


MISNA - A musical instrument segregation system from noisy audio with LPCC-S features and extreme learning

Himadri Mukherjee¹ · Sk Md Obaidullah² ·
Santanu Phadikar³ · Kaushik Roy¹ 

Received: 27 July 2017 / Revised: 1 February 2018 / Accepted: 9 April 2018 /

Published online: 26 April 2018

© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract Technology has developed a lot over the last decades and has made a profound impact in almost every field. The field of Music Information Retrieval (MIR) has not been an exception to this as well, one of its most promising applications being Automatic Music Transcription (AMT). It is important to identify the active regions of various Instruments in a piece before transcription and the challenge elevates even more when the audio clips are contaminated with noise. MISNA (Musical Instrument Segregation from Noisy Clips) is a system proposed towards the identification of isolated Instruments from noisy clips which can aid towards AMT in noisy environments. The system works using statistical features (LPCC-S) derived from raw Linear Predictive Cepstral Coefficient values on very short clips of lengths 1 and 2 seconds. The system has been tested for various SNR scenarios and highest accuracies of 98.63% and 97.42% for Individual Instruments and Instrument Family identification has been obtained with the aid of Extreme Learning based classifier for a highest of 2626 clips.

Keywords LPCC-S · Extreme learning · SNR · Mean · Standard deviation

✉ Himadri Mukherjee
himadrim027@gmail.com

Sk Md Obaidullah
sk.obaidullah@gmail.com

Santanu Phadikar
sphadikar@yahoo.com

Kaushik Roy
kaushik.mrg@gmail.com

¹ Department of Computer Science, West Bengal State University, Kolkata, India

² Department of Computer Science and Engineering, Aliah University, Kolkata, India

³ Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata, India

1 Introduction

The field of MIR has fascinated the research community for a long time and one of its most promising applications has come to us in the form of Automatic Music Transcription (AMT). AMT is the process of identification of the notes played by an Instrument from an audio clip. In a music piece, more than 1 Instrument is played at a time and not all the Instruments are played through the entire length of the piece. Thus it is essential to identify the active regions of the Instruments in a piece before transcription. The challenge of identifying such Instruments increases even more when a piece is accompanied by noise. It is important to be able to identify musical Instruments in isolation from noisy clips before identifying the same from a piece and MISNA is a system proposed towards such a task. The main contribution of our work includes the use of proposed lower dimensional features (LPCC-S) derived from standard LPCC values for minimizing computational overhead and overcoming uneven dimensionality issue, the use of Extreme Learning Machine based classification which is a faster version of standard neural network based classifier, experimentation with various levels and types of noisy environments and verification of the generalization capability of the proposed system for both Individual Instruments and Instrument families using clips of short durations.

2 Related works

Masood et al. [22] identified 5 different Instruments using MFCC and Timbral features with an accuracy of 89.17%. They used a neural network based classifier, which was trained using Conjugate gradient back propagation and Fletcher-Reeves updation technique. Patil et al. [25] classified 15 Instruments with an accuracy of 86.04% using a SVM and concept analysis based technique. Eronen et al. [7] used features based on temporal and spectral properties of sound to classify 30 orchestral Instruments from the bass, string and woodwind families and obtained an accuracy of 94% in identification of the correct Family. A system to identify 7 different Instruments was presented by Sturm et al. [30] using multiscale MFCC based features. A highest accuracy of 84.69% was obtained for the system using a SVM based classifier. Martin et al. [21] used a statistical pattern recognition based approach to identify 15 different orchestral Instruments using acoustic features related to physical properties of source excitation and resonance structure. Accuracies of 90% and 70% were obtained for Instrument Family and identification Individual Instrument identification using Gaussian models and Fisher multiple discriminant analysis. Takashi et al. [31] designed a system to identify 12 musical Instruments using zero crossing, pitch, brightness and spectral centroid based features. They obtained highest average accuracies of 82.1% and 56.2% for the University of Iowa musical Instrument database and RWC music databases using Random Forest and Linear Discriminant Analysis technique respectively. A system to classify 19 different musical Instruments was presented by Kitahara et al. [16] with the help of 18 dimensional features. The feature set was composed of F0 normalized covariance and mean which produced an accuracy of 79.73%. Benetos et al. [2] used various classification techniques to distinguish 20 Instruments with the help of MPEG-7 audio descriptors as well as zero crossing, spectrum flatness, MFCC, auto correlation, spectrum roll off frequency, specific loudness sensation and total loudness and produced accuracies as low as 88.7% and as high as 95.3%. Livshin et al. [19] presented a real time Instrument recognition technique from solos for 7 Instruments. Post 62 dimensional feature extraction, a dimension

reduction technique using Gradual Descriptor Elimination was applied to reduce the computational overhead. Accuracies of 88.13% and 85.24% were obtained respectively for the non reduced and reduced sets with the aid of KNN classification and LDA transformed learning set. Kaminskyj and Czaszejko [15] classified 19 Instruments from 9 major and sub families. They extracted 6 features for each namely cepstral coefficients, multidimensional scaling analysis trajectories, constant transform frequency spectrum, RMS amplitude envelope, presence of vibrato and spectral centroid. They obtained a highest accuracy of 97% using KNN classification technique for Family Identification. Lita et al. [17] presented a smart sound sensor based system for the identification of 3 musical Instruments in real time and obtained an average accuracy of 98.33%. Kaminskyj et al. [14] distinguished 4 different Instruments from 4 different families by employing various mechanisms in the pre processing stage including short term RMS energy envelope computation, Principal Component Analysis and Ratio or Product transformations of the same. Artificial Neural network and nearest neighbour based classifiers were applied for the same and accuracies in the range of 93.8% - 100% were obtained. Yu et al. [34] differentiated 14 Instruments from 4 Chinese folk Instrument families and obtained a highest accuracy of 89% by combining perceptron based features along with Mel Scale Cepstral Coefficients. Liu et al. [18] designed a system for identification of 4 Instrument families for both Chinese and Western Instruments. They experimented with various classifiers and features for both Chinese and Western genres and concluded that Spectral Flatness Measure coupled with KNN classifier produced the best result in the case of Chinese Instruments and the same feature coupled with SVM or MFCC coupled with KNN produced the highest accuracy for Western Instruments. They obtained a difference of 28% in the accuracy between the best and worst classification scheme. Agostini et al. [1] presented a system for the identification of 30 musical Instruments from the McGill University Master samples database using spectral features. Various classification techniques encompassing k-Nearest Neighbour, Canonical Discriminant Analysis, Quadratic Discriminant Analysis (QDA) and SVM were applied out of which highest accuracies of 80.2%, 78.6% and 69.7% were obtained for 17, 20 and 27 instruments respectively using SVM with a RBF kernel. They further obtained accuracies of 81% and 92.2% for the 27 instruments family and pizzicato-sustained discriminations respectively using QDA. They also highlighted obtained accuracies of 89%, 94% and 96% using QDA for rock strings, woodwind and brass families respectively as well. Livshin et al. [20] presented algorithms for outlier or bad sample Detection to improve musical Instrument identification. Sliding window of 60 ms along with a 66% overlap were used for calculation of features which helped in successfully discarding 70.1% of the bad samples which generally degrade Instrument recognition performance. Fragoulis et al. [8] designed a system to recognize 2 different Instruments namely guitar and piano using tonal spectral content for clips of average length of 1.8 sec. An accuracy of 100% was obtained for 926 isolated piano notes and 612 similar guitar notes. Röver et al. [29] presented a Hough transformation based approach to identify musical Instruments. They used a hybrid of Linear Discriminant Analysis and Quadratic Discriminant Analysis known as Regularised Discriminant Analysis to identify 25 Instruments and obtained a lowest misclassification rate of 26.1%. Donnelly et al. [6] used different Bayesian Networks to classify 24 different orchestral Instruments. Bayesian networks along with conditional dependencies in the frequency and time domain produced accuracies of 98% and 97% for Individual Instrument and Instrument Family identification. Yu et al. [33] proposed an improved matching pursuit algorithm for the identification of musical Instruments. They extracted atomic parameters for Instruments from the algorithm and fed it to a SVM in order to differentiate 10 musical Instruments and obtained an

accuracy of 87.44% in only $1/3^{\text{rd}}$ of the time as required by standard matching pursuit algorithm. Jadhav [12] obtained accuracies of 88%, 84% and 73.33% for 5, 10 and 15 different Instruments with the help of timbral audio descriptors and Binary Tree classifier. Accuracies of 90%, 77% and 75.33% were obtained for the same set using KNN classifier along with MFCC features.

3 Dataset development

One of the most important facets of any experiment is data collection. The database of our experiment was put together with the aid of synthesized tones of 7 different Instruments namely Flute, Grand Piano, Guitar, Saxophone, Harmonium, Violin and Santoor. The Instruments hailed from 3 families namely Wind (Flute and Saxophone), Keyboard (Grand Piano and Harmonium) and String (Violin, Nylon String Guitar and Santoor). Such Instruments were chosen to include both Indian as well as Western Instruments from the various families which are some of the most essential ingredients of melody. All the 22 natural notes in the scale of C from C2 to C5 were played 20 to 30 times for every Instrument in various playing styles including Fortississimo, Fortissimo, Mezzo forte, Forte, Marcato, Staccato, Legato, Pianissimo and Pianississimo. These clips were used to engender 2 datasets (D1 and D2) consisting of 2626 (1 second each) and 1311 (2 seconds each) clips respectively. The clips were stored in uncompressed .wav stereo format at a bitrate of 1411 kbps. The number of clips for both individual Instruments as well as for Instrument Families is presented in Table 1. Each of the presented datasets were used for both the recognition of Individual Instruments as well as Instrument families.

Data can be breathed upon by various kinds of noises in real life scenario. To test the performance of our proposed system, each of the datasets (D1 and D2) were contaminated with 4 types of noise sources namely Rain, Traffic, Vacuum Cleaner and Fan which produced $4 \times 2 = 8$ more datasets whose details are presented in Table 2.

The Instrument wise Signal to noise Ratios (SNRs) for the 1 second long clips (D3-D6) and 2 second long clips (D7-D10) in various noisy conditions for both Individual Instrument level as well as Instrument Family level is presented in Table 3.

3.1 Instruments in the dataset

A brief description of the instruments which were selected in our experiments is presented in the following paragraphs.

Flute: This is a wind instrument which is also known as Bansuri in India. Flutes can be either side blown or front blown. A flute is capable of producing sounds of different octaves in same fingering position if only the blowing pressure is varied. Flutes are mostly made from bamboo however many musicians use metallic flutes as well.

Guitar: It is a stringed instrument. Guitars are of various types like Nylon String, Steel String, Bass, etc. A musician needs to strike the strings either with fingers or with the help of a plectrum for producing sound.

Harmonium: It can be considered as a keyboard instrument due to the presence of keys. It also has reeds which play a vital part in tone production and thus can be considered as a reed instrument as well. The player needs to push and pull the front lever for air circulation within the instrument and at the same time press the keys for producing sound.

Table 1 Individual Instrument level and Instrument family level details of datasets D1 and D2 along with total (T) clips

Data Set / Type	Total (T) Clips									
	Wind		String			Keyboard				
Instrument	Flute	Saxophone	T	Guitar	Santoor	Violin	T	Piano	Harmonium	T
D1	461	360	821	331	324	328	821	441	381	822
D2	230	180	410	165	162	164	410	220	190	410

Table 2 Details of the noisy datasets

Original Dataset	Generated Noisy Dataset	Original Dataset	Generated Noisy Dataset	Type of Added Noise
D1	D3	D2	D7	Rain
	D4		D8	Traffic
	D5		D9	Vacuum Cleaner
	D6		D10	Fan

Piano: It is a keyboard instrument. There are various types of Pianos like acoustic grand pianos and the modern day electric piano. Pianos have evolved into modern day synthesizers which come with various tonal capabilities and other features which has made the task of music production a lot easier.

Santoor: It is a stringed musical instrument which is trapezoidal in shape. The Santoor is played by striking the strings with two wooden mallets. It is sensitive to glides as well as light strokes. The instrument has tuning pegs mostly on the right for tuning the strings in order to produce sounds of different frequencies.

Saxophone: It is a wind instrument which is mostly made of brass. A player needs to blow through the mouth piece located at the top and close the holes in various combinations with the help of a key system for producing music. There are various kinds of saxophones like alto, tenor, sorpano, etc.

Table 3 SNRs for individual instruments and instrument families

Noisy Set	Individual Instrument wise SNRs							Average SNR
	Fute	Saxophone	Guitar	Santoor	Violin	Piano	Harmonium	
D3	7.38	0.83	5.70	2.11	-1.73	0.64	6.10	3.00
D4	-0.44	-6.99	-2.12	-5.71	-9.55	-7.18	-1.72	-4.82
D5	3.46	-3.09	1.78	-1.81	-5.65	-3.27	2.18	-0.91
D6	4.98	-1.58	3.29	-0.30	-4.14	-1.76	3.70	0.60
D7	7.65	1.27	5.91	2.94	-1.16	1.40	6.39	3.49
D8	-0.82	-7.20	-2.57	-5.54	-9.63	-7.07	-2.08	-4.99
D9	3.99	-2.39	2.25	-0.73	-4.82	-2.26	2.73	-0.18
D10	5.96	-0.42	4.22	1.24	-2.85	-0.29	4.70	1.79
	Instrument Family wise SNRs							
	Wind		String		Keyboard			
D3	4.51		2.04		3.17			3.24
D4	-3.31		-5.79		-4.65			-4.58
D5	0.59		-1.88		-0.74			-0.68
D6	2.10		-0.37		0.77			0.83
D7	4.85		2.57		3.72			3.71
D8	-3.62		-5.91		-4.76			-4.76
D9	1.19		-1.10		0.05			0.47
D10	3.16		0.87		2.02			2.02

Violin: It is a stringed fret less instrument which is played by using a bow. A musician needs to bow with one hand and finger the fingerboard with another hand to produce sound Earlier violins were mostly acoustic but with advent of technology, electric violins are now also available which are mostly used in concerts and recordings.

4 Proposed methodology

The clips were first framed into short sections and then windowed as part of pre processing. Next standard LPCC features were extracted from the clips. In order to tackle the problem of uneven dimensionality, the LPCC-S features were generated for the clips which were then fed to an Extreme Learning based classifier. The proposed system is graphically illustrated in Fig. 1 whose details are presented in the subsequent paragraphs.

4.1 Pre-processing

4.1.1 Framing

The spectral properties of a sound signal vary a lot through its entire length thus posing difficulty in the task of analysis. To cope up with this problem, a clip is partitioned into small parts called frames. The spectral properties tend to be quasi stationary within such frames thereby facilitating in the task of analysis. A signal can be framed in 2 ways namely overlapped framing and non overlapped framing. In overlapped framing, a certain number of sample points towards the end intersect with the starting sample points of the next frame. This ensures continuity in between the frames and a smoother transition in between the same. In our experiment, sound signals were framed in overlapping mode with a frame size of 256 sample points and an overlap of 100 sample points. 2 consecutive overlapping frames are graphically illustrated in Fig. 2. The number of obtained frames (m) of size F for a signal consisting of n sample points with O overlapping points can be calculated using (1).

$$m = \left\lceil \frac{n - F}{O} + 1 \right\rceil \tag{1}$$

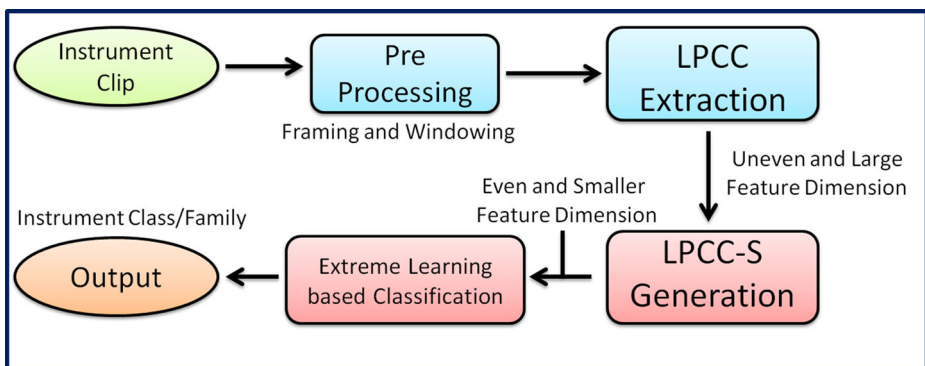


Fig. 1 Graphical illustration of the proposed system

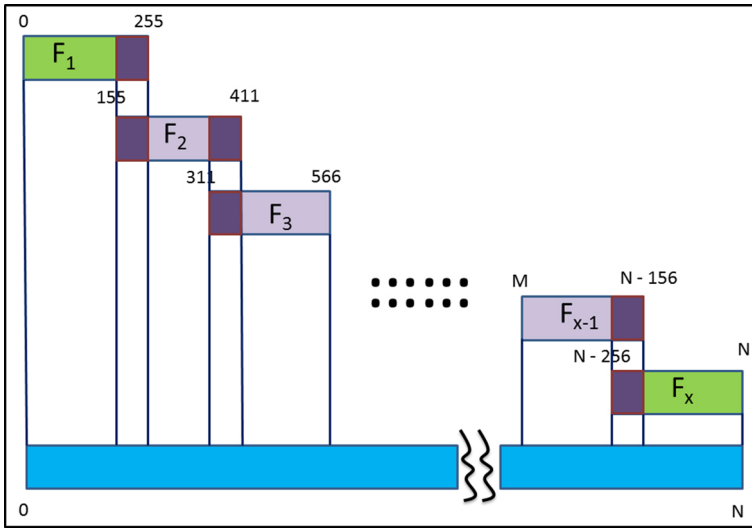


Fig. 2 Framing methodology

4.2 Windowing

Post framing, jitters might be observed in them which interfere with the Fourier Transformation of the same in the form of spectral leakage. In order to minimize such problems, the frames are usually multiplied with a windowing function which approaches 0 towards its ends and reaches its peak in the middle. Amidst various such windowing functions, Hamming Window function is one of the popularly used windowing functions whose utility has been presented in [23, 24] which inspired us to use the same in our experiment. The Hamming Window function is mathematically presented in (2) and graphically illustrated in Fig. 3.

$$w(x) = 0.54 - 0.46 \cos\left(\frac{2\pi x}{M - 1}\right) \tag{2}$$

Here $w(x)$ is the Hamming Window function where M is the frame size and x lies in between the start to end of the frame.

4.3 Feature extraction

Twelve standard Linear Prediction Cepstral Coefficients(LPCC) [5, 26] were obtained for every frame of every clip with the aid of Linear Predictive Analysis which predicts a present sound sample as a linear combination of previous sound samples. The mathematical representation of the n^{th} sample, estimated with previous J samples is presented in (3).

$$s(n) \approx A_1s(n - 1) + A_2s(n - 2) + A_3s(n - 3) + A_4s(n - 4) + \dots + A_Js(n - J) \tag{3}$$

Here, $A_1, A_2, A_3, A_4, \dots, A_J$ are assumed to be constants for an analysis frame which are also known as predictor or linear predictive coefficients which aid in predicting the present

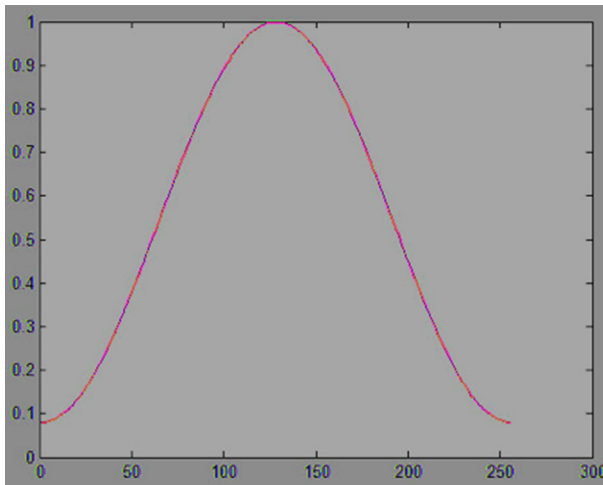


Fig. 3 Structure of hamming window

sample. The difference between the actual ($s(n)$) and predicted ($\hat{s}(n)$) samples is known as error $e(n)$, which is presented in (4) in terms of the predictor coefficients (A_k s)

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^J A_k s(n - k) \tag{4}$$

In order to engender a unique set of predictor coefficients, error minimization on the sum of squared differences is performed in accordance with (5), where m corresponds to the number of samples in a frame.

$$E_n = \sum_m \left[s_n(m) - \sum_{k=1}^J A_k s_n(m - k) \right]^2 \tag{5}$$

To solve (5) for the LPs, E_n is differentiated with respect to each of the A_k s as shown in (6)

$$\frac{\delta E_n}{\delta A_k} = 0, \quad \text{for } k = 1, 2, 3, \dots, J \tag{6}$$

Finally the Cepstral Coefficients are calculated with the recursive procedure as shown in (7).

$$\left. \begin{aligned} C_0 &= \log_e J \\ C_m &= A_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k A_{m-k}, \text{ for } 1 < m < J \text{ and} \\ C_m &= A_m + \sum_{k=m-J}^{m-1} \frac{k}{m} C_k A_{m-k}, \text{ for } m > J \end{aligned} \right\} \tag{7}$$

4.4 LPCC-S generation

Since clips of disparate lengths yielded disparate number of frames, so features of variable dimensions were obtained. A clip of 1 second, sampled at 44100 Hz produces 440 frames (256 points wide with 100 point overlap) according to (1). Since 12 LPCC features were extracted for every frame, so a total of 5280 (12 X 440) feature values were obtained. Clips

of larger length produced features of even larger dimension which heaped a serious burden on the system in terms of computation.

In order to deal with these 2 issues, LPCC-S (LPCC-Statistical) is proposed whose dimension does not vary with the length of a clip thereby attending to the uneven dimensionality problem and its lower dimension spares the system of computational overhead. Each of the 12 bands of the raw LPCC features for a clip were analysed and the mean for each of those bands were computed followed by Standard Deviation computation of the same. Finally, these values were added to yield a 24 dimensional feature. The LPCC-S generation methodology from LPCC representation of a clip is presented in Algorithm 1.

Algorithm 1 LPCC-S GENERATION

```

Input : LPCC[N][F]
Output: LPCC-S[N][2]

1 for  $i \leftarrow 0$  to  $N$  do
2   // all the LPCC Coefficients for a clip
3    $sum \leftarrow 0$ ;
4   for  $j \leftarrow 0$  to  $F$  do
5     // all the frames for a clip
6      $sum \leftarrow sum + LPCC(i, j)$ ;
7   end
8    $sum \leftarrow sum/F$ ;
9   LPCC-S( $i,1$ )  $\leftarrow sum$ ;
10   $temp \leftarrow 0$ ;
11  for  $j \leftarrow 0$  to  $F$  do
12     $temp \leftarrow temp + (LPCC(i,j) - sum) \wedge 2$ ;
13  end
14  LPCC-S( $i,2$ )  $\leftarrow (temp/F - 1) \wedge 0.5$ ;
15 end
    
```

A graphical representation of the features of the various Instruments for both 1 and 2 second long clips in Noise Free condition is presented in Fig. 4. It can be observed from the Figure that the feature values of the instruments have different trends which aids in classification.

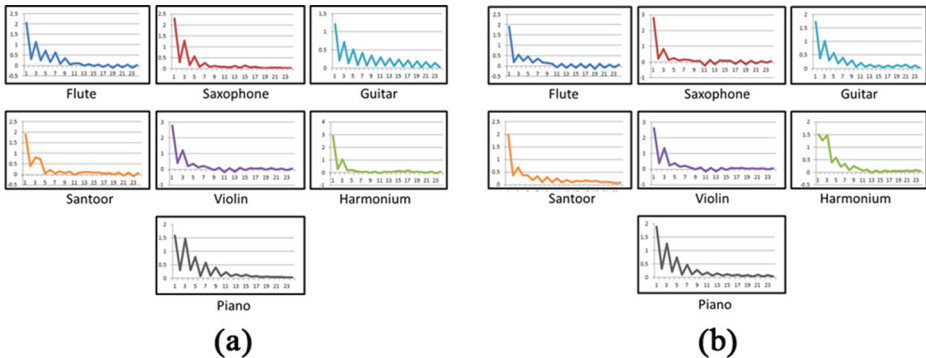


Fig. 4 **a** Feature values for the Instruments for 1 second long Clips in noise Free Condition. **b** Feature values for the Instruments for 2 second long Clips in noise Free Condition

The feature graphs for the 1 and 2 second clips in Rain and Fan Noise were analysed as well which is presented in the [Appendix](#). It can be observed from the Figure that the feature values for the various instruments appear to be very similar due to the effect of noise. Moreover not much change can be seen between the values of 1 and 2 second clips for a particular instrument.

The feature graphs for the 1 and 2 second clips in Traffic and Vacuum Cleaner Noise were also analysed which is presented in the [Appendix](#). It can be observed from the Figure that the feature values for the various instruments appear to show some deviations for different length of clips in the Traffic Noise condition. Moreover inter instrument differences are also visible for certain pairs. However, in the case of Vacuum Noise condition the feature values appear to be very close to one another for the various instruments and negligible changes are observed for different length of clips.

4.5 Classification with extreme learning machine (ELM)

Traditional Neural Networks trained using back propagation method have quite a few issues associated with them including a large number of steps involved in the gradient descent searching, local minima, slow convergence, etc. ELM however provides an efficient and unified learning framework by means of generalizing a feed forward neural network with 1 hidden layer only and that too with minimum human intervention for tuning the parameters like the number of nodes and hidden layers [9, 10, 32]. ELMs have the capability of solving an array of classification or regression problems by generating a random learning model which is very fast. In our experiment the number of output neurons was equal to the number of classes for various datasets. The number of hidden neurons was varied from 1 to 600 and was set to the value for which highest accuracy was obtained.

The learning method of ELM involves 2 major steps

Feature mapping In this stage, the ELM maps the input data to the hidden layer. The output function of this stage is shown in (8).

$$f(x) = \sum_{i=1}^L \beta_i h_i(x) = h(x)\beta \quad (8)$$

where $\beta = [\beta_1, \dots, \beta_L]^T$, is the generated weight vector between the hidden layer consisting of L nodes and the output layer consisting of $m \geq 1$ nodes. The vector corresponding to the output of the hidden layer is denoted by $h(x)=[h_1(x), \dots, h_L(x)]$. The value of $h_i(x)$ can be tabulated using (9).

$$h_i(x) = G(a_i, b_i, x), a_i \in \mathbb{R}^d, b_i \in \mathbb{R} \quad (9)$$

where, $G(a,b,x)$ corresponds to a continuous, piecewise, non linear function and (a_i, b_i) corresponds to the parameters of the i^{th} hidden node.

Among various activation functions, sigmoidal function was chosen based on trial runs as it out performed the rest. The sigmoidal function is represented in (10).

$$G(a, b, x) = \frac{1}{1 + \exp(-(a * x + b))} \quad (10)$$

Here, the parameters (a and b) of the output function $G(a, b, x)$ are generated randomly with continuous probability distribution. Thus it can be seen that unlike the feed forward neural networks where the hidden neurons require tuning, those of the ELM are randomly

generated. The function $h(x)$ does the work of mapping d -dimensional input data to the L -dimensional random hidden layer in which the parameters of the hidden nodes are generated randomly. So, this feature mapping ($h \rightarrow G$) is random in nature.

ELM learning In comparison to the various traditional learning techniques, the extreme learning technique states that no adjustment is required in terms of the hidden neurons. The target is to simultaneously achieve the smallest training error and smallest norm output weights.

The Universal approximation [11, 32] is satisfied by the ELM which is shown in (11). It holds with a probability of 1 for proper output weights (β). A 5 Fold cross validation technique was used in the current experiment for evaluating the system.

$$\lim_{L \rightarrow \infty} \left\| \sum_{i=1}^L \beta_i j_i(x) - f(x) \right\| = 0 \quad (11)$$

4.6 Statistical significance test

Statistical Significance Test was performed with the robust non-parametric Friedman test [4] for the purpose of comparing various popular classifiers for Pattern Recognition problems encompassing BayesNet [28], SVM [27], Naive Bayes [13] and RBF [3]. The number of datasets (N) and number of classifiers (k) were fixed at 3 and 5 respectively, which implies that each dataset was split into 3 parts. Since the noisy datasets for both 1 and 2 second clips for the Individual Instrument as well as Instrument Family levels were engendered by subjecting the clean datasets to various kinds of noises, so the tests were carried out on the 4 clean datasets (2 Individual Instrument level and 2 Instrument Family level) which are the base datasets of our experiment. The accuracies of each of the classifiers for each of the parts was recorded followed by assignment of a rank (R_j^i) in descending order. R_j^i signifies Rank of j^{th} classifier for i^{th} part). The mean rank of a classifier for the 3 parts was then calculated with the aid of (12).

$$R_j = \frac{1}{N} \sum_i R_j^i \quad (12)$$

Table 4 presents the accuracies and rank distributions for the various parts of the datasets D1 and D2 in the Individual Instrument level. It can be observed from the Table that highest accuracies of 98.29% and 98.70% were obtained for the 1st and 3rd parts of D1 and D2 respectively using ELM. Lowest accuracies of 64.53% and 54.02% were obtained for the 1st and 2nd parts of the respective datasets for LibSVM.

Table 5 presents the accuracies and rank distributions for the various parts of the datasets D1 and D2 in the Instrument Family level. It can be observed from the Table that highest accuracies of 99.77% and 100.00% were obtained for the 1st parts of D1 and D2 respectively using ELM. Lowest accuracies of 75.86% and 78.39% were obtained for the 2nd parts of the respective datasets for Naive Bayes based classification.

The Null hypothesis states that the equivalence of all classifiers ($\forall j, R_j$) is same. In order to verify the same for our experiment, the Friedman Statistic (χ_F^2) [4] was calculated with the help of (13). The set of critical values for (χ_F^2) (distributed in accordance with $k-1$ degrees of freedom) depicts that the value of (χ_F^2) for 4 ($k-1$) degrees of freedom along with significances (α) of 0.05 and 0.10 are 9.488 and 7.779 respectively. The calculated values of

Table 4 Rank Distribution (R) and accuracies (A) for the parts of D1 and D2 at Individual Instrument level

Classifiers		D1 (Partitions)			Mean Rank	D2 (Partitions)			Mean Rank
		1	2	3		1	2	3	
ELM	A	98.29	98.07	97.25	–	98.68	97.93	98.70	–
	R	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Bayes Net	A	83.3	81.81	79.96	–	81.61	79.31	82.54	–
	R	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Naive Bayes	A	71.51	72.20	67.88	–	74.71	74.02	79.59	–
	R	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
Lib SVM	A	64.53	70.14	67.54	–	60.46	54.02	54.86	–
	R	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
RBF	A	92.45	92.22	90.43	–	93.56	89.89	93.20	–
	R	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0

(χ^2_F) for the sets is shown in Table 6 which depicts that the value of (χ^2_F) varies significantly and thus rejects the Null Hypothesis.

$$\chi^2_F = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{13}$$

As per post hoc test, Nemenyi’s test [4] was carried out for comparing each of the classifier pairs. Any two classifiers can be regarded as significantly different performers if their average ranks differ by at least the critical difference (CD), which is calculated using (14). The values of $q_{0.05}$ and $q_{0.10}$ for 5 classifiers in the case of Nemenyi’s test are 2.728 and 2.459 respectively [4] which led to CDs of 3.52 and 3.17 respectively. It was found that similar CD values were obtained for both the datasets at Individual Instrument level which is presented in Table 7 (upper diagonal) with the significantly different pair CD value highlighted in green.

Table 5 Rank Distribution (R) and accuracies (A) for the parts of D1 and D2 at Instrument Family level

Classifiers		D1 (Partitions)			Mean Rank	D2 (Partitions)			Mean Rank
		1	2	3		1	2	3	
ELM	A	99.77	98.29	99.09	–	100.00	99.32	99.33	–
	R	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Bayes Net	A	98.40	85.24	90.43	–	96.78	84.37	93.20	–
	R	2.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0
Naive Bayes	A	95.89	75.86	76.65	–	94.71	78.39	83.67	–
	R	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
Lib SVM	A	96.68	89.24	90.21	–	95.40	78.62	85.26	–
	R	4.0	3.0	4.0	3.67	4.0	4.0	4.0	4.0
RBF	A	97.71	92.22	94.65	–	99.31	90.80	96.37	–
	R	3.0	2.0	2.0	2.33	2.0	2.0	2.0	2.0

Table 6 Values of Friedman’s Statistic for the Datasets

Dataset	Friedman’s Statistic	
Individual Instrument	D1	12
	D2	12
Instrument Family	D1	10.67
	D2	12

In the case of Instrument Family level slightly different CD values were obtained for the classifier pairs for D1 and D2 which is shown in Table 7 (lower diagonal in the order of D1/D2) with the significantly different pair CD value highlighted in blue.

$$CD = q_{\alpha} \sqrt{\frac{k(k + 1)}{6N}} \tag{14}$$

Bonferroni-Dunn [4] test was performed on the datasets to compare the performance of ELM (control classifier) along with the other classifiers. The computational and evaluation procedure of Bonferroni-Dunn’s test is similar to that of Nemenyi’s test. It is only the values of $q_{0.05}$ and $q_{0.10}$ which differ (2.498 and 2.241 respectively) which lead to CDs of 3.22 and 2.89 for the respective significance levels [4]. The calculated values of CD for the classifier pairs for the sets is presented in Table 8 for significance levels of 0.05 and 0.10 respectively. The CDs of the significantly different pairs are highlighted in blue and green for the respective significance values.

5 Result and discussion

The experiments were performed with the aid of a desktop having 16 GB of RAM, along with an I7 processor and Windows 10 operating system. In both the types of datasets, the highest accuracies were obtained in the case of noise Free scenarios. The results in the case of various noisy scenarios is presented and analysed in detail in the subsequent paragraphs. The analysis has been done in 2 phases for presenting a clear picture of the outcome of the experiments. In the 1st phase, the obtained results for the various datasets at Individual Instrument level is discussed. The 2nd phase casts light on the results obtained for Instrument Family level.

Table 7 Results of Nemenyi’s Test on D1 and D2 at Individual Instrument level and Instrument Family level for $q_{0.05}$ and $q_{0.10}$

Classifiers	ELM	Bayesnet	Naive Bayes	LibSVM	RBF
ELM		2.0	3.0	4.0	1.0
BayesNet	2.0/2.0		1.0	2.0	1.0
Naive Bayes	4.0/4.0	2.0/2.0		1.0	2.0
LibSVM	2.67/3.0	0.67/1.0	1.33/1.0		3.0
RBF	1.33/1.0	0.67/1.0	2.67/3.0	1.33/2.0	

Table 8 Results of Bonferroni-Dunn’s Test for the Datasets at $q_{0.05}$ and $q_{0.10}$

Significance Levels	Dataset		Classifier				
			BayesNet	Naive Bayes	LibSVM	RBF	
$q_{0.05}$	Individual Instrument	D1	2.0	3.0	4.0	1.0	
		D2	2.0	3.0	4.0	1.0	
	Instrument Family	D1	2.0	4.0	2.67	1.33	
		D2	2.0	4.0	3.0	1.0	
	$q_{0.10}$	Individual instrument	D1	2.0	3.0	4.0	1.0
			D2	2.0	3.0	4.0	1.0
Instrument Family		D1	2.0	4.0	2.67	1.33	
		D2	2.0	4.0	3.0	1.0	

5.1 Individual instrument level

The obtained accuracies for the various datasets along with the number of Hidden neurons is presented in Table 9. It can be observed from the Table that in the noise free scenario, the highest accuracy was obtained for D1 which is the overall highest among all the experiments. In case of the various noisy scenarios, the accuracies improved significantly on doubling the length of the clips (from D1 to D2). In case of the noisy sets, the highest and lowest accuracies were obtained for the Fan noise scenario and Vacuum Cleaner noise scenario respectively. In the case of 1 second long clip datasets, the performance of the system on the Traffic noise dataset was better than that of the Fan noise dataset which flipped in the case of the 2 second long datasets. An increase in the overall accuracy for all the noisy sets was observed from datasets of 1 second long clips to 2 second long clips. Analysis of the accuracies for those sets reveal that accuracy gains of 3.56%, 5.16%, 2.41% and 4.87% were obtained for the Fan noise, Rain noise, Traffic noise and Vacuum Cleaner noise sets respectively.

Table 9 Obtained Accuracies for various Datasets at Individual Instrument level as well as Instrument Family level using ELM along with number of neurons in the Hidden Layer

Dataset	Accuracy (%) Individual Instrument/ Instrument Family	Number of Hidden Neurons Individual Instrument/ Instrument Family
D1	98.63/96.95	392/355
D2	98.56/97.42	352/404
D3	87.89/83.28	397/470
D4	90.10/87.40	343/426
D5	86.20/78.81	292/197
D6	90.50/85.70	382/444
D7	93.05/87.60	177/234
D8	92.51/90.80	169/206
D9	91.07/83.12	163/221
D10	94.06/86.93	167/119

The Instrument wise accuracies for the various datasets encompassing both the 1 and 2 second long clips is presented in Table 10. It can be observed from the Table that a slightly better performance for Flute was obtained using 1 second long clips rather than 2 second long clips as observed in other instruments in Noise free scenario. One reason for this may be the sensitivity of the instrument to blowing technique as well as ambient air pressure. In the case of noisy sets, best results for Santoor, Violin and Harmonium were obtained for Fan Noise scenario while Flute, Guitar and Piano were most successfully identified in Rain Noise scenario. The best performance for Saxophone was obtained in Traffic Noise scenario.

The comparison of the confusions among the various Instrument pairs for the various datasets in the case of both 1 second (1s) and 2 second (2s) long clips was performed. The confusion matrices are available in the Appendix. The Instruments - Flute, Saxophone, Guitar, Santoor, Violin, Harmonium and Piano are numbered from 1-7 respectively for easier accommodation of the Tables.

It can be observed from the Tables that the highest misclassification for 1 second long clip datasets occurred in the case of Vacuum Cleaner noise scenario where Violin was classified as Piano. In the case of 2 second long clip sets, the highest misclassification was found in the case of Vacuum Cleaner and Rain noise scenarios where Flute was classified as Piano. The highest Individual accuracy in the case of noisy sets was obtained for Guitar in the case of both Fan and Rain noise scenarios for 1 second clip sets and Rain noise scenario among the 2 second long clip sets. The Lowest Individual accuracies were obtained for Santoor in the Rain noise scenario among the 1 second clip datasets and Harmonium in the case of Vacuum Cleaner noise scenario among the 2 second clip sets.

A comparison of the performance of MISNA with some of the systems reported in literature for the identification of Individual Instruments is presented in Fig. 5. Though the compared systems are heterogeneous in the thick of datasets but still they are compared for the sake of a graphical representation of their relative accuracies. The compared works are discussed in Section 2.

5.2 Instrument family level

The obtained accuracies for the various datasets along with the number of Hidden neurons is presented in Table 9. It can be observed from the Table that in the noise free scenario, the highest accuracy was obtained for D2. In case of the various noisy scenarios, the accuracies improved significantly on doubling the length of the clips (from D1 to D2). In case of the noisy sets, the highest and lowest accuracies were obtained for the Traffic noise scenario and Vacuum Cleaner noise scenario respectively. An increase in the overall accuracy for all the noisy sets was observed from datasets of 1 second long clips to 2 second long clips. Analysis of the accuracies for those sets reveal that accuracy gains of 1.23%, 4.32%, 3.40% and 4.31% were obtained for the Fan noise, Rain noise, Traffic noise and Vacuum Cleaner noise sets respectively.

The Instrument Family wise accuracies for the various datasets encompassing both the 1 and 2 second long clips is presented in Table 10. It can be observed from the Table that a fractionally higher accuracy was obtained for 1 second long clips in the case of Keyboard family in contrast to the others in Noise free condition. A probable reason for this could be the effect of fade out and fade in of the notes. In the case of noisy scenario, the best results for all the 3 families were obtained for the Traffic Noise dataset.

Table 10 Accuracy for the Individual Instruments and Instrument Families for 1 and 2 second long clips

Individual Instruments/ Family	Clip Lengths (seconds)	Various Scenarios						
		Noise Free	Fan Noise	Traffic Noise	Rain Noise	Vacuum Cleaner Noise		
Accuracies for Individual Instruments								
Flute	1	98.92	90.67	90.24	88.72	87.42		
	2	98.70	89.57	91.30	91.74	89.13		
Saxophone	1	98.06	91.39	90.83	87.78	84.72		
	2	97.22	95.00	96.11	91.11	90.56		
Guitar	1	99.70	96.98	94.26	96.98	95.77		
	2	100	97.58	94.55	98.79	96.97		
Santoor	1	99.07	84.26	85.19	78.70	85.49		
	2	100	93.83	87.04	93.21	90.74		
Violin	1	97.26	86.59	85.98	84.76	80.49		
	2	98.07	92.68	89.63	89.63	87.80		
Piano	1	99.21	87.66	88.98	87.14	82.41		
	2	99.47	91.58	92.11	93.68	86.84		
Harmonium	1	98.19	95.92	95.24	91.16	87.07		
	2	99.36	98.18	96.82	93.18	95.45		
Accuracies for Instrument Families								
Wind	1	87.69	84.90	88.55	87.21	75.40		
	2	97.32	83.90	92.44	88.54	81.71		
String	1	96.34	77.92	79.25	73.45	75.28		
	2	98.37	83.71	84.11	82.08	78.62		
Keyboard	1	96.84	94.28	94.40	89.17	85.77		
	2	96.59	93.17	95.85	92.20	89.02		

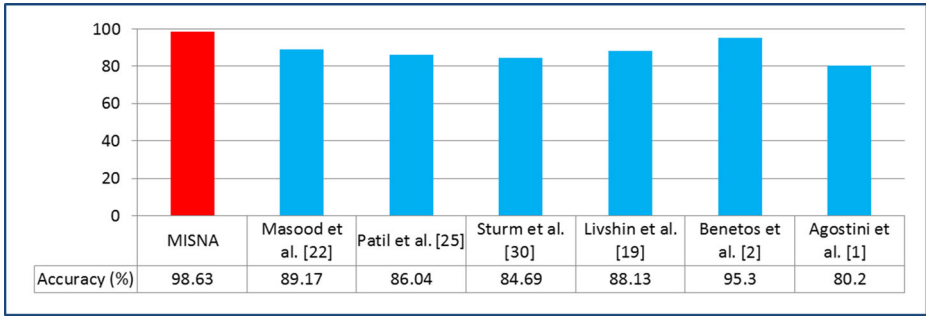


Fig. 5 Comparison of MISNA with some of the existing systems based on Individual Instrument Identification with the Highest Accuracy highlighted in Red

The comparison of the confusions among the Instrument Families for the various datasets in the case of both 1 and 2 second long clips was also performed. The confusion matrices are available in the [Appendix](#). The Families - Wind, String and Keyboard are numbered from 1-3 respectively for easier accommodation of the same.

It can be observed from Tables that the highest misclassification for both 1 and 2 second long clip datasets occurred where String Family was classified as Wind Family in the case of Rain noise scenario and Fan noise scenario respectively. The highest Individual accuracy for both the type of sets was obtained for Keyboard Family in the Traffic noise scenario. The Lowest Individual accuracies were obtained for String Family in the Rain noise scenario among the 1 second clip datasets and Keyboard Family in the case of Vacuum Cleaner noise scenario among the 2 second clip sets.

A comparison of the performance of MISNA with some of the systems reported in literature for the identification of Instrument Family is presented in Fig. 6. Though the compared systems are heterogeneous in the thick of datasets but still they are compared for the sake of a graphical representation of their relative accuracies. The compared works are discussed in Section 2.

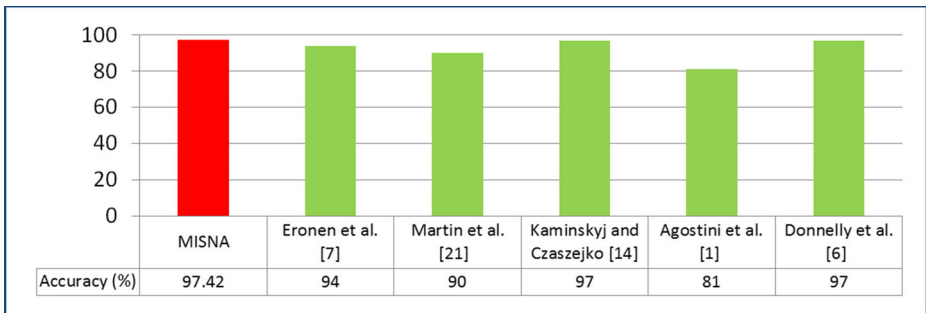


Fig. 6 Comparison of MISNA with some of the existing systems based on Instrument Family Identification with the Highest Accuracy highlighted in Red

6 Conclusion

MISNA is a system which is designed for identification of Individual Instruments as well as Instrument families from audio clips in both clean as well as noisy environments. The system has been tested for various Noisy scenarios with SNRs as low as -9.63 and encouraging accuracies for both type of identifications have been obtained. The system uses a new low dimensional feature namely LPCC-S which overcomes some of the shortcomings of standard LPCC features like uneven as well as large dimensionality. Extreme Learning based classification has also been used in the proposed work which makes the system lightweight in terms of computation due to its ability of generating randomised models. In future, we plan to use various pre processing techniques before feature extraction to filter out noise from the clips as well as for instrument activity detection. Various Feature Dimensionality Reduction techniques will also be experimented with for further dimensionality reduction of the proposed feature in future. We also plan to use other features and classification techniques and test our proposed system on a larger database comprising of a larger number of Instruments.

Appendix

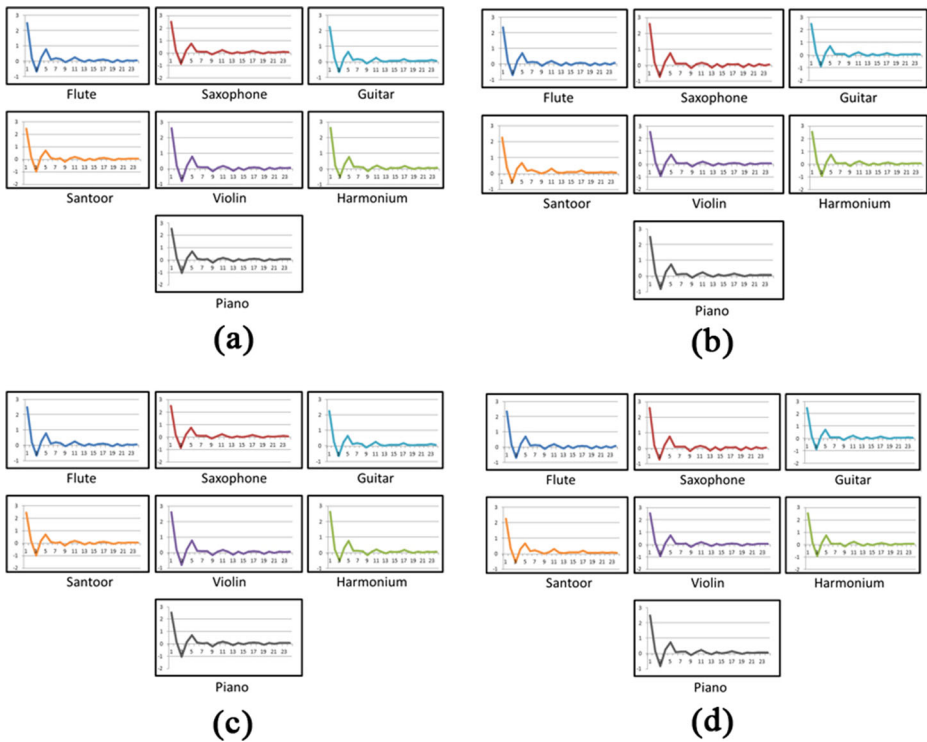


Fig. 7 **a** Feature values for the Instruments for 1 second long Clips in Fan Noise Condition. **b** Feature values for the Instruments for 2 second long Clips in Fan Noise Condition. **c** Feature values for the Instruments for 1 second long Clips in Rain Condition. **d** Feature values for the Instruments for 2 second long Clips in Rain Noise Condition

Table 11 Individual Instrument Confusions for both D1(1s) and D2(2s)

		1	2	3	4	5	6	7
1	1s	—	0	0	0	0.43	0	0.65
	2s	—	0	0.43	0	0	0	0.87
2	1s	0	—	0.56	0	0.28	0	1.11
	2s	0	—	0	0	0	0	2.78
3	1s	0.3	0	—	0	0	0	0
	2s	0	0	—	0	0	0	0
4	1s	0	0	0.31	—	0	0.31	0.31
	2s	0	0	0	—	0	0	0
5	1s	0	0.3	0	0.3	—	0.91	1.22
	2s	0	1.22	0.61	0	—	0	0
6	1s	0	0.26	0	0	0.26	—	0.26
	2s	0	0.53	0	0	0	—	0
7	1s	0.27	0.91	0	0	0.45	0.27	—
	2s	0.91	2.27	0	0	0.45	0	—

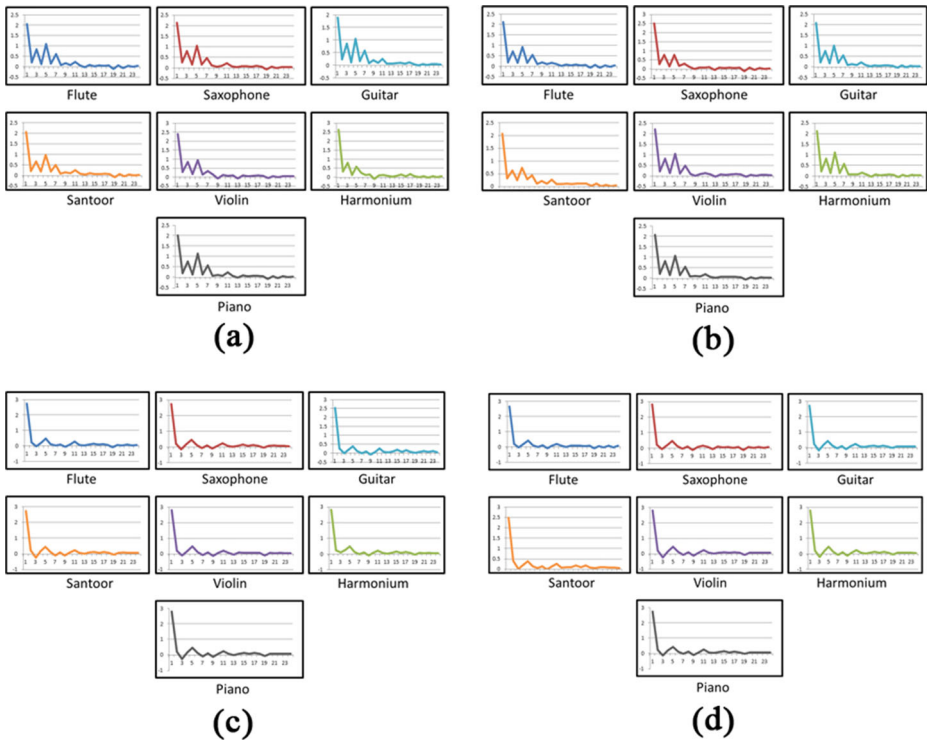


Fig. 8 **a** Feature values for the Instruments for 1 second long Clips in Traffic Noise Condition. **b** Feature values for the Instruments for 2 second long Clips in Traffic Noise Condition. **c** Feature values for the Instruments for 1 second long Clips in Vacuum Cleaner Condition. **d** Feature values for the Instruments for 2 second long Clips in Vacuum Cleaner Noise Condition

Table 12 Individual Instrument Confusions for both D3(1s) and D7(2s)

		1	2	3	4	5	6	7
1	1s	—	1.52	0.65	0	0	0	9.11
	2s	—	0.43	0.87	0	0	0	6.96
2	1s	0.28	—	0	0	1.39	0	10.56
	2s	0	—	0	0	2.22	0	6.67
3	1s	0	0.91	—	0	0	0	2.11
	2s	0	0	—	0	0	0	1.21
4	1s	2.16	5.86	4.63	—	2.78	0.93	4.94
	2s	1.23	1.23	0.62	—	0.62	0.62	2.47
5	1s	0.61	2.13	0.3	0	—	1.22	10.98
	2s	0	1.83	0	0	—	0	8.54
6	1s	0.52	5.25	0	0	3.67	—	3.41
	2s	0	1.05	0	0	1.58	—	3.68
7	1s	1.36	3.17	2.27	0	2.04	0	—
	2s	0.45	3.64	0.91	0	1.82	0	—

Table 13 Individual Instrument Confusions for both D4(1s) and D8(2s)

		1	2	3	4	5	6	7
1	1s	—	0.43	0.65	0	0	0	8.68
	2s	—	0.87	0.43	0	0	0	7.39
2	1s	0	—	0	0	0.28	0	8.89
	2s	0	—	0	0	0	0	3.89
3	1s	0	0	—	0	0	0	5.74
	2s	0	0	—	0	0	0	5.45
4	1s	0.93	2.47	5.86	—	0.62	0	4.94
	2s	0.62	0.62	9.88	—	0	0.62	1.23
5	1s	0	2.74	0.61	0	—	1.52	9.15
	2s	0	1.83	0	0	—	0.61	7.93
6	1s	0.79	0	0	0	3.41	—	6.82
	2s	0	0	0	0	2.11	—	5.79
7	1s	0.23	1.13	3.17	0	0.23	0	—
	2s	0	1.36	1.82	0	0	0	—

Table 14 Individual Instrument Confusions for both D5(1s) and D9(2s)

		1	2	3	4	5	6	7
1	1s	—	1.74	1.3	0	0.43	0	9.11
	2s	—	0.43	0.87	0	0	0	9.57
2	1s	0.28	—	2.78	0	3.06	0.56	8.61
	2s	0	—	0	0	2.22	0	7.22
3	1s	0	0.3	—	0	0.3	0	3.63
	2s	0	0	—	0	0	0	3.03
4	1s	0.31	2.16	5.25	—	1.54	0.62	4.63
	2s	0	0	3.7	—	0.62	1.23	3.7
5	1s	0	4.88	0	0	—	1.52	13.11
	2s	0	1.83	0.61	0	—	1.22	8.54
6	1s	0	6.82	0	0	4.2	—	6.56
	2s	0.53	5.79	0	0	1.05	—	5.79
7	1s	2.49	3.4	3.85	0	2.95	0.23	—
	2s	0.45	1.36	0	0	2.27	0.45	—

Table 15 Individual Instrument Confusions for both D6(1s) and D10(2s)

		1	2	3	4	5	6	7
1	1s	—	0	0.43	0	0	0	8.89
	2s	—	0.87	0	0	0	0	9.57
2	1s	0	—	0	0	0.28	0	8.33
	2s	0	—	0	0	0	0	5
3	1s	0	0	—	0	0	0	3.02
	2s	0	0	—	0	0	0	2.42
4	1s	0.31	2.16	5.86	—	0.31	0.62	6.48
	2s	0	1.85	2.47	—	0	0.62	1.23
5	1s	0	1.52	0.61	0	—	0.91	10.37
	2s	0	0	0.61	0	—	0	6.71
6	1s	0.52	0.26	0	0	3.41	—	8.14
	2s	0	0	0	0	3.16	—	5.26
7	1s	0	0.45	2.95	0	0.68	0	—
	2s	0	1.36	0.45	0	0	0	—

Table 16 (a) Instrument Family Confusions for both D1(1s) and D2(2s)

		1	2	3
1	1s	—	0.97	1.34
	2s	—	0.73	1.95
2	1s	0.92	—	2.75
	2s	0.41	—	1.22
3	1s	2.43	0.73	—
	2s	1.95	1.46	—

		1	2	3
1	1s	—	2.92	9.87
	2s	—	3.41	8.05
2	1s	5.09	—	21.46
	2s	7.13	—	10.79
3	1s	4.99	5.84	—
	2s	2.68	5.12	—

(a)

(b)

(b) Instrument Family Confusions for both D3(1s) and D7(2s). (c) Instrument Family Confusions for both D4(1s) and D8(2s). (d) Instrument Family Confusions for both D5(1s) and D9(2s). (e) Instrument Family Confusions for both D6(1s) and D10(2s)

		1	2	3
1	1s	—	0.85	10.6
	2s	—	0.49	7.07
2	1s	6	—	14.75
	2s	2.65	—	13.24
3	1s	2.07	3.53	—
	2s	1.95	2.2	—

		1	2	3
1	1s	—	6.82	17.78
	2s	—	5.85	12.44
2	1s	5.09	—	19.63
	2s	8.55	—	12.83
3	1s	7.3	6.93	—
	2s	5.85	5.12	—

(c)

(d)

		1	2	3
1	1s	—	0.49	14.62
	2s	—	2.2	13.9
2	1s	5.9	—	16.17
	2s	0.81	—	15.48
3	1s	1.58	4.14	—
	2s	2.68	4.15	—

(e)

References

- Agostini G, Longari M, Pollastri E (2003) Musical instrument timbres classification with spectral features. EURASIP J Appl Signal Process 2003:5–14
- Benetos E, Kotti M, Kotropoulos C (2007) Large scale musical instrument identification. In: 4th Sound and music computing conference, pp 283–286
- Biernacki A (2017) Analysis and modelling of traffic produced by adaptive HTTP-based video. Multimed Tools Appl 76(10):12347–12368
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7: 1–30
- Deshmukh S, Bhirud S (2014) Analysis and application of audio features extraction and classification method to be used for North Indian Classical Musics singer identification problem. Int J Adv Res Comput Commun Eng, 3(2)
- Donnelly PJ, Sheppard JW (2013) Classification of musical timbre using bayesian networks. Comput Music J 37(4):70–86
- Eronen A, Klapuri A (2000) Musical instrument recognition using cepstral coefficients and temporal features. In: 2000 IEEE International conference on acoustics, speech, and signal processing, 2000. ICASSP'00. Proceedings, vol 2. IEEE, pp II753–II756
- Fragoulis D, Papaodysseus C, Exarhos M, Roussopoulos G, Panagopoulos T, Kamarotos D (2006) Automated classification of piano-guitar notes. IEEE Trans Audio Speech Lang Process 14(3):1040–1050
- Huang GB (2014) An insight into extreme learning machines: random neurons, random features and kernels. Cogn Comput 6(3):376–390
- Huang GB, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern Part B (Cybern) 42(2):513–529
- Huang GB, Bai Z, Kasun LLC, Vong CM (2015) Local receptive fields based extreme learning machine. IEEE Comput Intell Mag 10(2):18–29

12. Jadhav PS (2015) Classification of musical instruments sounds by using MFCC and Timbral audio descriptors. *Int J Recent Innov Trends Comput Commun*, 3(7)
13. Jitpakdee P, Uyyanonvara B (2017) Computer-aided detection and quantification in glistenings on intraocular lenses. *Multimed Tools Appl*, 1–14
14. Kaminsky I, Materka A (1995) Automatic source identification of monophonic musical instrument sounds. In: *IEEE International Conference on neural networks*, 1995. Proceedings., vol 1. IEEE, pp 189–194
15. Kaminsky I, Czaszejko T (2005) Automatic recognition of isolated monophonic musical instrument sounds using kNNC. *J Intell Inf Syst* 24(2):199–221
16. Kitahara T, Goto M, Okuno HG (2005) Pitch-dependent identification of musical instrument sounds. *Appl Intell* 23(3):267–275
17. Lita AI, Ionescu LM, Mazare AG, Serban G, Lita I (2016) Real time system for instrumental sound extraction and recognition. In: *2016 39th International spring seminar on electronics technology (ISSE)*. IEEE, pp 456–461
18. Liu J, Xie L (2010) Comparison of performance in automatic classification between Chinese and Western musical instruments. In: *2010 WASE International conference on information engineering (ICIE)*, vol 1. IEEE, pp 3–6
19. Livshin A, Rodet X (2004) Musical instrument identification in continuous recordings. In: *Digital audio effects 2004*, pp 1–1
20. Livshin A, Rodet X (2009) Purging musical instrument sample databases using automatic musical instrument recognition methods. *IEEE Trans Audio Speech Lang Process* 17(5):1046–1051
21. Martin KD, Kim YE (1998) 2pMU9. Musical instrument identification: a pattern-recognition approach. In: *Presented at the 136th meeting of the acoustical society of America*
22. Masood S, Gupta S, Khan S (2015) Novel approach for musical instrument identification using neural network. In: *2015 Annual IEEE on India conference (INDICON)*. IEEE, pp 1–5
23. Mukherjee H, Rakshit P, Phadikar S, Roy K (2016) REARC-A Bangla phoneme recognizer. In: *2016 International conference on accessibility to digital World (ICADW)*. IEEE, pp 177–180
24. Mukherjee H, Halder C, Phadikar S, Roy K (2017) READ-A Bangla phoneme recognition system. In: *Proceedings of the 5th international conference on frontiers in intelligent computing: theory and applications*. Springer, Singapore, pp 599–607
25. Patil SD, Pattewar TM (2015) Musical instrument identification using SVM & MLP with formal concept analysis. In: *2015 International Conference on green computing and internet of things (ICGCIoT)*. IEEE, pp 936–939
26. Petruncio D, Hasegawa-Johnson MA (2002) Evaluation of various features for music genre classification with hidden Markov models. University of Illinois, Master's thesis
27. Rai A, Singh HV (2017) SVM based robust watermarking for enhanced medical image security. *Multimed Tools Appl*, 1–14
28. Ri CY, Yao M (2015) Bayesian network based semantic image classification with attributed relational graph. *Multimed Tools Appl* 74(13):4965–4986
29. Röver C, Klefenz F, Weihs C (2005) Identification of musical instruments by means of the Hough-transformation. In: *Classification—the ubiquitous challenge*. Springer, Berlin, pp 608–615
30. Sturm BL, Morvidone M, Daudet L (2010) Musical instrument identification using multiscale mel-frequency cepstral coefficients. In: *2010 18th European signal processing conference*. IEEE, pp 477–481
31. Takahashi Y, Kondo K (2014) Comparison of two classification methods for Musical Instrument identification. In: *2014 IEEE 3rd Global conference on consumer electronics (GCCE)*. IEEE, pp 67–68
32. Tang J, Deng C, Huang GB (2016) Extreme learning machine for multilayer perceptron. *IEEE Trans Neural Netw Learn Syst* 27(4):809–821
33. Yu F, Chen Y (2015) Musical instrument classification based on improved matching pursuit with instrument-specific atoms. In: *2015 IIAI 4th International congress on advanced applied informatics (IIAI-AAI)*. IEEE, pp 506–510
34. Yu J, Chen X, Yang D (2008) Chinese folk musical instruments recognition in polyphonic music. In: *International Conference on audio, language and image processing*, 2008. ICALIP 2008. IEEE, pp 1145–1152



Himadri Mukherjee has completed B.Sc. in Computer Science from Acharya Prafulla Chandra College, M.Sc. (Gold Medalist) in Computer Science from West Bengal State University in the year 2013 and 2015 respectively. He is currently a Research Scholar in the Department of Computer Science of the same University. He has published 4 Research papers in reputed conferences. His research interest includes Audio Signal Processing, Speech Recognition, Speaker Identification, Audio based Musical Information Retrieval, Pattern Recognition and Image Processing.



Sk Md Obaidullah has completed B.E in Computer Sc. & Engineering from Vidyasagar University, M.Tech in Computer Sc. & Application from University of Calcutta in the year 2004 and 2009 respectively. He has more than eleven years of professional experience including two years in industry and nine years in academia out of which five years of research. He is a registered PhD candidate in the Dept. of Computer Sc. & Engineering, Jadavpur University since Nov. 2014. Presently he is working as an Assistant Professor in the Dept. of Computer Science & Engineering, Aliah University, Kolkata. He has published more than 25 research papers in peer reviewed journal and national/international conferences. His research interests are document image analysis, medical image analysis and image processing applications for biometric, Audio Signal analysis. He is Life Member of IUPRAI (an unit of IAPR) and Associate Member The Institution of Engineers (India).



Dr. Santanu Phadikar has completed B.Sc. in Computer Science from Vidyasagar University, M.Sc. in Computer and Information Science, M. Tech. in Computer Science and Engineering from University of Calcutta and Ph.D. from IIST, Shibpur in the years 1999, 2001, 2003 and 2013 respectively. He has more than 14 years of teaching experience and is currently working as an Associate Professor in the Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology. He has published more than 30 Research papers and articles in reputed conferences and journals. His research interest includes Image Processing, Document Image Analysis, Speech Recognition, Audio based Musical Information Retrieval, Intelligent and Precision Farming.



Prof. Kaushik Roy has completed B.E in Computer Sc. & Engineering from NIT Silchar, M.E and PhD(Engg.) in Computer Sc. & Engg. from Jadavpur University in the year 1998, 2002 and 2008 respectively. He is currently working as a Professor and Head of the Department of Computer Science, West Bengal State University, Barasat, India. In 2004 he has received Young IT Professional award from Computer Society of India. He has published more than 100 research papers/book chapters in reputed conferences and journals. His research interest includes pattern recognition, document image processing, medical image analysis, online handwriting recognition, speech recognition and audio signal processing. He is Life Member of IUPRAI (an unit of IAPR) and Computer Society of India.