Egyptian Informatics Journal (2016) **17**, 1–10

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

FULL-LENGTH ARTICLE

Music emotion recognition: The combined evidence of MFCC and residual phase

N.J. Nalini *, **S. Palanivel**

Dept of Computer Science and Engineering, Annamalai University, Tamil Nadu 608002, India

Received 27 September 2014; accepted 8 May 2015

KEYWORDS

Music emotion recognition;
Mel frequency cepstral coefficient;
Residual phase;
Autoassociative neural network;
Support vector machine

Abstract The proposed work combines the evidence from mel frequency cepstral coefficients (MFCC) and residual phase (RP) features for emotion recognition in music. Emotion recognition in music considers the emotions namely anger, fear, happy, neutral and sad. Residual phase feature is an excitation source feature and it is used to exploit emotion specific information present in music signal. The residual phase is defined as the cosine of the phase function of the analytic signal derived from the linear prediction (LP) residual and also it is demonstrated that the residual phase signal contains emotion specific information that is complementary to the MFCC features. MFCC and residual phase features are used to create separate models for each emotion. The evidence from the models is combined at the score level for each emotion and it is used to recognize the emotion. The proposed method is evaluated using music files recorded from various websites and the method achieves a performance of 96.0%, 99.0%, 95.0% using AANN, SVM, RBFNN, respectively.

1. Introduction

Music plays an important role in human history and almost all music is created to convey emotion. Music emotion recognition (MER) got much development these years, and apparently, it will play an important role in digital entertainment and harmonious human–machine interaction [1]. However, the connection between music and feelings (emotion) is probably the simplest and profound. The emotion's crucial role in musical experience has long been the object of philosophical reflection and empirical study.

Many issues for music emotion recognition have been addressed by different disciplines such as psychology, physiology, musicology and cognitive science [2–4]. Emotion recognition from music signal is a difficult task due to the following reasons: First, emotion observation is basically subjective and people can recognize different emotions for the same song. Second, it is not easy to express emotion in a worldwide way because the adjectives used to describe emotions may be unclear, and the use of adjectives for the same emotion can vary from person to person. Third, it is still hard to know how music evokes emotion. Computational systems for music mood recognition may be based upon a model of emotion,

---

* Corresponding author.

E-mail addresses: njncse78@gmail.com (N.J. Nalini), spal_yughu@yahoo.com (S. Palanivel).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.05.004

although such representations remain an active topic of psychology research.

An important step in MER is feature extraction and classification. The importance in determination of feature extraction from music signals is in the sense that they represent the music well and computation can be carried out efficiently. Much of work on extraction of features from music devoted to timbre features. MFCC is the well known timbre texture feature or spectrum features which is the highest performing individual feature used in speech recognition, can be examined for modeling of music. Among regular features used in MER, in this work, there is a search for new features for recognizing the musical emotions.

Residual phase (RP) features is an excitation source feature used by very few researches for speaker recognition. In this work it can be used to exploit emotion specific information present in music signal. It is defined as the cosine of the phase function of analytic signal derived from LP residual. LP residual demonstrates the presence of audio specific information obtained after removing the predictable part of the signal [5]. It is known that the residual of a signal is less subject to degradations as compared to spectral information. So systems built using the residual may be robust against degradations. This emphasizes the importance of information present in the LP residual of music signal.

The MFCC features and residual phase features are combined at the score level to enhance the performance of the system and it is used with pattern classification techniques such as autoassociative neural network, support vector machine and radial basis function neural network for music emotion recognition. The recognition rates from the models, evidence that the emotion specific information is present in residual phase signal.

The paper is organized as follows. In Section 2, a review of some of the existing methods for music emotion recognition is described. The spectral and excitation source feature extraction for music emotion recognition is presented in Section 3. In Section 4 techniques used in MER is described. Experimental results and conclusion are discussed in Sections 5 and 6 respectively.

2. Related works

Determining the emotional content of music is not only through signal processing and machine learning, but also requiring an understanding of auditory perception, psychology, and music theory [6]. The related works are given based on two aspects: the features related to the music moods or emotions and the classification schemes.

The content-based acoustic features are categorized as timbre texture features, rhythmic content features, and pitch content features, harmony and temporal features [7]. The timbre features include cepstral features such as mel frequency cepstral coefficients (MFCC). Rhythmic features contain the regularity of the rhythm, the beat and tempo information, such as beat per minute (BPM) and rhythm histogram. Pitch features deal with the frequency information of the music signals. Harmony features include chromogram and temporal features include temporal centroid. In [8] author proposed a music classification based on spectral and cepstral features (fusion) achieves higher classification accuracy (86.83%) than the winner of the ISMIR 2004 music genre classification contest with a classification accuracy of 84.07%.

In [9], MFCC was used for their content based music similarity search and emotion detection based on SVM classifier and achieved the performance about 70.0%. In [10] Daubechies wavelet coefficient histograms (DWCH) was introduced for music feature extraction in music information retrieval. The histograms are computed from the coefficients of the db8 Daubechies wavelet filter applied to 3 s of music. A comparative study of sound features and classification algorithms on a dataset compiled by Tzanetakis shows that combining DWCH with timbral features (MFCC and FFT), with the use of multiclass extensions of support vector machine, achieves approximately 80.0% of accuracy. The chromagram is a well-established method for estimating the western pitch class components within a short time-interval [11]. Jonathan Foote's music retrieval database to do experiments and final results show that the retrieval accuracy can reach over 96.7% using chromagram as features even when the signal-to-noise ratio is 0 dB [12]. In [13] beat tracking and tempo analysis are measured for identifying the songs from large database and reported an accuracy of 51.0% and mean reciprocal ranking (MRR) of 0.53%. Overgoor [14] gives a nice overview of which beat detection algorithm can be used in what application.

3. Feature extraction for music emotion recognition

3.1. Mel frequency cepstral coefficients

Spectrum features are features computed from the short time Fourier transform (STFT) of an audio signal [4]. MFCC has proven to be one of the most successful spectrum features in speech and music related recognition tasks. The mel-cepstrum exploits auditory principles, as well as the decorrelating property of the cepstrum. The MFCC features for a segment of music file are computed using the following procedures [15]:

(1) The music waveform is first windowed with analysis window and the discrete short time Fourier transform (STFT) is computed.
(2) The magnitude is then weighted by a series of filter frequency responses whose center frequencies and bandwidths roughly match those of the auditory critical band filters. These filters follow the mel scale whereby band edges and center frequencies of the filters are linear for low frequency and logarithmically increase with increasing frequency as shown in these filters are mel-scale filters and collectively a mel-scale filter bank. This filter bank, with triangularly shaped frequency responses, is a rough approximation to actual auditory critical band filters covering a 4000 Hz range.
(3) The log energy in the STFT weighted by each mel-scale filter frequency response is computed.
(4) Finally discrete cosine transform (DCT) is applied to the filter bank output to produce the cepstral coefficients. This is represented by

$$Cepstrum(frame) = IDFT(log(|DFT(frame)|)) \tag{1}$$

For acoustic feature extraction, the music signal is divided into frames of 20 ms, with a shift of 10 ms. In this work 39th order MFCC is used to capture the static and dynamic features of music signal. It contains 13th order static coefficients, 13th order delta coefficients and 13th order acceleration (delta–delta) coefficients results in a 39 dimensional MFCC feature vector for each frame.

3.2. Residual phase

The basic idea behind the linear predictive analysis [16] is that a given music sample at n, $s(n)$ can be estimated as a linear combination of the past p music samples. The predicted samples $\dot{s}(n)$ is given by

$$\dot{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

where p is the order of prediction and the coefficients $\{a_k\}$ $k = 1, 2, \ldots, p$ is the set of linear prediction coefficients (LPCs). The prediction error $e(n)$ is defined as the difference between the actual value $s(n)$ and the predicted value and is given by

$$e(n) = s(n) - \dot{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{3}$$

The LPCs are obtained by minimizing the mean squared prediction error over the analysis frame. This error $e(n)$ is called the linear prediction (LP) residual $r(n)$ of the music signal. The LP residual contains much information about the excitation source. The phase of the analytic signal derived from the LP residual contains better speaker specific information [17]. In this work residual phase is used for extracting the emotion specific information present in the excitation source signal. The analytic signal $r_a(n)$ corresponding to $r(n)$ is given by

$$r_a(n) = r(n) + jr_h(n) \tag{4}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_a(n) = IFT[R_h(\omega)] \tag{5}$$

where

$$R_h(\omega) = \begin{cases} -jR(\omega), & 0 \leqslant \omega < \pi \\ jR(\omega), & -\pi \leqslant \omega < 0 \end{cases} \tag{6}$$

Here $R(\omega)$ is the Fourier transform of $r(n)$ and IFT denotes the inverse Fourier transform. The magnitude of the analytic signal $r_a(n)$ is given by

$$h_e(n) = |r_a(n)| = \sqrt{r^2(n) + r_h^2(n)} \tag{7}$$

and the cosine of the phase of the analytic signal $r_a(n)$ is given by

$$\cos(\theta(n)) = \frac{\mathrm{Re}(r_a(n))}{|r_a(n)|} = \frac{r(n)}{h_e(n)} \tag{8}$$

In [5] it is demonstrated that the residual phase signal contains speaker specific information that is complementary to the MFCC features. The residual phase is defined as the cosine of the phase function of the analytic signal derived from the linear prediction (LP) residual of a music signal. In this work, the emotion specific information from the residual phase features is analyzed using AANN, SVM and RBFNN. The recognition rates from the models, show that the emotion specific information is present in the music signal. The residual phase extraction is described in Fig. 1.

In this work music signal sampled at 44.1 kHz and the LP order 16 is used for deriving the LP residual. The LP residual is extracted from emotional music signal by pre-emphasizing the input music data using first-order digital filter and 20 ms frame size with an overlap of 50 percent between adjacent frames. The highest Hilbert envelope around 40 samples for each frame is extracted. A segment of music file from anger emotion, its LP residual, the Hilbert transform of the LP residual, the Hilbert envelope, and residual phase are shown in Fig. 2 and the 40 dimensional residual phase features extracted from music signal for a single frame for five emotions are shown in Fig. 3.

4. Techniques for music emotion recognition

4.1. Autoassociative neural network (AANN)

Neural network models can be trained to capture the nonlinear information present in the signal. In particular AANN models are basically feed forward neural network (FFNN) models which try to map an input vector onto itself. It consists of an input layer, an output layer and one or more hidden layers [18,19]. The number of units in the input and output layers is equal to the size of the input vectors. The number of nodes in the middle hidden layer is less than the number of units in the input or output layers. The middle layer is also the dimension compression hidden layer.

The activation function of the units in the input and output layers is linear (L), whereas the activation function of the units in hidden layer can be either linear or nonlinear (N). Studies on three layer AANN models show that the nonlinear activation function at the hidden units clusters the input data in a linear subspace [20]. Theoretically, it was shown that the weights of the network will produce small errors only for a set of points around the training data. When the constraints of the network are relaxed in terms of layers, the network is able to cluster the input data in the nonlinear subspace. Hence a five layer AANN model is used to capture the distribution of the feature vectors in this study.

Emotion recognition using AANN model is basically a two stage process namely, (i). Training phase and (ii). Testing phase. During training phase, the weights of the network are adjusted to minimize the mean square error obtained for each feature vector. If the adjustment of weights is done for all feature vectors once, then the network is said to be trained for one epoch. During testing phase (evaluation), the features extracted from the test data are given to the trained AANN model to find its match.

4.2. Support vector machines (SVM)

The support vector machine [21] is a useful statistic machine learning technique that has been successfully applied in the pattern recognition tasks [5,22]. If the data are linearly non-separable but nonlinearly separable, the nonlinear support vector classifier will be applied. The basic idea is to transform input vectors into a high-dimensional feature space using a nonlinear transformation ϕ and then to do a linear separation in feature space as shown in Fig. 4.

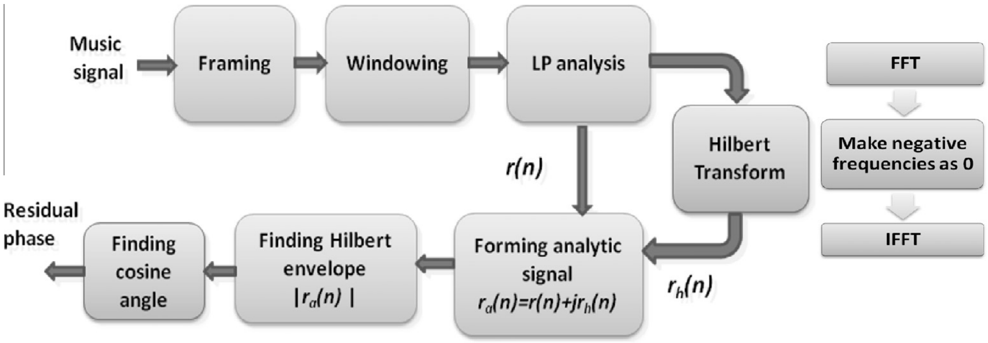

Music signal
Framing
Windowing
LP analysis
Hilbert Transform
FFT
Make negative frequencies as 0
IFFT
r(n)
Residual phase
Finding cosine angle
Finding Hilbert envelope |r_a(n) |
Forming analytic signal r_a(n)=r(n)+jr_h(n)
r_h(n)

Figure 1 Extraction of residual phase from music signal.

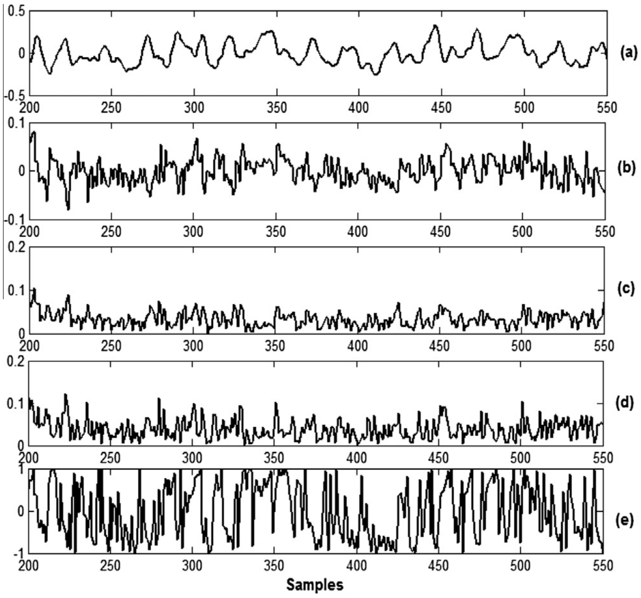


Figure 2 (a) Music signal. (b) LP residual signal. (c) Hilbert transform of residual signal. (d) Hilbert envelope. (e) Residual phase.

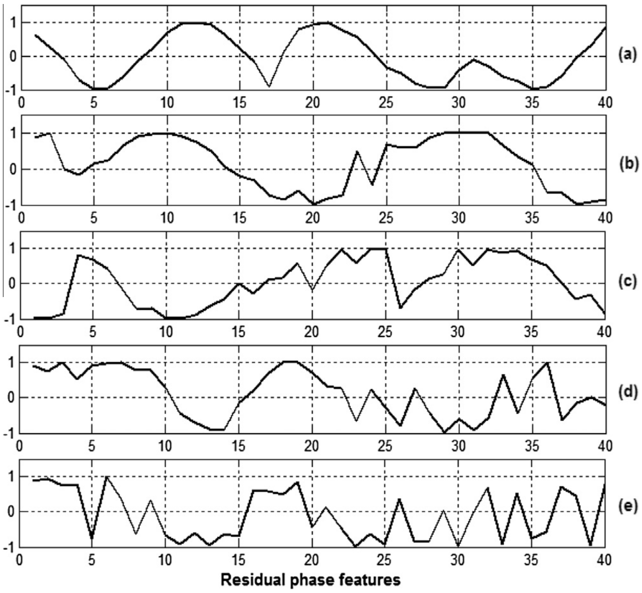


Figure 3 40 dimensional residual phase features for (a) Anger. (b) Fear. (c) Happy. (d) Neutral. (e) Sad emotions.

A nonlinear support vector classifier implementing the optimal separating hyperplane in the feature space with a kernel function $K(x, x_i)$ is given by

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N_{sv}} \alpha_i y_i K(x, x_i) + b\right) \tag{9}$$

where $x \in R^{N_f}$ is the N_f dimensional input vector, $\{X_i\}_{i=1}^{N_{SV}}$ are the support vectors that are obtained from the training set by an optimization process, N_{SV} are the total number of support vectors and $y_i \in \{-1, 1\}$ are their associated class labels. α_i and b are the model parameters. K (.,.) is the kernel function defining the inner product between x and x_i and is given by

$$K(x, x_i) = \Phi^T(x)\Phi(x_i) = (x, x_i) \tag{10}$$

The SVM has two layers. During the learning process, the first layer selects the basis $K(x, x_i)$ from the given set of bases defined by the kernel; the second layer constructs a linear function in this space. This is completely equivalent to constructing the optimal hyperplane in the corresponding feature space. SVM based supervised technique was proposed in [22] by labeling the music data of one emotion as (+1) class and the remaining other emotions as (−1) class for training an SVM hyperplane and then classified each emotion as (+1) or (−1).

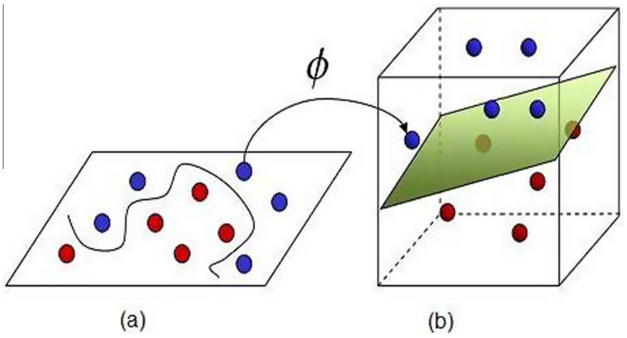


Figure 4 SVM kernel function $\phi(x)$ maps 2-Dimensional input space to higher 3-Dimensional feature space. (a) Nonlinear problem. (b) Linear problem.

4.3. Radial basis function neural network (RBFNN)

The radial basis function neural network (RBFNN) is a feed forward architecture with an input layer, a hidden layer and an output layer [23,25]. RBFNN solves the classification problem by transforming the input space into a high dimensional space in a nonlinear manner [24]. The RBFNN uses Gaussian basis function as the radial basis function in the hidden layer [25]. The significance of the radial basis function is that the output is a function of radial distance. The Gaussian basis function for the ith hidden unit for an input vector x_j is given by

$$g_i(x_j) = \exp\left(\frac{-\|x_j - \mu_i\|^2}{2\sigma_i^2}\right) \tag{11}$$

where m_i and σ_i^2 are the mean and standard deviation of ith cluster. A clustering algorithm such as k-means clustering is used to cluster the training vectors into N_h clusters, where N_h is the number of units in the hidden layer. k-means clustering algorithm is employed to determine the centers. The algorithm is composed of the following steps:

1. Randomly initialize the samples to k means (clusters) $\mu_1, \ldots, \mu_\mathrm{k}$
2. Classify n samples according to nearest μ_k.
3. Recompute μ_k.
4. Repeat the steps 2 and 3 until no change in μ_k.

5. Experimental results

5.1. Music database

In this work MER categorizes emotions into a number of classes such as anger, fear, happy, neutral and sad. The database used in this work is perceived emotion of music signals with 44.1 kHz sampling frequency and 16 bit monophonic PCM wave format. For each emotion 20 music files are collected of 10 s duration from various websites in .wav format, totally 100 files are collected for this work. Ten music files are randomly selected for training and the remaining ten music files are used for testing.

During testing phase, each 2 s duration of music signal is used for testing the model. The experiment is repeated once to generate 500 test cases. The performance of emotion recognition is assessed in terms of equal error rate (EER) and in terms of accuracy. For acoustic feature extraction, the music signal is divided into frames of 20 ms, with a shift of 10 ms.

5.2. Performance measures

The performance of emotion recognition is assessed in terms of equal error rate (EER) and accuracy or recognition rate (RR).

5.2.1. Equal error rate (EER)

The EER is the result obtained by adjusting the system threshold such that FAR and FRR are equal. A false acceptance rate (FAR) is defined as the rate at which an emotion model gives high confidence score when compared to the test emotion model. A false rejection rate (FRR) is defined as the rate at which the respective model for the test emotion gives low confidence score when compared to one or more other emotion models. Generally, FAR and FRR depend on the system threshold t. The error at which the two curves intersect represents the EER.

5.2.2. Accuracy

Accuracy or recognition rate is defined as

$$\text{Accuracy} = \frac{\text{Number of correctly predicted testing}}{\text{Total number of testing}}$$

The emotion recognition performance using MFCC, residual phase and combined MFCC and residual phase features with AANN, SVM and RBFNN is shown as a consolidated table in Table 2.

5.3. MER using AANN

5.3.1. MFCC

During the training phase a single AANN is trained separately for each emotion. The AANN structure 39Lu 50Nu 16Nu 50Nu 39Lu achieves an optimal performance in training and testing the MFCC features for each emotion. The structure is obtained from the experimental studies. The performance of emotion recognition is obtained by varying the second (expansion layer) and third layer (compression layer) of AANN model. The effect of changing the neurons in the compression layer N_c on the performance of recognizing emotional music is studied. There is no major change in the performance if N_c is between 15 and 20 and the emotion recognition performance (ERP) decreases if it less than 15 or more than 20. The results are shown in Table 1. Similarly the performance is obtained by varying the number of units in the second layer (expansion layer) keeping the number of units in the compression layer to 16.

The MFCC feature vectors are given as both input and output. The weights are adjusted to transform input feature vector into the output. The number of epochs needed depends upon the training error. In this work the network is trained for 500 epochs, but there is no major change in training error after 400 epochs and it is shown in Fig. 5.

During testing phase the MFCC features extracted from test samples are given as input to the AANN and the output is computed. The anger emotion test sample of 5 s duration is tested against all the five models and the result is shown in Fig. 6. It shows that anger emotion model gives high confidence scores compared to other models. The output of each model is compared with the input to compute the normalized squared error. The normalized squared error (e) for the feature vector y is given by, $e = \frac{\|y-o\|^2}{\|y\|^2}$ where o the output vector is given by the model. The error (e) is transformed into a confidence score (c) using $c = \exp(-e)$. The average confidence score is calculated for each model.

The category of the emotion is decided based on the highest confidence score. The performance of the music emotion

Table 1 Emotion recognition performance in terms of number of units in compression layer.

	$N_c = 10$	$N_c = 15$	$N_c = 20$	$N_c = 25$
ERP (%)	88.0	92.0	91.3	89.0

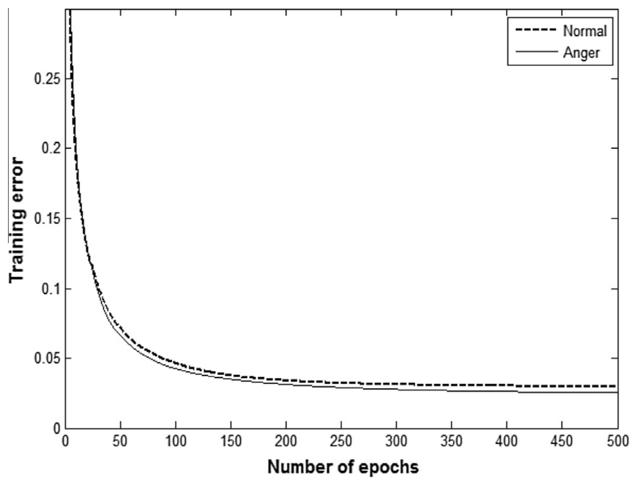


Figure 5 AANN training error Vs number of epochs for MER.

recognition using MFCC features is evaluated in terms of accuracy and equal error rate. Accuracy is shown in terms of confusion matrix.

The average emotion recognition performance for the five emotions is about 92.0%. An equal error rate (EER) of 8.0% is obtained by evaluating the performance in terms of FAR and FRR at threshold 0.69 and is shown in Fig. 7.

5.3.2. Residual phase

The AANN structure 40Lu 60Nu 20Nu 60Nu 40Lu achieves an optimal performance in training and testing the residual phase features for each emotion. The residual phase feature vectors are given as both input and output. The weights are adjusted to transform input feature vector into the output. The number of epochs needed depends upon the training error.

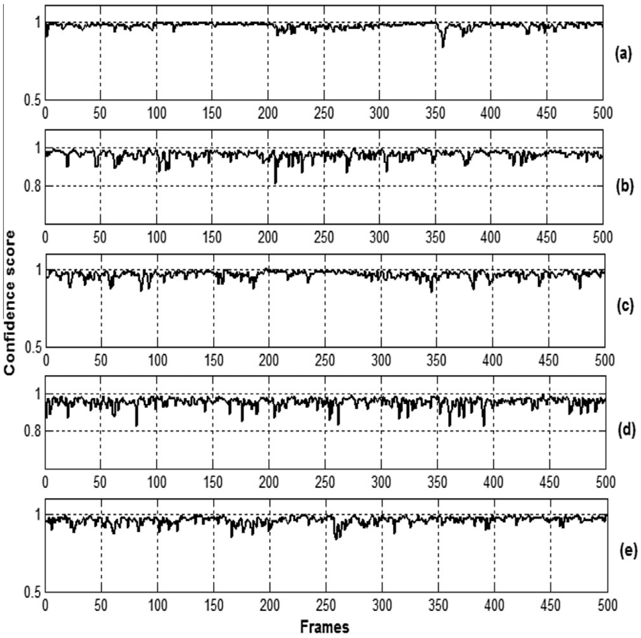


Figure 6 MFCC features extracted from anger music emotion are tested against all the five AANNs. (a) Anger. (b) Fear. (c) Happy. (d) Neutral. (e) Sad.

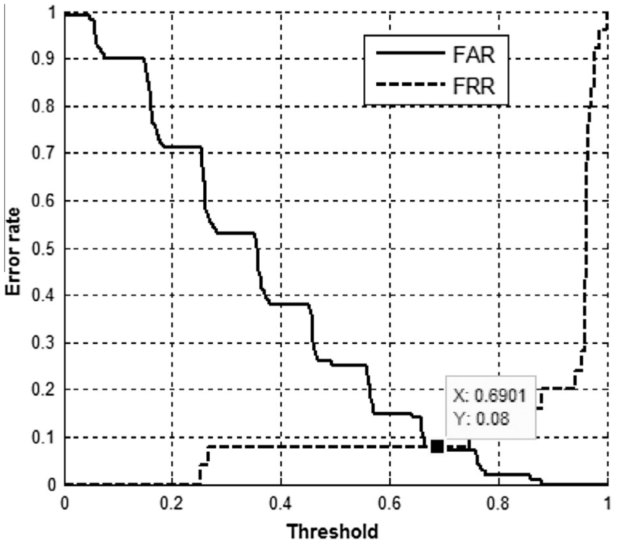


Figure 7 FAR and FRR curves using MFCC features and AANN.

During testing phase the residual phase features of test samples are given as input to the AANN and the output is computed. The output of each model is compared with the input to compute the normalized squared error. The error (e) is transformed into a confidence score (c) using $c = \exp(-e)$. The average confidence score is calculated for each model.

In Fig. 8 the test samples of 5 s duration of anger emotion music signal are tested against all other AANN models. By evaluating the performance in terms of FAR and FRR, at threshold of 0.63, an equal error rate (EER) of 40.0% is obtained and it is shown in Fig. 9(a). The average emotion recognition performance is about 60.0%.

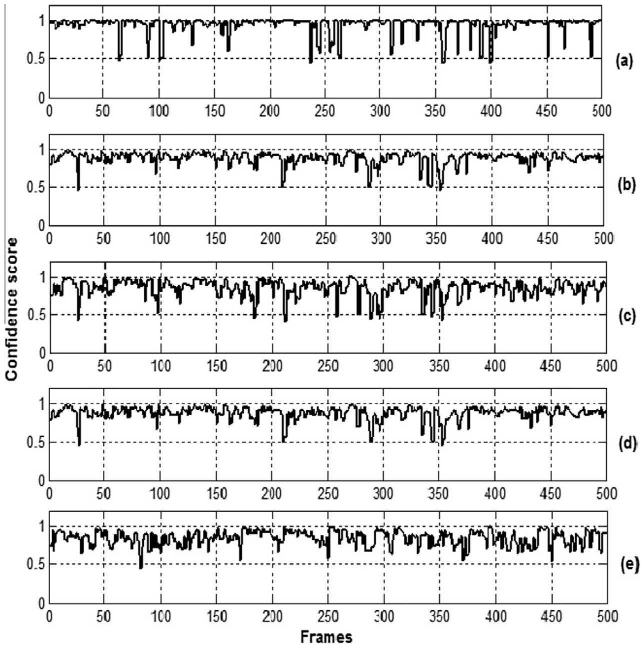


Figure 8 Residual phase features extracted from anger emotion are tested against all the five AANNs. (a) Anger. (b) Fear. (c) Happy. (d) Neutral. (e) Sad.

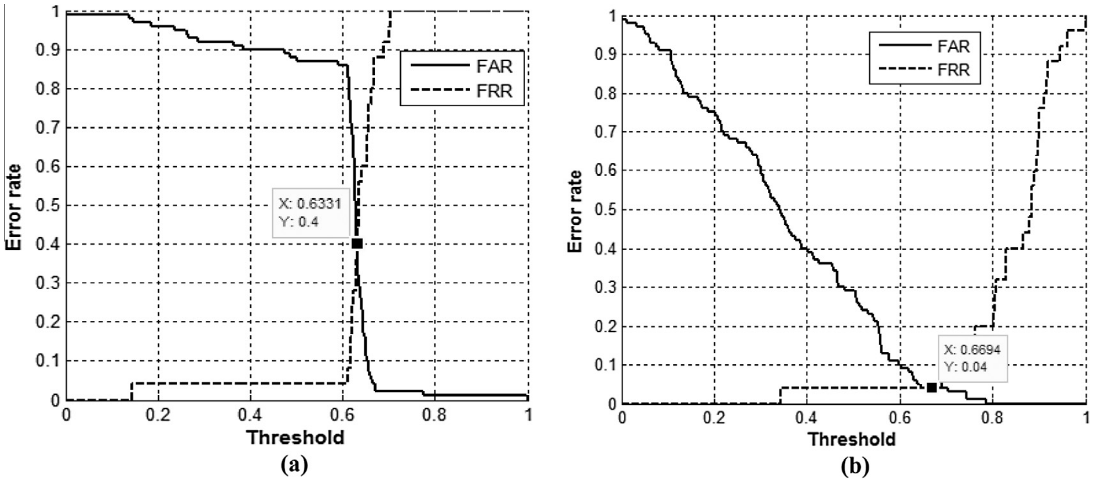


Figure 9 FAR and FRR curves for (a) residual phase and (b) combined features using AANN.

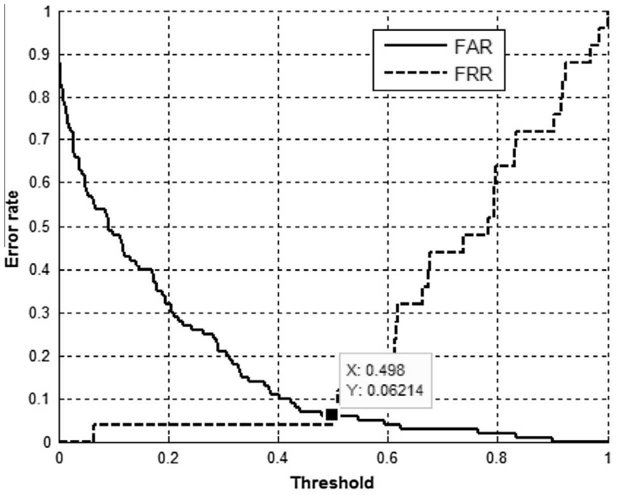


Figure 10 FAR and FRR curves using MFCC features and SVM.

5.3.3. Combined features

The excitation (residual phase) and spectral features (MFCC) are combined because of its complementary nature. The two features are combined at score level using

$$c = ws_1 + (1 - w)s_2 \tag{12}$$

where s_1 and s_2 are the scores or output of the model for MFCC and residual phase features, respectively and w is the weight, $0 \leqslant w \leqslant 1$. In this work it is observed that for the weight 0.6 ($w = 0.6$) an EER of about 4.0% is achieved using AANN by combining the features and it is shown in Fig. 9(b).

The overall recognition performance for the combined MFCC and residual phase features is about 96.0%. The emotion recognition performance of the combined system is increased than individual system due to the complementary nature of residual phase with MFCC and also it shows that residual phase contained emotion specific information.

5.4. MER using SVM

SVM is trained to distinguish the acoustic features (class label ’+1’) of an emotion and all other acoustic features (class label

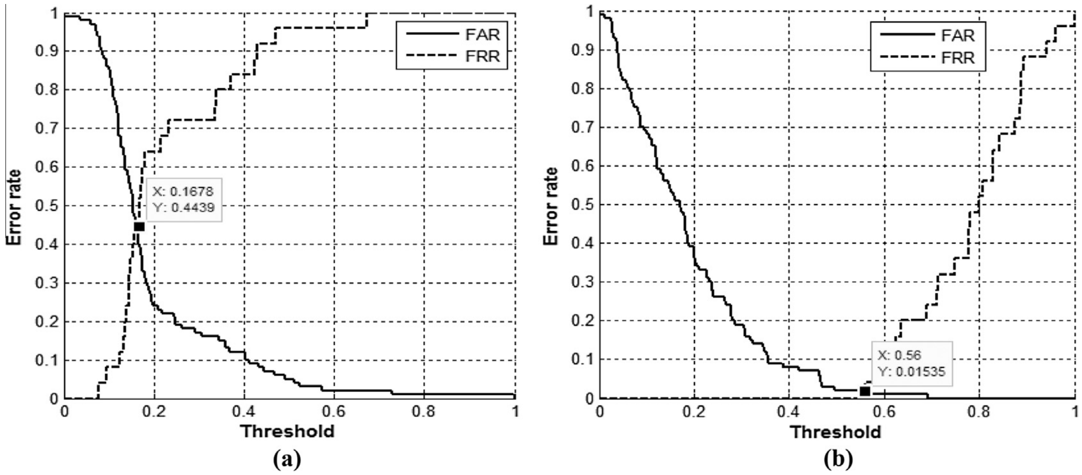


Figure 11 FAR and FRR curves for (a) residual phase and (b) combined features using SVM.

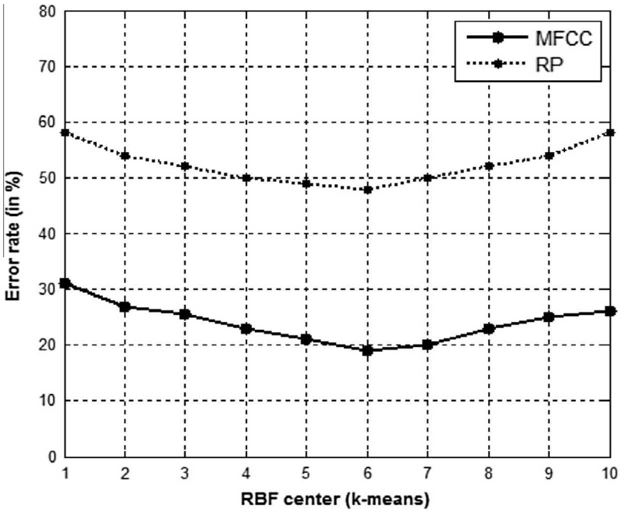


Figure 12 Music emotion recognition error rate with respect to number of mean centers for each emotion.

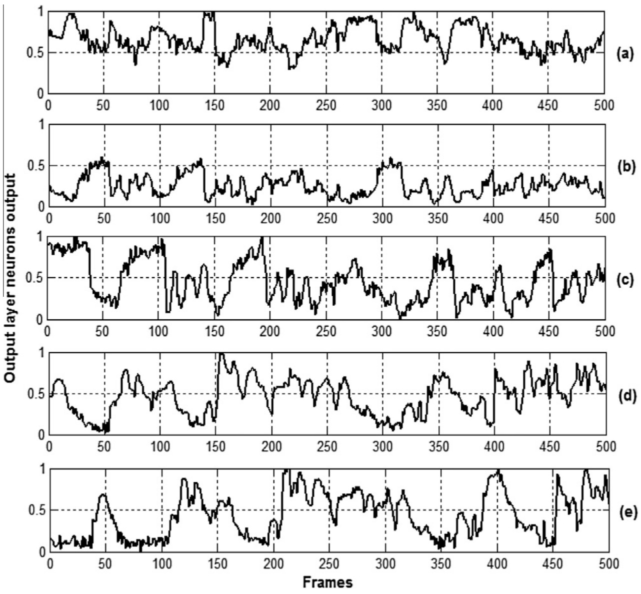


Figure 13 Testing MFCC features of anger music emotion using RBFNN. (a) Anger. (b) Fear. (c) Happy. (d) Neutral. (e) Sad (five outputs of RBFNN).

'−1') in the training set. The MFCC features extracted from each frame are of 39 dimensional and residual phase feature extracted from each frame is of 40 dimensional is given as input to SVM model. For each music emotion a single SVM model is created. All music emotions belong to same category is merged into a single category. Totally five SVM models are created. In the training phase, SVM is trained to distinguish acoustic features of one emotion (+1) and acoustic features of all other emotions (−1). The SVM is trained with Gaussian, polynomial and sigmoidal kernel functions, among which Gaussian kernel gives better emotion recognition performance.

5.4.1. MFCC

During testing the MFCC features extracted from the test utterances are given as input to the SVM model. The distance between each of the feature vectors and the SVM hyperplane is obtained. The number of positive score is calculated for each emotion model. Based on the score, the category of emotion is decided. This process is repeated for all the test utterances of emotions. The average emotion recognition performance of 94.0% is obtained as a result

By varying the threshold FAR and FRR are obtained from the scores. The EER of 6.0% is obtained at threshold 0.49 and it is shown in Fig. 10. A false acceptance rate (FAR) is defined as the rate at which an emotion model gives high score when compared to the test emotion model. A false rejection rate (FRR) is defined as the rate at which the respective model for the test emotion gives low score when compared to other emotion models.

5.4.2. Residual phase

In this work LP order 16 is used for deriving the LP residual. The LP residual is extracted from emotional music signal by pre-emphasizing the input music data using first-order digital filter and 20 ms frame size with an overlap of 50 percent between adjacent frames. The highest Hilbert envelopes around 40 samples for each frame are extracted. A single SVM is developed for one emotion. Five SVM models are created for residual phase features. The average emotion recognition performance of about 56.0% is obtained after testing all the test utterances.

By evaluating the performance in terms of FAR and FRR from the scores, EER of 44.0% is obtained at threshold 0.16 and it is shown in Fig. 11(a).

5.4.3. Combined features

It is observed that an EER of about 1.0% for the combined features by evaluating the FAR and FRR values and it is shown in Fig. 11(b). The overall emotion recognition system is obtained by combining the evidence of MFCC and residual phase features and their average performance is about 99.0% is obtained.

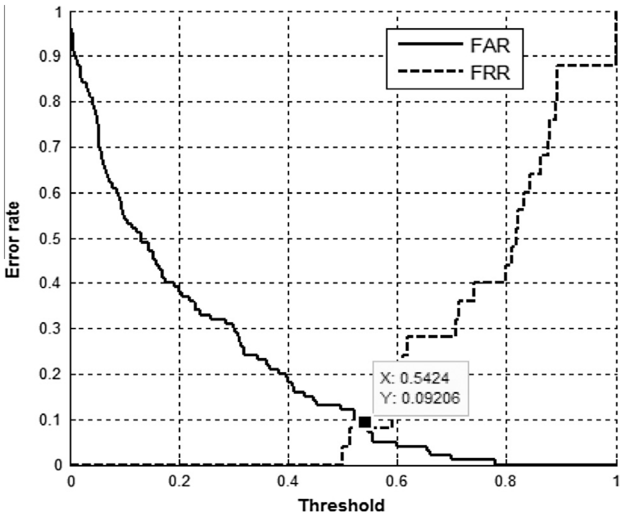


Figure 14 FAR and FRR curves using MFCC features and RBFNN.

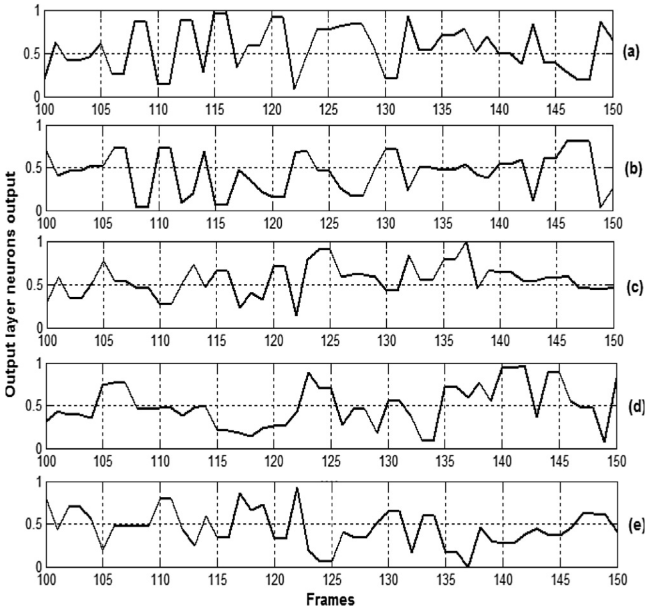


Figure 15 Testing residual phase features of anger music signal using RBFNN. (a) Anger. (b) Fear. (c) Happy. (d) Neutral. (e) Sad (five outputs of RBFNN).

5.5. MER using RBFNN

The music emotion recognition system is evaluated in terms of radial basis function neural network model. Five types of emotions namely anger, fear, happy, neutral and sad are studied in this work. All five emotions are recorded for 10 s at 44,100 samples per second. For music emotion recognition, different music emotion signals are analyzed by dividing it into 20 ms with a shift of 10 ms. A 39 dimensional MFCC feature vector and 40 dimensional residual phase feature vectors are extracted from each frame of the music signal. All the feature vectors (MFCC/residual phase) belong to a same category are merged into a single file. The file is given as input to the RBFNN model. k-means algorithm is used to find the RBF centers and the weights are calculated using least square algorithm. The value of k varies from 1 to 10 in this work for each emotion. The system gives an optimal performance for $k = 6$ for MFCC and residual phase feature as shown in Fig. 12. For training 50,000 feature vectors are extracted from input file (10,000 × 5). The size of the G matrix is 50,000 × 31 (last column of G matrix is bias value of RBFNN). The size of the weight matrix is 31 × 5 is calculated using least square algorithm.

5.5.1. MFCC

During testing phase the MFCC features extracted from test samples are given as input to the RBFNN and output is computed. The anger emotion test sample of 5 s duration is tested and the result is shown in Fig. 13.

It shows that anger emotion is recognized better than other emotional music samples. By evaluating FAR and FRR for various values of threshold, EER is obtained. In Fig. 14 an EER of about 9.0% at threshold of 0.54 is obtained.

5.5.2. Residual phase

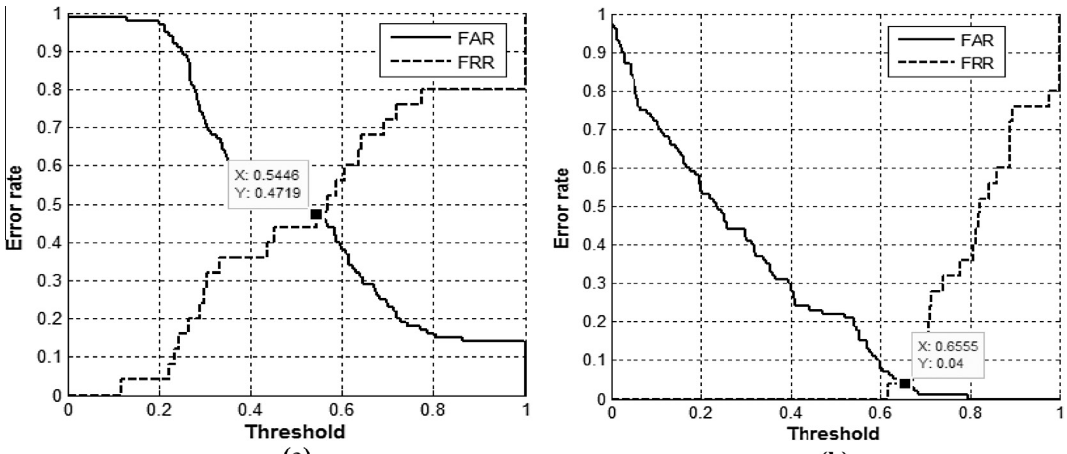
Other than the spectral features, excitation source feature is used to evaluate the performance of RBFNN. The weight matrix of size 31 × 5 is calculated using the least square algorithm. For evaluating the performance of the system the test data of 5 s music signal is used. The 500 frames anger music signal is tested against other emotions using RBFNN and it is shown in Fig. 15.

The overall recognition is reported about 53.0% and another performance measure EER of about 47.0% is obtained from FAR and FRR by varying the threshold and it is shown in Fig. 16(a).

5.5.3. Combined features

For evaluating the RBFNN model, it is observed that an EER of about 4.0% by evaluating the FAR and FRR values is shown in Fig. 16(b). The overall recognition performance of 95.0% is obtained by combining the evidence of MFCC and residual phase features.

The consolidated music emotion recognition performance using AANN, SVM and RBFNN is shown in Table 2. The experiment results show that MFCC with residual phase features gives an optimal performance of 99.0% and EER of 1.0% using SVM, when compared to other models. It indicates that residual phase contains emotion specific information in music and the combined MFCC and residual phase increases the rate of recognition.



(a) **(b)**

Figure 16 FAR and FRR curves for (a) residual phase and (b) combined features using RBFNN.

Table 2 Emotion recognition performance (in%) using AANN, SVM and RBFNN.

Model	Measures	Features		
		MFCC	RP	MFCC + RP
AANN	RR	92.0	60.0	96.0
	ERR	8.0	40.0	4.0
SVM	RR	94.0	56.0	99.0
	EER	6.0	44.0	1.0
RBFNN	RR	92.0	53.0	95.0
	EER	9.0	47.0	4.0

6. Conclusion

In this work, methods were proposed for combined mel frequency cepstral coefficients (MFCC) and residual phase (RP) features for emotion recognition in music (audio). Emotion recognition in music considers the emotions namely anger, fear, happy, neutral and sad. For music emotion recognition, MFCC (spectral features) and residual phase features (excitation source) were extracted from the music, and were used to create models for each emotion using AANN, SVM and RBFNN. The MFCC and residual phase features were combined, due to the complementary nature of emotion specific information present in the residual phase in comparison with the information present in the conventional MFCC which increase the emotion recognition performance. AANN is used to capture the distribution of the acoustic feature vectors and is very effective in recognition of emotion in music with an accuracy of 96.0%. A nonlinear support vector machine learning algorithm is applied to obtain the optimal class boundary between the various emotions namely anger, fear, happy, neutral and sad and obtained the recognition accuracy of 99.0%. The distribution of the acoustic features of music was captured using a Gaussian distribution in RBFNN shows the recognition accuracy of 95.0%. SVM shows the highest performance for MER with MFCC and residual phase features.

References

[1] Yang Yi-Hsuan, Lin Yu-Ching, Ya-Fan Su, Homer H. A regression approach to music emotion recognition. IEEE Trans Audio Speech Lang Process 2008;16(2):448–57.

[2] Thayer RE. The biopsychology of mood and arousal. New York: Oxford University Press; 1989.

[3] Yang Yi-Hsuan, Chen Homer H. Music emotion recognition. CRC Press; 2011.

[4] Zhu Bin, Zhang Kejun. Music emotion recognition system based on improved GA-BP. Comput Des Appl 2010:409–11.

[5] Koolagudi Shashidhar G, Sreenivasa Rao K. Emotion recognition from speech: a review. Int'l J Speech Technol 2012;15:99–117.

[6] Kim Youngmoo E. Music emotion recognition: a state of the art review. In: Eleventh international society for music information retrieval conference; 2010. p. 255–66.

[7] Kaminskas Marius, Ricci Francesco. Contextual music information retrieval and recommendation: state of the art and challenges. Comput Sci Rev 2012;6:89–119.

[8] Lee Chang-Hsing, Shih Jau-Ling. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. IEEE Multimedia 2009;11(4).

[9] Li Tao, Ogihara Mitsunori. Content-based music similarity search and emotion detection. In: Int'l conference on acoustics, speech and signal processing; 2004. p. 705–8.

[10] Li Tao, Ogihara Mitsunori. Toward intelligent music information retrieval. IEEE Multimedia 2008;8(3).

[11] Schmidt EM, Turnbull D, Kim YE. Feature selection for content-based, time varying musical regression. In: Int'l conference on multimedia information retrival; 2010. p. 267–74.

[12] Yu Xiaoging, Zhang Jing, Yang Wei. An audio retrieval method based on chroma-gram and distance metrics. In: Int'l conf on audio language and image processing; 2010. p. 23–5.

[13] Ellis D, Poliner PW, Graham E. Identifying cover songs with chroma features and dynamic programming beat tracking. In: IEEE conf on acoustic, speech, and signal processing, vol. 4; 2007. p. 1429–32.

[14] Overgoor JM. An evaluation method for audio beat detectors. In: Fourth twente student conference on IT; 2006. p. 23–9.

[15] Jothilakshmi S, Ramalingam Vennila, Palanivel S. Unsupervised speaker segmentation with residual phase and MFCC features. Expert Syst Appl 2009;36(6):9799–804.

[16] Rabiner Lawrence, Schafer Ronald. Theory and applications of digital speech processing. England: Pearson Education; 2010.

[17] Sri Rama Murty K, Yegnanarayana B. Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process Lett 2006;13(1):52–5.

[18] Palanivel S. Person authentication using speech, face and visual speech, Ph.D thesis, Department of Computer Science and Engg. Madras: Indian Institute of Technology; 2004.

[19] Yegnanarayana B, Kishore SP. AANN: an alternative to GMM for pattern recognition. Neural Networks 2002;15:459–569.

[20] Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks. Neural Comput Appl 2000;9:290–6.

[21] Vapnik V. Statistical learning theory. New York: John Wiley and Sons; 1998.

[22] Shen Peipei, Pan Yixiong, Shen Liping. Speech emotion recognition using support vector machine. Int'l J Smart Home 2012;6(2).

[23] Haykin S. Neural networks, a comprehensive foundation. Singapore: Pearson Education; 2001.

[24] Yegnanarayana B. Artificial neural networks. New-Delhi: Prentice-Hall of India; 1999.

[25] Koh LH, Ranganath S, Venkatesh YV. An integrated automatic face detection and recognition system. Pattern Recogn 2002;35:1259–73.