Circuits and Systems, 2016, 7, 255-279
Published Online April 2016 in SciRes. http://www.scirp.org/journal/cs
http://dx.doi.org/10.4236/cs.2016.74024

An Efficient Approach for Segmentation, Feature Extraction and Classification of Audio Signals

Muthumari Arumugam[1*], Mala Kaliappan[2]

[1]Department of Computer Science and Engineering, University College of Engineering, Ramanathapuram, India
[2]Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India
Email: *muthu_ru@yahoo.com

Received 14 March 2016; accepted 23 April 2016; published 26 April 2016

Abstract

Due to the presence of non-stationarities and discontinuities in the audio signal, segmentation and classification of audio signal is a really challenging task. Automatic music classification and annotation is still considered as a challenging task due to the difficulty of extracting and selecting the optimal audio features. Hence, this paper proposes an efficient approach for segmentation, feature extraction and classification of audio signals. Enhanced Mel Frequency Cepstral Coefficient (EMFCC)-Enhanced Power Normalized Cepstral Coefficients (EPNCC) based feature extraction is applied for the extraction of features from the audio signal. Then, multi-level classification is done to classify the audio signal as a musical or non-musical signal. The proposed approach achieves better performance in terms of precision, Normalized Mutual Information (NMI), F-score and entropy. The PNN classifier shows high False Rejection Rate (FRR), False Acceptance Rate (FAR), Genuine Acceptance rate (GAR), sensitivity, specificity and accuracy with respect to the number of classes.

Keywords

Audio Signal, Enhanced Mel Frequency Cepstral Coefficient (EMFCC), Enhanced Power Normalized Cepstral Coefficients (EPNCC), Probabilistic Neural Network (PNN) Classifier

1. Introduction

In this paper, an efficient approach for segmentation, EMFCC-EPNCC based feature extraction and PNN-based classification of audio signal is proposed. The background section presents a brief overview of the existing audio

[*] Corresponding author.

How to cite this paper: Muthumari, A. and Mala, K. (2016) An Efficient Approach for Segmentation, Feature Extraction and Classification of Audio Signals. *Circuits and Systems*, **7**, 255-279. http://dx.doi.org/10.4236/cs.2016.74024

segmentation, feature extraction and classification techniques, along with its drawbacks. The proposed work is illustrated in the contribution section.

1.1. Background

Segmentation and classification [1] of audio signals play a major role in the audio signal processing application. Audio segmentation is an essential preprocessing step for audio signal processing utilized in various applications such as medical applications, broadcast applications, etc. During the recent years, there have been many researches on the automatic audio segmentation and classification by using various features and techniques. The segmentation algorithms are classified into decoder-based [2], model-based [3] and metric-based [4] [5] segmentation algorithms.

Feature extraction techniques are classified as temporal and spectral feature extraction techniques [6] [7]. Temporal feature extraction [7] utilizes the waveform of the audio signal for analysis process. Spectral feature extraction [7] utilizes the spectral representation of the audio signal for analysis process. Accurate classification of the musical and non-musical segments in the audio signal is the most common problem [8]. Classification of audio signal has become an interesting topic with respect to the increase in the growth and availability of audio database. Automatic classification of audio information has gained more popularity for organizing the large number of audio files in the database.

1.2. Drawbacks of Existing Techniques and Merits of Proposed Work

However, the conventional audio segmentation techniques are usually quite simple and do not consider all possible scenarios. The decoder-based segmentation approaches only place the boundaries at the silence locations. This do not have any connection with the acoustic changes in the audio data. The model-based approaches do not generalize to the data conditions, as the models are not compatible with the new data conditions. The metric-based approaches generally require a threshold to make decisions. These thresholds are set empirically and require an additional development data. Hence, there arises a need in the development of efficient segmentation technique. The computational complexity of the traditional feature extraction approaches is increased with respect to the increase in the number of audio signals. The traditional classification techniques applied directly on the feature-vectors yielded poor results. Therefore, classification of audio signal is done without depending on the feature vectors. However, the existing audio classification systems do not represent the perceptual similarity of audio signals, as they mainly depend on the single similarity measure.

To overcome the challenges in the existing techniques, an efficient approach for segmentation, feature extraction and classification of audio signals is introduced in this paper. The proposed work involves the combination of new objective function of peak and pitch extraction and EPNCC and EMFCC based feature extraction. The presence of silence and irrelevant frequency details in the audio signal is eliminated. The clear features of the speech signal filtered from other background signals are obtained. PNN-based classification provides a better prediction of the classified label using the probability estimation.

1.3. Contribution of the Proposed Work

This paper proposes an efficient approach for segmentation, feature extraction and classification of audio signals. In our proposed work, mean filtering is utilized for filtering the audio signal. Better reduction in the Gaussian noise is achieved than the traditional filtering techniques. Segmentation of the audio signal is performed using the peak estimation and pitch extraction. Then, the spectral difference in the audio signal pattern is estimated. Feature extraction is performed by using the combination of EMFCC-EPNCC, peak and pitch feature extraction for collecting the testing features of the audio signal. Multi-label and multi-level classification is performed for classifying the audio signal as a musical or non-musical signal. The category of the audio signal is extracted from the classification result. Finally, the proposed approach is compared with existing algorithms. The PNN-based classification approach achieves better performance in terms of sensitivity, specificity, accuracy, FAR, FRR and GAR. The proposed approach achieves high precision, NMI, F-score and entropy.

The remaining sections of the paper are organized as follows: Section II describes about the conventional works related to the audio segmentation and classification process. Section III explains the proposed approach including mean filtering, segmentation, feature extraction and PNN-based classification processes. The performance evaluation result of the proposed approach is illustrated in the Section IV. Section V discusses about the conclusion and future work of the proposed approach.

2. Related Work

This section presents the conventional research works related to the automatic segmentation and classification of audio signals and feature extraction using various techniques and approaches. Haque and Kim proposed a correlation intensive FCM (CIFCM) algorithm for the segmentation and classification of audio data. The audio-cuts were detected efficiently irrespective of the presence of the fading effects in the audio data. The boundaries between different types of sounds were detected and classified into clusters. The conventional FCM approach was outperformed by the proposed CIFCM approach [9]. An automatic segmentation approach combining the SVM classification and audio self-similarity segmentation was introduced for separating the sung clips and supplement clips from the pop music. The heuristic rules were utilized for filtering and integrating the classification result to determine the potential boundaries for the audio segment. The segmentation boundaries were determined accurately by the proposed approach [10]. Lefèvre and Vincent proposed a two level segmentation process by computing numerous features for each audio sequence. Initial classification was performed by using the k-means classifier and segment-related features. Final classification was done by using the Multidimensional Hidden Markov Models and frame-related features [11].

The usage of the audio signals in the identification of bird species was outperformed by using the short audio segments having high amplitude called as pulses. Training of the Support Vector Machine (SVM) classifiers was performed by using a previously labeled database of bird songs. Best results can be obtained by using the automatically obtained pulses and SVM classifier [12]. Dhanalakshmi *et al.* proposed effective algorithms for the automatic classification of the audio clips into several classes. The Auto Associative Neural Network (AANN) model was used to acquire the acoustic feature vector distribution of the classes. The weights of the network were adjusted by using the back propagation learning algorithm, for reducing the mean square error of each feature vector. The Gaussian Mixture Model (GMM) for those classes was trained by using the feature vectors [13]. Haque and Kim proposed an efficient approach for classifying the audio signals into broad categories by using a fuzzy c-means (FCM) algorithm. Different characteristic features of the audio signals were analyzed and an optimal feature vector was selected using an analytical scoring technique. The FCM-based classification scheme was applied on the optimal feature vector to achieve efficient classification performance [14].

Dhanalakshmi *et al.* proposed effective algorithms for the automatic classification of the audio clips into six classes. The audio content was characterized by extracting the acoustic features such as Linear Prediction Cepstral Coefficients (LPCC) and MFCC. A method for indexing the classified audio was proposed by utilizing the k-means clustering algorithm and LPCC features [15]. The spatial distribution of microphone from ad-hoc microphone arrays was utilized for the accurate classification of disturbed signals. The proposed algorithm was evaluated in the simulated reverberant scenarios and multichannel recordings of microphone setup in the real-time environment. The cluster based classification accuracy of the proposed algorithm was found to be high [16]. Bhat *et al.* proposed an automated and efficient method for observing the mood of music or the emotions. The songs were classified according to the mood, based on the Thayer's model. Various different features of the music were analyzed before the classification of the music. From a database of over 100 songs, the western and Indian Hindi film music were classified. The classification efficiency of the proposed method was improved [17].

Gergen and Martin introduced various data combination strategies for the efficient classification of audio signal. The audio classification performance was analyzed based on the simulations and audio recordings. High classification accuracy was achieved [18]. The automatic estimation of audio chord was addressed using stacked generalization of multiple classifiers over Hidden Markov model (HMM) estimators. A new compositional hierarchical model and standard chroma feature vectors was modelled with the HMMs, for estimating the chords in the music recordings. A binary decision tree and SVM were proposed for binding the HMM estimations into a new feature vector. The classification efficiency was improved with the additional stacking of the classifiers [19]. Murthy and Koolagudi [20] employed machine learning algorithms and signal processing techniques to identify the vocal and non-vocal regions of the songs. The characteristics of vocal and non-vocal segments were obtained by using Artificial Neural Networks (ANN). The classification accuracy of the vocal and non-vocal segments was improved. Koolagudi and Krothapalli [21] performed recognition of emotions from speech signal by using the spectral features including LPCC and MFCC. Vowel onset points were used to determine consonant, vowel and transition regions of each syllable. The emotions in the speech signal were identified by exploring the sub-syllabic regions.

Ludeña-Choez and Gallardo-Antolín [10] studied the spectral characteristics of various acoustic events along with the speech spectra. A novel parameter for Acoustic Event Classification (AEC) process was proposed. The performance of the proposed approach in the clean and noisy conditions was higher than the conventional MFCC in an AEC task. Geiger *et al.* [22] presented an acoustic scene classification system using audio feature extraction. The spectral, energy, voice-related and cepstral audio features were extracted from the recordings of acoustic scenes. The shorter and longer recordings were classified using SVM and majority voting scheme. From the feature analysis, Mel spectra was found as the most relevant feature. Higher accuracy was achieved when compared to the existing classification approaches. Oh and Chung [23] used a method for extracting features from the speech signal using a non-parametric correlation coefficient. The performance of the proposed method was better than the selective feature extraction using cross correlation. Gajšek *et al.* [24] presented an efficient approach for modeling the acoustic features to recognize various paralinguistic phenomena. The Universal Background Model (UBM) was represented by building a monophone-based Hidden Markov Model (HMM). The proposed method has achieved better results than the state-of-the-art systems.

Anguera [25] combined K-means clustering algorithm and GMM posterior grams to obtain highly discriminant features. The evaluation results have shown that the standard MFCC features were outperformed by the GMM posterior grams. Salamon *et al.* [26] presented a novel method for the classification of musical genre based on high-level melodic features extracted directly from the audio signal of polyphonic music. The melodic features were used for the classification of excerpts into different musical genres by using the machine learning algorithms. The proposed method was compared with a standard approach using low-level timbre features.

Alam *et al.* [27] analyzed the performance of the multi-taper MFCC and Perceptual Linear Prediction (PLP) features. The robust PLP features were computed by using multitapers. The recognition accuracy was improved significantly by using the MFCC and PLP features computed through multitapers. Muthumari and Mala [28] presented a study of the existing audio segmentation and classification techniques and comparison of the performance of the existing approaches. In this article, typical feature extraction techniques used in audio information retrieval for different music elements were reviewed. Two main paradigms for audio classification were presented with their advantages and drawbacks. The drawbacks of the existing approaches and merits of the proposed work are depicted in **Table 1**.

3. Efficient Approach for Segmentation, Feature Extraction and Classification of Audio Signals

The proposed approach is clearly explained in this section. Smoothening of the audio signal is performed by using mean filter. Segmentation of the audio signal is performed by using peak estimation and pitch extraction process. Peak estimation is applied to identify the variation in signal amplitude with previous and present values of the signal amplitude with respect to the sampling time. The pitch extraction is performed, based on the frequency difference of the audio signal. Then, it is determined whether the pitch satisfies the segmentation of signal sample, based on the pitch frequency deviation.

Table 1. Drawbacks of existing approaches and merits of proposed work.

Drawbacks of existing segmentation, classification and feature extraction approaches	Merits of our proposed work
● It can be applied only on discrete audio segments. ● Generation of extra overhead during the computation of MFCC features. ● Distinguishing of the speech from the music signals is poor. ● The accuracy of the existing classification techniques is low. ● High computational complexity and cost.	● The clear features of the speech signal filtered from other background signals are obtained. ● PNN-based classification provides a better prediction of the classified label using the probability estimation. ● The presence of silence and irrelevant frequency details in the audio signal is eliminated. ● PNN classification approach achieves efficient classification of the musical and non-musical signal. ● The accuracy of the proposed approach is high.

The index region of the audio sample is extracted and represented as a projection line over the audio signal. Segmentation of the audio signal is done by extracting the signal amplitude according to the window selection of the sampling time. EMFCC-EPNCC is applied to extract testing feature for the classification stage with the combination of peak estimated signal feature. Classification of audio signal into musical or non-musical signal is done by using PNN classifier. From this classification result, the category of the audio signal is specified. This is done to extract index of audio input for retrieving the audio signal. The overall flow diagram of the proposed approach is shown in the **Figure 1**. The main stages of the proposed work are

- Mean filter
- Segmentation
- Peak Estimation
- Peak extraction
- Pitch extraction
- Feature Extraction
- EMFCC-EPNCC based feature extraction
- PNN-based classification

3.1. Mean Filter

Filtering of the audio signal is performed by using the mean filter. The mean filter is applied directly to the input audio signal, without the need to know about the statistical characteristics of the audio signal. This filter operates by using small movable window for each sample duration of the audio signal. Smoothing signal is obtained by considering the mean values of the side window and replacing the central window element with the mean value. The amplitude of the audio signal is normalized and the Gaussian noise present in the audio signal is reduced. This filtered signal is then applied to the segmentation process. The plot of the input audio signal is depicted in **Figure 2(a)** and the filtered audio signal is shown in **Figure 2(b)**.

3.2. Segmentation

The main purpose of the segmentation process is to divide the input audio signal into homogeneous segments. This is done by evaluating the similarity between two contiguous windows of fixed length, in the cepstral domain. The audio segmentation is performed by using three processes:

❖ Peak Estimation
❖ Peak Extraction
❖ Pitch Extraction

1) *Peak Estimation*

During the peak estimation, peaks are calculated from amplitude and frequency of input signal from the parameters of α, β and γ. The threshold peak value is calculated based on the average value of the signal. The interpolated peak location is calculated and the condition of peak from the peak magnitude is checked with the

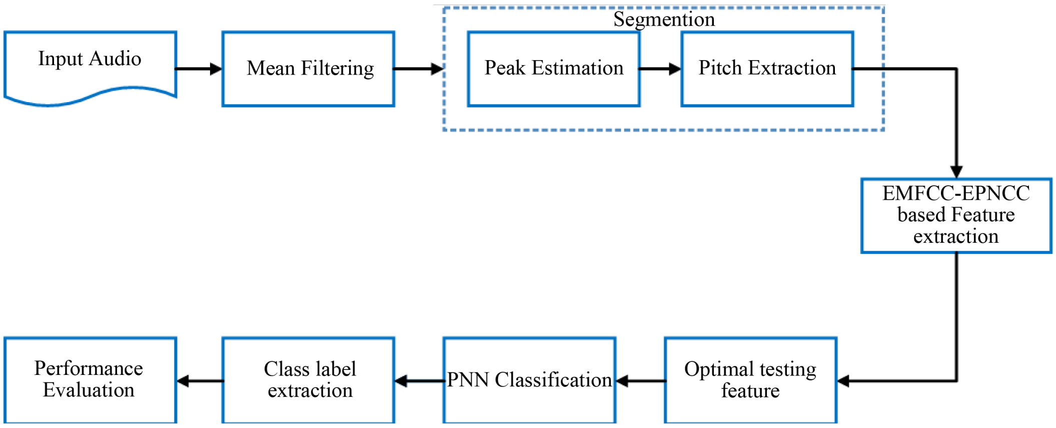
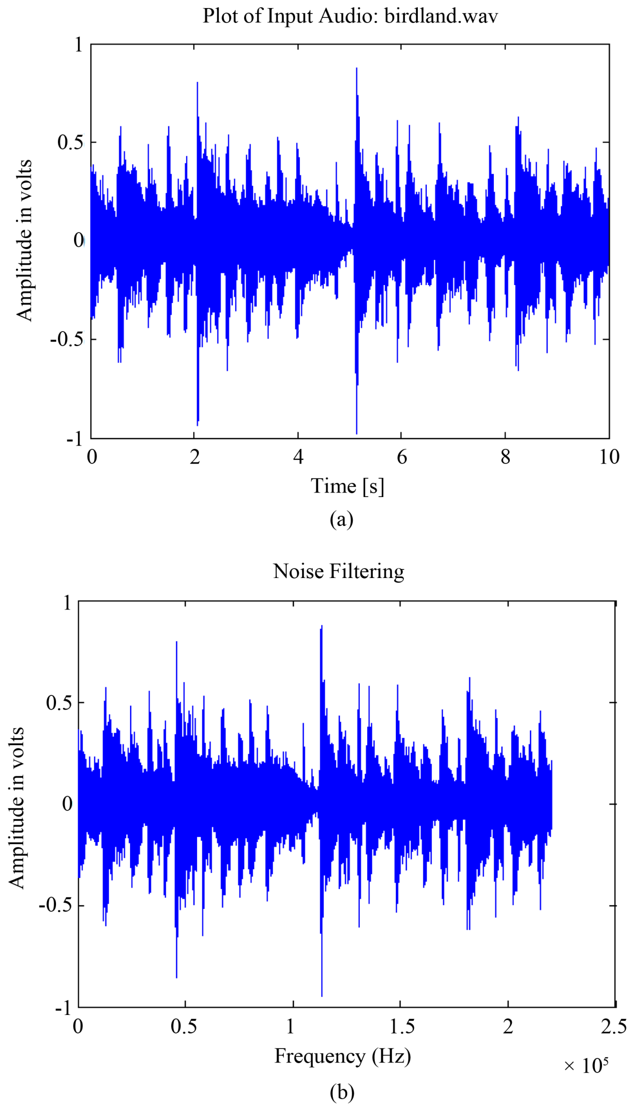


Figure 1. Overall flow diagram of the efficient approach for segmentation, feature extraction and classification of audio signals.



(a)

(b)

Figure 2. (a) Plot of input audio signal; (b) Filtered signal.

threshold peak magnitude estimate value. If the peak magnitude is greater than the estimate, then it is noted as a peak range in the sampled size of signal. Interpolated peak location is given as,

$$P = \left| \left(\frac{1}{2}\right) * \left(\frac{(\propto - \gamma)}{\alpha - 2\beta + \gamma}\right) \right|_{\left[\frac{-1}{2},\frac{1}{2}\right]} \tag{1}$$

The peak magnitude estimate is given as,

$$T_1(P) = \beta - \frac{1}{4}(\alpha - \gamma)\, p \tag{2}$$

where "α" is the starting edge of parabola of the signal, "β" is the peak amplitude edge of signal and "γ" is the finishing edge of parabola of the signal. The above parameters are calculated from the transformation signal obtained as the result of MFCC method. **Figure 3(a)** shows the input audio signal, **Figure 3(b)** shows the audio signal after cancellation of Direct Current (DC) drift and normalization, **Figure 3(c)** shows the audio signal after applying the derivative function, **Figure 3(d)** shows the integrated signal and **Figure 3(e)** shows the audio signal with peak points.

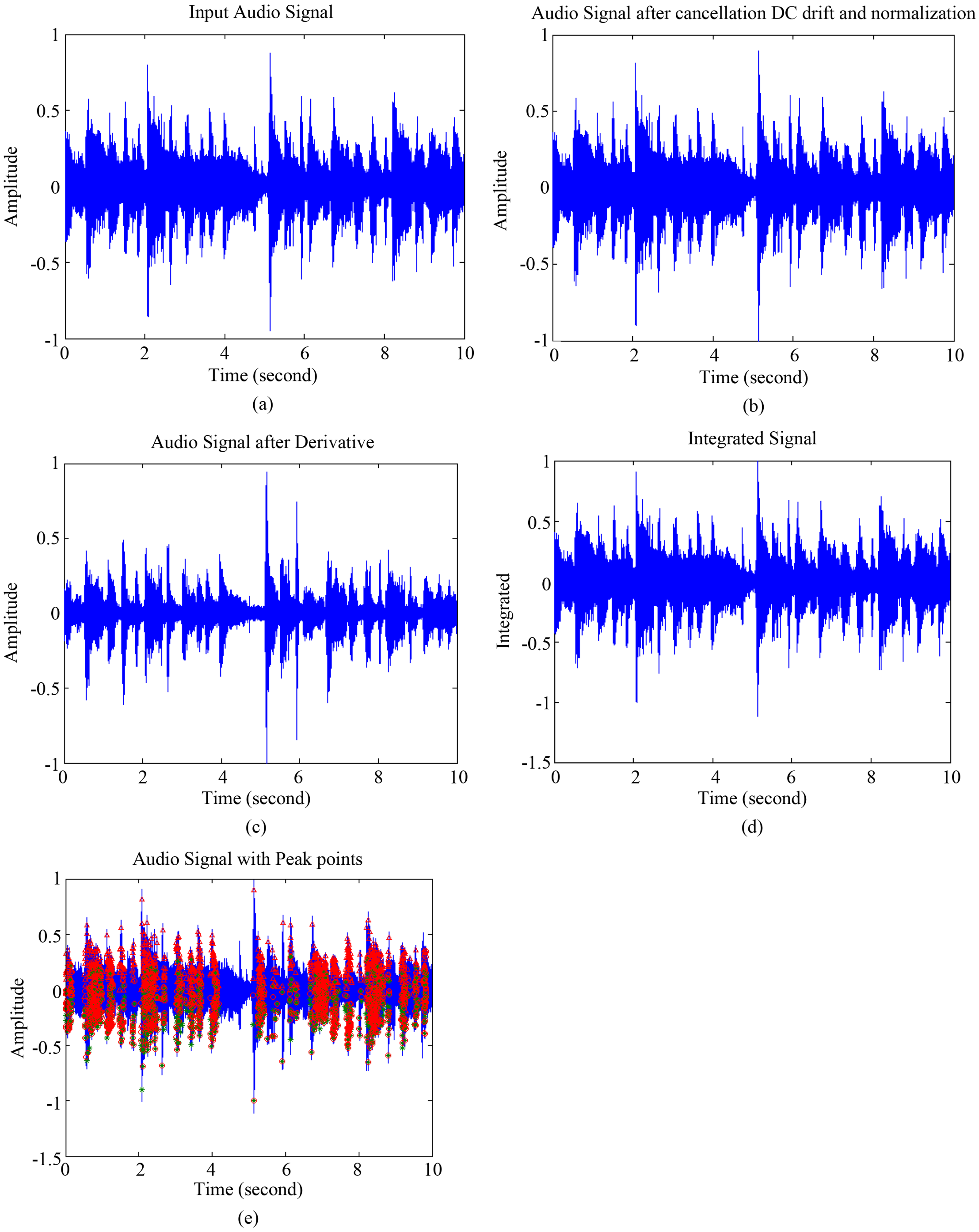


Figure 3. (a) Input audio signal; (b) Audio signal after cancellation of DC drift and normalization; (c) audio signal after applying the derivative function; (d) Integrated signal; (e) Audio signal with peak points.

2) *Peak Extraction*

In this feature extraction stage, R_Loc represents the feature at which the wave is in high peak Positive and Q_Loc represents the features at small signal difference at negative edge of the audio signal and S_Loc represents the feature values [29] at maximum signal difference at negative point of input audio signal. For each stage of convolution process, low pass filter and high pass filter are used for calculating difference in peak extraction using transfer function as represented by $X_1, X_2, \cdots, X_6$. This is updated by the threshold value extracted from input audio signal. Here, left and right specifies the left position and right position of sampled input signal. **Figure 3** shows the extraction process of the testing features.

Feature Vector is formed as,

a) Max (Q_loc),
b) Max (R_loc),
c) Max (S_loc),
d) Length (Q_loc > 0),
e) Length (R_loc > 0),
f) Length (S_loc > 0),
g) Sum (Q_loc > 0),
h) Sum (R_loc > 0),
i) Sum (S_loc > 0).

where,

$$R_{loc} = indx_{\max}\left(X_{1(Left \to Right)}\right) - 1 + Left\,; \tag{3}$$

$$Q_{loc} = indx_{\min}\left(X_{1(Left \to R_{loc})}\right) - 1 + Left\,; \tag{4}$$

$$S_{loc} = indx_{\min}\left(X_{1(Left \to Right)}\right) - 1 + Left\,; \tag{5}$$

where X_1 -Input Audio Signal.

$$Left = \left\{Pos_{Reg} == 1\right\}; \tag{6}$$

$$Right = \left\{Pos_{Reg} == -1\right\}; \tag{7}$$

$$Pos_{Reg} = \left\{X_6 > \left(Thres * Max_h\right)\right\}; \tag{8}$$

$$Thres = Mean\left(X_6\right); \tag{9}$$

$$Max_h = Max\left(X_6\right); \tag{10}$$

$$X_6 = \frac{\left(\left(\sum_{m=0}^{N}\left(x_5\left(m\right) * h\left(m\right)\right)\right) * \left(T_s\right)\right)}{\left(\max\left(\left(\sum_{m=0}^{N}\left(x_5\left(m\right) * h\left(m\right)\right)\right) * \left(T_s\right)\right)\right)} \tag{11}$$

$$X_5 = \frac{\left(X_4\right)^2}{\max\left(X_4\right)} \tag{12}$$

$$X_4 = \frac{\left(\left(\sum_{m=0}^{N}\left(x_3\left(m\right) * h\left(m\right)\right)\right) * \left(T_s\right)\right)}{\left(\max\left(\left(\sum_{m=0}^{N}\left(x_3\left(m\right) * h\left(m\right)\right)\right) * \left(T_s\right)\right)\right)} \tag{13}$$

$$X_3 = \frac{\left(\left(\sum_{m=0}^{N}\left(x_2\left(m\right) * h_2\left(m\right)\right)\right) * \left(T_s\right)\right)}{\left(\max\left(\left(\sum_{m=0}^{N}\left(x_2\left(m\right) * h_2\left(m\right)\right)\right) * \left(T_s\right)\right)\right)} \tag{14}$$

$$X_2 = \frac{\left(\left(\sum_{m=0}^{N}\left(x_1\left(m\right) * h_1\left(m\right)\right)\right) * \left(T_s\right)\right)}{\left(\max\left(\left(\sum_{m=0}^{N}\left(x_1\left(m\right) * h_1\left(m\right)\right)\right) * \left(T_s\right)\right)\right)} \tag{15}$$

$$X_1 = \frac{S_1}{\max\left(S_1 - mean\left(S_1\right)\right)} \tag{16}$$

where "N" is the sample size of input audio, $h_1(m)$ is the transfer function of Low pass filter, $h_2(m)$ is the transfer function of the High pass filter and $h(m)$ is the transfer function of convolution.

3) *Pitch Extraction*

In this pitch extraction, initially the objective function is implemented to perform weight calculation from the input audio signal based on the cosine angle difference of the signal amplitude. The pitch angle variation for each pre-allocated time samples calculated from the length of input signal (X_i) is extracted based on the objective function from [29]. Then, difference in the limitation of time sequence with the $Pitch(t,\tau)$ calculation and extracted pitch angle is checked. The pitch of signal is estimated by using time domain based detection method. There are several methods used for signal pitch estimation.

a) Zero Crossing
b) Autocorrelation
c) Maximum Likelihood
d) Adaptive filter using FFT
e) Super Resolution pitch detection

In the proposed method, Maximum Likelihood based Pitch extraction is implemented. This is represented as,

$$Pitch(t,\tau)=\begin{cases}\frac{1}{N+1}\sum_{n=0}^{N}P(t+n\tau) & 0\le t\le b\\ \frac{1}{N}\sum_{n=0}^{N}P(t+n\tau) & b\le t\le \tau\end{cases} \tag{17}$$

where, "τ" is the frame size of audio signal, "t" is the sampling time and "N" is the total size of audio signal. This is updated by using the objective function as,

$$P(t+n\tau)=10*size(X_i)+\sum_{i=1}^{N}X_i^2-(10*\cos(2*pi*X_i)) \tag{18}$$

The pitch frequency in each frame of the audio signal is calculated. The threshold value of the amplitude of the segmented audio signal is calculated by using

$$Th=\max(S)*25/100 \tag{19}$$

Then, the minimum and maximum peak values of the segmented signal are checked based on the threshold value. The peak value is estimated based on the positive and negative peak values lying on the left and right sides of the segmented signal. The positive small and large pitches and negative small and large pitches are obtained based on the peak values. **Figure 4** flow diagram of the pitch feature extraction process.

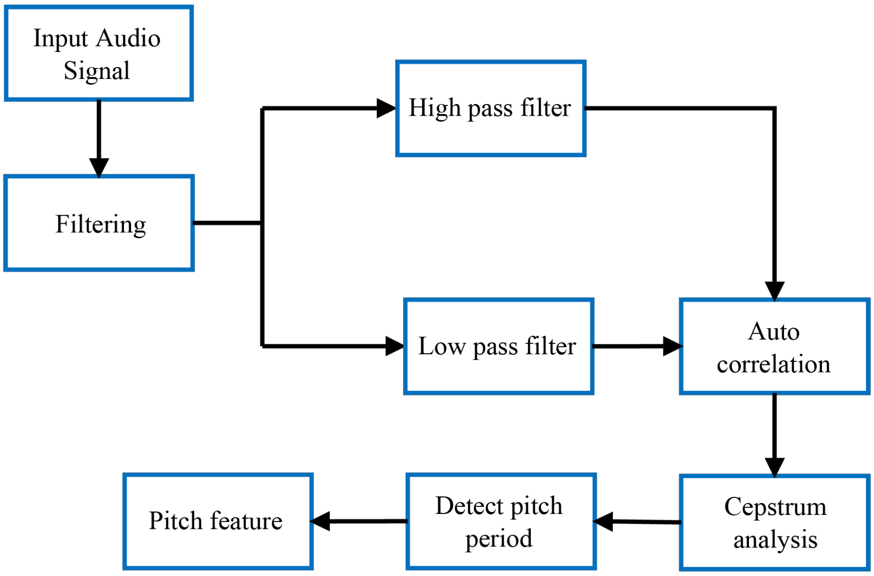


Figure 4. Flow diagram of the pitch feature extraction process.

Pitch Feature extraction Algorithm

```
Input: Pitch result, $P(t+n\tau)$ and Segmented signal, ‘S’
Output: Pitch Feature, ‘PT’
Step1: initialize threshold value by using equation (25)
Step2: For i = (1 to size(S))
Step3:     if (S (i) > Th)
Step4:             M (i) = 1;
Step5:     else if (S (i) < -Th)
Step6:         M (i) = -1;
Step7:     else
Step8:         M (i) = 0;
Step9:         end if
Step10: end ‘i’ for
Step10: S1 = find (M = 0);
Step11: S2 = find (M=1 or M=-1)
Step12: PS = {mean (S2 > 0), variance (S2>0), SD (S2>0), Max (S2>0)};
Step13: NS = {mean (S2 < 0), variance (S2<0), SD (S2<0), Max (S2<0)};
Step14: LargeP = find (S2 < Th || S2 > -Th);
Step15: SmallP = find (S2 > Th || S2 < -Th);
Step16: SS = {mean (SmallP), variance (SmallP), SD (SmallP), Max (SmallP)};
Step17: LS = {mean (LargeP), variance (LargeP), SD (LargeP), Max (LargeP)};
Step18: PLS = {mean (LargeP>0), variance (LargeP>0), SD (LargeP>0), Max (LargeP>0)};
Step19: NLS = {mean (LargeP<0), variance (LargeP<0), SD (LargeP<0), Max (LargeP<0)};
Step20: PSS = {mean (SmallP>0), variance (SmallP>0), SD (SmallP>0), Max (SmallP>0)};
Step20: NSS = {mean (SmallP<0), variance (SmallP<0), SD (SmallP<0), Max (SmallP<0)};
Step21: PT ={LS, SS, PS, NS, PLS, NLS, PSS, NSS};
```

Figure 5 shows the input waveform and pitch track. The segmented audio result is shown in the **Figure 6**.

3.3. Feature Extraction

EMFCC-EPNCC is applied for the extraction of features from the audio signal. In several feature analysis techniques, the signal intensity is estimated based on spectrum depth variation only. In our proposed work, we implement both Mel-function with Power normalized Cepstral Coefficients for speech signal analysis. This method filters other signals present in the speech data with Gamma tone frequency integration. By using this method, the feature of signal is clear than other feature extraction types. Representation of the audio signal is performed by using a set of features.

1) *EMFCC-EPNCC*

Feature extraction is performed based on the EMFCC and EPNCC to return the feature values computed from the audio signal and sampled at fs (Hz). In the EMFCC process, 20 frame size is chosen from the sample size of input audio signal. The audio signal is subjected to the windowing process to divide it into frames and perform spectrum analysis for each and every frame of the signal. Then, Discrete Fourier Transform (DFT) is applied to the frames. The Mel frequency warping is applied to the DFT output. Logarithm is applied to the filter bank of the Mel frequency warping output. Inverse DFT is applied to obtain the Mel cepstrum coefficients.

In the EMFCC based audio feature extraction, the Mel Cepstrum is extracted from the transformation output. The input audio is divided into frames by applying the windowing function at fixed intervals. The distribution function for the window is defined as

$$P_i(k) = \left(\frac{1}{N}\right)\left|X_i(k)\right|^2 \tag{20}$$

Windowing involves multiplication of the time record using a finite-length window with a smoothly varying amplitude. This results in the continuous waveforms without sharp transitions. Windowing process minimizes the disruptions at the starting and end point of the frame. The output of the window is given as

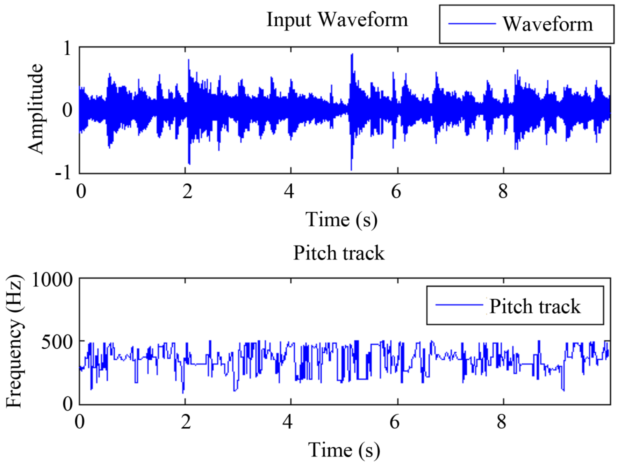


Figure 5. Input waveform and pitch track.

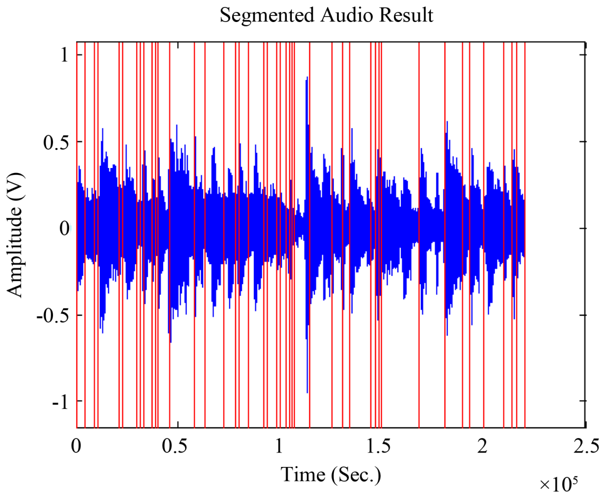


Figure 6. Segmented audio result.

$$Y(n) = X(n)W(n) \tag{21}$$

where $0 \le n \le N-1$. Here, "N" denotes the quantity of samples within every frame, $Y(n)$ represents the output signal obtained after multiplying the input signal $X(n)$ with the window $W(n)$. Figure 7(a) shows the output window plot of the audio signal and Figure 7(b) shows the reduction in the spectral leakage effect by applying window.

A cepstral feature vector is generated for each frame and the DFT is applied to each frame. Mel frequency warping represented by the cosine transformation is applied to the DFT output. The cosine transform is described as

$$C_n = \sum_{k=1}^{N} \log\left(X_i(k)\right)\cos\left(n\left(k-\left(\frac{1}{2}\right)\right)*\left(\frac{\pi}{k}\right)\right) \tag{22}$$

where, $n = 1, 2, \cdots, N$ and "N" is the sample size of audio input.

The cosine transform is used to convert the log Mel cepstrum back into the spatial domain. The FFT is applied to calculate the coefficients from the log Mel cepstrum. The main advantage of the Mel frequency warping is the uniform placement of the triangular filter on the Mel scale between the lower and upper frequency limits of the Mel-warped spectrum.

The Mel frequency warping is calculated using the formula

$$\mu(\omega) = 2595 \cdot \log\left(1 + \frac{\omega f_s}{2\pi \cdot 700\ \text{Hz}}\right) \tag{23}$$

Here “ f_s ” is the sampling frequency and “ω” is the warping function. To be integrated with the cosine transformation, the Mel-warping function is to be normalized to satisfy the specific criterion $\tilde{\mu}(\pi) = \pi$.

$$\tilde{\mu}(\omega) = \frac{\pi}{\mu(\pi)} \cdot \mu(\omega) \tag{24}$$

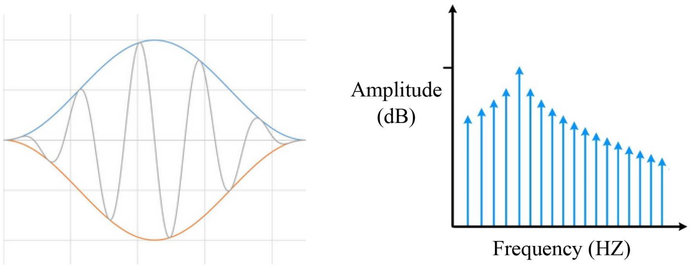
$$= d \cdot \log\left(1 + \frac{\omega f_s}{2\pi \cdot 700\ \text{Hz}}\right) \tag{25}$$

The output of the Mel-frequency warping is shown in **Figure 8**. The phase information is omitted and the amplitude of the audio signal is considered. The logarithmic value of the amplitude is taken. Then, the inverse DFT is applied to extract the Mel Cepstrum output as feature of audio signal from EMFCC. The Log filter bank energies and Mel frequency cepstrum are shown in **Figure 9**.

The EMFCC-EPNCC is applied for extracting the audio features. **Figure 10** shows the flow diagram of the EMFCC-EPNCC process.

2) *EMFCC-EPNCC Algorithm*

The EPNCC extraction process involves frequency-to-Mel conversion, Mel-to-frequency conversion and cosine transform process. In the EPNCC-EMFCC method, the frequency to Mel is performed for extracting spectral data of signal based on the peak and pitch variation. Mel to frequency conversion is performed to filter out



(a)

(b)

Figure 7. (a) Output Window Plot of audio signal; (b) Reduction in the spectral leakage effect by applying window.

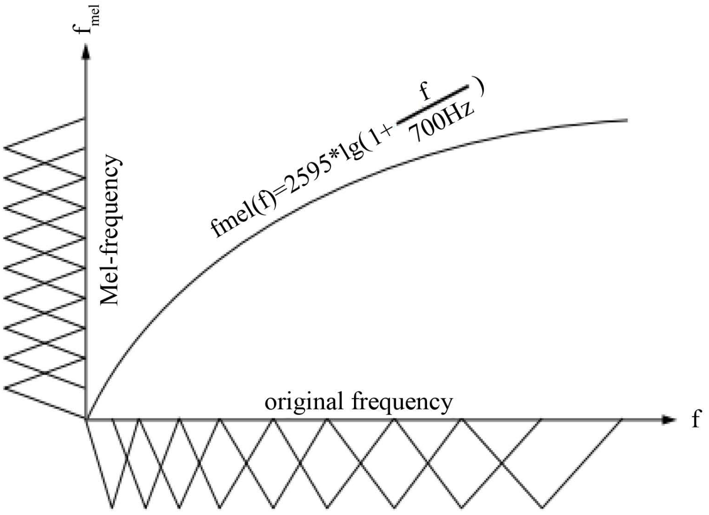


Figure 8. Output plot of Mel-frequency warping.

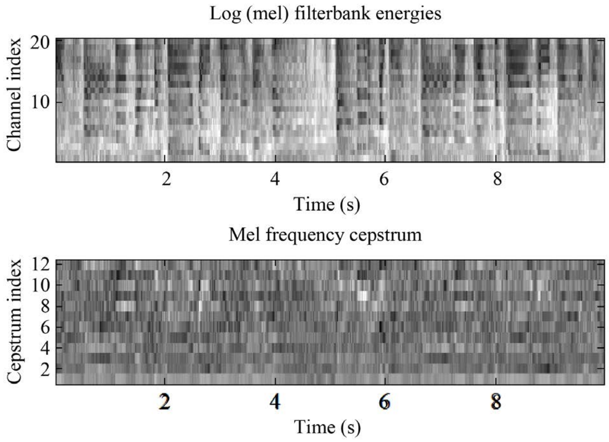


Figure 9. Log filter bank energies and output Mel frequency cepstrum.

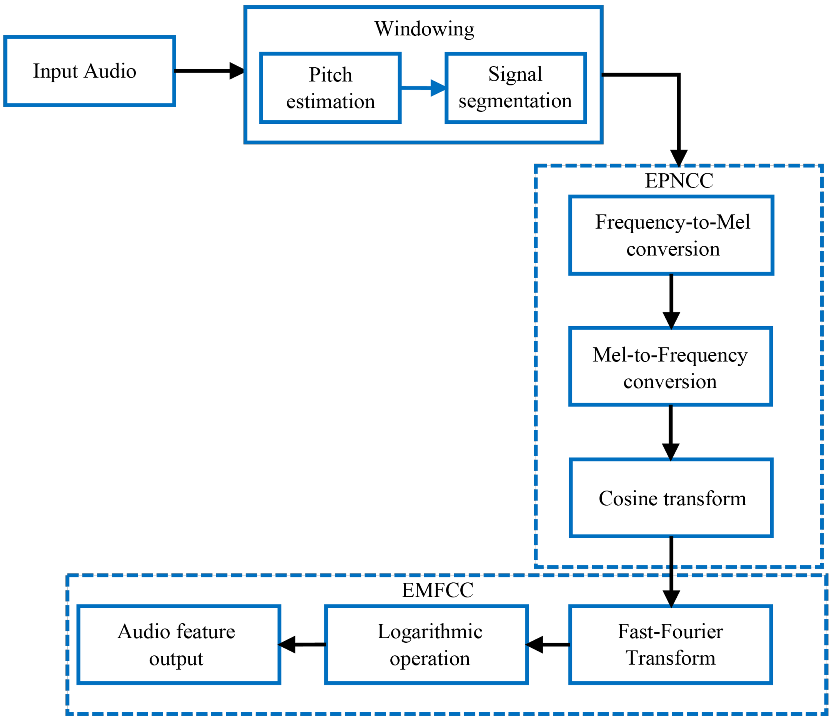


Figure 10. Flow diagram of the EMFCC-EPNCC process.

other frequency signals by the frequency domain. The window size of the filtered signal is initialized by using the equations

$$N_w = 1e^{-3} * T_w * F_s, \tag{26}$$

$$N_s = 1e^{-3} * T_s * F_s \tag{27}$$

where, "T_w" denotes the time division of window, "F_s" represents the frequency of the audio signal and "F_s" indicates the time samples of audio signal. The frequency of the audio signal is converted to Mel frames by using the formula,

$$H = 1127 * \log\left(\frac{1 + HZ}{700}\right) \tag{28}$$

Then, the Mel frames are converted into frequency by using the equation

$$M = 700 * \mathrm{e}^{\frac{M}{1127} - 700} \tag{29}$$

The cosine transform is applied by using

$$CT = \sqrt{\frac{2}{M}} \cos\left(f\left(x_F\right) * f\left(\frac{\pi x_F}{M}\right)\right) \tag{30}$$

The output of the DCT process is shown in **Figure 11**. Then, FFT is applied for deconstructing a time domain representation of audio signal into the frequency domain representation. This is done by using the exponential of radian value for each sampling difference in the input signal with K^{th} iteration.

$$X_K = \sum_{n=0}^{N-1} x_n \mathrm{e}^{-\frac{i 2\pi K n}{N}} \tag{31}$$

where, $K = 0, \cdots, N-1$, "N" is the sample size of audio signal and "x_n" is the input signal. The magnitude of the spectrum is extracted by applying the FFT transform to the filtered signal and multiplying it with the Mel frames

$$Mag = FFT\left(X_F\right) * M \tag{32}$$

Figure 12 shows the single-sided Amplitude Spectrum of output signal of FFT process y(t).

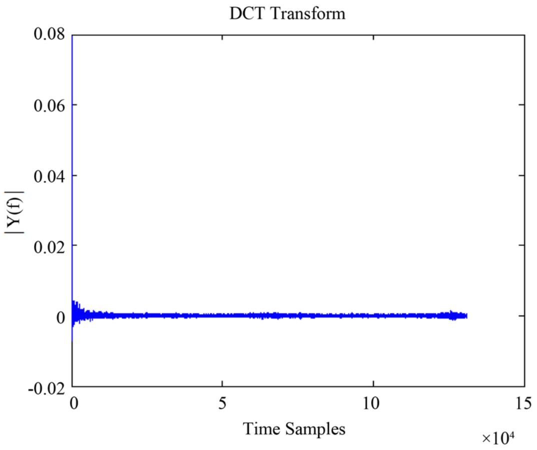


Figure 11. DCT output plot.

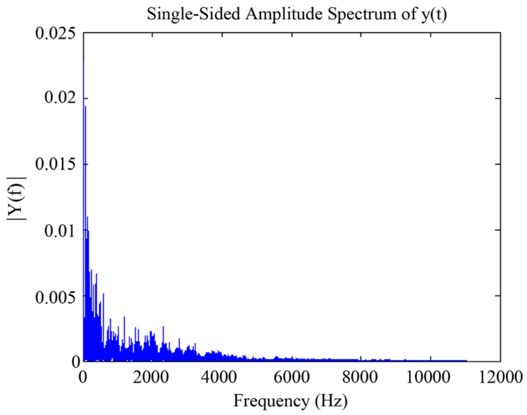


Figure 12. Single-sided amplitude spectrum of *y*(*t*).

The Filter Coefficient "FL" is extracted by using the equation

$$FL = 1 + \frac{1}{2} L * \sin\left(\pi * \frac{f(x_F)}{L}\right) \tag{33}$$

The audio feature output is obtained from the product of the cosine transformation value, logarithmic value of the magnitude and filter coefficient.

$$M_F = CT * \log(Mag) * diagonal(FL) \tag{34}$$

```
EMFCC-EPNCC Algorithm
Input: Filtered Signal, "$X_F$" and Window, "$H$"
Output: Audio feature output, "$M_F$".
Step1: Initialize window size based on Equations (26) and (27)
Step2: Convert frequency to Mel using Equation (28)
Step3: Convert Mel to Frequency using Equation (29)
Step4: Apply Cosine Transform given by Equation (30)
Step5: Extract Magnitude by applying FFT transform using Equation (32)
Step6: Extract Filter Coefficient using Equation (33)
Step7: Obtain audio feature output by using the Equation (34)
```

The index difference plot is shown in **Figure 13**.

3.4. Classification

The audio feature from the selected feature vectors obtained from the segmentation based on peak and pitch estimation is applied to the classification process. Classification of audio signal is performed using the PNN classifier, based on the testing features. Multi-label feature analysis is presented in the proposed work. Hence, a multi-class classifier model is implemented. Compared with other types of classifier, PNN provides a better prediction of the classified label using the probability estimation based on the neural network function.

The neural network is frequently used for the classification of the signals. The PNN is the quick learning model than the other neural network models. Hence it is used for classification of audio signal. The Probability Density Function (PDF) for a single sample is calculated as the output of the neuron of the pattern layer. This is given as

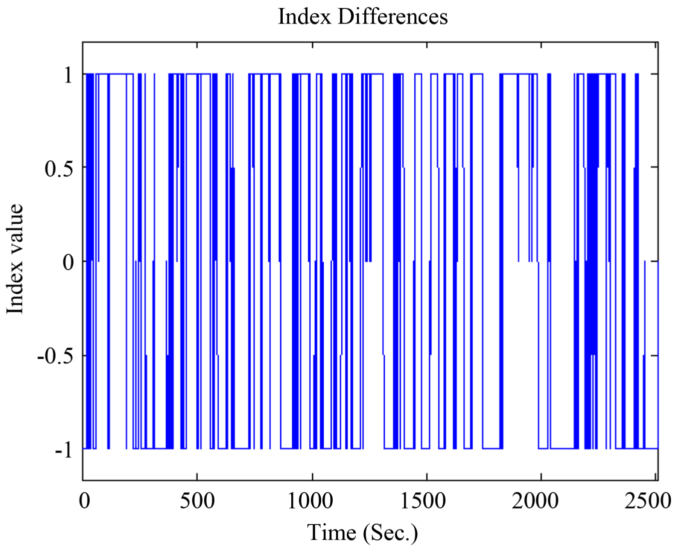


Figure 13. Shows the index difference plot.

$$F\left(Y\right)=\frac{1}{\left(2\pi\right)^{\frac{d}{2}}\cdot\sigma^{d}}\mathrm{e}^{\frac{-\left\|Y-Y_{j}\right\|^{2}}{2\sigma^{2}}}\tag{35}$$

where "*Y*" denotes the unknown input vector. "Y_j" represents the j^th^ sample input vector. "*k*" denotes the smoothing parameter and "*d*" denotes the dimension of the input vector. The output of the neuron of the summation layer is calculated as the PDF for a single pattern by using the equation

$$G_{i}\left(Y\right)=\frac{1}{\left(2\pi\right)^{\frac{d}{2}}\cdot\sigma^{d}}\frac{1}{N_{k}}\sum_{j=1}^{N_{k}}\mathrm{e}^{\frac{-\left\|Y-Y_{j}\right\|^{2}}{2\sigma^{2}}}\tag{36}$$

where N_k is the total number of samples in the kth population. The decision layer performs classification of the pattern according to the Bayes decision rule, based on the output of the neurons of the summation layer. This is done, when the apriori probabilities for each classes are similar and the losses associated with incorrect decision making for each class are similar.

$$E\left(Y\right)=\arg\max\left\{G_{i}\left(Y\right)\right\}\tag{37}$$

where $i=1,2,3,\cdots,N$. "*E*(*Y*)" denotes the estimated class of the pattern and "*N*" is the total number of classes in the training samples. The performance of the PNN classifier is more reliable than the Back Propagation Neural Network (BPN). The convergence rate of the PNN classifier is faster with respect to the increase in the size of the training set. Addition and removal of the training samples are performed without the need for extensive retraining. By using the PNN classifier, gradient vector of the proposed Kernel function is implemented for the selected optimal testing features. This is described as,

$$g_{i}=\sum_{p=1}^{P}\sum_{m=1}^{M}\left(\frac{\partial X_{p,m}}{\partial t_{i}}e_{p,m}\right)\tag{38}$$

where, "*e*" is the feature vector of input signal, and "$X_{p,m}$" represents feature matrix of dataset. The category of the audio signal is determined based on the classification result. Initially, it is detected whether the given testing feature is musical or non-musical. If it is detected as a musical signal, the label is classified as Piano, Guitar, etc. Hence, the presence of silence and irrelevant frequency details in the audio signal is eliminated.

4. Performance Analysis

This section illustrates the performance evaluation and comparative analysis of the proposed approach with the existing techniques. The datasets obtained from Ffuhrmann [30] and Marsyasweb [31] are used for the performance evaluation of the proposed approach. For the musical classes, Ffuhrmann includes 11 classes of pitched instruments including cello (cel), clarinet (cla), flute (flu), acoustic guitar (gac), electric guitar (gel), organ (org), piano (pia), saxophone (sax), trumpet (tru), violin (vio), and Band (ban) and hundred number of files and Marsyasweb includes 5 number of classes and 64 number of files. The data for the 11 pitched instruments is obtained from the pre-selected music tracks, with the objective of extracting excerpts containing a continuous presence of a single predominant target instrument. The dataset contains a total number of 220 pieces of Western music including various musical genres and instrumentations. For Non-Musical classes, Marsyasweb includes 64 number of files. The dataset consists of 120 audio tracks each 30 seconds long. Each has 60 examples. The tracks are all 22,050 Hz Mono 16-bit audio files in .wav format. The performance of the proposed approach is evaluated using the metrics such as

- Precision
- NMI
- F-Score
- Entropy

The comparison of the Precision, NMI, F-score and Entropy of the proposed approach and existing features is shown in **Table 2**. The proposed approach for segmentation, feature extraction and classification of audio signals (ASFEC) is compared with the acoustic features of spectral clustering and rotation, functional Magnetic Resonance Imaging (fMRI)-measured features of Support Vector Regression (SVR), Improved Twin Gaussian

Table 2. Comparative Analysis of Precision, NMI, F-score and Entropy of the proposed approach and existing features.

Methods	Precision	NMI	F-Score	Entropy
Acoustic features (spectral clustering)	0.373	0.129	0.502	1.405
Acoustic features (spectral rotation)	0.384	0.127	0.485	1.389
fMRI-measured features of SVR	0.406	0.213	0.539	1.304
fMRI-measured features using high-level features	0.423	0.21	0.543	1.262
fMRI-measured features of ITGP	0.485	0.294	0.585	1.155
Integrated features of kernel addition	0.52	0.323	0.61	1.083
Integrated features of kernel product	0.499	0.324	0.583	1.117
Integrated features of CCA	0.51	0.317	0.599	1.1
Integrated features of ITGP	0.541	0.337	0.623	1.079
Proposed ASFEC approach	0.718	0.412	0.7195	1.6875

Process (ITGP), integrated features of kernel addition, kernel product, Canonical Correlation analysis (CCA) and ITGP. The precision, NMI, F-score and Entropy of the proposed approach are found to be relatively higher than the acoustic and integrated features [32]. From the comparison result, it is clearly evident that the proposed approach outperforms the existing acoustic and integrated features.

Acoustic features: Only the acoustic features are used in this experiment for the audio signal clustering. Spectral rotation is used to replace the K-means in the spectral algorithm. The performance of the spectral rotation has proven to be better than the spectral clustering approach.

fMRI-measured features of SVR: First, the SVR model is trained by adopting the fMRI-measured features and acoustic features of audio selections and applied to predict the fMRI-features of the audio samples.

fMRI-measured features of ITGP: The ITGP model is trained with fMRI-measured features and acoustic features of audio selections and applied to predict the fMRI-features of the audio samples.

Integrated features of Kernel addition: The kernel addition method is applied on the fMRI-measured features and acoustic features of testing audio samples. First, the kernels are integrated by adding them and the Eigen vectors of the Laplacian of the integrated kernel are computed. Then, a matrix is generated by using the Eigen vectors as columns. Finally, each row of this matrix is considered as an integrated feature.

Integrated features of kernel product: The corresponding elements of kernels of different views are multiplied with each other to form the integrated kernel.

Integrated features of CCA: The correlated features are extracted from the fMRI-measured features and acoustic features.

4.1. Precision

Precision is defined as the ratio of the number of correct results to the number of predicted results.

$$\text{Precision} = \frac{\text{TP}}{\left(\text{TP} + \text{FP}\right)} \tag{39}$$

4.2. NMI

NMI is one of the rapidly prevalent measures to evaluate the agreement level between two affinity matrices formed by the predicted labels and true labels of the audio samples.

$$NMI = \sum\nolimits_{x,y} p\left(x,y\right) \ln \frac{p\left(x,y\right)}{p\left(x\right) p\left(y\right)} \tag{40}$$

where $p(x)$ and $p(y)$ are the marginal probabilities and $p(x,y)$ is the joint probabilities.

4.3. F-Score

F-score is taken as a weighted average of the precision and recall values. Recall is defined as the ratio of number of correct results to the number of returned results. Higher values of Precision, NMI and F-score indicate the improved efficiency for segmentation and classification of audio signal.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{41}$$

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}} \tag{42}$$

4.4. Entropy

The entropy is the sum of the individual entropies for the classification process weighted according to the classification quality. Higher entropy values indicate better classification results.

$$H(X) = \sum_j P(x_j) I(x_j) \tag{43}$$

where, "*H*" is the entropy of the discrete random variable "*X*". "*P*" is the probability of *X* and "*I*" is the information content of "*X*". *I*(*X*) is a random variable.

Figure 14 shows the graph illustrating the comparative analysis of Precision, NMI, F-Score and Entropy for the proposed approach and existing acoustic features, fMRI and integrated features with respect to the prediction rate. It is clearly observed that the precision, NMI, F-Score and entropy for the proposed approach with respect to the prediction rate are higher than the existing fMRI and integrated features.

4.5. ROC Plot for Classification

The ROC curve is a graphical plot that shows the performance of the PNN classifier for the classification of audio signal. The true positive rate is plotted with respect to the false positive rate at various threshold settings. The ROC curve is generated by plotting the cumulative distribution function of the true detection probability versus the false-alarm probability. Each point on the ROC plot represents a pair of the sensitivity/specificity values corresponding to the specific decision threshold value. The proximity of the ROC plot to the upper left corner indicates the higher accuracy of the classification process. **Figure 15** shows the ROC curve for classification. From the figure, it is clearly evident that the proposed approach achieves high classification result.

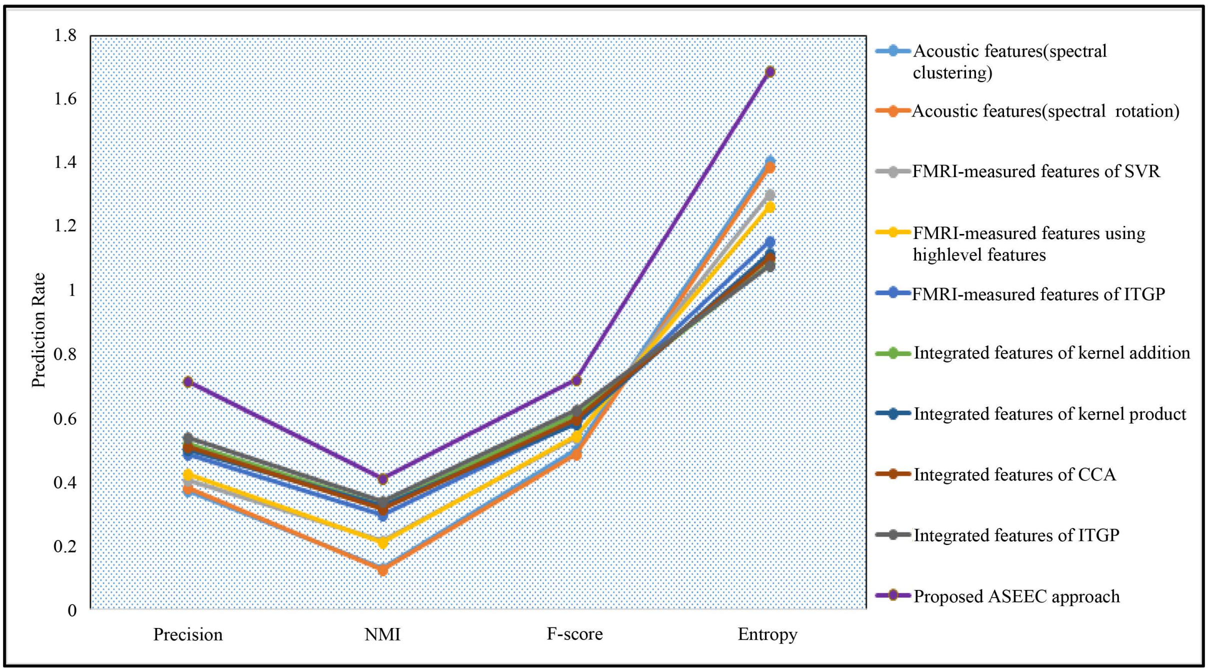


Figure 14. Comparative analysis of precision, NMI, F-score and entropy for the existing features and proposed approach.

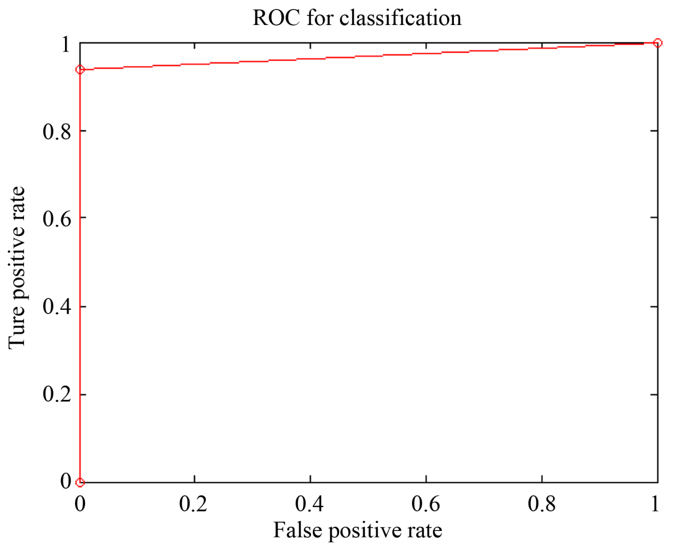


Figure 15. ROC for classification.

4.6. FRR Graph

The FRR is defined as the ratio of the number of false rejections to the number of the classified signals.

$$\text{FRR} = \frac{\text{Number of false rejections}}{\text{Total Number of classified signals}} \tag{44}$$

Figure 16 shows the FRR graph showing the relationship of the FRR with respect to the number of class. The FRR reduces with the increase in the number of classes. Reduction in the rejection rate of the incorrectly predicted sample indicates the effective classification of audio signal.

4.7. FAR Graph

FAR typically is defined as the ratio of the number of false acceptances to the number of classified signals.

$$\text{FAR} = \frac{\text{Number of false acceptances}}{\text{Total Number of classified signals}} \tag{45}$$

Figure 17 is the comparison graph between the FAR and number of classes. The FAR seems to increase with the increase in the number of classes. Hence, the incorrect classification of audio signal is prevented.

4.8. GAR Graph

The GAR is the fraction of the genuine scores exceeding the threshold value. Higher the GAR value, higher is the classification efficiency. **Figure 18** shows the comparison of the GAR with respect to the number of classes.

$$\text{GAR} = 1 - \text{FRR} \tag{46}$$

4.9. Sensitivity

The sensitivity is a measure of the actual members of the class that are correctly identified. It is defined as the ratio of the positively classified instances that are predicted correctly by the PNN classifier.

$$\text{Sensitivity}(\%) = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{47}$$

Here, True Positive (TP) is the number of audio signals that are correctly classified as a music or non-musical signal and False Negative (FN) is the number of music signals that are incorrectly classified as non-musical signal.

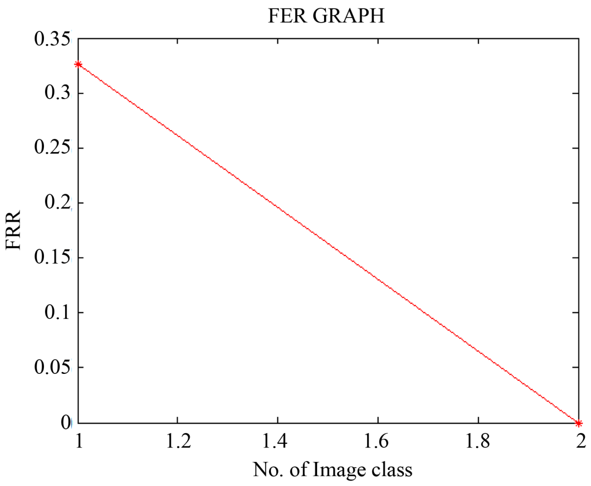


Figure 16. FRR graph.

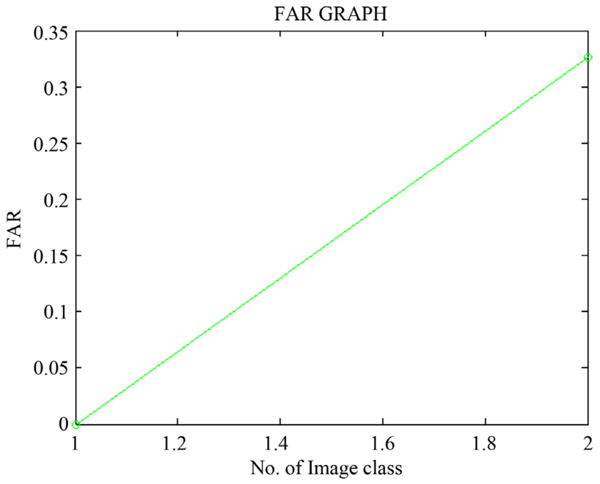


Figure 17. FAR graph.

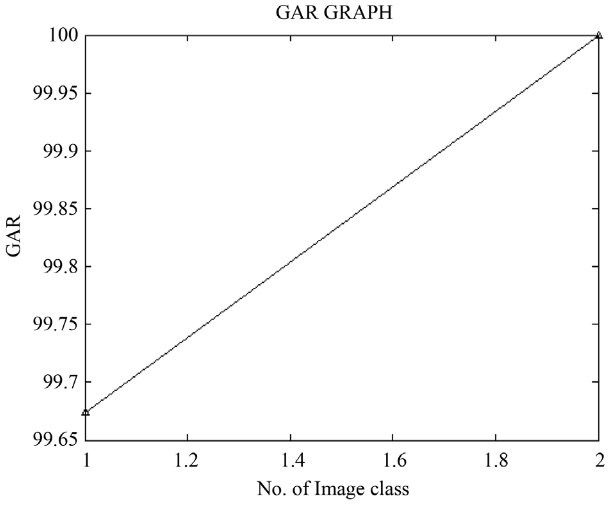


Figure 18. GAR graph.

4.10. Specificity

Specificity is referred as a true negative rate. It is defined as the ratio of the negatively classified instances that are predicted correctly by the PNN classifier.

$$\text{Specificity}(\%) = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100 \tag{48}$$

Here, True Negative (TN) is the number of audio signals that are incorrectly classified as a music or non-musical signal and False Positive (FP) is the number of music signals that are incorrectly classified as non-musical signal.

4.11. Accuracy

Accuracy is defined as the ratio of number of correctly classified results to the total number of the classified results. The performance of the classifier is determined based on the number of samples that are correctly and incorrectly predicted by the classifier.

$$\text{Accuracy}(\%) = \frac{\text{Number of correctly classified results}}{\text{Total number of classified results}} \times 100 \tag{49}$$

The comparative analysis of the sensitivity, specificity and accuracy with respect to the prediction rate is shown in the **Figure 19**. From the figure, it is observed that the PNN-based classification approach achieves high sensitivity, specificity and accuracy.

Table 3 shows the average GAR, FAR, FRR, Accuracy and Error rate values of the proposed approach. The proposed approach achieves high average GAR and accuracy and low FAR, FRR and error rate. Hence, the segmentation and classification efficiency of the proposed approach are improved.

Figure 20 shows the comparative analysis of the classification rate of the musical data for five different classes. The correct rate of the proposed approach for the five classes of the musical data is found to be higher than the error rate.

Figure 21 shows the comparative analysis of the classification rate of the musical data and non-musical data. The correct rate of the proposed approach for the musical and non-musical data is found to be higher than the error rate. This implies that the PNN classification approach achieves efficient classification of the musical and non-musical signal. **Table 4** shows the overall accuracy analysis for Online Dictionary Learning (ODL), K-means, Exemplar [33] and proposed EMFCC-EPNCC with PNN. **Figure 22** shows the overall accuracy graph for GTZAN dataset [34] and Music Technology Group (MTG) dataset [35]. The proposed EMFCC-EPNCC with PNN classifier achieves higher accuracy of 96.2% and 97.3% for both GTZAN dataset and MTG dataset.

GTZAN dataset: It is composed of 1000 30-second clips covering 10 genres such as blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock, with 100 clips per genre.

MTG dataset: It consists of approximately 2500 excerpts of Western music labeled into 11 classes of pitched

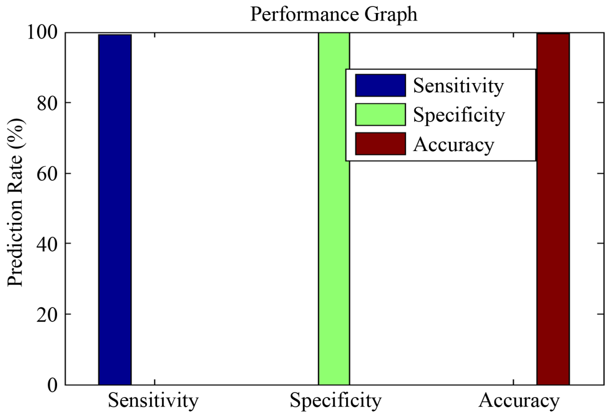


Figure 19. Comparative analysis of sensitivity, specificity and accuracy.

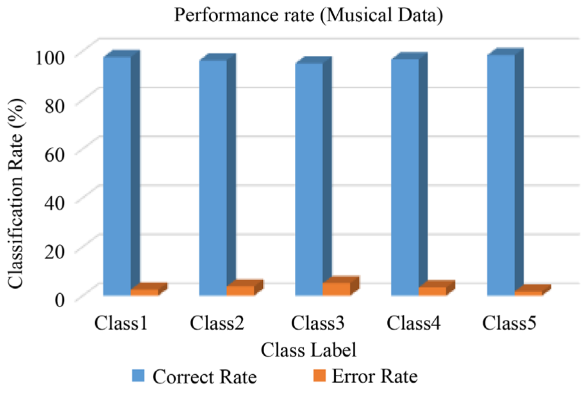


Figure 20. Comparative analysis of classification rate of musical data.

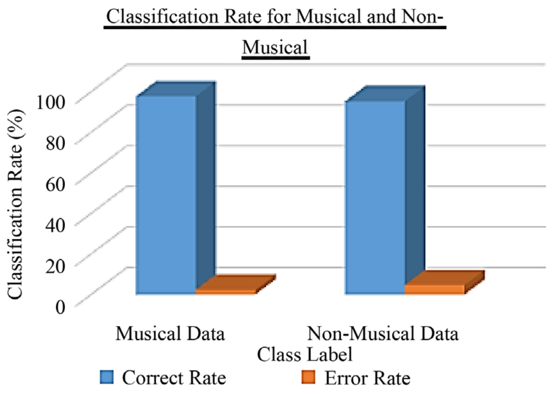


Figure 21. Comparative analysis of classification rate of musical and non-musical data.

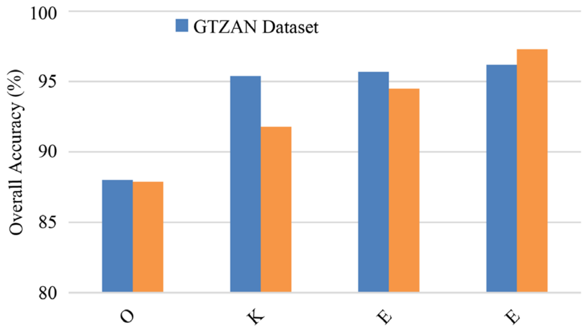


Figure 22. Overall accuracy graph for ODL, K-means, Exemplar and proposed EMFCC-EPNCC with PNN.

Table 3. Average GAR, FAR,FRR, Accuracy and Error rate values of the proposed approach.

Parameters	Value
Average GAR	99.67%
Average FAR	0.33%
Average FRR	0.33%
Average Accuracy	96.50%
Average Error Rate	3.50%

Table 4. Overall accuracy analysis for ODL, K-means, Exemplar and proposed EMFCC-EPNCC with PNN.

	Overall Accuracy (%)			
Dataset	**ODL**	**K-means**	**Exemplar**	**EMFCC-EPNCC with PNN**
GTZAN Dataset	88	95.4	95.7	96.2
MTG Dataset	87.9	91.8	94.5	97.3

instruments such as cello, clarinet, flute, acoustic guitar, electric guitar, Hammond organ, piano, saxophone, trumpet, violin and singing voice and two classes of drums and no-drums. The class labels are applied to the predominant instrument over a 3-second snippet of polyphony music.

5. Conclusion and Future Work

The conclusion and future work of the proposed approach are discussed in this section. An efficient approach for segmentation, feature extraction and classification of audio signals is presented in this paper. Audio segmentation is performed by extracting the signal amplitude between the lengths of sample time. From this segmented output, EMFCC is applied to extract testing feature for the classification process, along with the combination of peak estimated signal feature. This extracts 41 number of feature vectors for the audio signal. PNN classifier is used for classification of audio signal. From this classification result, the category of given audio input is specified. The audio signal is classified as a musical or non-musical signal, based on the testing feature. If it is detected as a musical signal, the label is classified as Piano, Guitar, etc.

The proposed approach achieves better performance in terms of precision, NMI, F-score and entropy. The FRR, FAR, GAR, sensitivity, specificity and accuracy of the PNN classifier are higher with respect to the number of classes. In future, the audio signal is segmented from the given input and various frequencies presented in single audio input are separated. Then, the separated frequency is retrieved by classifying features of segmented signal frequency.

References

[1] Castán, D., Tavarez, D., Lopez-Otero, P., Franco-Pedroso, J., Delgado, H., Navas, E., *et al.* (2015) Albayzín-2014 evaluation: Audio segmentation and classification in broadcast news domains. *EURASIP Journal on Audio*, *Speech*, *and Music Processing*, **2015**, 33. http://dx.doi.org/10.1186/s13636-015-0076-3

[2] Kubala, F., Jin, H., Matsoukas, S., Nguyen, L., Schwartz, R. and Makhoul, J. (1997) The 1996 BBN Byblos HUB-4 Transcription System. *Proceedings of the* 1997 *DARPA Speech Recognition Workshop*, Chantilly, VA, 2-5 February 1997, 90-93.

[3] Bakis, R., Chen, S., Gopalakrishnan, P., Gopinath, R., Maes, S., Polymenakos, L. and Franz, M. (1997) Transcription of Broadcast News Shows with the IBM Large Vocabulary Speech Recognition System. *Proceedings of the Speech Recognition Workshop*, Chantilly, February 1997, 67-72.

[4] Beigi, H.S. and Maes, S. (1998) Speaker, Channel and Environment Change Detection. *Proceedings of the World Congress on Automation*, Anchorage, AK, 18 May 1998, 18-22.

[5] Siegler, M.A., Jain, U., Raj, B. and Stern, R.M. (1997) Automatic Segmentation, Classification and Clustering of Broadcast News Audio. *Proceedings of DARPA Speech Recognition Workshop*, Chantilly, VA, 2-5 February 1997, 97-99.

[6] Zhang, X., Su, Z., Lin, P., He, Q. and Yang, J. (2014) An Audio Feature Extraction Scheme Based on Spectral Decomposition. *International Conference on Audio*, *Language and Image Processing* (*ICALIP*), Shanghai, 7-9 July 2014, 730-733.

[7] Patil, H.A., Madhavi, M.C., Jain, R. and Jain, A.K. (2012) Combining Evidence from Temporal and Spectral Features for Person Recognition Using Humming. In: Kundu, M.K., Mitra, S., Mazumdar, D. and Pal, S.K., Eds., *Perception and Machine Intelligence*, Springer, Berlin Heidelberg, 321-328. http://dx.doi.org/10.1007/978-3-642-27387-2_40

[8] Bhalke, D., Rao, C. and Bormane, D.S. (2014) Musical Instrument Classification Using Higher Order Spectra. *International Conference on Signal Processing and Integrated Networks* (*SPIN*), Noida, 20-21 February 2014, 40-45. http://dx.doi.org/10.1109/spin.2014.6776918

[9] Haque, M.A. and Kim, J.-M. (2013) An Enhanced Fuzzy C-Means Algorithm for Audio Segmentation and Classification. *Multimedia Tools and Applications*, **63**, 485-500. http://dx.doi.org/10.1007/s11042-011-0921-z

[10] Ludeña-Choez, J. and Gallardo-Antolín, A. (2015) Feature Extraction Based on the High-Pass Filtering of Audio Signals for Acoustic Event Classification. *Computer Speech & Language*, **30**, 32-42. http://dx.doi.org/10.1016/j.csl.2014.04.001

[11] Lefèvre, S. and Vincent, N. (2011) A Two Level Strategy for Audio Segmentation. *Digital Signal Processing*, **21**, 270-277. http://dx.doi.org/10.1016/j.dsp.2010.07.003

[12] Evangelista, T.L., Priolli, T.M., Silla, C.N., Angelico, B. and Kaestner, C. (2014) Automatic Segmentation of Audio Signals for Bird Species Identification. *IEEE International Symposium on Multimedia* (*ISM*), Taichung, 10-12 December 2014, 223-228. http://dx.doi.org/10.1109/ism.2014.46

[13] Dhanalakshmi, P., Palanivel, S. and Ramalingam, V. (2011) Classification of Audio Signals Using AANN and GMM. *Applied Soft Computing*, **11**, 716-723. http://dx.doi.org/10.1016/j.asoc.2009.12.033

[14] Haque, M.A. and Kim, J.-M. (2013) An Analysis of Content-Based Classification of Audio Signals Using a Fuzzy C-Means Algorithm. *Multimedia Tools and Applications*, **63**, 77-92. http://dx.doi.org/10.1007/s11042-012-1019-y

[15] Dhanalakshmi, P., Palanivel, S. and Ramalingam, V. (2011) Pattern Classification Models for Classifying and Indexing Audio Signals. *Engineering Applications of Artificial Intelligence*, **24**, 350-357. http://dx.doi.org/10.1016/j.engappai.2010.10.011

[16] Gergen, S., Nagathil, A. and Martin, R. (2015) Classification of Reverberant Audio Signals Using Clustered Ad Hoc Distributed Microphones. *Signal Processing*, **107**, 21-32. http://dx.doi.org/10.1016/j.sigpro.2014.04.034

[17] Bhat, A.S., Amith, V., Prasad, N.S. and Mohan, D.M. (2014) An Efficient Classification Algorithm for Music Mood Detection in Western and Hindi Music Using Audio Feature Extraction. 5*th International Conference on Signal and Image Processing* (*ICSIP*), Jeju Island, 8-10 January 2014, 359-364. http://dx.doi.org/10.1109/icsip.2014.63

[18] Gergen, S. and Martin, R. (2014) Linear Combining of Audio Features for Signal Classification in Ad-Hoc Microphone Arrays. 11 *ITG Symposium*; *Proceedings of Speech Communication*, Erlangen, 24-26 September 2014, 1-4.

[19] Pesek, M., Leonardis, A. and Marolt, M. (2014) Boosting Audio Chord Estimation Using Multiple Classifiers. *International Conference on Systems, Signals and Image Processing* (*IWSSIP*), Dubrovnik, 12-15 May 2014, 107-110.

[20] Srinivasa Murthy, Y. and Koolagudi, S.G. (2015) Classification of Vocal and Non-Vocal Regions from Audio Songs Using Spectral Features and Pitch Variations. *IEEE* 28*th Canadian Conference on Electrical and Computer Engineering* (*CCECE*), Halifax, 3-6 May 2015, 1271-1276. http://dx.doi.org/10.1109/ccece.2015.7129461

[21] Koolagudi, S.G. and Krothapalli, S.R. (2012) Emotion Recognition from Speech Using Sub-Syllabic and Pitch Synchronous Spectral Features. *International Journal of Speech Technology*, **15**, 495-511. http://dx.doi.org/10.1007/s10772-012-9150-8

[22] Geiger, J.T., Schuller, B. and Rigoll, G. (2013) Large-Scale Audio Feature Extraction and SVM for Acoustic Scene Classification. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), New Paltz, 20-23 October 2013, 1-4. http://dx.doi.org/10.1109/waspaa.2013.6701857

[23] Oh, S.Y. and Chung, K.-Y. (2014) Target speech Feature Extraction Using Non-Parametric Correlation Coefficient. *Cluster Computing*, **17**, 893-899. http://dx.doi.org/10.1007/s10586-013-0284-5

[24] Gajšek, R., Mihelič, F. and Dobrišek, S. (2013) Speaker State Recognition Using an HMM-Based Feature Extraction Method. *Computer Speech & Language*, **27**, 135-150. http://dx.doi.org/10.1016/j.csl.2012.01.007

[25] Anguera, X. (2012) Speaker Independent Discriminant Feature Extraction for Acoustic Pattern-Matching. *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Kyoto, 25-30 March 2012, 485-488. http://dx.doi.org/10.1109/icassp.2012.6287922

[26] Salamon, J., Rocha, B. and Gómez, E. (2012) Musical Genre Classification Using Melody Features Extracted from Polyphonic Music Signals. *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Kyoto, 25-30 March 2012, 81-84. http://dx.doi.org/10.1109/icassp.2012.6287822

[27] Alam, M.J., Kinnunen, T., Kenny, P., Ouellet, P. and O'Shaughnessy, D. (2013) Multitaper MFCC and PLP Features for Speaker Verification Using i-Vectors. *Speech Communication*, **55**, 237-251. http://dx.doi.org/10.1016/j.specom.2012.08.007

[28] Muthumari, A. and Mala, K. (2015) Computerized Methods for Audio Segmentation and Classification: Survey. *International Journal of Applied Engineering Research*, **10**, 26857-26870.

[29] Rajeswari, K.C. and Uma Maheswari, P. (2015) Feature Extraction and Analysis of Speech Quality for Tamil Text System using Fast Fourier Transform. *Australian Journal of Basic and Applied Sciences*, **9**, 349-356.

[30] (2015) 21 August 2015. http://www.dtic.upf.edu/~ffuhrmann/PhD/data/

[31] (2015) 21 August 2015. http://marsyasweb.appspot.com/download/data_sets/

[32] Ji, X., Han, J., Jiang, X., Hu, X., Guo, L., Han, J., *et al.* (2015) Analysis of Music/Speech via Integration of Audio Content and Functional Brain Response. *Information Sciences*, **297**, 271-282.

http://dx.doi.org/10.1016/j.ins.2014.11.020

[33] Su, L., Yeh, C.-C.M., Liu, J.-Y., Wang, J.-C. and Yang, Y.-H. (2014) A Systematic Evaluation of the Bag-of-Frames Representation for Music Information Retrieval. *IEEE Transactions on Multimedia*, **16**, 1188-1200. http://dx.doi.org/10.1109/TMM.2014.2311016

[34] Fu, Z., Lu, G., Ting, K.M. and Zhang, D. (2011) Music Classification via the Bag-of-Features Approach. *Pattern Recognition Letters*, **32**, 1768-1777. http://dx.doi.org/10.1016/j.patrec.2011.06.026

[35] Fuhrmann, F. (2012) Automatic Musical Instrument Recognition from Polyphonic Music Audio Signals. PhD Thesis, Universitat Pompeu Fabra, Barcelona.