A Novel Gender Recognition in Emotional Environment using GMM

A. Shamim Banu
Saranathan College of Engineering, Tiruchirappalli
Tamilnadu, India

Suganthi Venkatachalam
Saranathan College of Engineering, Tiruchirappalli
Tamilnadu, India

V. Kavitha
Saranathan College of Engineering, Tiruchirappalli
Tamilnadu, India

ABSTRACT

The emotional speech is considered as a prominent aspect in human speech communication. Emotional Speech based gender recognition has its efficacy in real world applications like biometrics, human computer related fields, voice synthesis etc. The goal of gender recognition is to extract the information from speech signal depending on speaker identity. The recognition of the gender of the speaker in emotional state is main premise of this proposed work. Both prosodic features and spectral domain features are used. The parameters of prosodic features include pitch, energy and formant; spatial features include Mel Frequency Cepstral Co efficient (MFCC). Gaussian Mixture Model is used as classifier to train and classify the gender from the features extracted from Berlin database. The accuracy of the proposed techniques is measured using metrics like Sensitivity, Specificity and likelihood ratio.

Keywords

Gender recognition, GMM, short time energy, formant.

1. INTRODUCTION

Gender based models are more perfect than gender independent models. Gender recognition is identifying the gender of the speaker from the utterances. Emotions represent physical state of a speaker. Emotions are classified as happiness, anger, sad, boredom, and disgust .There are many application based on gender based information- human machine interaction, call centres to analyse the gender of the customer to create strategies to serve them better. In the previous works, Zeng, Z. Wu, T. Falk, and W. Y. Chan[1] have identified the gender from non emotional speeches with GMM classifier trained with pitch and relative spectral linear predictive co-efficient which achieved an accuracy of 98% for clean speech and 95% for noisy speech. Gender classification prior to emotional state classification enhances the accuracy of emotion recognition.

Considering emotion recognition without prior gender information, Chung-Hsien Wu and Wei-Bin Liangif [3] obtained an accuracy of 85.7% and when gender information is available the accuracy is improved significantly to 94.5% [2] .The results reveal that gender information is useful in emotional classification.

In the first fold of the proposed work, features are extracted from emotional speech. The choice of feature vectors is one of the important tasks in the recognition system. Features are broadly classified into two classes, spectral features and prosodic features. The spectral features [14] are Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral coefficients (LPCCs) and Perceptual linear prediction coefficients (PLPCs). The prosodic features [5] are pitch, short term energy, energy entropy, formant, duration and their derivatives. Prosodic and spectral features which are generated by vocal tract and excitation source respectively are considered for analysis. In this paper, MFCC from spectral domain and pitch, energy and formant features from spatial domain are used for recognition.

The feature vectors which are extracted from the utterances are used to train the model. The model used for classification is Gaussian Mixture Model (GMM). GMM is a parametric method used for modelling [9]. GMM identifies the gender of the speaker in emotional circumstances, with significant accuracy. They calculate the probability density function of input speech features using a multivariate Gaussian mixture density. GMM refine the weights of each distribution through Expectation- Maximization (EM) algorithm. During testing conditional probabilities are calculated for the given input test patterns.

2. SPEECH CORPUS

Emotional speech database used is EMO DB [6] which is a Berlin emotional database with 500 utterances spoken by 5 male and 5 female actors in 6 emotions namely anger, happiness, sadness, fear, boredom and neutral in 10 different texts. For the purpose of analysis and testing of gender recognition, the database is sorted out according to the gender.

3. METHODOLOGY

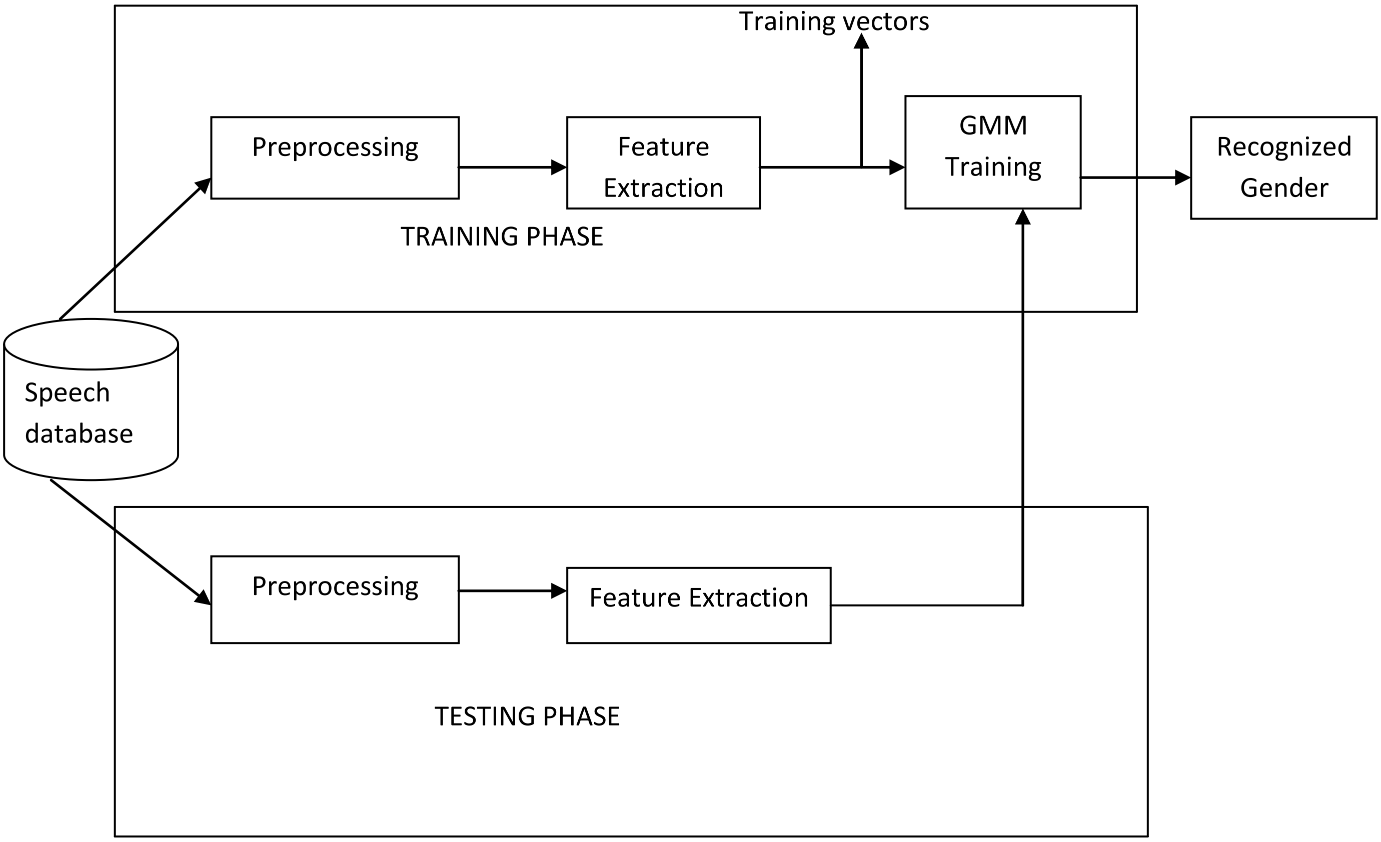

Fig 1: Overview of gender recognition

The steps involved in gender recognition system are

1. Pre processing
2. Feature Extraction
3. Training phase
4. Testing phase

3.1 Pre Processing

To facilitate short time analysis of speech signals, Pre Processing is done where the signal is decomposed into equal overlapping frames which are quasi stationary. To achieve this, Pre Emphasis, Frame blocking and Windowing are executed in sequence.

3.1.1 Pre Emphasis

A pre-emphasis filter compresses the dynamic range of the speech signal's power spectrum by flattening the magnitude spectrum and balance the low and high frequency components. Typically, the filter is in the form

$$P(z) = 1 - az^{-1} \quad (1)$$

where variable 'a' is filter co-efficient with value ranging from 0.9 to 1.0.

3.1.2 Frame Blocking

When speech signal is considered to be of small fragments it is assumed to be time invariant i.e. signal characteristics does not changes with respect to time. Framing is done to convert the speech into statistically stationary. The speech signal after pre emphasis is segmented into frames with frame size between 20~30 ms and with an overlapping of 1/3~1/2 ms. For the frame length smaller than the 20 ms, there are no enough samples for analysis and if it is longer than 30 ms, the properties of the signal changes too much which is considered as non stationary signal. Hence, the specified frame length is appropriate.

3.1.3 Windowing

To obtain a short time stationary signal, windowing is needed. Hamming window is used as a windowing function. Windowing integrates all the closest frequency lines. Hamming window is defined using the equation (2)

$$W(n) = (1-\alpha) - \alpha 0.46 \cos[2\pi n / (N-1)] \quad (2)$$

Where N is the number of samples in a frame, n ranges between 0 and N-1. If a signal is denoted as S (n) and window as W (n) then the signal after hamming windowing is

$$S'(n) = S(n) * W(n) \quad (3)$$

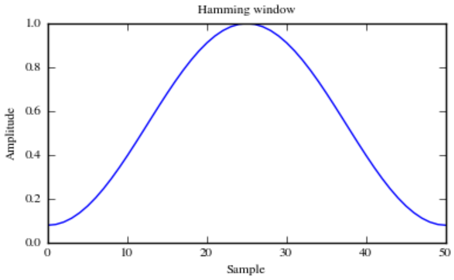

Fig 2: Hamming Window

3.2 Feature Extraction

Feature extraction is the process of converting original speech signal into parametric representation of vectors. Both spectral and prosodic features are analysed in this study

3.2.1 Specral Features

From each frame, periodogram of the power spectrum is calculated. Mel filter bank is applied to the power spectra and energy in each filter is added. The logarithm of the sum of the filter bank energies is taken and Discrete Cosine Transform is applied to extract the MFCC. These basic steps in feature extraction are explained in this section.

- Fast Fourier Transform (FFT)

After pre processing, the speech in form of small frames are applied to FFT blocks to convert the signal in time domain to frequency domain. FFT estimates the frequencies present in the signal.

FFT of length 512 is applied to the windowed sequence of the signal which finds the log magnitude spectrum of the signal. The length of FFT can be calculated using the equation (4)

$$L = 2^{\frac{\log(W)}{\log(2)}} \qquad (4)$$

Where W is window time*sampling rate

$$S(k) = \sum_{n=1}^{N} S'(n)W(n)\, e^{-\frac{j2\pi kn}{N}} \qquad (5)$$

$$P(k) = \frac{1}{N}(|S(k)|^2 \qquad (6)$$

N is the no of samples in each frame, L is the length of FFT. P (k) is the resulting periodogram (6)

- Mel Frequency Warping

Mel-frequency relates the perceived frequency which is proportional to the logarithm of the linear frequency. Lumps of periodogram bins which are calculated using FFT are summed up to analyse the energy content in various frequency regions. For this process, a filter bank with triangular shaped band-pass filter, which is centred on equally spaced frequencies ranging between 0Hz and 8 kHz in the Mel domain are used. Magnitude frequency response is multiplied by a set of triangular band pass filters to get the log energy at each triangular band pass filter. The positions of these filters are spaced based on Mel frequency. The relation between Mel frequency and common linear frequency is given by the following equation.

$$Melf = 2595 * \log10(1 + \frac{f}{700}) \qquad (7)$$

The resulting filter bank energies are subjected to logarithmic operation for channel normalization.

- Cepstrum Extraction

Discrete Cosine Transform is applied to log filter bank energies to extract the MFCC. 12 MFCC features are extracted by using above steps and considered for the model training and testing

3.2.2 Prosodic Feature Extraction

Prosodic features are the features in time domain. In the proposed method, three features that are considered are

- Short Time Energy
- Pitch
- Formant

Short time energy is the amplitude variation of speech which can be used to distinguish voiced and unvoiced segments in the speech. Energy can be mathematically represented as

$$E = \sum_{M=-\infty}^{\infty} |S(m)w(n-m)|^2 \qquad (8)$$

Pitch is the fundamental frequency of speech signal perceived by ear. It depends on vibrations of vocal cords. Pitch detection algorithm [11][13] used in the proposed work is based on autocorrelation method. In this method pitch is estimated based on highest value of auto correlation function using the equation (9)

$$R_x(m) = \frac{1}{N}[x(n+l)w(n)][x(n+l+m)\, w(n+m)] \qquad (9)$$

w(n) is appropriate window for analysis. N is the segment length. M is the number of autocorrelation points to be computed and l is the index of starting sample of the frame.

Formant is the range of frequencies over which there are spectral peaks or resonance. Formant is estimated by LPC analysis [10][13]. The signal is converted into frequency domain using FFT and passed through Linear Predictive Filter (LPC) to obtain the absolute values of the coefficients. The polynomial roots are obtained with real and imaginary parts. The formant frequencies are extracted from phase spectrum. Four formant frequencies are extracted and considered for analysis.

3.3 Training Phase

The above extracted features are used to train Gaussian mixture model. A Gaussian Mixture Model is a parametric probability density function with weighted sum of Gaussian component densities.GMM is commonly used for continuous evaluation of vocal tract related features. The parameters like mean, covariance matrix and mixture weights are estimated from the input data using iterative EM algorithm of GMM model. The parameters obtained are stored as a reference for matching.

3.4 Testing Phase

Parameters are estimated for the test data using the same iterative steps as in training phase. The test data parameters are compared with estimated reference values and best match is found.

4. RESULTS AND DISSCUSSION

All the experimental simulation is performed in MATLAB. Here 100 speech signals with emotional states such as happiness, anger, neutral and sadness from Berlin Emotional database are taken. The features are extracted from the dataset and 80 % of the extracted features are used to train the GMM and the remaining speech features are used for testing. Performance is evaluated using three different models which are

Model1- spectral domain features (MFCC).

Model 2- prosodic features (Pitch, Formant, and Energy).

Model3- Prosodic and spectral features (MFCC, Pitch, Formant, and Energy).

The performance is evaluated using the parameters which includes Specificity (SP), Sensitivity (SE), False Positive Rate (α), False Negative Rate (β), Likelihood Ratio Positive (LRP), Likelihood Ratio Negative(LRN) and Accuracy in percentage which are calculated by using the equations (10-16).

$$SP = \frac{TN}{(FP+TN)} \quad (10)$$

$$SE = \frac{TP}{TP+FN} \quad (11)$$

$$\alpha = \frac{FP}{FP+TN} \quad (12)$$

$$\beta = \frac{FN}{TP+FN} \quad (13)$$

$$LRP = \frac{SE}{1-SP} \quad (14)$$

$$LRN = \frac{1-SE}{SP} \quad (15)$$

$$ACCURACY = \frac{TP+TN}{TP+FP+TN+FN} \quad (16)$$

The calculated parameter values are tabulated in table I. The comparison of the parameter values for the three models is depicted in figure 3 and figure 4.

Table 1 Performance analysis

PARAMETERS	MODEL 1	MODEL 2	MODEL 3
A	0.2	0.161	0.101
B	0.08	0.299	0.122
SE	0.54	0.517	0.85
SP	0.79	0.838	0.89
LBR	2.57	3.19	7.72
LRN	0.58	0.57	0.16
Accuracy in %	84.5%	75%	88.7%

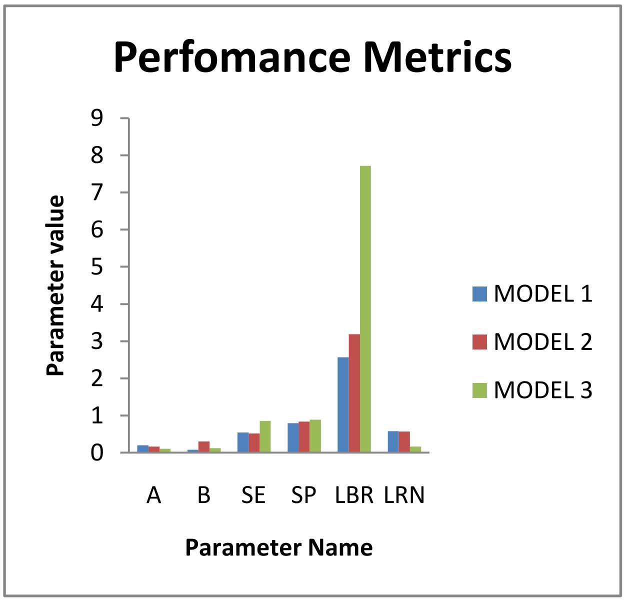

Fig:3 Performance analysis parameters

Figure 3 explains the comparision of various parameters measured in model1 , model 2, model 3. The plotted results reveal that model 3 performs better compared to other models.

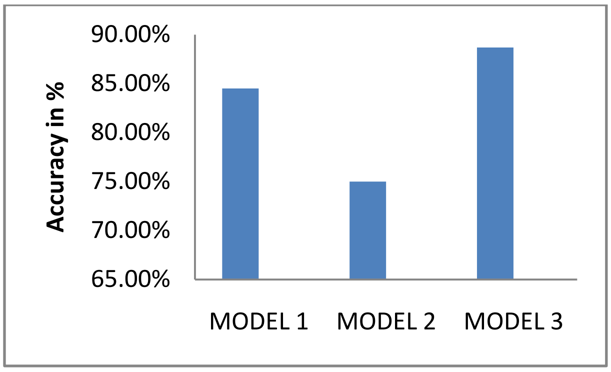

Fig 4: Accuracy comparison of 3 models

Figure 4 depicts the analysis of accuracies of the above mentioned model. The mathematical results reveal that model 3 recognises the gender more accurately than model 1, 2.

5. CONCLUSION

In the proposed work the model 3 which combines spectral and prosodic features performs with high accuracy of 88.7% compared to the accuracy of 84.5% in model 1 and 75% in model 2. As a future enhancement the proposed work can be implemented with various training models like SVM, HMM, fuzzy Clustering and Neural Network. Analysis can also be made with the above mentioned models by extraction various feature vectors like PLP, MF-PLP, Delta features.

6. REFERENCES

[1] Y. Zeng, Z. Wu, T. Falk, and W. Y. Chan. “Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech”. In Proc.5th. IEEE Int. Conf. Machine Learning and Cybernetics, pages 3376–3379.Dalian, China, 2006.

[2] D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. In Proc. of 12th European Signal Processing Conf., pages 341–344. Vienna, Austria, September 2004.

[3] Chung-Hsien Wu and Wei-Bin Liang “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels.IEEE Transaction on affective computing ,Issue no 1 jan-march 2011

[4] D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. In Proc. of 12th European Signal Processing Conf., pages 341–344. Vienna, Austria, September 2004.

[5] K. Meena, , Subramaniam, Muthusamy and Gomathy, “Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network” The International Arab Journal of Information Technology, Vol. 10, No. 5, September 2013

[6] Rajeshwara Rao et al ” Source Feature Based Gende Identification System Using GMM” International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397, Vol 3 No. 2 Feb

[7] DB- Berlin Emotional Database http://emodb.bilderbar.info

[8] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal. “A comparative performance study of several pitch detection algorithms”. IEEE Transactions on Audio,Signal, and Speech Processing 24, 399-417 1976

[9] Rakesh K., Dutta S., and Shama K., “Gender Recognition using Speech Processing Techniques” in LABVIEW,” International Journal of Advances in Engineering & Technology, vol. 1, no. 2, pp. 51-63, 2011

[10] Roberts, William J.J., Willmore “.Automatic speaker recognition using Gaussian mixture model” Information, Decision and Control, 1999. IDC 99. Proceedings. 1999, 10.1109/IDC.1999.754201, pages 465 – 470. ISBN:-7803-5256-4.

[11] Juan-Luis García Zapata, Juan Carlos Díaz Martín,, Pedro Gómez Vilda, “Fast formant estimation by complex analysis of LPC coefficients” Proceedings/Eusipco, Eusipco 2004, pages 737-740.

[12] Lawrence R. Rabiner,” On the Use of Autocorelation Analysis for Pitch Detection” IEEE transactions on acoustics, speech, and signal processing, vol. asp-25, no. 1, february 1977

[13] Jakub Karwan, Khalid Saeed, “A New Algorithm for Speech and Gender Recognition on the Basis of Voiced Parts of Speech” Communications in Computer and Information Science Volume 245, 2011, pp 113-120.

[14] Bageshree V. Sathe-Pathak, Ashish R. Panat “Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person identify 3 different emotional states of a person ”ISSN (Online): 1694-0814’