LONDON JOURNAL OF SOCIAL SCIENCES

Musical Instruments Sound Classification using GMM

P. Aurchana
Department of Master of Computer Applications
Sri Manakula Vinayagar Engineering College
Puducherry, India

S. Prabavathy
Department of Computer Science
A. P. C. Mahalaxmi College for Women Thoothukudi
Tamilnadu, India

Abstract

Classification is the task of assigning objects to one of several predefined categories. In today’s decade classifying the musical signal from large data is a major task; the proposed work classifies the music into their respective classes. In this paper, the sound of the musical instruments classified automatically from the musical signals. Mel frequency cepstral coefficient is used as a feature extractor and the machine learning model namely Gaussian Mixture Model is used for classification. This system tested in ten different classes of musical instrument sound from two different instrument families such as Woodwind and Brass instruments. In this proposed work, the result yields satisfactory accuracy in the classification of musical instruments sound.

Keywords: Musical Instrument Sound Classification (MISC), mel frequency cepstral coefficient (MFCC), Gaussian Mixture Model (GMM).

1. Introduction

Classification is more useful in the context of constructing vast audio collections that have been investigated, because the assigned class labels are directly displayed to the user and applied as a filter. As a result, it is used in an indirect way for music recommendation, where similarity may be determined for all labels based on advanced listening habits, and pick songs to listen to from the classes where a user has the most similarity. The task of automatically classifying the musical instruments is difficult. In the digital era, musical data classification has become a very prominent research topic. The classification of musical instruments was a lengthy manual process. This approach divides musical instruments into categories based on acoustic characteristics such as MFCC, Sonogram, and MFCC combined with Sonogram. The characteristics are classified using two modelling techniques: SVM and kNN. In this research, we use modern algorithms to classify musical instruments based on their attributes that are retrieved from diverse instruments. The suggested research compares and contrasts the performance of kNN and SVM. SVM and kNN classifiers are used to identify musical instruments and compute their accuracy [4].

Everyone in the present era listens to and plays music. Music is diverse all around the world. It is the fulcrum of all the arts and a language that speaks for itself. We might argue that this immaculate art's vast history extends to infinity and beyond. It would be more interesting if there was a method for us to learn about the instruments that are used in the song. As a result, may categorise music depending on certain instruments. Researchers have been actively involved in human perception towards the study of Musical Instruments for the past two decades [5].

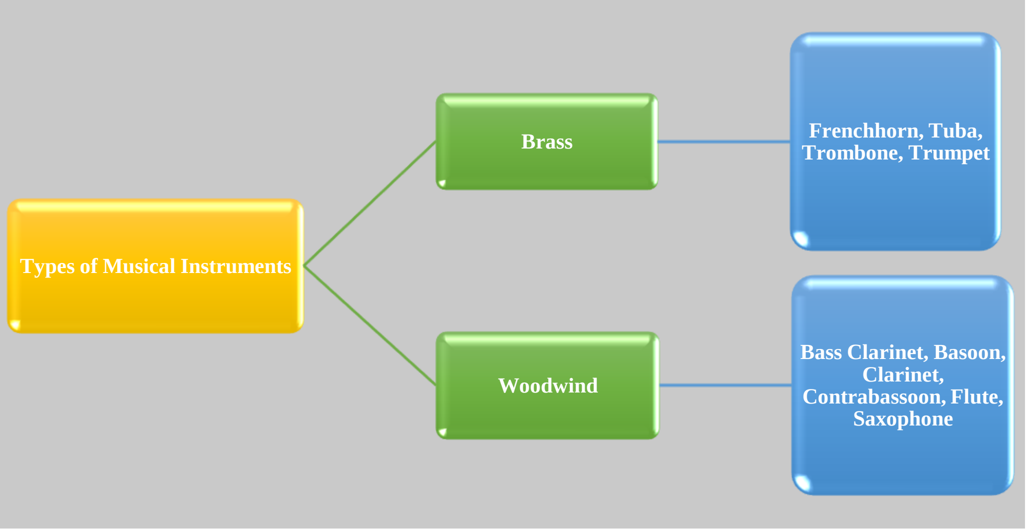
Types of Musical Instruments
Brass
Woodwind
Frenchhorn, Tuba, Trombone, Trumpet
Bass Clarinet, Basoon, Clarinet, Contrabassoon, Flute, Saxophone

Fig. 1 Musical Instruments Types

In this proposed work, the musical instruments sound classification is done in three steps, first is the preprocessing of musical data, then the features are extracted and finally the classification process. Fig. 1 shows the types of musical instrument sound which is going to be categorized. The sound of the musical instruments used in this paper are French Horn, Tuba, Trombone, Trumpet from Brass and Bass Clarinet, Bassoon, Clarinet, Contrabassoon, Flute, Saxophone from Woodwind instrumentsFig.2 shows the proposed work of the system.

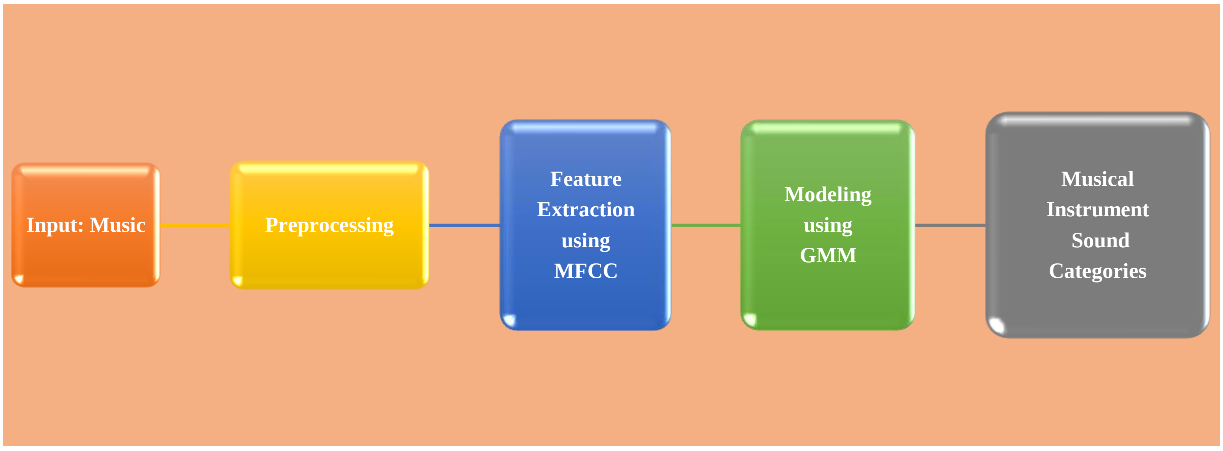
Input: Music
Preprocessing
Feature Extraction using MFCC
Modeling using GMM
Musical Instrument Sound Categories

Fig. 2 Block diagram of the proposed work

2. Review of Literature

Digital audio applications are already a common place aspect of our lives. Audio classification can be a useful method for managing content. If an audio clip can be automatically categorized, it may be saved in an organized database, greatly improving audio management. We present effective algorithms in this research for automatically classifying audio recordings into one of six categories: music, news, sports, advertisement, cartoon, and movie. To characterize the audio content for these categories, a number of acoustic features such as linear predictive coefficients, linear predictive cepstral coefficients, and mel-frequency cepstral coefficients are retrieved. The distribution of auditory feature vectors is captured using the Auto Associative Neural Network model (AANN). The AANN model captures the distribution of a class's acoustic features, and the weights of the network are adjusted using the backpropagation learning algorithm to minimise the mean square error for each feature vector. The suggested technique additionally compares the performance of AANN with that of a GMM, in which feature vectors from each class were utilized to train GMM models for those classes. The likelihood of a test sample belonging to each model is calculated during testing, and the sample is assigned to the class whose model produces the highest likelihood.

In music indexing, automatic music genre classification is quite useful. Tempogram is a feature extraction method based on the temporal structure of music data that is used in the classification of musical genres. The major aspects of today's music genre classification system are searching and arranging. This research describes a new methodology for classifying music that employs support vector machines. The Gaussian mixture model learns from training data to classify music audio into its appropriate categories. The suggested feature extraction and classification methods improve music genre classification accuracy [2].

This research provides a unique feature extraction approach based on Fractional Fourier Transform (FrFT)-based Mel Frequency Cepstral Coefficient (MFCC) characteristics for autonomous musical instrument categorization. The proposed system's classifier model was created using a Counter Propagation Neural Network (CPNN). When compared to other traditional features, the proposed features' discriminating capability has been maximised for between-class instruments and minimised for within-class instruments. In addition, when compared to other conventional features, the proposed features show a significant improvement in classification accuracy and robustness against Additive White Gaussian Noise (AWGN). The sound database from McGill University Master Sample (MUMS) was utilised to evaluate the system's performance [3].

To solve the challenge of music instrument recognition, speech and audio processing techniques are combined with statistical pattern recognition concepts. The suggested approach is scalable from isolation notes to solo instrumental phrases without the necessity for temporal segmentation of solo music because only non temporal, frame level information are utilized. Line Spectral Frequencies (LSF) are presented as features for music instrument recognition based on their usefulness in speech. MFCC and LPCC features have also been used to evaluate the proposed system and for classification, Gaussian Mixture Models (GMM) and the K-Nearest Neighbor (KNN) model are utilized. The experimental dataset includes databases from the University of Iowa's MIS and the C Music Corporation's RWC. When identifying 14 musical instruments, the best scores were around 95% at the musical instrument family level and 90% at the musical instrument level [10].

3. Preprocessing

A musical audio signal preprocessing takes place is as follows. Pre-emphasis, segmentation, and windowing are the steps in the preprocessing of raw musical data. The original music signal is first pre-processed, with the main purpose of unifying the music format, applying pre-emphasis, and segmenting the musical signal. Windowing and framing are then applied to all audio parts of the music.

3.1 Preemphasis

The digitized music signal is processed through a low order digital system to spectrally flatten it and make it less sensitive to fixed precision effects later in the music signal processing. It is usual practice to use the first order difference equation to preemphasis the music signal.

$$\boldsymbol{s'_n = s_n - k\, s_n - 1} \qquad (\mathbf{1})$$

to the samples {sn, n = 1, N} in each window. Here k is the preemphasis coefficient which should be in the range 0 ≤k< 1.

3.2 Frame blocking

The continuous music signal is then divided into N frames of musical audio samples, with neighbouring frames separated by M (M < N). The first N audio samples make up the first frame. After the first frame, the second frame starts with M samples and overlaps it with N - M samples, and so on. This process is repeated until all of the music data is accounted for one or more frames. A frame rate of 160 frames per second is employed throughout this paper, with each frame lasting 20 milliseconds and a 50% overlap between subsequent frames. Fig. 3 depicts the overall process, which displays the sampled audio waveform being converted into a sequence of parameter blocks. The waveform segment used to determine each parameter vector is commonly referred to as a window, and the volume of the window is referred to as the window size. Frame rate and window size are unrelated. In general, the window size will be larger than the frame rate, causing succeeding windows to overlap.

3.3 Windowing

Finally, tapering the samples in each window to reduce discontinuities at the window's edges is beneficial. It performs the following transformation on the samples in the window sn, n = 1,N

$$s'_n = \left\{0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} s_n \qquad (2)$$

In practice, all the above three steps are typically followed.

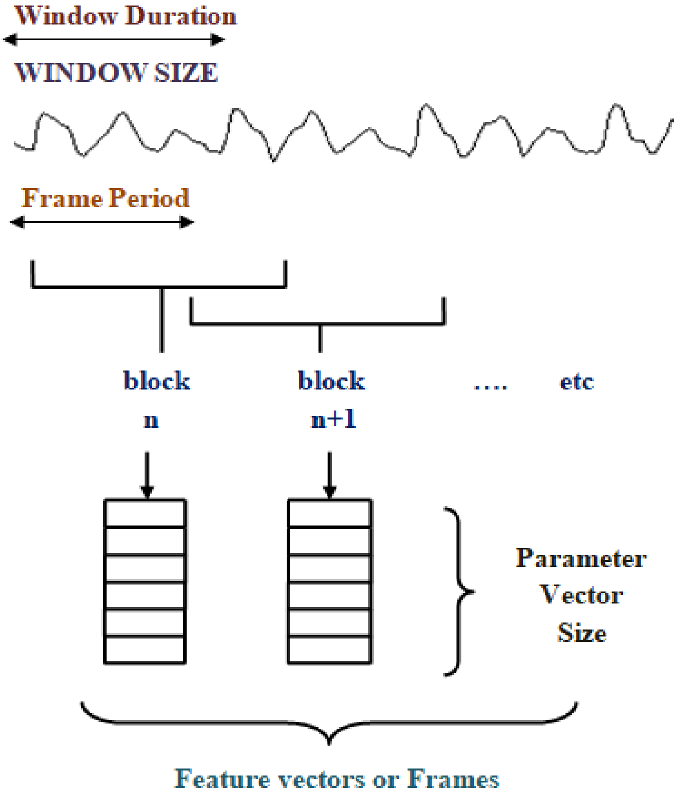

Fig. 3 Frame blocking

4. Feature Extraction Techniques

Once the music signal gets preprocessed, MFCC features are extracted.

Mel frequency Cepstral Coefficients

In the area of music, audio and speech processing the Mel frequency Cepstral Coefficients (MFCC) are the short-term spectral features that are commonly used. The mel frequency cepstrum has been shown to be extremely effective at identifying the composition of musical audio signals as well as modelling the subjective frequency and pitch content of those signals.

The MFCCs have been used in a variety of sound mining activities and have demonstrated superior performance when compared to other features. MFCC was calculated in a variety of ways by different writers. As a result, the goal of this research is to figure out how many coefficients are appropriate for sound classification of musical instruments. The MFCC features are employed to analyse the musical instrument signal in this study.

The phonetically important elements of audio, music, and speech are captured using MFCCs, which are based on the well-known change of the human ear's essential bandwidths with frequency, logarithmically at high frequencies and linearly at low frequencies. The musical audio signals are segmented and windowed into short frames of 20 ms to obtain MFCCs. Figure 1 shows a block diagram for extracting MFCC features. in Fig 4.

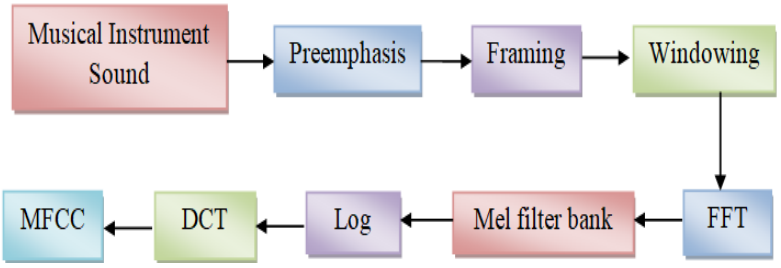

Fig. 4 Extraction the MFCC features from Music Signal

Mel frequency wrapping

For each of these frames, the magnitude spectrum is calculated using the Fast Fourier Transform (FFT) and converted into a collection of outputs from the mel scale filter bank. The filter bank analysis makes obtaining the appropriate non-linear frequency resolution a lot easier. Filter bank amplitudes, on the other hand, are highly correlated, making the employment of a cepstral transformation nearly mandatory in this scenario. On a mel-scale, a simple Fourier transform based filter bank is designed to provide about equivalent resolution.

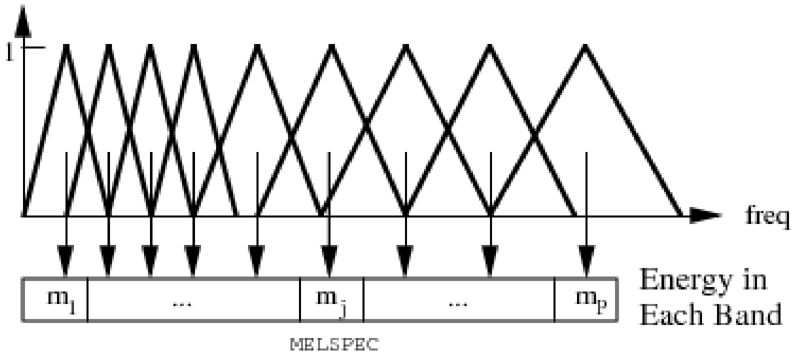

Fig. 5 Mel scale filter bank

Fig. 5 illustrates the general structure of this filter bank. As can be seen, the filters used are triangular and they are equally spaced along the mel-scale which is defined by

$$Mel(f) = 2595 \ln\left(1 + \frac{f}{700}\right) \tag{3}$$

To create this filter bank, a Fourier transform is applied to a window of musical audio data, and the magnitude is calculated. By correlating the magnitude coefficients with the complete triangle filter, the magnitude coefficients are then binned. In this case, binning implies multiplying each FFT magnitude coefficient by the relevant filter grow and collecting the results. As a result, each bin contains a weighted sum that represents the spectral magnitude in that filter bank channel.

Triangular filters are often stretched across the entire frequency range from zero to Nyquist frequency. Band-limiting, on the other hand, is frequently used to reject undesired frequencies or avoid assigning filters to frequency zones with no useful signal energy. Lower and upper frequency cut-offs can be specified for filter bank analysis. The stated number of filter bank channels are spread evenly on the mel-scale over the resulting pass-band when low and high pass cut-offs are adjusted in this manner.

Cepstrum

To obtain the MFCCs, the logarithm is applied to the filter bank outputs, followed by Discrete Cosine Transformation (DCT). Because the mel spectrum coefficients (and their logarithms) are real numbers, the DCT can be used to transform them to the time domain. In practice, for computing efficiency, the final step of inverse Discrete Fourier Transform (DFT) is replaced by DCT. For the stated frame analysis, the cepstral representation of the music spectrum provides a superior representation of the signal's nearby spectral features. The first 13 MFCCs are typically used as features [6] [7].

At the segmental level, 12 MFCC coefficients (c1, c2, c3, …, c12) for each frame are retrieved from the musical audio signal. The DCT excludes the 'null' MFCC coefficient c0, which represents the mean value of the input musical audio stream, which contains minimal information. Dynamic parameters generated from 13th order static cepstral coefficients (c0, c1, c2, c3, …, c12) have been proposed and proved to improve audio categorization system performance. The delta-cepstrum (first-order difference of the short-time static cepstrum), the delta-delta-cepstrum (second-order difference of the static cepstrum), and delta- and delta-delta-energy are examples of dynamic characteristics. In noisy environments, dynamic features have been shown to be more durable than static features. The dynamic and static aspects of a musical audio signal spectrum with 13th order static coefficients, 13th order delta coefficients, and 13th order acceleration (delta-delta) coefficients are captured using a 39th order MFCC. For each frame, this yields a 39-dimensional MFCC feature vector.

The MFCC features are extracted as described for ten categories of sound of the musical instruments namely Bass Clarinet, Bassoon, Clarinet, Contrabassoon, Flute, French Horn, Saxophone, Trombone, Trumpet, and Tuba respectively.

5. Modeling the Features

Gaussian Mixture Model

Parametric or nonparametric approaches are used to model the probability distribution of feature vectors. Parametric models are those that assume the shape of a probability density function. In nonparametric modelling, the probability distribution of feature vectors is assumed to be minimum or nonexistent. The Gaussian mixing model (GMM) is briefly discussed in this section. The rationale for employing GMM is that a mixture of Gaussian densities can be used to describe the distribution of feature vectors derived from a class, as illustrated in Fig. 6.

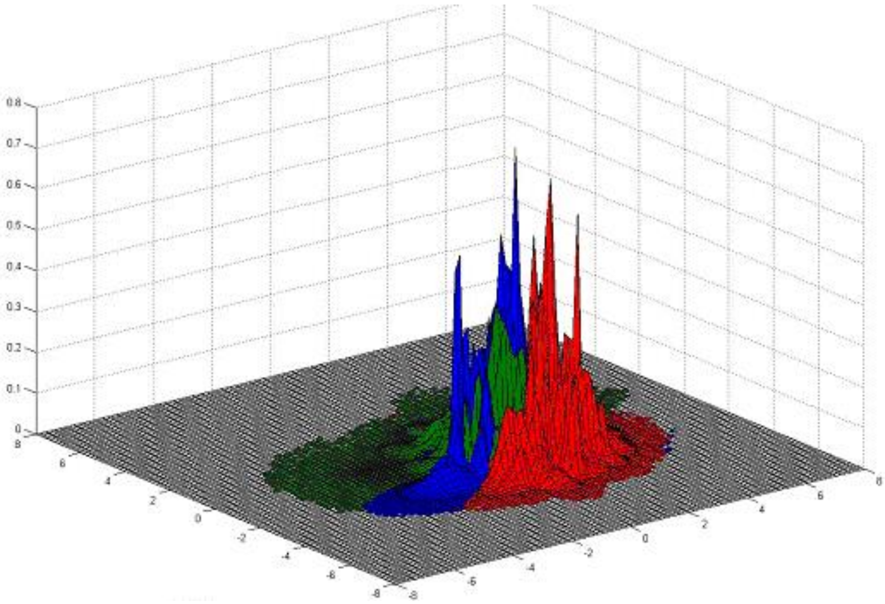

Fig. 6 Gaussian Mixture Model

GMMs use Gaussian components to describe feature vectors and are defined by the mean vector and co-variance matrix [8]. GMM models have the potential to construct an arbitrary shaped observation density even in the absence of other information [9].

6. Performance Measures

This study used a set of assessment metrics, including Accuracy, Precision, Recall, and F-score, to assess the performance of the classifier GMM using MFCC. The confusion matrix, which is obtained from the classification process's output, is used to determine these metrics. The confusion matrix is a 2 x 2 matrix with four elements: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), where TP indicates that the prediction is correct, TN indicates that the prediction is incorrect, FP indicates that the correct value is predicted incorrectly, and FN indicates that the wrong value is predicted correctly.

7. Experimental Results

7.1. Dataset

The data was collected from the online musical database Instrument Recognition in Musical Audio Signals (IRMAS). Bass Clarinet, Bassoon, Clarinet, Contrabassoon, Flute, French Horn, Saxophone, Trombone, Trumpet, and Tuba were among the 1000 musical audio clips. Each clip contains music data with a duration ranging from 1 to 10 seconds, sampled at 8 kHz and encoded in 16-bit. For training, 800 musical audio data samples were employed, and 200 for testing. For further implementation, the music clips are preprocessed using pre-emphasis, segmentation, and windowing.

7.2. Acoustic Feature Extraction

Fixed-length and overlapping frames are created from the training data (in our work 20 ms frames with 10 ms overlapping). The temporal features of music data can be processed in the training phase when the surrounding frames are overlapped. Because of the 8 kHz sampling rate, 20 ms frames contain 160 values. These 160 values are translated into 39 MFCC coefficients, each of which represents a single frame. For 1 second of music data, there are 100 such frames. For each of the ten categories, the feature extraction method is performed for musical audio samples of varied durations such as 1 second, 3 seconds, 5 seconds, and 10 seconds. Experiments are carried out to test acoustic feature MFCC, as well as the performance of GMM.

7.3. Gaussian Mixture Model

Gaussian mixtures for the ten classes are modeled for the MFCC feature. For classification the feature vectors are extracted and each of the feature vector is given as input to the GMM model. The distribution of the acoustic features are captured using GMM. We have chosen a mixture of 1, 3, 5, 10 mixture models. Audio classification using GMM gives an accuracy of 86.98%. The performance of GMM for different mixtures as shown in Fig. 7 shows the accuracy of different data durations, the best performance was achieved with 10 Gaussian mixtures compared to the other mixtures. Fig. 8 shows the performance of GMM for musical audio classification.

Table 1 Overall accuracy of GMM with MFCC

Model	1 second	2 seconds	3 seconds	10 seconds
GMM	86.90	84.11	88.24	89.35

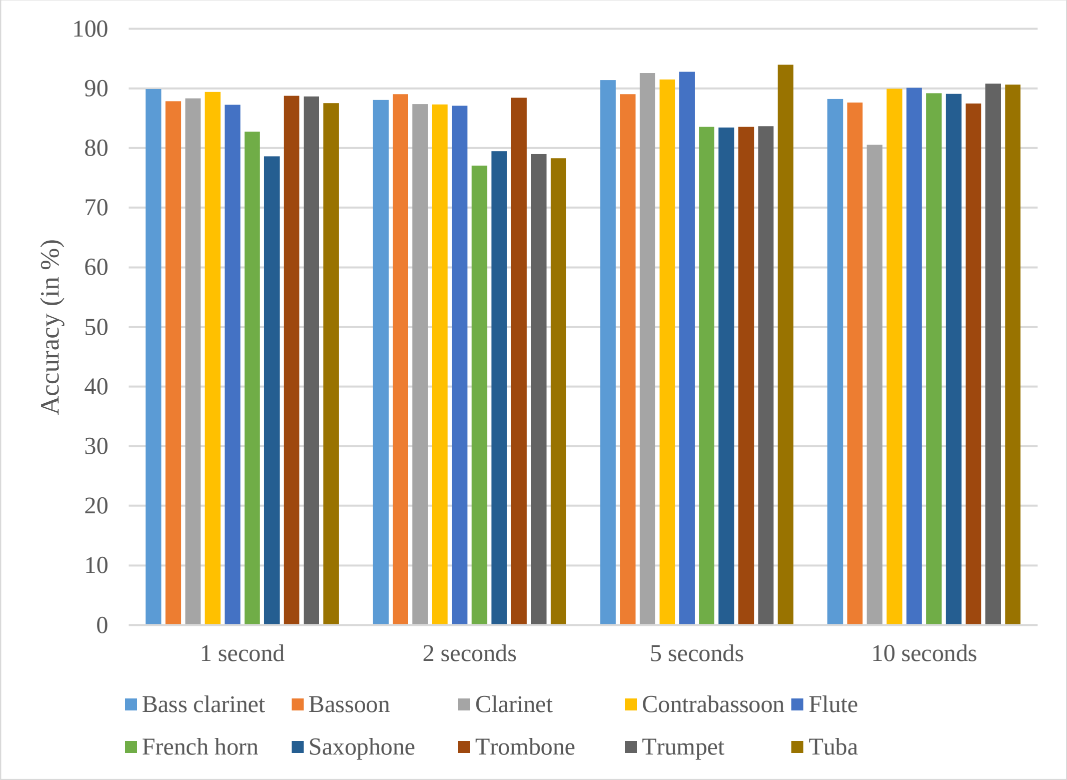

Fig. 7 Accuracy of the proposed system for various duration

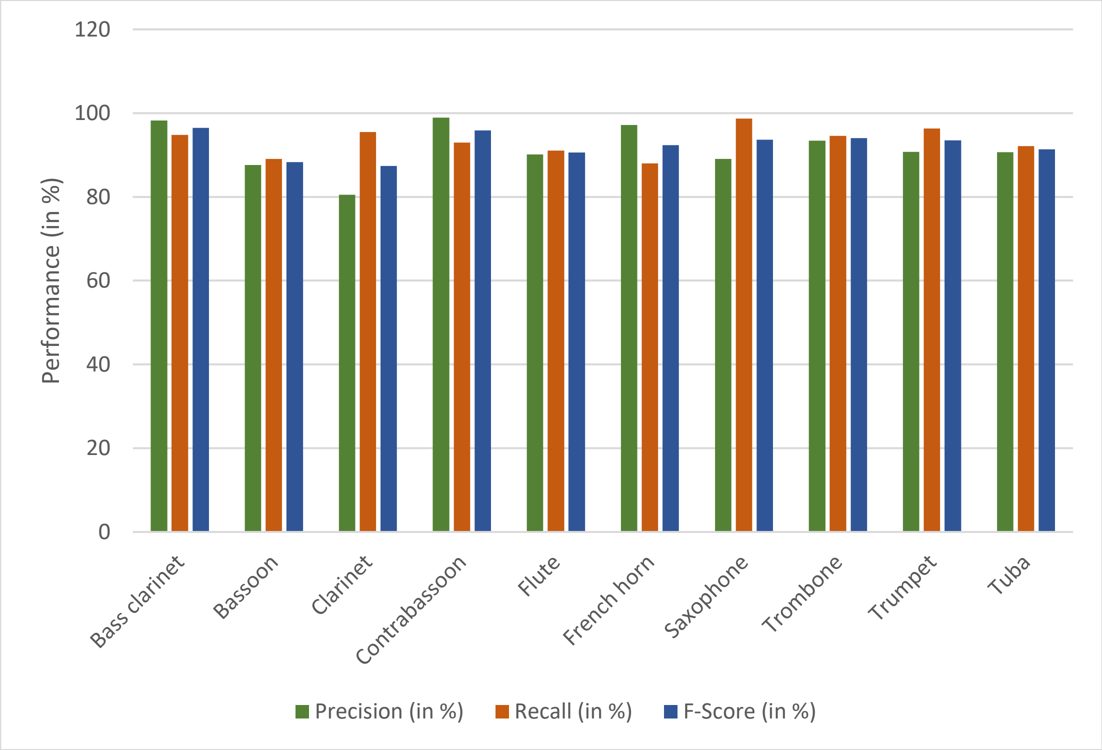

Fig. 8 Performance of the proposed work

8. Conclusion

In this work, a GMM classifier with MFCC features used to classify musical instrument sounds automatically. The performance of GMM yields a satisfactory accuracy of 86.98 %. When compared to 1 second, 3 seconds, and 5 seconds data the 10 seconds data provides the maximum accuracy from the training data. The sound classification model for musical instruments will be updated in the future to discriminate more classes.

References

[1] Dhanalakshmi, P., Palanivel, Sengottayan & Ramalingam, Vivekanandan. (2011). Classification of audio signals using AANN and GMM. *Applied Soft Computing.* 11. Pp. 716-723, 10.1016/j.asoc.2009.12.033.

[2] Thiruvengatanadhan, R., Music Genre Classification using GMM (2018). *International Research Journal of Engineering and Technology (IRJET),* Volume: 05 Issue: 10.

[3] Bhalke, D.G., Rao, C.B.R. & Bormane, D.S. Automatic Musical Instrument Classification using Fractional Fourier Transform based- MFCC features and Counter Propagation Neural Network. *J Intell Inf Syst* 46, 425–446 (2016). https://doi.org/10.1007/s10844-015-0360-9

[4] Prabavathy, S, Rathikarani, V., & Dhanalakshmi, P, Classification of Musical Instruments using SVM and KNN, *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* ISSN: 2278-3075, Volume-9 Issue-7, May 2020.

[5] Shreevathsa, P. K., Harshith, M., A. R. M. & Ashwini, "Music Instrument Recognition using Machine Learning Algorithms," 2020 *International Conference on Computation, Automation and Knowledge Management (ICCAKM),* 2020, pp. 161-166, doi: 10.1109/ICCAKM46823.2020.9051514.

[6] Dhanalakshmi, P. (2010). *Classification of Audio for Retrieval Applications.* [Doctoral Dissertation]. Faculty of Computer Science and Engineering, Annamalai University, Chidambaram, Tamilnadu.

[7] Zokaee, Sara & Faez, Karim. (2012). *Human Identification Based on Electrocardiogram and Palmprint.* 2. 261-266.

[8] Rafael Iriya & Miguel Arjona Ramírez. Gaussian Mixture Models with Class-Dependent Features for Speech Emotion Recognition. *IEEE Workshop on Statistical Signal Processing,* pp. 480-483, 2014.

[9] Tang, H., Chu, S. M., Hasegawa-Johnson, M. & Huang, T. S., Partially Supervised Speaker Clustering. *IEEE transactions on Pattern Analysis and Machine Intelligence.* vol. 34, no. 5, pp. 959-971, 2012.

[10] Krishna, A.G. & Sreenivas, Tv. (2004). Music instrument recognition: from isolated notes to solo phrases. *International Conference on Acoustics, Speech, and Signal Processing,* ICASSP.