Gamma_g(\lambda_g) = \frac{1}{2}\sum_{u=1}^{l_g}\sum_{v=1}^{l_g} y_g^u y_g^v \phi\left(x_g^u x_g^v\right)\lambda_g^u\lambda_g^v - \sum_{u=1}^{l_g}\lambda_g^u \ ,$$

$$\sum_{u=1}^{l_g}\lambda_g^u\lambda_g^v = 0, \quad 0 < \lambda_g^u < C, \ \forall\, u, \qquad (12)$$

where $\lambda_g = \{\lambda_g^1, \lambda_g^2 \dots \lambda_g^{l_g}\}$ represents the well-known Lagrangian Multipliers vectors of the Quadratic Programming problem. Vectors $x_g^1, \dots, x_g^u, \dots, x_g^{l_g}$ are feature vectors while scalars $y_g^1 \dots y_g^u \dots y_g^{l_g}$ are the related labels (emotions). They represent vectors of the gender g. $\left(x_g^u, y_g^u\right), \forall\ u\ \varepsilon\ [1, l_g]$ is the related observation between the u-th input feature vector x_g^u and its label y_g^u. The quantity l_g is the total amount of related observations composing the training set. The quantity C is the complexity constant. It determines the trade off between the flatness and the tolerance level of the misclassified samples.

Equation 4.11 represents a non-linear SVM and $\phi\left(x_g^u x_g^v\right)$ is the kernel function. Here,$\phi\left(x_g^u x_g^v\right) = \left(x_g^u\right)^T\left(x_g^v\right) + 1$. The QP problems are solved by the Sequential Minimal Optimization approach that provides an optimal point if and only if Karush-Kuhn-Tucker conditions are verified and the matrices $y_g^u y_g^v \phi\left(x_g^u x_g^v\right)$ are positive semi-definite.

EXPERIMENTAL RESULTS AND DISCUSSION

Emotional speech samples from the Berlin Emotional Speech database was taken as input. The database covers six emotions (anger, boredom, disgust, fear, happiness, sadness) and neutral state. On the basis of this database, analysis of distinguishable features of the specific emotion is done for the system. The simulation is performed in MATLAB platform.

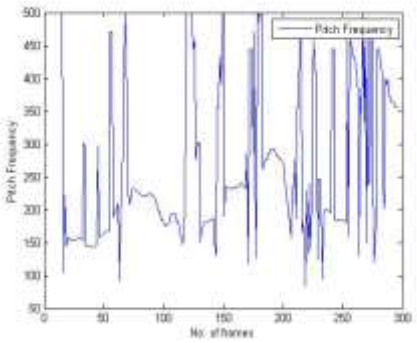
In the feature extraction stage, MFCC is used as the spectral feature. Twelve MFCC features are extracted from the speech signal per frame and it is observed that the energy of the mel cepstral coefficients are concentrated on the low coefficients.The twelve coefficients for the first five frames of boredom and anger emotional speech samples is shown. From the values, it can be inferred that MFCC values are higher for emotions such as boredom.

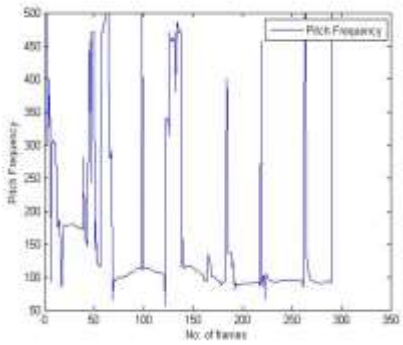
Table.1. MFCC values for boredom emotion

-8.37723	-7.68792	-7.12035	-6.67017	-6.33219
7.4545	7.554883	7.65206	7.748418	7.844383
0.133857	0.025286	-0.07813	-0.17683	-0.27085
3.023304	3.066898	3.111834	3.160341	3.211321
0.855775	0.787344	0.717601	0.647857	0.578762
-0.73739	-0.75556	-0.77106	-0.78358	-0.79347
-1.05734	-1.05867	-1.06036	-1.06224	-1.06388
-0.43238	-0.45502	-0.48328	-0.51693	-0.55504
-0.27731	-0.24259	-0.21089	-0.18111	-0.15318
0.136801	0.076234	0.008143	-0.06599	-0.14539
0.968227	0.991324	1.009311	1.024247	1.036314
-1.04307	-1.04349	-1.05024	-1.06283	-1.08092

Table.2. MFCC values for anger emotion

-4.57554	-3.30112	-2.12181	-1.0205	0.137838
7.081219	7.171192	7.205796	7.154948	6.954543
0.475364	0.464574	0.46827	0.511032	0.468277
3.211215	3.271585	3.323582	3.371205	3.586068
0.577336	0.59909	0.616361	0.613685	0.561465
1.240631	1.233302	1.220487	1.226955	1.130098
0.461068	0.526293	0.572269	0.596179	0.711834
-0.46154	-0.51125	-0.55754	-0.58136	-0.56341
-1.06869	-1.03066	-0.99802	-0.99924	-1.14086
-0.95181	-1.08595	-1.17808	-1.21024	-1.12262
-0.30872	-0.30046	-0.30279	-0.33583	-0.34234
0.693816	0.733691	0.761882	0.793632	0.74331





To analyse the performance of the classifier, confusion matrix showing emotion classification is obtained.

Table.3. Confusion matrix for male speakers

EMOTIONS	AN	BO	HA	SA
AN	4	1	0	0
BO	0	4	0	1
HA	0	0	5	0
SA	0	0	0	5

Table.4.Confusion matrix for male speakers

EMOTIONS	AN	BO	HA	SA
AN	5	0	0	0
BO	0	5	0	0
HA	0	0	4	1
SA	0	1	2	2

CONCLUSION

Emotions from the speech is represented by large number of parameters in the speech signal. Due to the change in these parameters, there will be corresponding change in emotions. The system is successfully implemented in real time. The system is language and text independent. Based on gender, there are corresponding changes in speech parameters and emotions. Emotions can be recognized from the spectral feature itself. For more accuracy of the system prosodic features are added to the system. Simulation results are presented to demonstrate the effectiveness of the proposed scheme.

REFERENCES:

[1] Bisio, Delfino, Lavagetto, Marchese and Sciarrone, "Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Application", IEEE Trans. on emerging topics in computing, vol.1, no.2, pp.244–257, December 2013.

[2] Syed Abbas Ali, Sitwat Zehra, Mohsin Khan and Faisal Wahab, "Development and Analysis of Speech Emotion Corpus Using Prosodic Features for Cross Linguistics", International Journal of Scientific & Engineering Research, vol.4, no.1, pp.1–8, January 2013.

[3] Shashidhar G.K, R. Reddy, J. Yadav and K. Sreenivasa Rao, "IITKGPSEHSC: Hindi Speech corpus for emotion analysis", IEEE Conf. on Devices and Communications, January 2011.

[4] Jagvir Kaur and Abhilash Sharma, "Emotion Detection Independent of User Using MFCC Feature Extraction", International Journal of Advanced Research in Computer Science and Software Engineering, vol.4, no.6, pp. 230–234, June 2014.

[5] Roy C Snell and F. milinazzo, "Formant location from LPC analysis data*", IEEE transac. on audio and speech processing*, vol.1, no.2, April 1993.

[6] B.Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll and Andreas Wendemuth,, "Acoustic Emotion Recognition", Proc. ASRU, pp.552–557, 2009.

[7] M.EI Ayadi and F.Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, vol.44, no.3, pp.572-587, 2011.

[8] Yixiong Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition using Support Vector Machines" International Journal of Smart Home, vol.6, no.2, pp.101–108, April 2012.

[9] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization", Advances in Kernel Methods, pp.185–208, 1999.

[10] Biswajit Nayak, Mitali Madhusmita and Debendra Ku Sahu, "Speech Emotion Recognition using Different Centered GMM", International Journal of Advanced Research in Computer Science and Software Engineering, vol.3, no.9, pp.646-649, September 2013.

[11] Tin Lay Nwe,Say Wei Foo and Liyanage C.De Silva, "Speech emotion recognition using hidden Markov models", Elsevier Speech Communications Journal , vol.41, no.4, pp.603–623, November 2003.

[12] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh and Yuan-Hao Chang, "Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification", [online] Available: *http://www.aclweb.org/anthology/O05-1013*.

[13] Firoz Shah.A, Raji Sukumar.A and Babu Anto.P, "Automatic Emotion Recognition from Speech using Artificial Neural Networks with Gender- Dependent Databases", Proc. IEEE ICACCT , pp.162–164, 2009.

[14] Rabiner L.R. and Juang B.H, "Fundamentals of Speech Recognition", Prentice-Hall, Englewood Cliffs, NJ.