International Journal of Speech Technology (2018) 21:723–739
https://doi.org/10.1007/s10772-018-9546-1

Robust emotion recognition from speech: Gamma tone features and models

A. Revathi[1] · N. Sasikaladevi[2] · R. Nagakrishnan[1] · C. Jeyalakshmi[3]

Received: 17 April 2018 / Accepted: 30 July 2018 / Published online: 4 August 2018

Abstract
Affective computing is gaining paramount importance in ensuring the better and effective human–machine interaction. As glottal and speech signals depict the characteristics of the emotional nature of the speaker in addition to the linguistic information, speaker's emotions are needed to be recognised to give meaningful response by the system. This paper emphasises the effectiveness and efficiency in selecting the energy features by passing the speech through the Gamma tone filters spaced in Equivalent rectangular bandwidth (ERB), MEL and BARK scale. Various modelling techniques are used to develop the robust multi-speaker independent speaker's emotion/stress recognition system. Since EMO-DB Berlin database and SAVEE emotional audio-visual database used in this work contain the only limited set of speech utterances uttered by 10/4 actors/speakers in different emotions, it has become challenging to improve the performance of the stress/emotion recognition system. Speaker independent emotion recognition is done by extracting the Gamma tone energy features and cepstral features by passing the concatenated speech considered for training through the Gamma tone filters spaced in ERB, MEL and BARK scales. Subsequently, VQ/Fuzzy clustering models and continuous density hidden Markov models are created for all emotions and evaluation is done with the utterances of a speaker independent of speeches considered for training. The proposed features for test utterances are captured and applied to the VQ/Fuzzy/MHMM/SVM models and testing is performed by using minimum distance criterion/maximum log-likelihood criterion. The proposed Gamma tone energy/cepstral features and modelling techniques provide complementary evidence in assessing the performance of the system. This algorithm offers 96%, 79%, and 95.3% as weighted accuracy recall for the stress recognition system with respect to the classification done on emotion-specific group VQ/Fuzzy/MHMM/SVM models for GTF energy features with Gamma tone filters spaced in ERB, MEL and BARK scale respectively for the system evaluated for the EMO-DB database. Weighted accuracy recall is found to be 91%, 93% and 94% for the classification done on emotion-specific group models for GTF energy features with Gamma tone filters spaced in ERB, MEL and BARK scale respectively for the evaluation done on the utterances chosen from the SAVEE database. Gamma tone Cepstral features provide the overall accuracy of 92%, 90% and 92% for filters spaced in ERB, MEL and BARK scale for Berlin EMO-DB. Decision level fusion classification based on GTF energy features and modelling techniques provides the overall accuracy as 99.8% for EO-DB database and 100% for SAVEE database.

Keywords Emotion recognition system (ERS) · Gamma tone features · Vector quantization (VQ) · Fuzzy C means clustering (FCM) · Multi variate hidden Markov models (MHMM) · Support vector machine (SVM)

1 Introduction

Acoustic signals are perceived as a signal obtained by convolving excitation and vocal tract information. These signals carry the information regarding age, gender, social status, accent and emotional state of a speaker in addition to the linguistic data. The database containing the set of utterances conveying the same information uttered by the limited set of speakers becomes a bottleneck to develop a robust emotion recognition system. Emotions are expressed in different ways

✉ N. Sasikaladevi
sasikalade@gmail.com

[1] Department of ECE/SEEE, SASTRA Deemed University, Thanjavur, India

[2] Department of CSE/SoC, SASTRA Deemed University, Thanjavur, India

[3] K Ramakrishnan College of Engineering, Samayapuram, Trichy, India

by the speakers. Business outsourcing call centres would find speech recognition on emotional utterances useful, because people working in their offices may not behave in the same manner at all times when attending calls of the customers. For example, people in these offices may fall sick or feeling depressed occasionally. They may not be able to give a helpful and correct reply to the customers' technical query. Then, the system has to adjust itself to the needs of the customer or pass the control to the human agents for giving other convenient and correct reply to the customers. These emotion-specific recognition systems would be useful in power plants/industries/nuclear plants where the physical presence of humans is not possible. These systems would find applications in health care systems and the patients with depression and anxiety can be treated properly with proper diagnosis of the problems. It also finds applications in interactive web services, information retrieval, medical analysis and text to speech synthesis. These systems would find applications in the human–robot interaction where robots will behave according to the emotional state of the operator. Further, ERS can be used in the forensic department to obtain the truth from the convicts, development of flexible learning environment in education. The automatic speech emotion recognition aims to make the automated system to understand the people's emotional state by analysing the parameters extracted from the speech utterances of a speaker in different emotions. Then, the human–machine interaction would be eventually real, natural and friendly. This system could find applications in the medical field to treat and diagnose the mentally challenged people. So, the Doctors could understand the patient's feelings and offer appropriate medical care to them. Based on the emotional state of the patients, music therapy can be given to mitigating stress, anxiety, and depression. It also finds applications in mentoring/tutoring system to improve the learning quest of the students by adjusting the presentation style of the online tutor. So, the teaching–learning process will be more effective and interactive. Further, emotion recognition system plays a significant role in accident prevention by checking an emotional state of the car/vehicle drivers and other road users are appropriately alerted. To improve the sales in marketing the products, this ERS would play a pivotal role.

Performance of the emotion recognition system is evaluated (Nwe et al. 2003) with short-time log frequency power coefficient as a feature and discrete HMM as a classifier. Various modeling methods are used (Morrison et al. 2007) to compare the accuracy of emotion recognition system. Modulation spectral feature is used as a new feature (Wua et al. 2011) for implementing the emotion recognition system. Emotion recognition system is implemented (Lee et al. 2011) by using hierarchical binary classifier and acoustic and statistical feature. Combination of the pitch, energy, and MFCC (Vogt and Andr 2006) is used as a feature for emotion recognition and gender detection. MFCC is used as a feature and GMM as a classifier (Sreenivasa Rao et al. 2012) for recognizing emotions. Modified MFCC is used as a feature and NN is used as a classifier (Sapra et al. 2013) for emotion recognition. Speaker identification in the emotional environment has been done (Shahin 2009) by using log frequency power coefficients as a feature and evaluated the system using HMM, CHMM, and SPHMM. Speaker recognition in the emotional environment (Koolagudi et al. 2012) is done by using MFCC and GMM. Pitch, energy and duration are used as local prosodic features and mean median, and standard deviation of the prosodic contours are used as global prosodic features with SVM for classification (Rao et al. 2013). Dysfluency features (Moore et al. 2014) are used for emotion classification. Details of the feature selection, classifiers, and databases available are summarized (Anagnostopoulos et al. 2015). Pitch, short-time energy, zero crossing rate, Formants, Mel frequency Cepstral coefficients (MFCC) with multi-class SVM as modeling technique (Zhang et al. 2015) are used for emotion recognition. Pitch, log-energy, zero crossing rate, formants and MFCCs with SVM as a modeling technique (Sharma and Anderson 2015) are used for deep emotion recognition. Convolutional neural network (Trigeorgis et al. 2016) is used for emotion recognition from the speech signal. Pitch, energy, zero crossing rate, Formants, MFCC and LPCC with SVM (Pervaiz and Khan 2016) for classification of emotions. Optimization algorithms (Yogesh et al. 2017) are used for feature selection to evaluate emotion and stress recognition system. Emotion recognition (Patel et al. 2009) is done by using pitch, loudness, and formants as the feature and boosted GMM for creating models. By introducing a novel speaker feature (Babu et al. 2014), Gamma tone Cepstral coefficient (GTCC), based on an auditory periphery model, accuracy of the emotion recognition system is improved under noisy conditions. A speech emotion recognition system (Garg and Bahl 2014) is implemented using Gamma tone Cepstral coefficient (GTCC), and the performance of the system is compared with the system using MFCC as a feature. As features of the speech signals, the Cochleagram Model through the robust Gamma tone frequency Cepstral coefficients (GFCC) (Mohanty 2016) is calculated using the ERB filters, and the emotions in the speech signals are recognized successfully. This paper (Kaur et al. 2017) implements a combined system for detection of speech emotion by feature extraction using Gamma tone Cepstral coefficients, and enhanced genetic algorithm is used to extract features based on pitch, energy, filtered data, and frequency. Classification is used to detect emotion in by supervised approach implemented by using K-mean clustering algorithm and DNN algorithms. Whispered speech recognition (Marković et al. 2017) is based on dynamic time warping and hidden Markov models methods. Generally, the use of Cepstral mean subtraction for

whispered speech leads to a significant improvement in performance especially when mixed scenarios (normal/whisper and whisper/normal) are used for both methods, DTW and HMM. Speech emotion recognition (Peng et al. 2017) using deep learning methods based on computational auditory models of the human auditory system is a new way to identify the emotional state. This paper proposes to utilize multichannel parallel convolutional recurrent neural networks (MPCRNN) to extract salient features based on Gamma tone auditory filter bank from raw waveform and reveal that this method is effective for speech emotion recognition. First, divide the speech signal into segments, and then get multichannel data using Gamma tone auditory filter bank, which is used as a first stage before applying MPCRNN to get the optimum features for emotion recognition from speech. We subsequently obtain emotion state probability distribution for each speech segment. Eventually, utterance-level features are constructed from segment-level probability distributions and fed into support vector machine (SVM) to identify the emotions. In this paper, Speaker independent emotion recognition system is developed by using GTF energy and GTFCC as features, VQ/FCM/MHMM/SVM as modeling techniques and decision level fusion classifier. Performance of the system is evaluated by applying the features of the speech of the test emotion to the emotion-specific group models, and subsequently, emotion classification is done by making the comparison with the individual models specific to the emotion in a group. The system is evaluated by extracting GTF energy and GTFCC features with Gamma tone filters spaced in ERB/MEL/BARK scale and applying the features to the models and fusion level score calculates the accuracy concerning the features and modeling techniques. This work is mainly done to ascertain the performance of the emotion recognition system by using Gamma tone features with filters spaced in different frequency scales, modeling techniques and decision level fusion classifier used for evaluation.

2 Feature extraction

In the human ear hearing mechanism, the cochlea plays a pivotal role which has basilar membrane used to receive sounds. Acoustic waves are said to be propagating as a traveling wave through the basilar membrane. From the frequency selective characteristics of the basilar membrane, it is understood that for a high-frequency pure tone input, bottom of the basilar membrane gives the location of the largest amplitude and for the low-frequency, pure tone input top of the basilar membrane indicates the maximum amplitude. Thus, the basilar membrane has the feature of analyzing the sound signals in the frequency domain, and it converts the different frequencies into different locations where there is a high amplitude for complex sound signals with the conversion of the intensity into the amplitude of the vibration in the basilar membrane. Thus, different frequency components can be separated according to the amplitudes of acoustic signals vibrating the basilar membrane in ears. The basilar membrane responds to sounds with vibration in different positions according to the intensity of sounds and this process is analogous to the filtering action on input signals containing different frequency components. Gamma tone filter bank better simulates the mechanism of the basilar membrane as a cochlear model to perceive acoustic signals in the form of sounds.

2.1 Gamma tone filter banks

The Gamma tone filter can be a standard cochlea auditory filter based on the psychophysical observations of the auditory periphery of the ears. The characteristic of the filter is described by the impulse response function which is physically realizable. The impulse response of a Gamma tone filter with centre frequency f_i distributed between 50 and 8000 Hz is given in Eq. (1)

$$g_i(t) = At^{n-1}e^{-(2\pi b_i t)}\cos(2\pi f_i t + \Phi_i) \text{ for } t \geq 0 \,\&\, 1 \leq i \leq N \quad (1)$$

where A is the loudness based gain of the filter, n is the order of the filter and is taken as 4 (Li and Gao 2016), N is the number of channels and is considered as 128 (Li and Gao 2016).

The attenuation factor of each Gamma tone filter spaced in ERB, MEL and BARK scale is given as in Eq. (2)

$$b_i = 1.019 * ERB(f_i) \text{ or } MEL(f_i) \text{ or } BARK(f_i) \quad (2)$$

The bandwidth of the Gamma tone filters spaced in ERB, MEL, and BARK scale is given as in Eqs. (3, 4 5)

$$ERB(f_i) = 24.7 * ((4.37 * f_i(Hz)/1000) + 1) \quad (3)$$

$$MEL(f_i) = 2595 * \log_{10}(1 + f_i(Hz)/700) \quad (4)$$

$$BARK(f_i) = 6 * \sinh^{-1}(f_i(Hz)/600) \quad (5)$$

The distribution of the Gamma tone filter banks' frequency response is shown in Figs. 1, 2 and 3.

It is easy to observe that the filter's frequency response is narrower for low-frequency band and it is more extensive for the high-frequency band for the Gamma tone filters spaced in ERB and MEL scale, and this fact is indicated in Figs. 1 and 2. For the Gamma tone filters spaced in BARK scale, the frequency response of the screens in the high-frequency band suggests the variation as compared to the filters spaced in ERB and MEL scale. So, the Gamma tone filters simulate the non-linear behaviour of the human ear mechanism in perceiving and discriminating between low and high-frequency acoustic signals. The more significant gradient on both sides of the centre frequency indicates the closeness

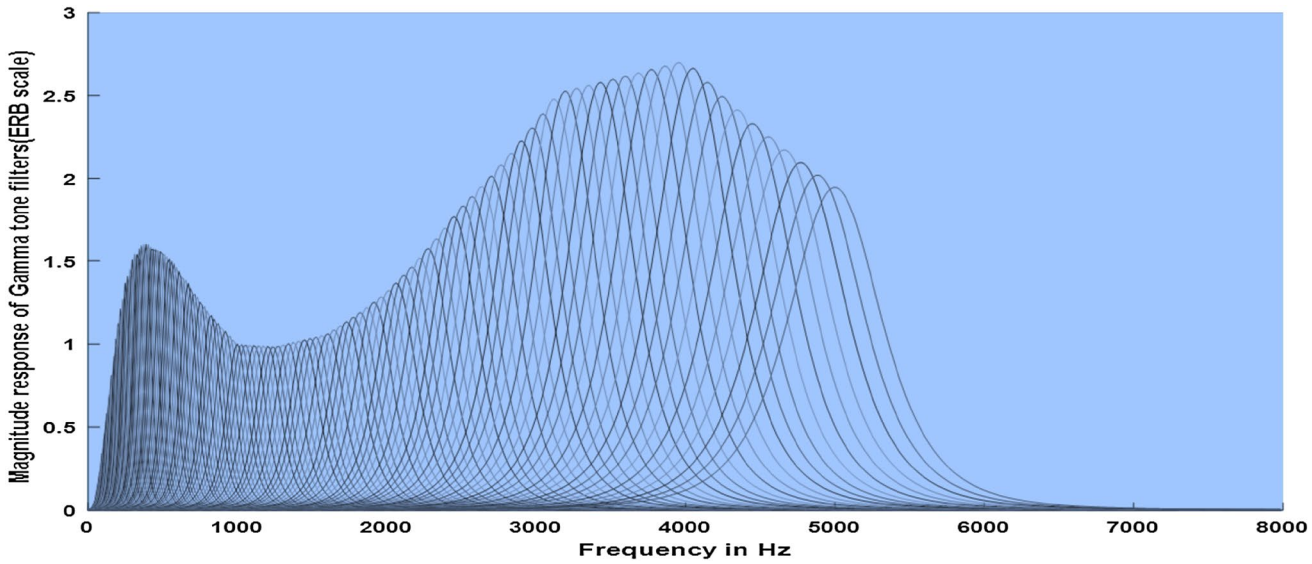

Fig. 1 Frequency response of Gamma tone filters spaced in ERB scale

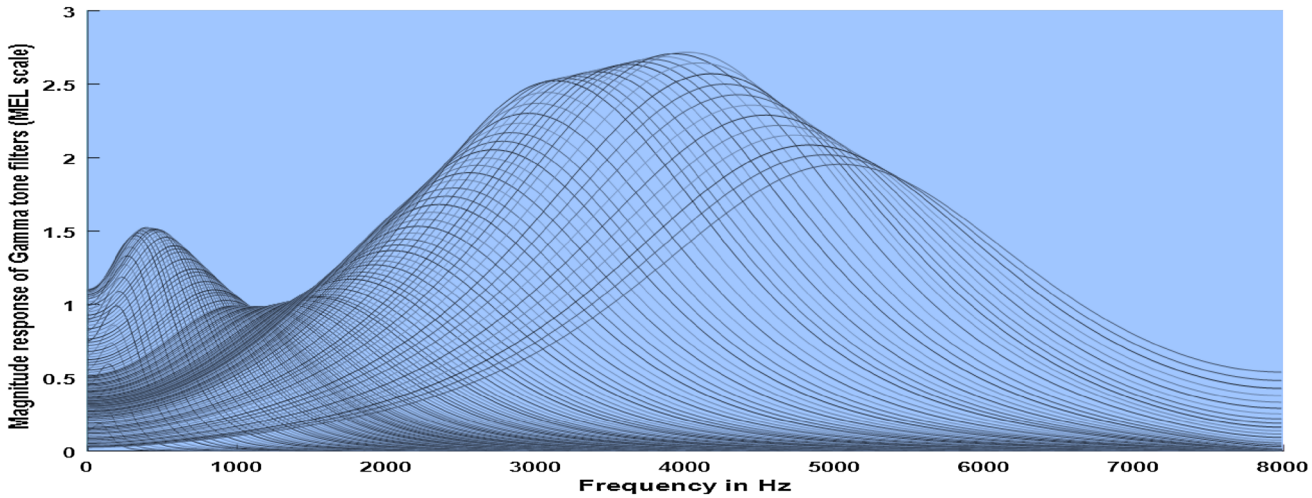

Fig. 2 Frequency response of Gamma tone filters spaced in MEL scale

between the performance of the Gamma tone filters in simulating the mechanism and behaviour of the basilar membrane in perceiving sounds. Loudness based gain adjustments for 128 channel Gamma tone filters spaced in ERB, MEL and BARK scales is plotted in Fig. 4.

3 Emotion recognition based on features and modeling techniques

EMO-DB Berlin emotional speech database (Burkhardt et al.2005) considered in this work contains about 500 utterances spoken by actors in seven different emotions such as happy, angry, disgust, boredom, fear, neutral and sad. Ten actors in all seven emotions utter statements of ten different sentences. Five male and female actors are participating in recording the ten utterances in seven emotions, and they are in the age group of 21–35 years. Developing the robust system for recognising the feelings from the limited set utterances uttered by the limited set of actors is more challenging regarding getting a good accuracy of the system. Emotion recognition algorithm is also tested on the statements chosen from database Surrey Audio-Visual Expressed Emotion (SAVEE) (http://kahlan.eps.surrey.ac.uk/savee/) database. This database is created by recording the speeches of four male actors in seven different emotions with sentences

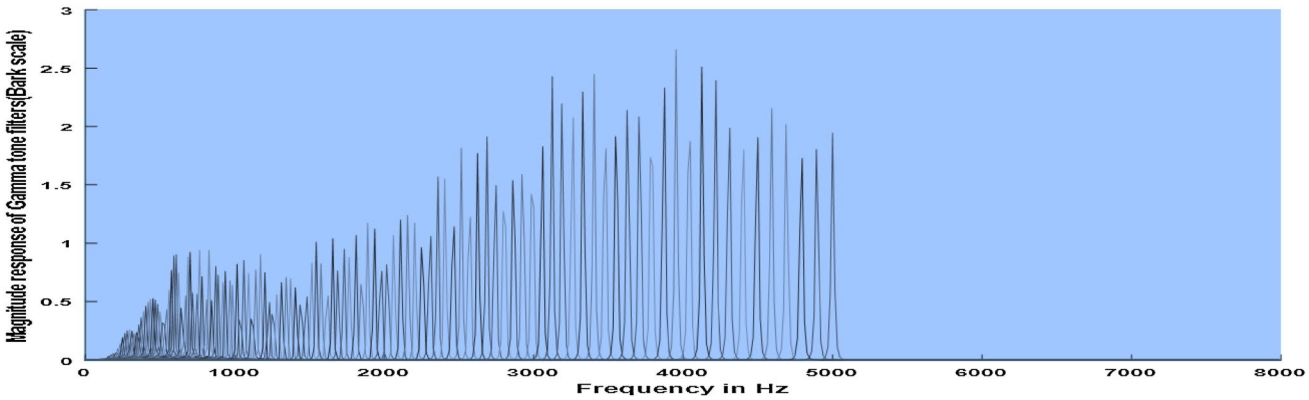

Fig. 3 Frequency response of Gamma tone filters spaced in BARK scale

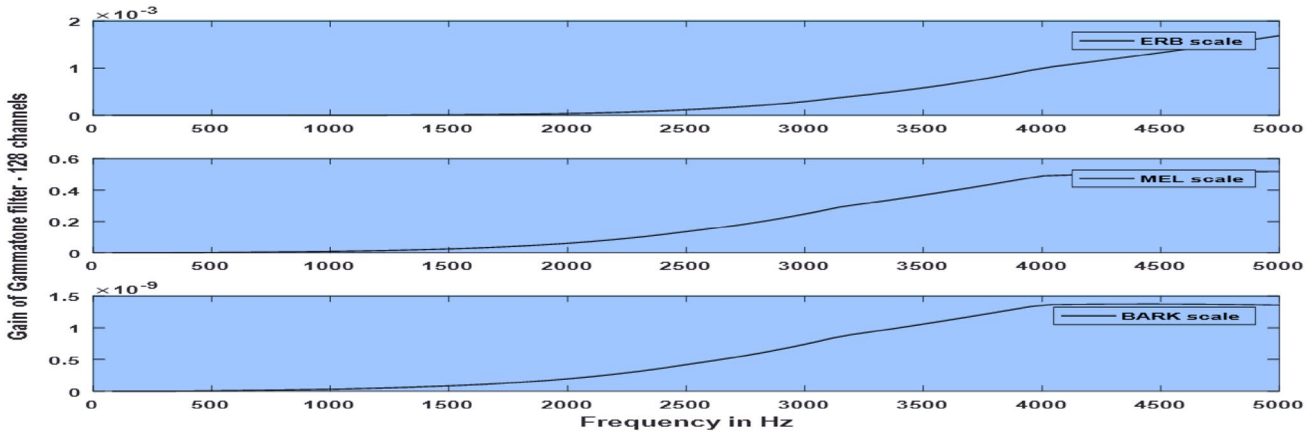

Fig. 4 Gain adjustment plot for Gamma tone filters spaced in ERB, MEL and BARK scale

chosen from the standard TIMIT corpus and phonetically-balanced for each emotion. High-quality audio processing equipment is used to record the utterances and processed to be free from noise, and the speech files are appropriately labelled for using them to evaluate the performance of the emotion recognition algorithm/system. In our work, templates for each emotion are created by feature extraction from the preprocessed speech and the extracted features are applied to the different training algorithms. Preprocessing stages include pre-emphasis block to spectrally flatten the speech signal, followed by, frame blocking which converts the speech signal into overlapped speech frames with 16 ms duration for the frames and 8 ms overlapping. Each speech frame is windowed by the standard Hamming window to remove the signal discontinuities at the beginning and end of the speech segment. For each frame, GTF energy and GTFCC features with the Gamma tone filters spaced in ERB, MEL and BARK scale and number of coefficients in each feature vector is 13. Normalization is done on each coefficient vector and thus feature vectors are formed. Feature vectors are applied to develop the set of clusters using VQ and FCM modelling techniques for each emotion. Templates are created with MHMM modelling techniques based on the expectation maximisation algorithm and SVM structures which contain information about the trained classifier, including the support vectors. For emotion recognition system using EMO-DB database, during training, set of ten utterances uttered by nine speakers are used, and the utterances are constituting the observation sequence with predefined characteristics in time and frequency domain. For the system using SAVEE database, statements of three speakers in a particular emotion are considered for developing training models/templates. Statements of tenth speaker/fourth speaker in the specific feeling have been used for testing on using EMO-DB database and SAVEE database respectively. Testing is done by applying the feature vectors of the GTF

energy and GTFCC features considered in our work to the group models and subsequently to the emotion-specific models in a group. Decision level fusion is performed to augment the overall accuracy of the system.

3.1 Experimental analysis based on VQ/FCM/MHMM/SVM technique

VQ based clustering algorithm (Juang 1993) is used to convert the set of training vectors into a set of clusters as representative templates/models of the emotions. It partitions the points in the N-by-P training data matrix X into M-by-P clusters with P data coefficients for each centroid. This partition is done by finding the distance between the training vectors and cluster centroids and the training vectors are grouped into cluster set based on minimum distance criterion. Classification procedure deals with computation of the Euclidean distance between each of the test vectors and M cluster centroids of all models. Cluster centroid is appropriately chosen which produces the minimum distance between the test vector and cluster centroids of each model. This is repeated for all the test vectors. Model is chosen to be associated with the test speech which has the minimum of an average of these minimum distances. The spectral distance measure for comparing features v_i and v_j is as in (6).

$$d(v_i, v_j) = d_{ij} = 0 \quad \text{when} \quad v_i = v_j \tag{6}$$

If codebook vectors of an M-vector codebook are taken as y_m, $1 \le m \le M$ and new spectral vector to be classified is denoted as v, then the index $m^*$of the best codebook entry is as in (7)

$$m^* = \arg\left(\min\left(d(v, y_m)\right)\right) \quad \text{for} \quad 1 \le m \le M \tag{7}$$

The formation of clusters better captures characteristic of the training data distribution. It is observed that the most frequently occurring test vectors have small distance and the least frequently occurring ones have considerable distance. In FCM, training data size is considered as M-by-N with M and N indicating number of training vectors and number of coordinates for each data vector respectively. The coordinates for each cluster centre are returned in the rows of the matrix cluster centroid with membership function indicating full and no membership for the values 1 and 0 respectively and partial membership for the values between 0 and 1. Clustering process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of development specified. Training procedure using MHMM modelling technique computes the maximum likelihood parameters of an HMM with (mixtures of) Gaussians output using the expectation–maximization algorithm. Testing the data using the MHMM technique deals with the application of test features to the MHMM models, and likelihood values are computed. The identification of the model which is closely associated with the test data is done by choosing the model whose likelihood value is the largest. For using SVM for training and classification, the training algorithm is used to train the model to create SVM structure containing support vectors in the appropriate emotion which will be later used for classification of the test data for the particular passion.

3.2 Experimental evaluation: results and discussion

GTF and GTFCC feature based emotion recognition system are evaluated by applying test speech vectors to the training models corresponding to the emotions. Training and recognition phase of the of emotion recognition system involves the extraction of features, development of templates and testing the speech data for the appropriate feeling. The method used for extraction of GTF Energy and GTFCC features from the speech signal is shown in Fig. 5. First, the set of speech utterances are concatenated to form a single vector of pertinent emotion. Pre-emphasis is done on the speech vector to flatten the signal spectrally. Then, the speech vector is converted into frames of 16 ms duration with 50% overlap. Each frame is windowed by using Hamming window which has been universally used for speech processing applications for removing the signal discontinuities at the beginning and end of the speech frames. GTF and GTFCC feature for the Gamma tone filters spaced in ERB, MEL and BARK scale are extracted for each speech frame. Then, feature vectors are applied to the training algorithm, and VQ/FCM/MHMM/SVM models are developed for GTF Energy features, and VQ/FCM/SVM models are created for GTFCC features.

Block diagram of a parallel group classifier and decision level fusion classifier is shown in Figs. 6 and 7 for GTF energy and GTFCC features. Speech utterances considered for testing in particular emotion are concatenated and after initial pre-processing stages such as pre-emphasis, frame blocking and windowing, GTF energy and GTFCC features are extracted. A group of 100 vectors (Reynolds et al. 1995) can form features of a test segment. Feature vectors of the speech segment of the test emotion are applied to the group models, and the group is identified based on the comparison concerning the minimum of an average of minimum distances. Then, subsequently testing is done with models of emotions in a group. Speech segment of the test emotion is identified by first computing average of minimum distances for all the models corresponding to the group and classification is done with emotion-specific models, among

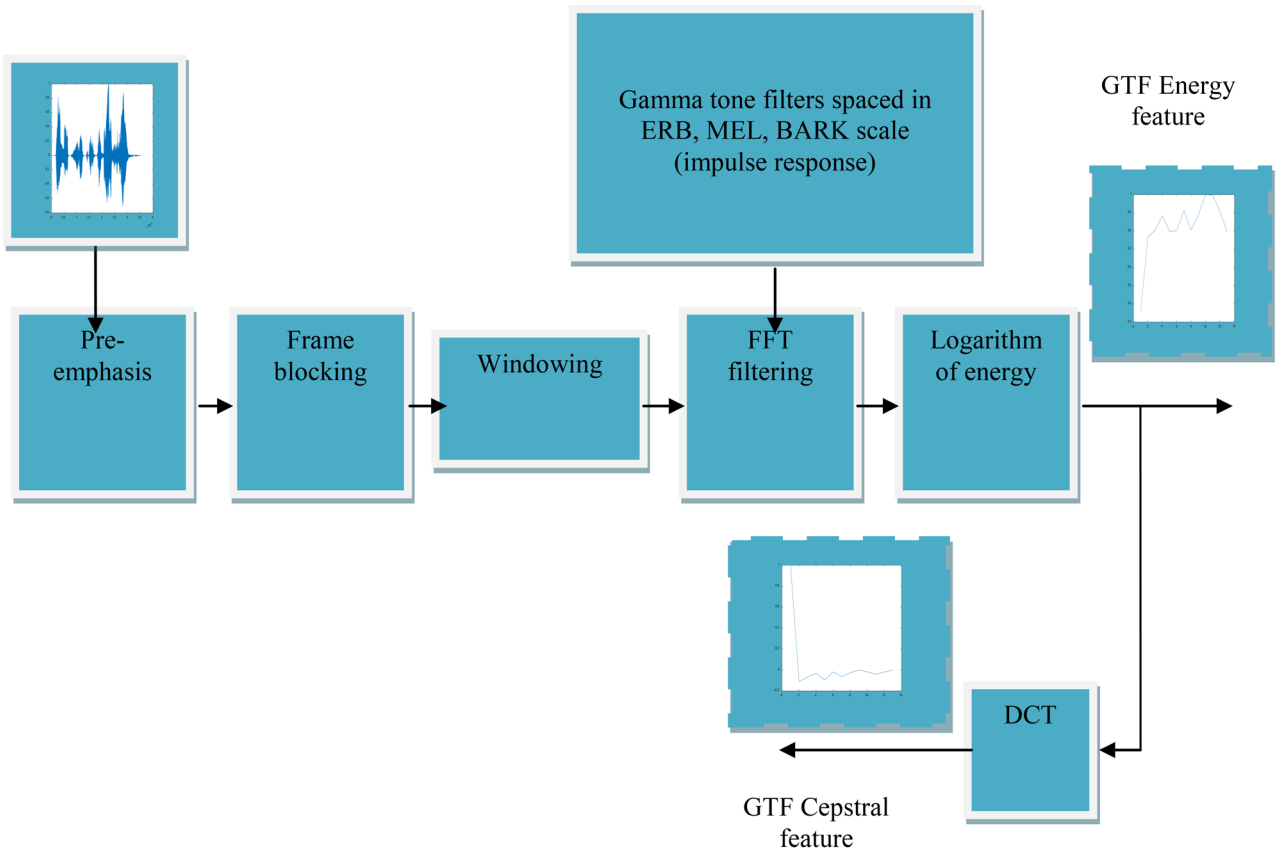

Fig. 5 Emotion recognition—training phase

which the model produces the minimum of averages is the identified model corresponding to the emotion for VQ and fuzzy VQ based algorithms. SVM based emotion recognition is done by performing classification based on each row of the Feature vectors of the test data using the information in a SVM structure created during training. The number of feature vectors matched with each group is calculated, and the group is identified which a maximum number of feature vectors has associated. Continuous density HMM technique (MHMM) uses the computation of the log-likelihood values for each model with the feature vectors of the test speech segment, and correct classification is done concerning the model whose log-likelihood value is maximum.

Classification is done based on minimum distance for VQ and FCM clustering approaches, and the weighted average recall is the number of correct choices over the total number of test speech segments considered for each emotion. Figures 8 and 9 indicate the speech waveforms of the same utterance spoken by the same speaker in arousal emotions such as anger, fear, and happy emotions and soft emotions such as boredom, disgust, neutral and sad emotions respectively.

Figures 10 and 11 indicate the GTF energy feature variation between the arousal emotions anger and fear and GTFCC feature variation between soft emotions sad and neutral.

Figures 12 and 13 indicate the spectral variation for the speech uttered by the same speaker in different emotions corresponding to the group of arousal and soft emotions.

The plot is shown in Fig. 14 which depicts the performance evaluation of the system for EMO-DB database for the application of the GTF energy features with Gamma tone filters spaced in ERB scale to the emotion-specific models in a group for VQ/FCM/MHMM/SVM modelling and classification techniques.

The plot is shown in Fig. 15 which indicates the performance of the emotion recognition system for EMO-DB database for the application of the GTF energy features with Gamma tone filters spaced in MEL scale to the emotion-specific models in a group for VQ/FCM/MHMM/SVM modelling and classification techniques.

The plot is shown in Fig. 16 which provides the performance evaluation of the system for EMO-DB database for the application of the GTF energy features with Gamma tone

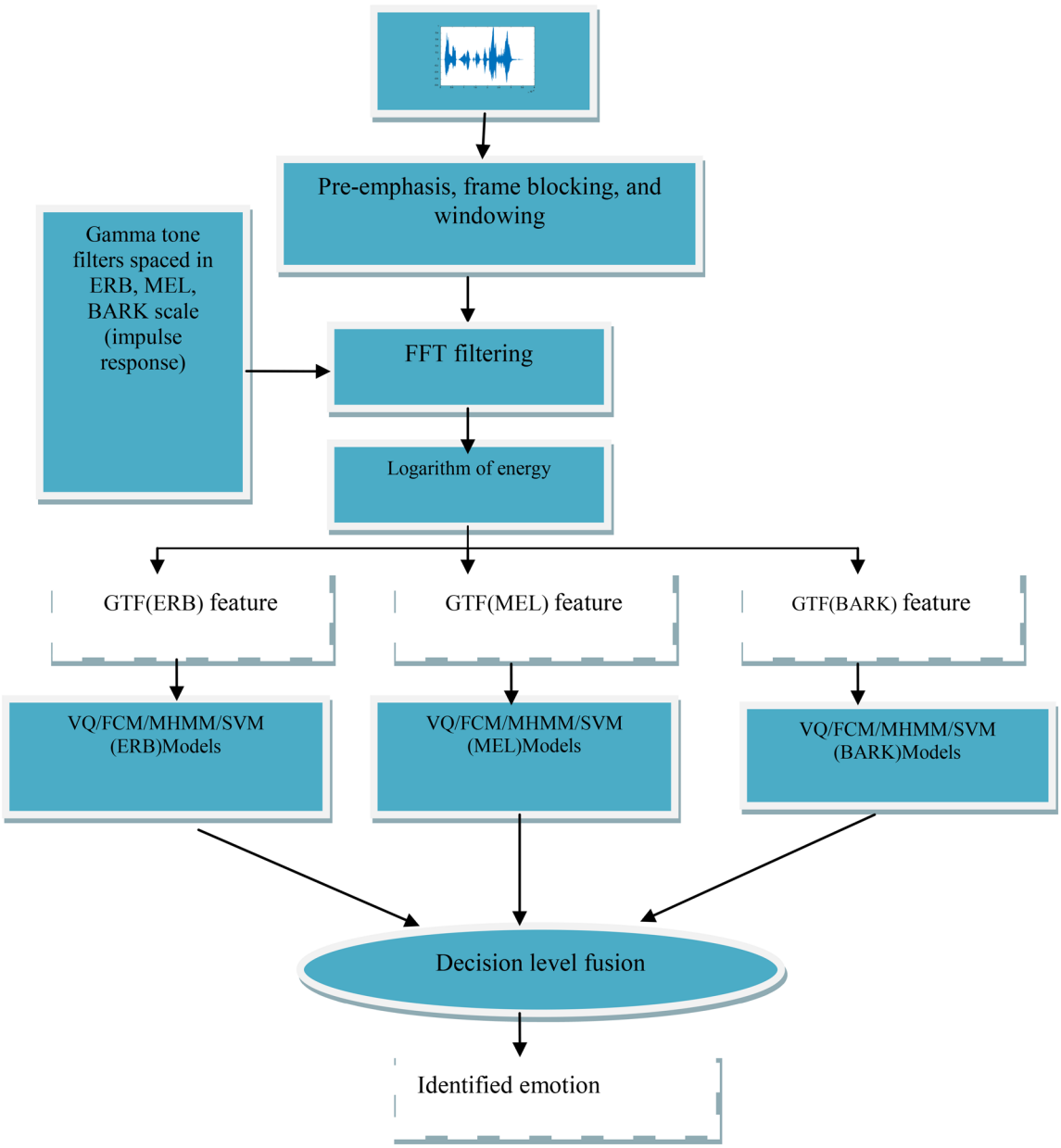

Fig. 6 Emotion recognition—testing phase—GTF energy features

filters spaced in BARK scale to the emotion-specific models in a group for VQ/FCM/MHMM/SVM modelling and classification techniques.

Figure 17 depicts the average performance of the emotion recognition system by considering GTF energy features with Gamma tone filters spaced in ERB or MEL or BARK scale and 'true' state decides the correct identification of the emotion regarding any one technique among VQ./FCM/MHMM/SVM modelling techniques used. Overall accuracy is 100% except for boredom.

From the Fig. 17, it is evident that decision level fusion classification among the modelling techniques produces better accuracy for all emotions except 'boredom'. Figures 18, 19 and 20 provide the details about the performance of the system for GTFCC with filters spaced in ERB/MEL/BARK scale for the three modelling techniques such as VQ, FCM, and SVM.

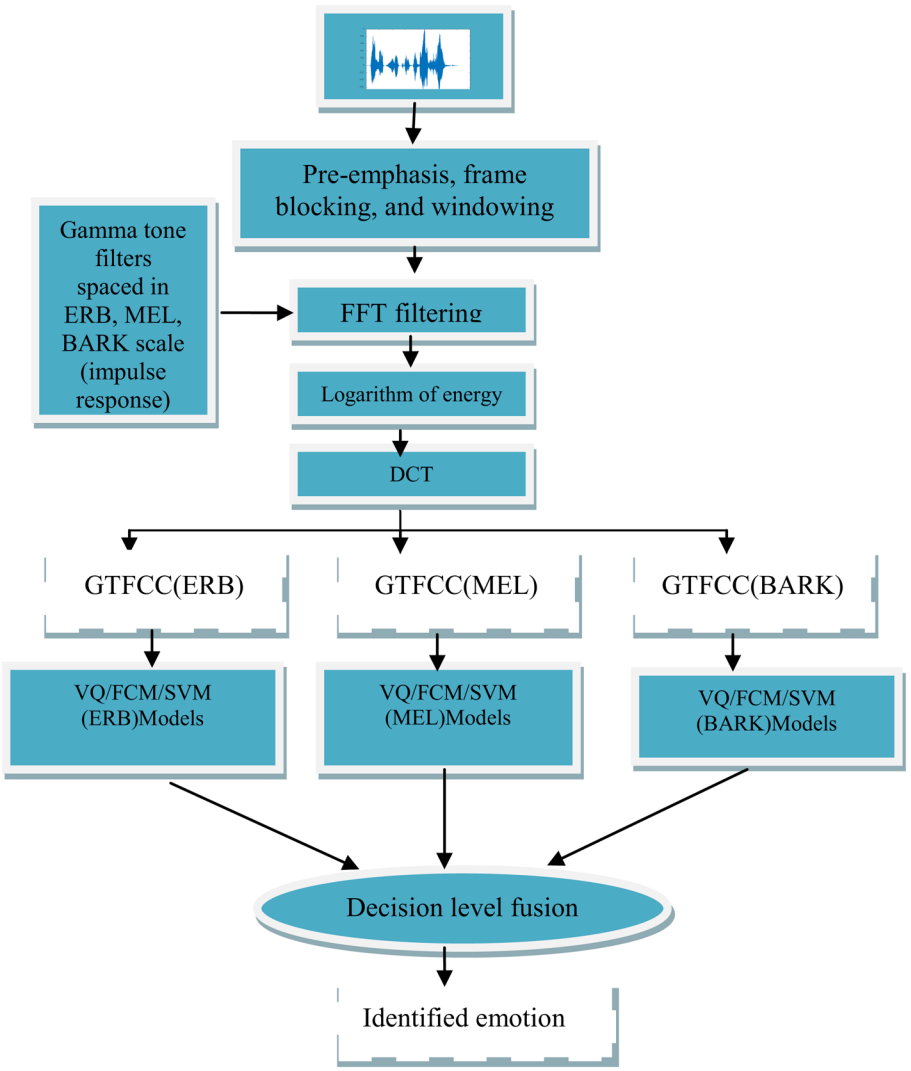

Fig. 7 Parallel group and decision level fusion classifier—GTFCC feature

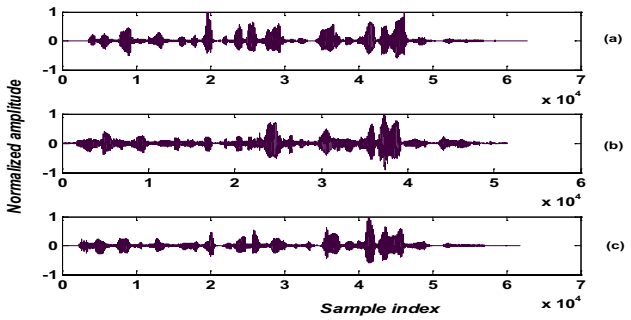

Fig. 8 Signal **a** anger, **b** fear, **c** happy

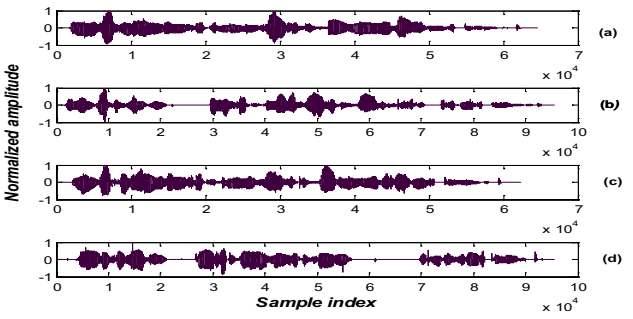

Fig. 9 Signal **a** boredom, **b** disgust, **c** neutral, **d** sad

Figure 21 depicts the robust performance of the system using the decision level classifier concerning GTFCC feature with filters spaced in ERB/MEL/BARK scale and the modelling techniques VQ/FCM/SVM.

From the Fig. 21, it is evident that the accuracy is 100% for all emotions for the decision level fusion classifier implemented by using the GTFCC features with Gamma tone filters spaced in ERB/MEL/BARK scale and modelling

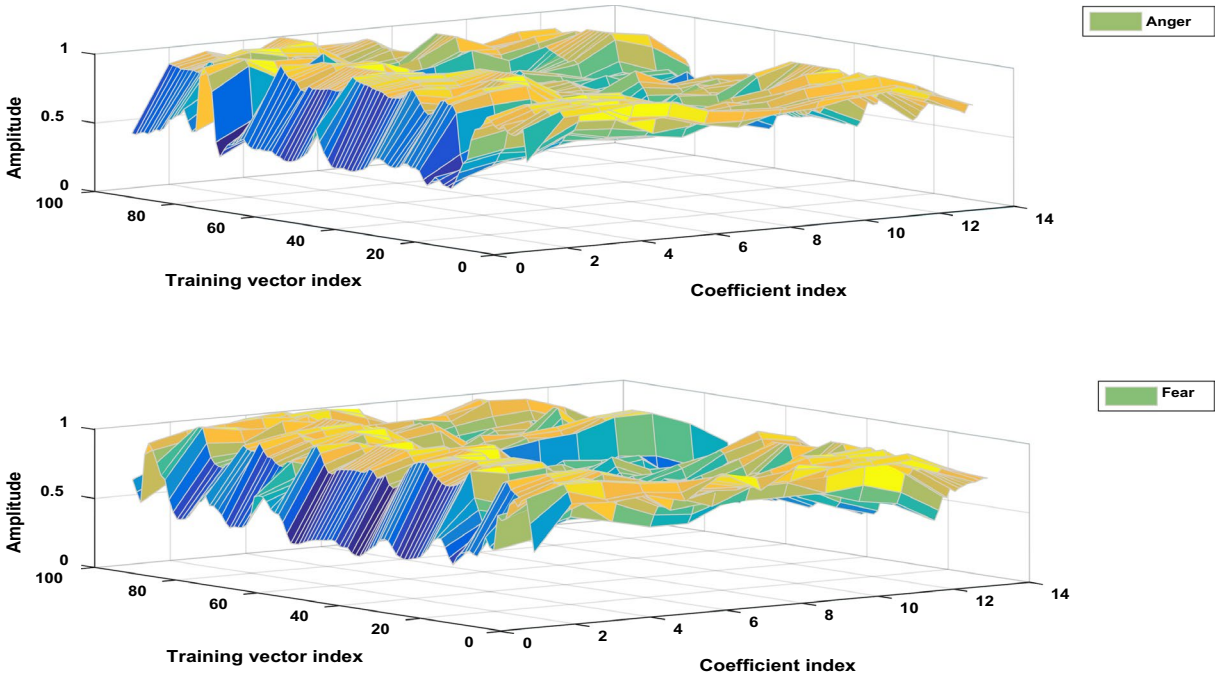

Fig. 10 GTF Energy feature variation—arousal emotions—anger and fear

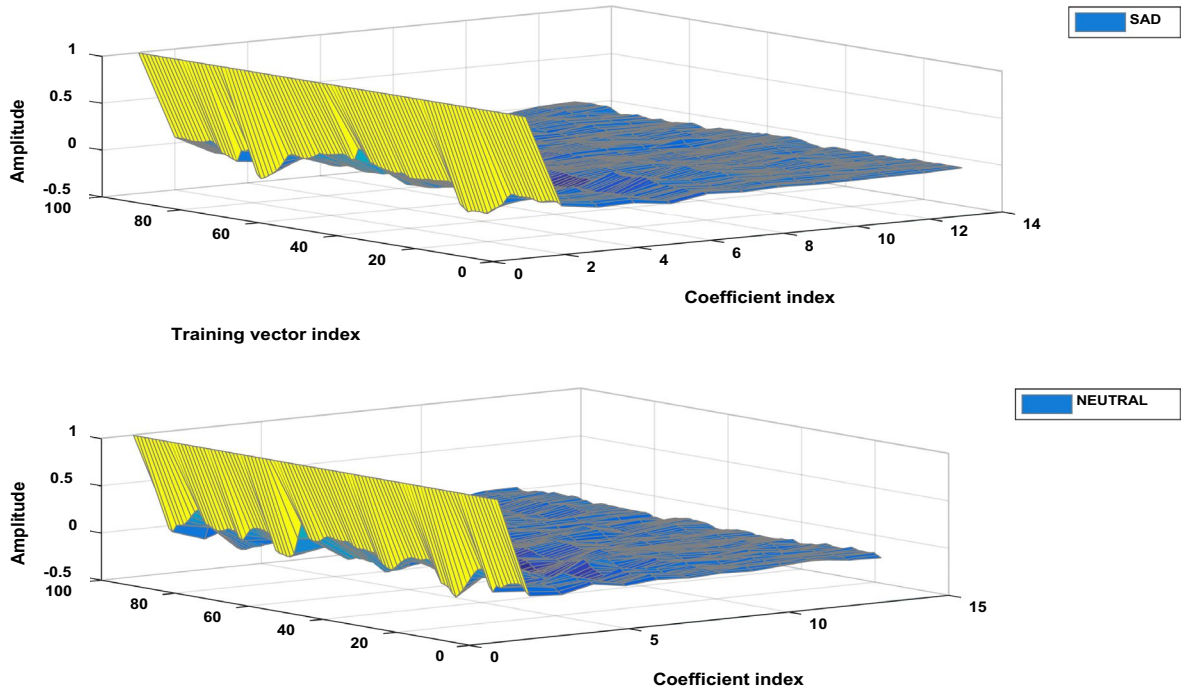

Fig. 11 GTFCC feature variation—soft emotions—sad and neutral

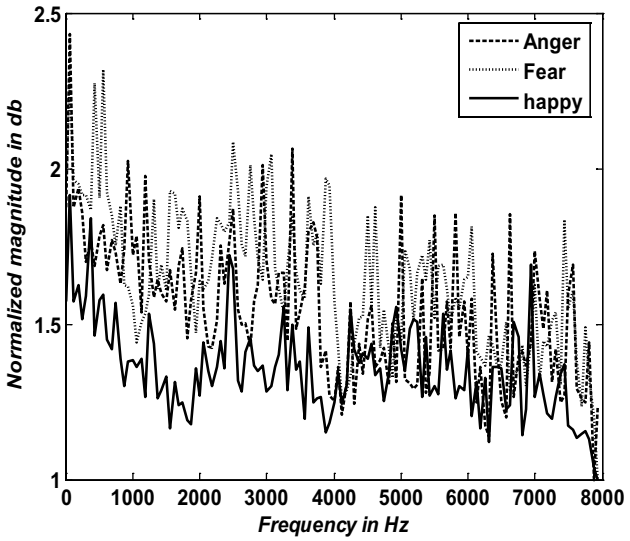

Fig. 12 Spectral variation—arousal emotions

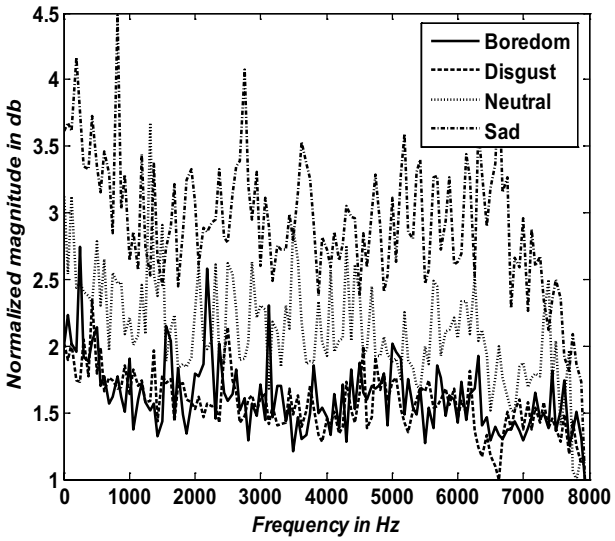

Fig. 13 Spectral variation—soft emotions

techniques VQ/FCM/SVM. Performance of the system is also evaluated for the emotional speech utterances from the "SAVEE" database. GTF energy features with Gamma tone filters spaced in ERB/MEL/BARK scale are assessed by using the modelling techniques VQ/FCM/MHMM, and Figs. 22, 23 and 24 indicate the performance of the system.

Figure 25 indicates the Average performance of the system for using decision level fusion classifier using features with Gamma tone filters spaced in ERB/MEL/BARK scale and modelling techniques VQ/FCM/MHMM.

The plot shown in Fig. 25 reveals that the average accuracy is 100% for all emotions except disgust, sad and surprise emotions. Figure 26 depicts the average performance of the system for EMO-DB database for GTF energy features with Gamma tone filters spaced in ERB/MEL/BARK scale by using VQ/FCM/MHMM/SVM as modelling techniques. Gamma tone filters spaced in ERB and BARK scale give the better performance consistently for all modelling techniques.

Figure 27 indicates the average performance of the system for SAVEE database for GTF energy features with filters spaced in ERB/MEL/BARK scale by using VQ/FCM/MHMM as modelling techniques. It is revealed that Gamma tone filters spaced in ERB and MEL perform better for VQ/MHMM techniques and FCM technique respectively. However, filters spaced in BARK scale play better for the decision level fusion of the modelling techniques.

4 Conclusions

This paper proposes the extraction of GTF energy and GTF Cepstral features with the Gamma tone filters spaced in ERB/MEL/BARK scale from the speech signal and modelling techniques to evaluate the performance of the emotion recognition system for the utterances chosen from EMO-DB database and SAVEE database. Speech utterances considered for training are concatenated and converted into frames after pre-emphasis and windowing. GTF energy and GTFCC features with Gamma tone filters spaced in ERB/MEL/BARK scale are extracted. This feature set is applied to the VQ/ FCM/MHMM/SVM template production techniques, and templates/models are created. GTF energy and GTFCC features extracted from the test utterances and applied to the emotion-specific models in a group and group classification is done between arousal and soft emotions. Subsequently, identification of emotion is done with emotion specific models in a group, and the performance is assessed with computation of recognition accuracy. Fusion level classification provides the overall efficiency of 99.8% for EMO-DB database for GTF energy feature. GTFCC feature offers the overall accuracy of 100% for the decision level fusion classification by considering the features with Gamma tone filters spaced in ERB/MEL/BARK scale and VQ/FCM/SVM modelling techniques for EMO-DB database. For SAVEE database, decision level fusion classification is done using GTF energy features with Gamma tone filters spaced in ERB/MEL/BARK scale and VQ/FCM/MHMM modelling techniques provide the overall accuracy of 99.4%. These results indicate the effectiveness of the GTF energy and

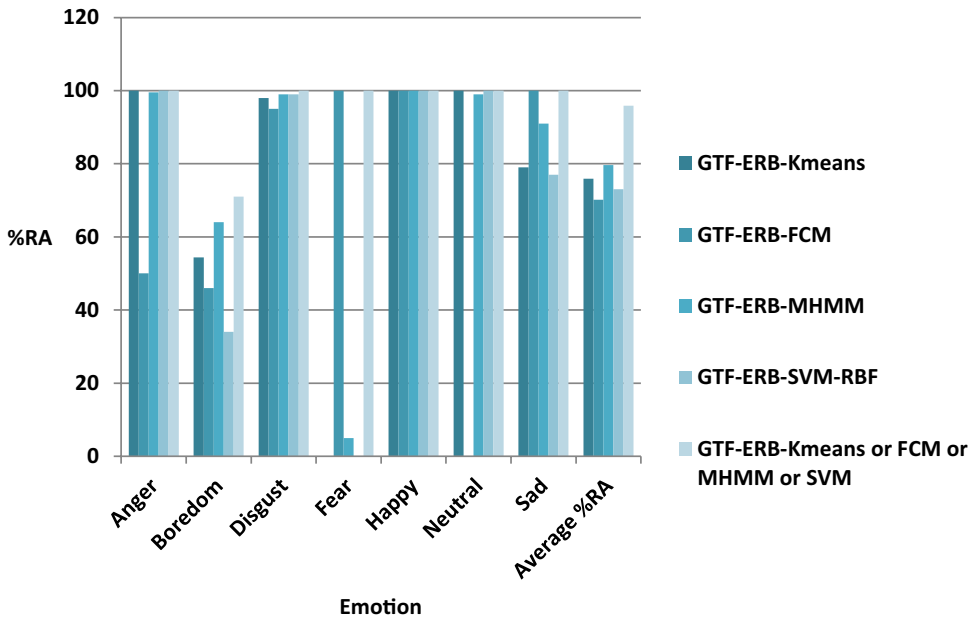

Fig. 14 Performance evaluation of the system for EMO-DB database—GTF energy feature (ERB Scale)

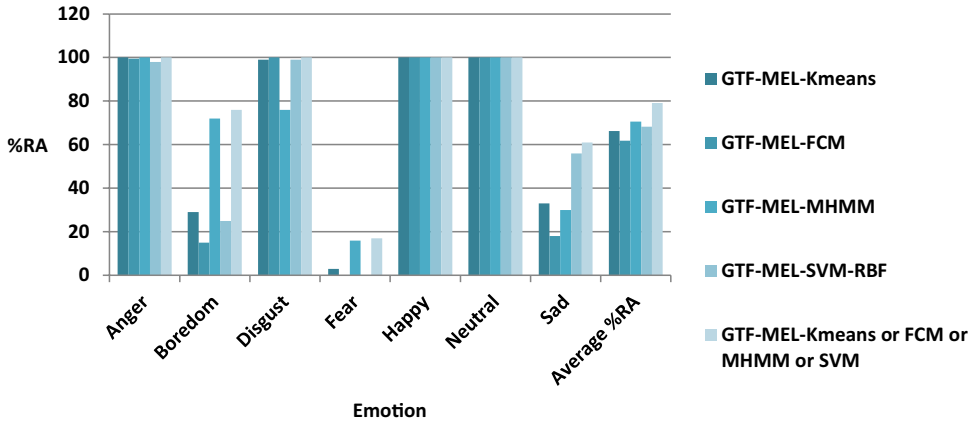

Fig. 15 Performance evaluation of the system for EMO-DB database—GTF energy feature (MEL Scale)

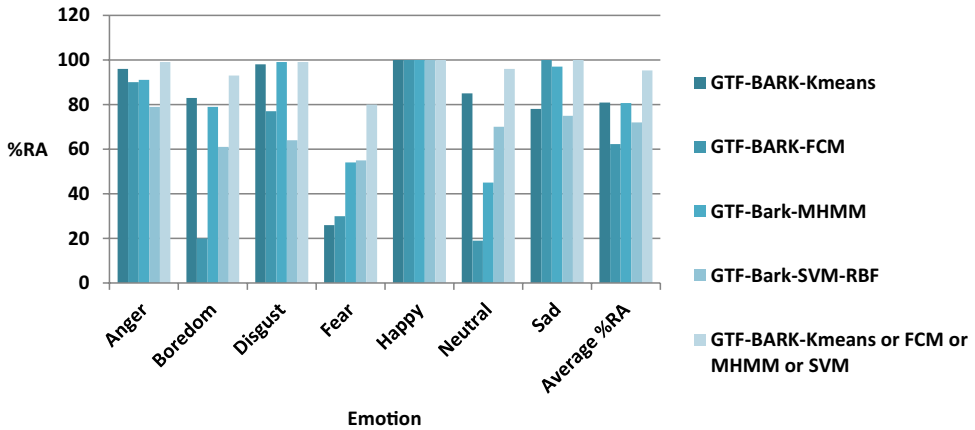

Fig. 16 Performance evaluation of the system for EMO-DB database—GTF energy feature (BARK Scale)

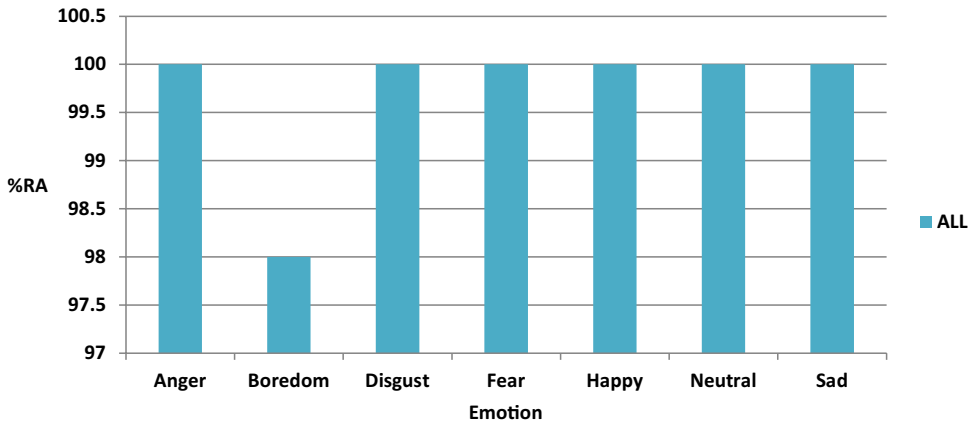

Fig. 17 Average performance for emotions in EMO-DB database—GTF energy features (ERB or MEL or BARK) for VQ/FCM/MHMM/SVM techniques

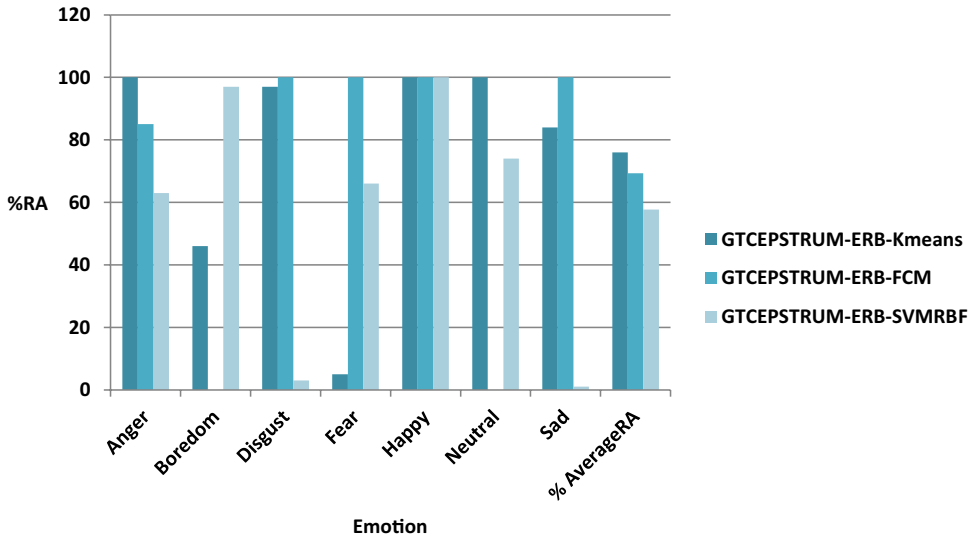

Fig. 18 Performance evaluation of the system for EMO-DB database—GTFCC feature (ERB Scale)

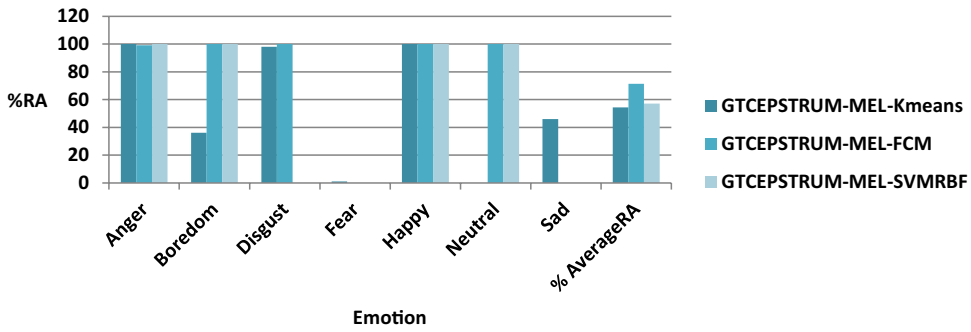

Fig. 19 Performance evaluation of the system for EMO-DB database—GTFCC feature (MEL Scale)

GTFCC features, modelling techniques and decision level fusion classification for recognising emotions from the emotional speeches uttered by the limited set of speakers/actors. A speaker independent and different style of speech ERS finds applications in human–machine interactions, healthcare systems and investigation of criminal activities. The extention of these proposed features and decision level fusion techniques to real-time emotion recognition

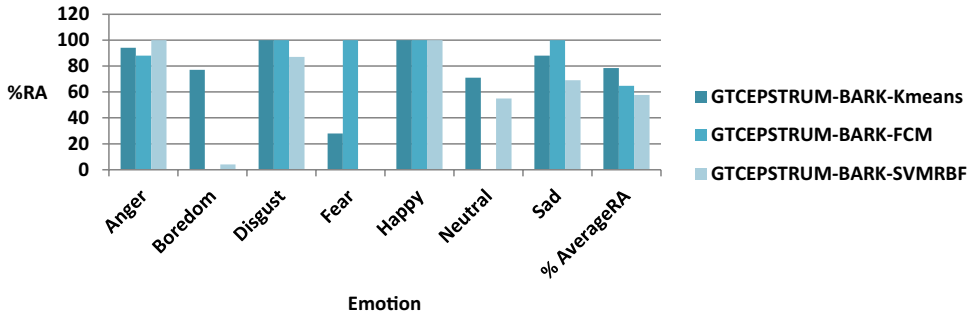

Fig. 20 Performance evaluation of the system for EMO-DB database—GTFCC feature (BARK Scale)

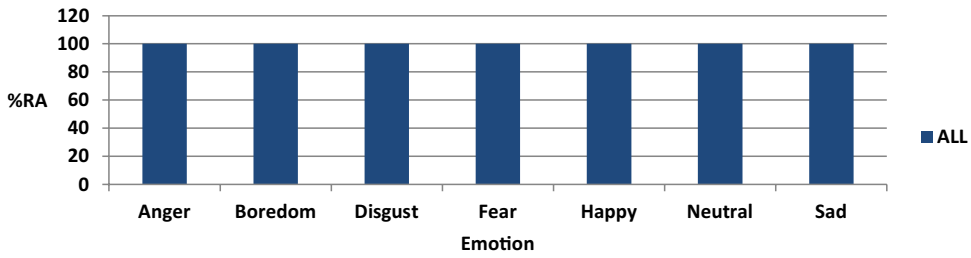

Fig. 21 Average performance for emotions in EMO-DB database—GTFCC features (ERB or MEL or BARK) for VQ/FCM/SVM techniques

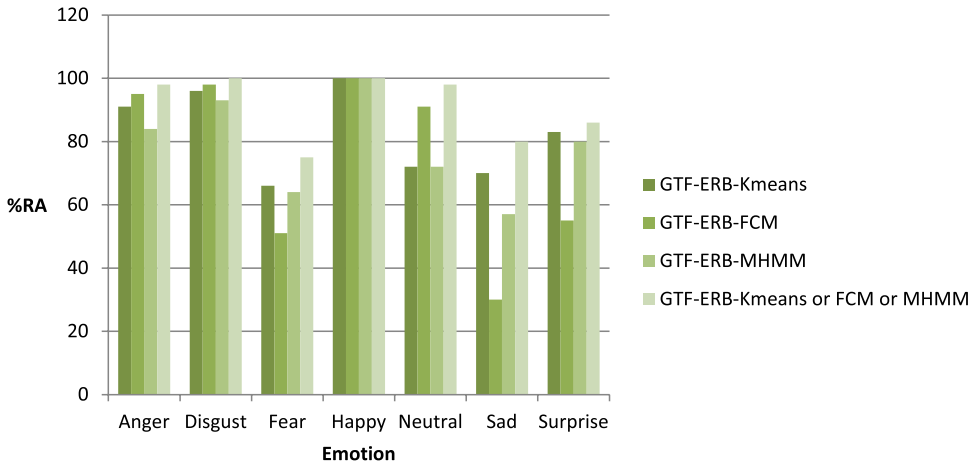

Fig. 22 Performance evaluation of the system for SAVEE database—GTF energy feature (ERB Scale)

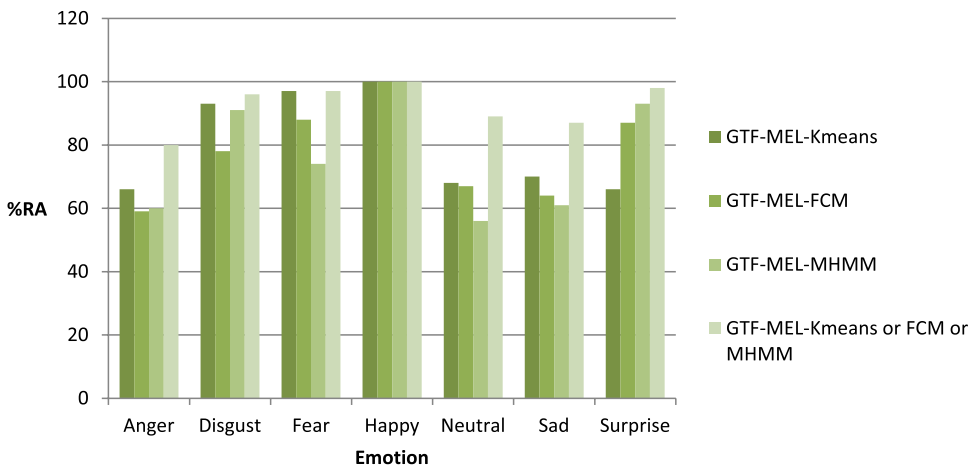

Fig. 23 Performance evaluation of the system for SAVEE database—GTF energy feature (MEL Scale)

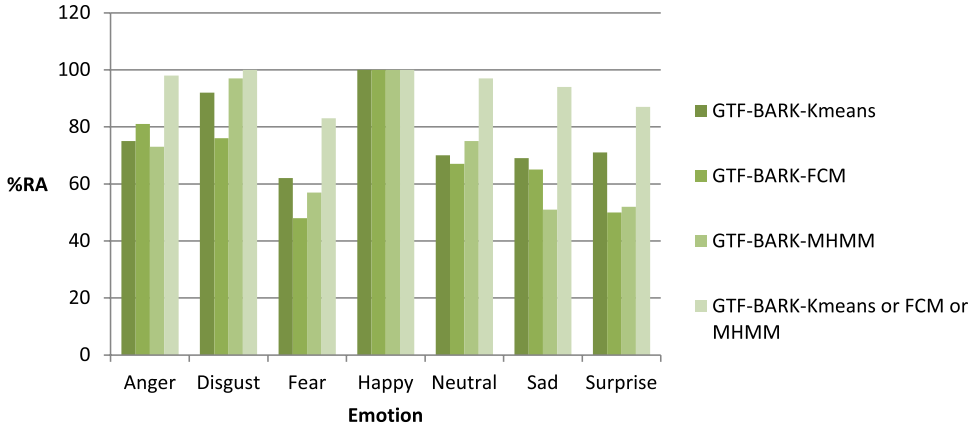

Fig. 24 Performance evaluation of the system for SAVEE database—GTF energy feature (BARK Scale)

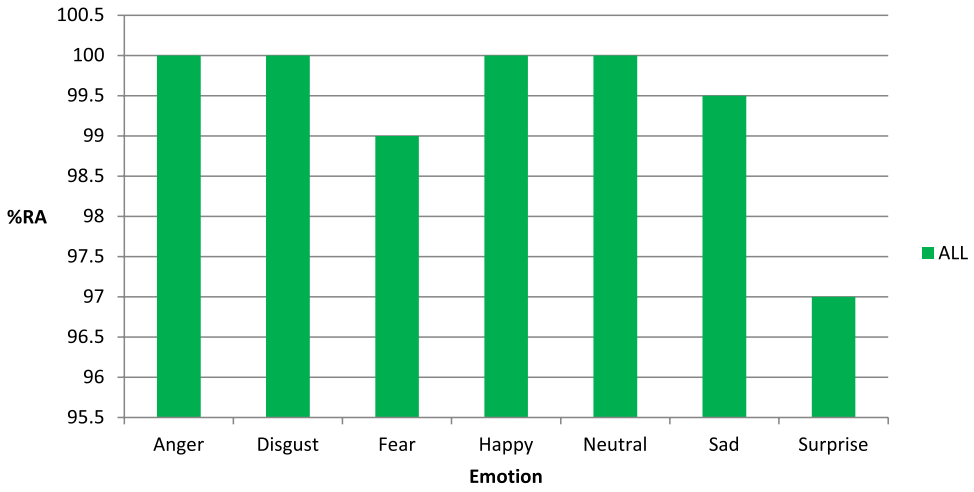

Fig. 25 Average performance of the system for SAVEE database—GTF Energy feature—decision level fusion classifier

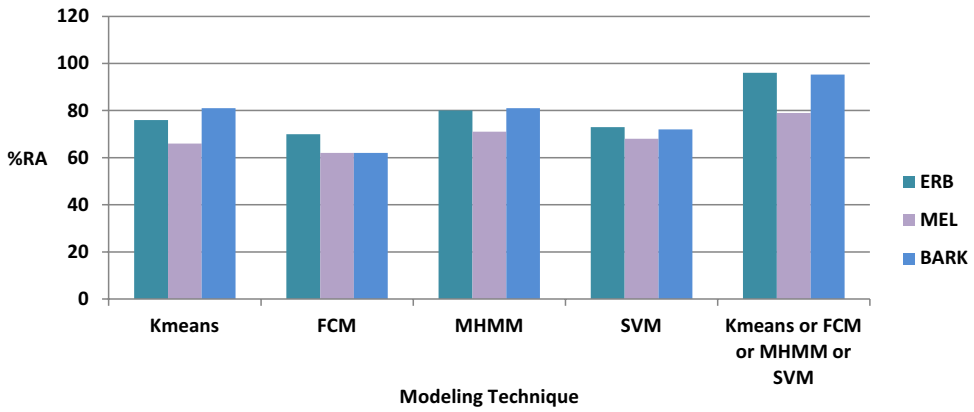

Fig. 26 Average performance of the system for EMO-DB database for the Gamma tone filters spaced in ERB/MEL/BARK scale based on modelling techniques used

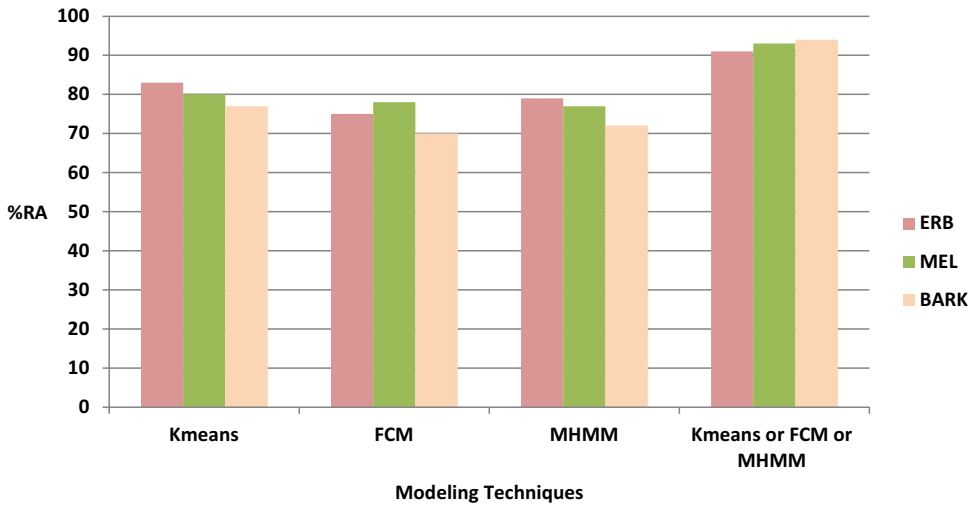

Fig. 27 Average performance of the system for SAVEE database for the Gamma tone filters spaced in ERB/MEL/BARK scale based on the modelling techniques used

systems for automatically detecting the emotions from the natural speech.

References

Anagnostopoulos, C.-N.,·Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review, 43*, 155–177.

Babu, M., Arun Kumar, M. N., & Santhosh, S. M. (2014). Extracting MFCC AND GTCC features for emotion recognition from audio speech signals. *International Journal of Research in Computer Applications and Robotics, 2*(8), 46–63.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of german emotional speech (EMO-DB). Proceedings Interspeech. Lissabon, Portugal. http://emodb.bilderbar.info/start.html.

Garg, E., & Bahl, M. (2014). Emotion recognition in speech using gammatone cepstral coefficients. *International Journal of Application or Innovation in Engineering & Management (IJAIEM), 3*(10), 285–291.

Kaur, I., Kumar, R., Kaur, P. (2017). Speech emotion detection based on optimistic—DNN (Deep Neural Network) approach. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 6*(4), 150–156.

Koolagudi, S. G., Sharma, K., & Sreenivasa Rao, K. (2012). Speaker recognition in emotional environment. *Communications in Computer and Information Science, 305*, 117–124.

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication, 53*, 1162–1171.

Li, Z., & Gao, Y. (2016). Acoustic feature extraction method for robust speaker identification. *International Journal of Multimedia Tools and Applications, 75*, 7391–7406.

Marković, B., Galić, J., Grozdić, Đ, Jovičić, S. T., & Mijić, M. (2017). Whispered speech recognition based on gammatone filterbank cepstral coefficients. *Journal of Communications Technology and Electronics, 62*(11), 1255–1261.

Mohanty, S. (2016). Language independent emotion recognition in speech signals. *International Journal of Advanced Research in Computer Science and Software Engineering, 6*(10), 299–301.

Moore, J. D., Tian, L., Lai, C. (2014). *Word-level emotion recognition using high-level features*, LNCS. Berlin: Springer. https://doi.org/10.1007/978-3-642-54903-8_2.

Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication, 49*, 98–112.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*, 603–623.

Patel, P., Chaudhari, A., Kale, R., & Pund, M. A. (2009). Emotion recognition from speech with gaussian mixture models & via boosted GMM. *International Journal of Research In Science & Engineering, 3*(2), 47–53.

Peng, Z., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2017). Speech emotion recognition using multichannel parallel convolutional recurrent neural networks based on Gammatone Auditory Filterbank. *Proceedings of APSIPA Annual Summit and Conference*, pp 1750–1755. https://ieeexplore.ieee.org/document/8282316/.

Pervaiz, M., & Khan, T. A. (2016). Emotion recognition from speech using prosodic and linguistic features. *International Journal of Advanced Computer Science and Applications, 7*(8), 84–90.

Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology, 16*, 143–160.

Rabiner, L. & Juang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice Hall.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72–83.

Sapra, A., Panwar, N., & Panwar, S. (2013). Emotion recognition from speech. *International Journal of Emerging Technology and Advanced Engineering, 3*(2), 341–345.

Shahin, I. (2009). Speaker identification in emotional environments. *Iranian Journal of Electrical and Computer Engineering, 8*(1, Winter-Spring), 41–46.

Sharma, A., Anderson, D. V. (2015). Deep emotion recognition using prosodic and spectral feature extraction and classification based on cross-validation and bootstrap. *IEEE Signal Processing and Signal Processing Education Workshop*. https://ieeexplore.ieee.org/document/7369591/.

Sreenivasa Rao, K., Kumar, T. P., Anusha, K., Leela, B., Bhavana, I., & Gowtham, S. V. S. K. (2012). Emotion recognition from speech. *International Journal of Computer Science and Information Technologies, 3*(2), 3603–3607.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE ICASSP*, pp. 5200–5204. https://ieeexplore.ieee.org/document/7472669/.

Vogt, T., Andr, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. *Proceedings Language Resources and Evaluation Conference*, pp. 1123–1126. https://www.informatik.uni-augsburg.de/lehrstuehle/hcm/publications/2006-LREC/.

Wua, S., Falk b, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication, 53*, 768–785.

Yogesh, C. K., Hariharan, M., Ngadiran, R., Adom, A. H., Yaacob, S., Berkai, C., Polat, K. (2017). A new hybrid PSO assisted biogeography-based optimisation for emotion and stress recognition from speech signal. *Expert Systems with Applications, 69*, 149–158.

Zhang, W., Meng, X., Li, Z., Lu, Q., & Tan, S. (2015). Emotion recognition in speech using multi-classification SVM. *UIC-ATC-IEEE ScalCom-CBDCom-IoP*, pp. 1181–1186. https://ieeexplore.ieee.org/document/7518394/.