Improved Musical Instrument Classification Using Cepstral Coefficients and Neural Networks

Shruti Sarika Chakraborty and Ranjan Parekh

1 Introduction

A recognition problem deals with the identification of a test data, comparing it against a set of train data stored in a training database, with which the system has been modeled. Classification is the categorization of objects (in this case, audio test samples) based on their similarities. Musical instrument identification is the process where a test audio signal derived from a musical instrument is matched to one of the sets of predefined classes of musical instruments which are trained with their respective sounds during a training phase. For robustness, many train samples encompassing variations in sound are used during the training phase. If there exist a number of test samples, derived from multiple instruments of different families, the system classifies each of them to their respective families based on what the machine has learnt during the training phase. This is known as musical instrument classification [1]. As the domain of musical instruments is very wide and expanding, manual cataloging is difficult and prone to errors. Automated musical instrument classification aids in extraction of melody from musical sound, identification and separation of sound sources in polyphonic audio, identification of solo musical instruments, automatic music transcription, beat tracking, musical information retrieval, and much more application of similar types [2]. The focus of this article is to improve on existing techniques available for musical instrument classification using multiple features and multiple classifiers and observing which produces optimal results. The organization of this paper is as follows: Sect. 2 highlights previous approaches, Sect. 3 outlines the various features and classifiers used in the proposed approach, Sect. 4 tabulates the experimental procedures and results obtained, Sect. 5 provides a comparative analysis of the present approach vis-à-vis other approaches, and Sect. 6 brings up the conclusions and future scopes of work.

S. S. Chakraborty (✉) · R. Parekh
School of Education Technology, Jadavpur University, Kolkata, India
e-mail: shrutisarikachakraborty@gmail.com

J. K. Mandal et al. (eds.), *Methodologies and Application Issues of Contemporary Computing Framework*, https://doi.org/10.1007/978-981-13-2345-4_10

2 Previous Work

There is a number of works related to musical instrument classification. MFCC features are proposed for identification tasks along with delta MFCC features which are obtained by taking the time derivative of MFCCs. SVM algorithm is popularly used for classification [2]. In other works, also, MFCC is used to extract the features of audio signals arising from musical instruments and are paired with K-NN classifier for classification into five classes which are Cello, Piano, Trumpet, Flute, and Violin. The overall accuracy was more than 80% with 90 samples in train set and 60 for the test set. The sound samples were obtained from Electronic Music Studio, University of Iowa, and accuracy is prone to decrease with increase in the number of instruments and instruments in polyphonic recordings will not be distinguished by this method [1]. An algorithm proposed for classification task involved SVM, MLP, and AdaBoost where AdaBoost gave the best result with an accuracy of more than 90% [3]. The use of MFCC- and Timbral-related audio descriptors for the identification of musical instrument is very common. K-nearest neighbor, support vector machine, and binary tree are used for classification purpose. The accuracy was found to decrease from 90 to 75% with an increase in the number of instruments from 5 to 15 [4]. The use of MFCCs for the musical instrument identification is also overviewed in signal processing techniques for music analysis [5]. Musical instrument classification using wavelet dependent timescale features has been proposed. In this, at first, the continuous wavelet transform of the signal frame is taken and then features related to temporal variation and bandwidth are considered for feature extraction [6]. A preliminary work on ontology designed for musical instruments has been proposed. The paper also provided the investigation of heterogeneity and limitations in existing instrument classification schemes [7]. A method utilizing convolutional neural networks has been proposed. It produced high performances in their confusion matrix. The experiment used 11 instruments including cello and clarinet. The accuracy obtained was better than previous approaches [8]. Mel scale cepstral coefficient (MFCC)-based features coupled with a multilayer perceptron (MLP)-based classifier has been used for categorization of 2716 clips from seven different instruments obtaining an average accuracy of 98.38% accuracy [9]. Recent journals mostly dealt with instrument identification from polyphonic sources which is beyond the scope of this study. However, extraction of the timbre of an instrument utilizing Bayesian network achieved an accuracy of over 95%. The timbre was extracted using a set of features. The feature is the amplitude of the frequency peak with the highest magnitude within each window. The FFT of the signal was partitioned into 10 exponentially increasing windows [10]. A method was proposed for musical instrument identification using short-term energy and ZCR. ZCR is the count of how many times signal changes the sign [11].

3 Proposed Approach

The classification of musical instruments has many challenges due to their multidimensional characteristics. The musical instruments also vary in shape, sizes, types, geographical locations, cultures, and playing style all of which makes it extremely difficult for a common person to categorize them reliably. As the machine learns from its train data, hence, it is required to input a large number of sound samples as train samples ranging in all variety for each instrument to the system in order to attain maximum accuracy for major problem faced in this case that is low accuracy as number of musical instrument increases in train set, the accuracy decreases. Musical instrument classification is a very important task for musical information retrieval system, audio source separation, automatic music transcription, and genre classification as the domain of musical instrument is very wide. The goal of this paper is to maximize the accuracy of musical instrument classification.

For the musical instrument classification task, in the initial stage after preprocessing, the feature vectors or acoustic vectors of all six audio descriptors are extracted. MFCC, LPC, and spectral centroids are spectral descriptors, while pitch salience, HPCP, is tonal descriptors. The CC is categorized into spectral descriptors according to its characteristic nature. Then, all the features are individually inputted to four different classifiers which are K-NN or K-nearest neighbors (number of neighbors ranging between 1 and 5 and metric Euclidian distance), support vector machine (SVM) (linear), artificial neural network (ANN) (with Softmax and Rectifier activation function), and random forest (number of decision trees equal to 10,000). The performance of six features against all four classifiers is studied, and the set of best feature and classifier is chosen to be cepstral coefficients and ANN as the accuracy of cepstral coefficient as more than all others in maximum cases. The reason behind the choice of the best classifier is discussed later. Three sets of experiments were performed with varied conditions for each case and the maximum overall accuracy ranged between 90 and 93%. The overall block diagram of two processes is given in Fig. 1.

3.1 Preprocessing

The sound samples used in musical instrument classification are sampled at 44,000 kHz. They are noise free. Ten instruments are used which belongs to five different families. The instruments are organ, French horn, cello, clarinet, tambourine drum flute, trumpet, violin, and piano. The train samples are of duration 5–10 s each while the test samples range from 10 to 20 s. Fourteen to fifteen train samples are used in train set per instrument or class and three test samples are used per instrument in the test set in each of three experiments.

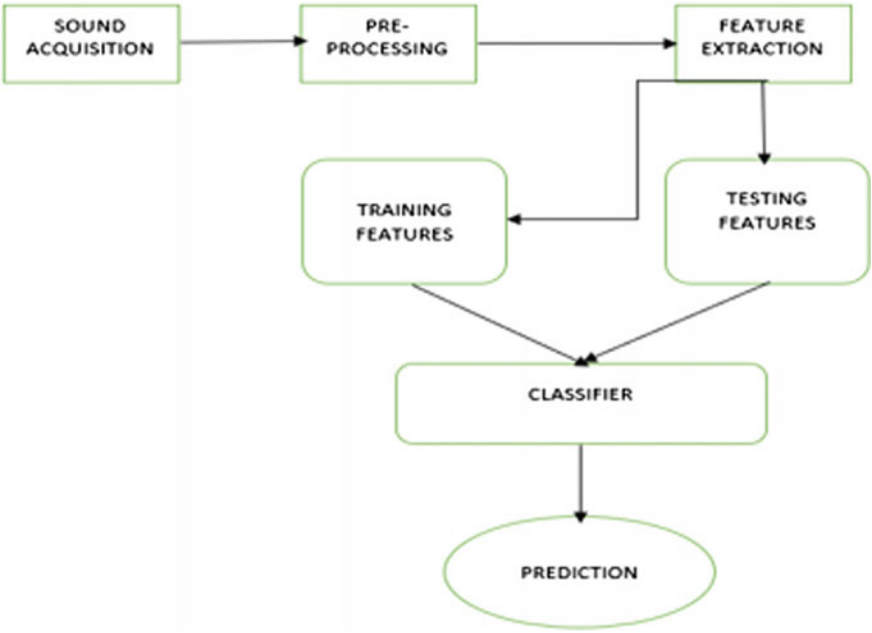


Fig. 1 Overall block diagram of overall process

3.2 Feature Extraction

The features extracted for musical instrument classification are harmonic pitch class profile (HPCP), linear predictive coding (LPC) coefficients, Mel frequency cepstral coefficients (MFCC), spectral centroids, pitch salience peaks, and cepstral coefficients (CC). A brief description of their characteristics is listed below.

Mel frequency cepstral coefficients (MFCC): MFCCs are useful in identifying the linguistic content and timbre of the sound discarding the background noise, emotion, etc. MFCCs are commonly used in speech recognition and are finding increased use in music information recognition and genre classification systems and also speaker recognition. It is evident from psychological studies that the human perception of the contents of the frequency of sound for speech signals does not follow a linear scale. For each tone with an actual frequency f measured in Hz, a subjective pitch on a scale being measured is called the "Mel" scale. This Mel frequency scale M follows a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [1]. See Eq. (1).

$$M = 2595.\log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

(Linear Predictive Coding- LPC) coefficients: A filter that approximates spectral characteristics of a sound is the filter which is approximated by LPC coefficients. The LPC coefficients give a set of filtered coefficients (a_k), and the frequency response of the resulting filter approximates spectrum of the input sound. So, a signal can be approximated by LPC model as the linear combination of past samples which is basically the expression of IR filter (infinite response filter) that is a linear combination of previous samples. The goal of LPC is to find the coefficients that best approximates the signal in question.

Spectral Centroids: The centroid, of the frequency spectrum, is the human perception of "brightness". It is derived by multiplying the value of each frequency by its magnitude in the spectrum, then taking the sum of all these and again dividing the whole numerator by the magnitude in the spectrum of the signal. The centroid is the descriptor feature that helps in the characterization of the spectral shape of a particular sound. See Eq. (2) where k = frequency, X[k] = magnitude in the spectrum.

$$\text{Centroid} = \frac{\sum_{k=0}^{N/2} k|X[k]|}{\sum_{k=0}^{N/2}|\text{X}[k]|} \tag{2}$$

Pitch Salience Peaks: The salience function is given by Eq. (3) where S[b] = salience at bin frequency b, e() = magnitude threshold function, g() = weighting function applied to peak p, β = magnitude compression value, A_p = amplitude of peak, P = number of peaks, and H = number of harmonics. The spectral peaks which are extracted from the spectrum of the signal are used to construct a salience function which is a representation of pitch salience over timescale. The function contains peaks which are F0 candidates for the main melody. The idea of peak salience relates to how much of a peak is present at a particular frame of the sound sample.

$$\text{S[b]} = \sum_{\text{h=1}}^{\text{H}} \sum_{\text{p=1}}^{\text{P}} \text{e(Ap)g(b, h, f p)(Ap)}^{\beta} \tag{3}$$

Harmonic Pitch Class Profile(HPCP): HPCP is a vector which represents the intensities of the twelve semitone pitch classes (corresponding to notes from A to G#). It is a group of features which is extracted from a sound signal, based on a pitch class profile. HPCP are features that are pitch distributed and are sequences of feature vectors which describes tonality and measures the relative intensity of each of the 12 pitch classes of the equal-tempered scale within the frame under analysis. They are often referred to as chroma. This is explained using Eq. (4) where A_p = amplitude of spectral peak p, P = total number of peaks, w (k, f_p) = weight of the peak frequency f_p for bin k, and k = spectral bin locations of the chosen HPCP frequencies. HPCP is the sum of the weighted square of amplitudes of peaks along all peaks. Twelve HPCP filters are taken to represent a sound sample.

$$\text{HPCP[k]} = \sum_{p=1}^{P} w\left(k, f_p\right) A_p^2 \tag{4}$$

Cepstral Coefficients(CC): These features are not used extensively but only in few research articles [12]. These features are extracted with a view to extracting the overall spectral characteristics of the signal which unlike MFCC does not obstruct spectral regions and in fact this is an improved version of MFCC for overall sound recognition problems as MFCC was made especially for speech recognition problems as the phonemes uttered by human are better captured by application of Mel scale which gives more emphasis to areas of lower frequencies. The CC represents the timbre of the sound by the envelope of the spectrum. Again, the short-time power spectrum enhances the process of timber extraction from the sound. The timbre is the property which works independently of tone or pitch to identify sounds. Hence, correct representation of it will lead to better identification. The cepstral coefficients are simply DCT of the log of short-time power spectrum given by Eq. (5).

$$\text{CC} = \text{DCT}\left(\log\left(\frac{1}{N}|X_k|^2\right)\right) \text{ where } X_k = \sum_{i=0}^{N-1} x_i \cdot e^{-j \cdot \frac{2\pi ki}{N}} \tag{5}$$

Here, X_k is the FFT of a hamming windowed signal applied on frame size 25 ms overlapped by 10 ms. In a typical case, the signal in a frame, denoted by (n), where $n = \{0, \ldots, N-1\}$, and after windowing the signal is given by $s(n) * t(n)$, where $t(n)$ is the representation of Hamming window which is given by Eq. (6)

$$t(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right); 0 \leq n \leq N-1 \tag{6}$$

3.3 Classification

This is a vital step in any recognition problem. In this step, the features from the test data and train data are compared against each other to measure similarity in order to achieve classification. There are many classifiers present but the choice of most effective one for a specific problem helps in achieving the goal of the system and has better accuracy compared to others while also preventing overfitting. In this study, four major and most prevalent machine learning classifiers are used which are K-NN, SVM, ANN, and random forest. After observing the effectiveness, efficiency, robustness, and reproducibility, the choice of the best classifier is made.

K-nearest neighbors (K-NN)

The K-nearest neighbors algorithm (k-NN) is a popular nonparametric algorithm used for solving classification as well as regression problems. Here, k-NN is used as a classifier. The input to k-NN is two vectors each deriving from test set and train

set. For CC, each vector is of size [1 × 2560]. As 40 coefficients are taken and 64 centroids are taken in Vector quantization step for codebook formation, [40 × 64 = 2560] is the size. The output of a k-NN classifier is a class membership. An object is classified by a majority vote of its surrounding neighbors. That means an object is assigned to the class that consists of the highest number of common elements among its nearest neighbors. The number of neighbors of K-NN was between 1 and 5 with metric as the Euclidian distance in all cases. For two n-dimensional vectors $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, Euclidean distance is defined as

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{7}$$

Support Vector Machine (SVM)
Support vector machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression problems. However, it is mostly used in solving classification tasks. Each data item is plotted as a point in an n-dimensional space (where n is the number of features) with the value of each feature being the value of a particular coordinate. Classification is performed by finding the hyperplane that differentiates the two classes with the maximum margin. The sum of distances between two boundary points of two classes has to be maximum in order to choose that hyperplane as an optimal hyperplane. One versus rest method is used for multi-class classification. In this case, we choose the class which classifies the test datum with greatest margin. The SVM used here was linear.

Random Forest
Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks, which operate by constructing many decision trees at training time and outputting the class based on maximum votes achieved from the number of decision trees used. The number of trees chosen in this case is 10,000.

Artificial Neural Networks (ANN)
An ANN is based on a collection of connected units or nodes called artificial neurons. Each connection between artificial neurons is capable of transmitting a signal from one to another. The artificial neuron that receives the signal processes it and then it signals artificial neurons which are connected to it. The input nodes connected to the layer of neurons connect again to the next hidden layer of neurons (if the number of hidden layers>1) and after connecting to n hidden layers (specified by the user) at last the network connects to the output layer. Now there is weight associated to each connection or synapse and that gets modified due to backpropagation in order to minimize error. So, in this study, multilayer perceptron (MLP) is used. A multilayer perceptron (MLP) consists of more than two layers of nodes and also utilizes a supervised learning technique known as backpropagation for training purposes. It is distinguished from linear perceptron by its multiple layers and nonlinear activation function. In this study, four hidden layers are used apart from input and output layer.

The rectifier activation function is used for input and hidden layers while Softmax activation function which can be vaguely said as the categorical version of sigmoid function where sigmoid is used for binary classification and Softmax function is used for multi-class classification is used at the output layer. The loss function used here is categorical cross entropy and Adam optimizer is used. The loss in case of CC and MFCC for all three cases was in the order of e^{-6}. The rectifier function is given by Eq. (8), where x is the independent variable of input from input/hidden layers.

$$\Phi_{\mathrm{x}} = max(x, 0) \tag{8}$$

The Sigmoid function is given by

$$\Phi_x \frac{1}{1 + e^{-x}} \tag{9}$$

The Softmax function is given by Eq. (10):

$$\sigma(\mathrm{z})_{\mathrm{j}} = \frac{\mathrm{e}^{\mathrm{zj}}}{\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{e}^{\mathrm{zk}}} \mathrm{for} j = 1, \ldots, K \tag{10}$$

Here

$$\mathrm{z} = w_0 \chi_0 + w_1 x_1 + \cdots w_m x_m \tag{11}$$

where w is the weights of the input variable and j = 1, 2…K are the categories or classes. Softmax function calculates the distribution of probability of the event over "n" different events. Later, the calculated probabilities serve helpful for the purpose of determining the target class for the given inputs.

The accuracy obtained with the Cepstral coefficients (40 coefficients are taken, as it provides best accuracy and reproducibility) along with ANN surpassed the accuracy of previous approaches conducted with MFCCs and that is discussed in next section. The decrease of accuracy with an increase in a number of train data is found to be less and the feature is also robust for it provided good accuracy in all cases. ANN and K-NN both provided very good accuracy but ANN is chosen over K-NN for it can be tuned and accuracy can be improved further.

4 Experimentations and Results

This section is a major part of this study as it consists of all the results and plots that has led us to the conclusion. It is this part, which has helped us in choosing appropriate feature and classifier for the system. The experiments are mostly done with Python code in Spyder environment. The dataset for musical instrument classification is derived from Philharmonia Orchestra instruments [13] and Freesound

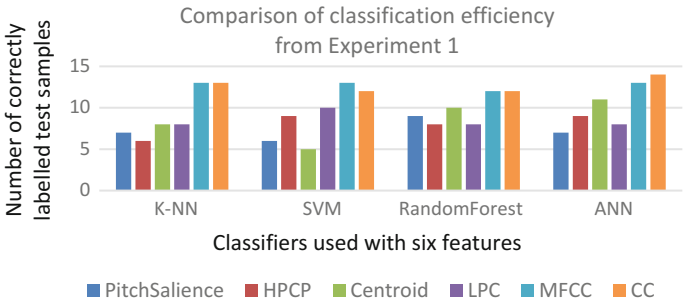


Fig. 2 Comparison of classification efficiency from experiment 1

music database [14]. The pitch of the instruments ranged from A#3 to G5, and the dynamic was of two types, Forte and Piano. The articulation was normal. Dynamics investigates how loud or quiet the sound is. In music, articulation is the performance technique which affects the transition or continuity on a single note or between multiple notes or sounds. Pitch may be quantified as a frequency. From the whole set, experimentation is done only with ten instruments. They are organ, French horn, cello, clarinet, tambourine drum, flute, trumpet, violin, and piano. They belong to five families of instruments. Three experiments have been carried out varying the characteristics of sounds. Two experiments are carried out with the same set of five instruments with varying conditions to study the effect of change of pitch and dynamics in the classification process and the third experiment is made with 10 instruments. The instruments belong to the following families.

Keyboards and Harp:	Organ, Piano
Woodwind Family:	Flute, Clarinet
Brass Family:	Trumpet, French Horn
String Family:	Violin, Cello
Percussion Family:	Drum, Tambourine

Experiment 1:
The first experiment was carried out with drum, flute, trumpet, violin, and piano which belongs to five different families of musical instrument mentioned earlier. For each instrument or class, 15 samples were taken during training and 3 samples were taken for testing, so a total of 75 samples were taken during train phase and 15 samples were used in the test phase to check the accuracy and robustness of the system. The samples belonged to random pitch and dynamics. The experiment was carried out with six features and four classifiers as specified in the previous section. The features are MFCC, CC, LPC coefficients, Centroid, HPCP, and Pitch salience peaks. The classifiers are K-NN, ANN, random forest, and SVM. The result obtained showed maximum accuracy with CC. The result is graphically plotted in Fig. 2.

The results show that the accuracy is quite low for Pitch salience peaks, HPCP, LPC coefficients, and centroid. It can be observed that maximum accuracy is obtained

Table 1 The confusion matrix obtained with CC and ANN for experiment 1

Actual	Predicted				
	Drum	Flute	Trumpet	Violin	Piano
Drum	3	0	0	0	0
Flute	0	2	1	0	0
Trumpet	0	0	3	0	0
Violin	0	0	0	3	0
Piano	0	0	0	0	3

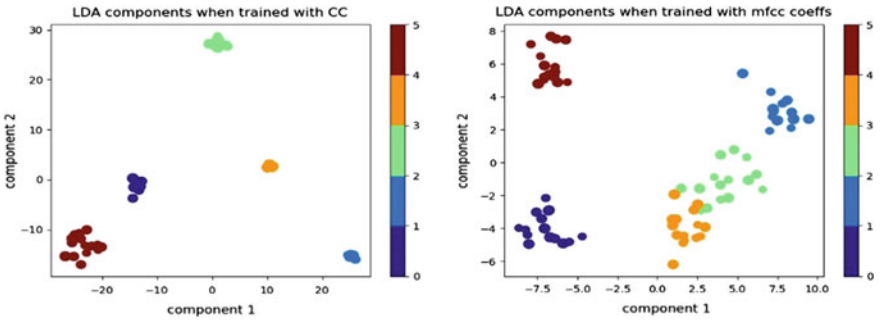


Fig. 3 LDA plots depicting clustering and overlapping of train samples for CC and MFCC

against ANN followed by K-NN for CC and MFCC. The highest number of the correctly labeled test sample is 14 with CC, while it is 13 for MFCC (Table 1).

Since the accuracy obtained for MFCC and CC is very close, hence, the linear discriminant analysis (LDA) plots are used to decide the best classifier to observe the degree of uniqueness imparted by each during the training phase (Fig. 3).

The above plots clearly show that the cluster formed by CC is tighter and farther apart from each other than MFCC. There is some overlapping in case of MFCC but there is no overlapping in case of CC.

Experiment 2:
The second experiment was carried out with the same instruments but each musical instrument had random dynamics but a fixed pitch with six features and four classifiers: Drum—Unpitched, Flute—B4 and B5, Trumpet—C5 and C6, Violin—A3 and A4, and Piano—D4. The pitch information of each instrument is different. The number of train samples used here is 14 per class and 3 test samples per class are used. This was done deliberately to observe the capability of classification of HPCP and pitch salience especially. If CC and MFCC give good accuracy under this condition also, it can be inferred that they are robust features as the train samples and test samples are different from the previous case. The result obtained showed maximum accuracy with CC. The result is graphically plotted in Fig. 4.

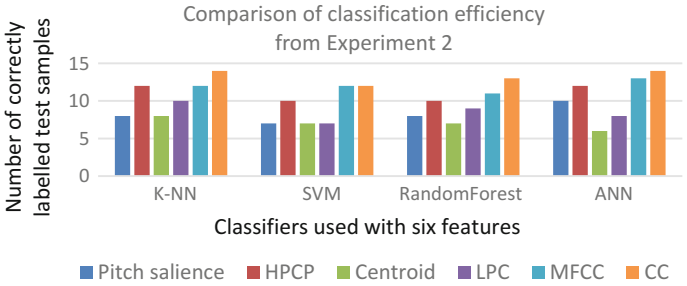


Fig. 4 Comparison of classification efficiency from experiment 2

Table 2 The confusion matrix obtained with CC and ANN for experiment 2

Actual	Predicted				
	Drum	Flute	Trumpet	Violin	Piano
Drum	3	0	0	0	0
Flute	0	3	0	0	0
Trumpet	0	0	3	0	0
Violin	0	0	0	2	1
Piano	0	0	0	0	3

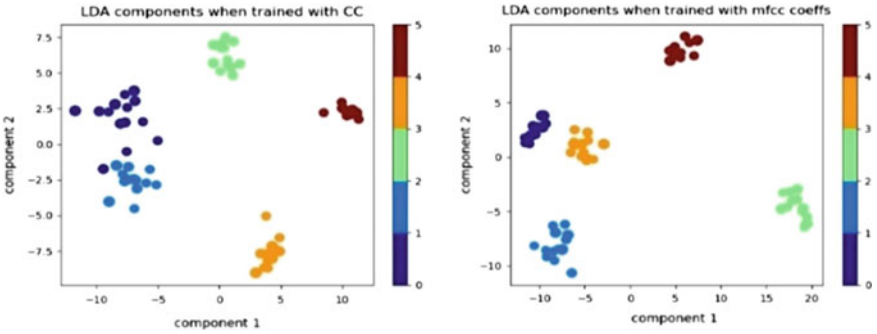


Fig. 5 LDA plots depicting clustering and overlapping of train samples for CC and MFCC

As observed from this plot, the maximum accuracy is obtained for CC with ANN. The highest number of the correctly labeled test sample is 14 with CC, while it is 13 for MFCC. The other features have much lesser accuracy (Table 2).

Since the accuracy obtained for MFCC and CC are very close, hence, the LDA plots are used to decide the best classifier to observe the degree of uniqueness imparted by each during the training phase (Fig. 5).

The above plot clearly shows that the cluster formed by CC is tight and far apart from each other and so does MFCC. In this experiment due to varying pitch characteristics, the classification is even better for MFCC. In order to come to a conclusion

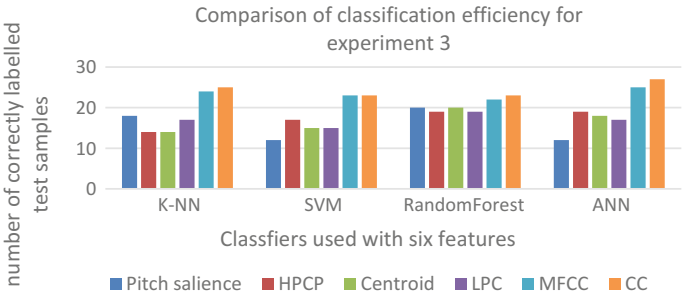


Fig. 6 Comparison of classification efficiency from experiment 3

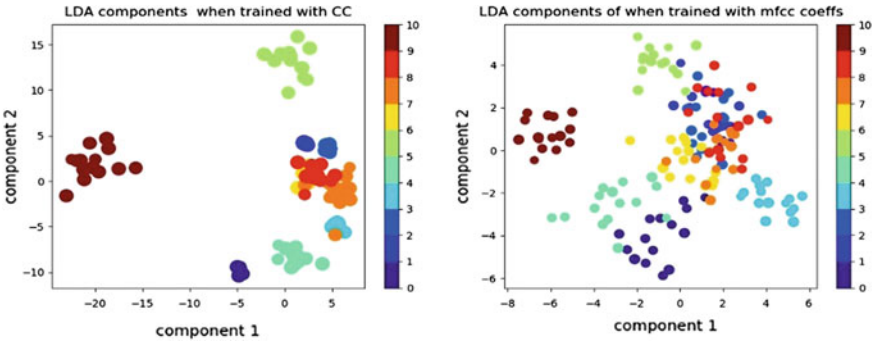


Fig. 7 LDA plots depicting clustering and overlapping of train samples for CC and MFCC

of the best feature in next experiment, number of the instrument has been increased to ten in order to observe differences.

Experiment 3:
The third experiment is carried out with ten instruments which are organ, French horn, cello clarinet, tambourine drum flute, trumpet, violin, and piano. For each instrument or class, 15 samples were taken during training and 3 samples were taken for testing, so a total of 150 samples were taken during train phase and 30 samples were used in the test phase to check the accuracy and robustness of the system. The pitch and dynamics are random. The result is graphically plotted in Fig. 6.

As observed from this plot, the maximum accuracy is obtained for CC with ANN. The highest number of the correctly labeled test sample is 27 with CC, while it is 25 for MFCC. The other features have much lesser accuracy. This shows that the accuracy was maximum in case of CC which is followed by MFCC when compared with ANN and K-NN. The LDA plots used for further study are given in Fig. 7.

Some overlapping is observed for MFCC, and the cluster boundary is not well defined so there are chances of false mismatch. Classes 9, 1, 2, and 8 are overlapping quite a lot (in case of MFCC) and hence with an increase in a number of instruments accuracy will fall much steeply. Clearly, CC is the best feature as the overlapping is

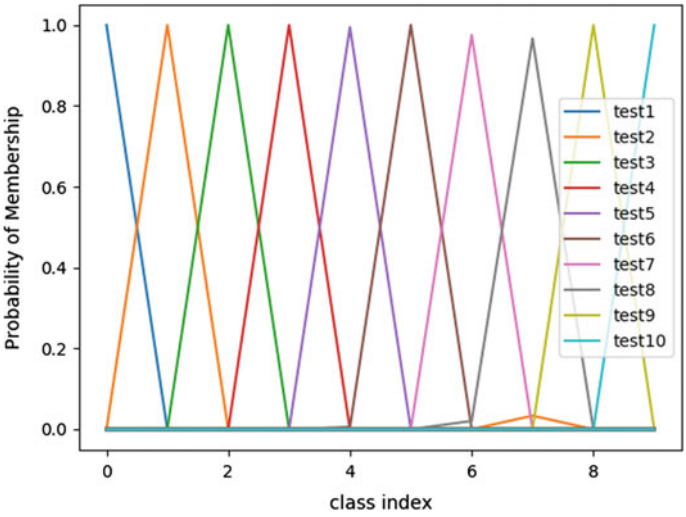


Fig. 8 Classification plot for test samples

much less compared to MFCC and the clusters are tight and well defined. Only classes 9 and 8 have some overlapping; otherwise, other classes are well distinguished. So, the accuracy will fall much less steeply when compared to MFCC and it has the potential for better results. Hence, observing the above results CC is chosen as the final feature and ANN as the final classifier for this set provided best accuracy and the feature is robust as can be observed from LDA plots and accuracy with an increase in train data decreases less steeply than MFCC which provided second best result in terms of accuracy. The test samples are classified into the class with which it received the highest probability of likeliness. The plot below shows how each of ten test samples belonging to ten different classes are correctly classified based on the maximum probability obtained when cepstral coefficients are used as features and ANN is used as a classifier (Fig. 8).

It can be observed that the test samples are matched to the class with which it has maximum probability. In the above case, the accuracy obtained is 100% and hence each test class is classified into its respective train classes.

5 Analysis

The proposed approach uses cepstral coefficients as features and ANN as the classifier for the musical instrument classification task. The overall accuracy obtained in experiment 1 and 2 is 14/15 as 14 test samples from a total of 15 samples are correctly classified. Hence, the accuracy is given by Eq. (12)

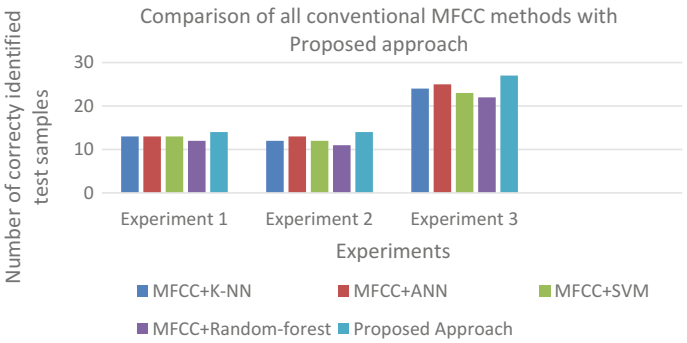
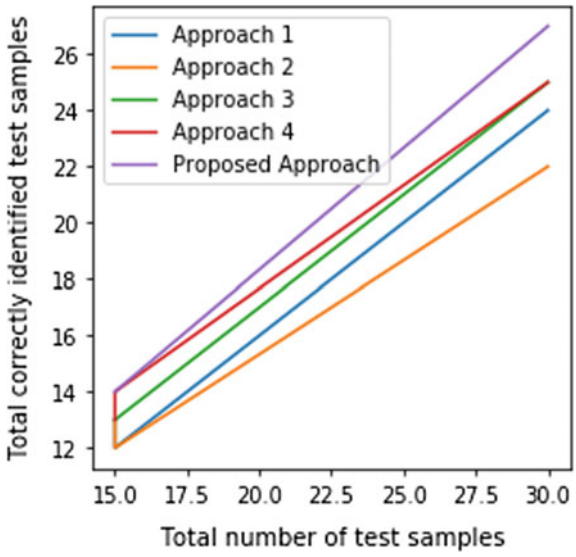


Fig. 9 Comparison of all conventional MFCC methods with proposed approach

Fig. 10 Comparison of earlier approaches with proposed approach



$$\text{Accuracy} = \frac{\textit{Number of correctly classified test samples}}{\textit{Total number of test samples}} \times 100 \qquad (12)$$

Hence, in the first and second case, overall accuracy is 93% (14/15) while in third case 27 out of 30 samples are correctly classified, and hence overall accuracy is 90% for the proposed approach. For MFCC, in first two cases, overall accuracy is 86% as 13 out of 15 test samples are correctly classified, while in third case it is 83% as 25 out of 30 samples were correctly classified by it. The accuracy of proposed approach is larger than all other combinations of MFCC (Fig. 10).

The above plot shows the comparison of earlier approaches [1, 4, 9, 10] with the proposed approach for three experiments. The proposed method obtained with CC and ANN has the following advantages over other methods.

1. The accuracy obtained with MFCC across all classifiers used in related papers and timbre related other feature used in [10] was less than the accuracy obtained with the proposed approach [1, 4, 9] (Fig. 9).
2. ANN is chosen as final classifier as with better tuning, improvement of accuracy is possible. K-NN can be used as an alternative to avoid tuning.
3. The rate of decrease of accuracy with CC is much less compared to MFCC, as observed from graph and LDA plots. Hence, this limitation is somewhat reduced [1, 4].
4. The feature is robust as across three experiments in varied conditions, it provided least overlapping of train samples and maximum accuracy of test samples.
5. For CNN or convolutional neural network [8], the number of train and test data required is very high. The method will fail in situations where there is a limited number of data. This is not the problem with the proposed approach.
6. CC is the best classifier for musical instrument classification among LPC, HPCP, MFCC, pitch salience, and centroid as observed from experimental results and LDA plots.

6 Conclusions and Future Scopes

It is hence concluded that the proposed approach is not only capable of classifying ten aforementioned musical instruments of five different musical families better than conventional approaches dealing with MFCC but also better than other features like LPC coefficients, HPCP, pitch salience peaks, and spectral centroids. It can work in situations where data is limited unlike CNN-based approaches and the decrease of accuracy with an increase in the number of instruments is less than all other features mentioned above. The approach is robust to changes in pitch and dynamics of musical instruments. The limitations of this method are that it cannot distinguish instruments playing in polyphonic audio which will be dealt with in future.

References

1. M.S. Nagawade, V.R. Ratnaparkhe, Musical instrument identification using MFCC, in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT),* Bangalore (2017), pp. 2198–2202
2. S. Essid, G. Richard, B. David, Instrument recognition in polyphonic music based on automatic taxonomies. IEEE Trans. Audio Speech Lang. Process. **14**(1), 68–80 (2006)
3. S.D. Patil, P.S. Sanjekar, Musical instrument identification using SVM, MLP & AdaBoost with formal concept analysis, in *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, Aurangabad (2017), pp. 105–109
4. P.S. Jadhav, Classification of musical instruments sounds by using MFCC and timbral audio descriptors. Int. J. Recent Innov. Trends Comput. Commun (IJRITCC) **3**(7), 5001–5006 (2015)
5. M. Muller, D.P.W. Ellis, A. Klapuri, G. Richard, Signal processing for music analysis. IEEE J. Sel. Top. Signal Process. **5**(6), 1088–1110 (2011)

6. F.H. Foomany, K. Umapathy, Classification of music instruments using wavelet-based time-scale features, in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, San Jose, CA (2013), pp. 1–4
7. M. Abulaish, *Ontology Engineering for Imprecise Knowledge Management* (Lambert Academic Publishing, Saarbrucken, Germany, 2009)
8. Y. Han, J. Kim, K. Lee, Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 208–221 (2017)
9. H. Mukherjee, S.M. Obaidullah, S. Phadikar, K. Roy, SMIL—a musical instrument identification system, in *Computational Intelligence, Communications, and Business Analytics. CICBA 2017. Communications in Computer and Information Science*, ed. by J. Mandal, P. Dutta, S. Mukhopadhyay, vol. 775 (Springer, Singapore, 2017)
10. P.J. Donnelly, J.W. Sheppard, Classification of musical timbre using Bayesian networks. Comput. Music J. **37**(4), 70–86 (2013)
11. S.K. Banchhor, A. Khan, Musical instrument recognition using zero crossing rate and short-time energy. Int. J. Appl. Inf. Syst. (IJAIS) **1**(3), 16–19 (2012)
12. J.L.C. Loong, K.S. Subari, M.K. Abdullah, N.N. Ahmad, R. Besar, Comparison of MFCC and cepstral coefficients as a feature set for PCG biometric systems. World Acad. Sci. Eng. Technol. Int. J. Biomed. Biol. Eng. **4**(8) (2010)
13. Philharmonia Orchestra Sound Samples, www.philharmonia.co.uk/explore/sound_samples
14. Freesound, http://freesound.org/