Journal of Scientific & Industrial Research
Vol. 82, January 2023, pp. 26-37
DOI: 10.56042/jsir.v82i1.70251

# A Comprehensive Review on Audio based Musical Instrument Recognition: Human-Machine Interaction towards Industry 4.0

Sukanta Kumar Dash[1]*, S S Solanki[1] & Soubhik Chakraborty[2]

[1]Department of Electronics and Communication Engineering, [2]Department of Mathematics, Birla Institute of Technology, Mesra, Ranchi 835 215, Jharkhand, India

*Received 30 May 2022; revised 20 September 2022; accepted 07 October 2022*

Over the last two decades, the application of machine technology has shifted from industrial to residential use. Further, advances in hardware and software sectors have led machine technology to its utmost application, the human-machine interaction, a multimodal communication. Multimodal communication refers to the integration of various modalities of information like speech, image, music, gesture, and facial expressions. Music is the non-verbal type of communication that humans often use to express their minds. Thus, Music Information Retrieval (MIR) has become a booming field of research and has gained a lot of interest from the academic community, music industry, and vast multimedia users. The problem in MIR is accessing and retrieving a specific type of music as demanded from the extensive music data. The most inherent problem in MIR is music classification. The essential MIR tasks are artist identification, genre classification, mood classification, music annotation, and instrument recognition. Among these, instrument recognition is a vital sub-task in MIR for various reasons, including retrieval of music information, sound source separation, and automatic music transcription. In recent past years, many researchers have reported different machine learning techniques for musical instrument recognition and proved some of them to be good ones. This article provides a systematic, comprehensive review of the advanced machine learning techniques used for musical instrument recognition. We have stressed on different audio feature descriptors of common choices of classifier learning used for musical instrument recognition. This review article emphasizes on the recent developments in music classification techniques and discusses a few associated future research problems.

**Keywords:** Classifier learning, Feature descriptors, Instrument recognition, Multimodal communication, Music information retrieval

## Introduction

Music is one of the natural forms of art that spreads its essence over our minds. It has a substantial social and physiological impact. With the advent of technology and the industry 4.0 revolution, researchers have now focused on the human-machine interaction, a multimodal communication. In response, there has been an enormous growth in the music industry. Distribution and capturing of music data have become more accessible now. All these lead to a massive repository of music data. Automatic identification of music characteristics can act as a fundamental step towards organized storage and retrieval of music data. In the context of Music Information Retrieval (MIR), proper organization of the extensive collection of music data is very important. Music data can be archived in a structured manner based on metadata. The extraction of such metadata may be manual, where a domain expert annotates the piece of music or a text-based metadata-oriented query. The annotation problem may be less severe nowadays as different music formats embedded metadata in them.[1] But music recorded from other sources lack this information. A primary concern arises when the user does not provide the metadata as a music query rather submit the music clip as the query and expect the music with similar characteristics. Thus, a content-based music retrieval system becomes essential, automatically extracting the properties from the query signal and comparing them with the same obtained from the music signal in the database. Automatic classification of music signals based on genre, singer, instrumental, etc.[2–6] has gained impetus over the last decades. It can be crucial in various applications like music retrieval and recommendation systems, archival and indexing of music database, and annotating a music database. Nowadays, Automatic Musical Instrument Recognition (AMIR) systems are more accurately presented through the industry 4.0 revolution.

*Author for Correspondence
E-mail:skdash@bitmesra.ac.in

Automatic musical instrument recognition is vital as a high-level subtask in MIR. The goal of AMIR is to identify the different types of instruments in the raw music at different time intervals. Unlike genre, mood, and artist classification, instrument recognition is a sequence of labelling tasks based on which the music classification may be tagged to monophonic or polyphonic instrument labels. Early instrument recognition research was focused on recognizing instruments from solo played music with one instrument at a time. Hence, this is far from real-world music performances. Now the researcher's focus has switched from solo to polyphonic music to deal with several instruments simultaneously.[6] Instrument recognition in polyphonic music is more complicated than its monophonic counterpart.[7] The fact that the source to be detected often corresponds to a very minimal proportion of the overall energy of the mixture signal makes polyphonic music instrument recognition extremely difficult. Also, the identities of the other instruments are frequently unknown; the interference induced by them is very non-stationary and unpredictable.[8] Although there have been a lot of pioneer works on music classification, it is worth mentioning that AMIR is now becoming an evolving task in music classification. Automatic Musical Instrument Recognition has two impacts on music recording. One is the retrieval of musical instruments played in the music, and the other is managing the music audio. As a result, musical instrument identification and retrieval is a critical step in organizing a database to allow for faster access to needed data, automatic indexing of musical data and database retrieval applications. Therefore, developing an AMIR system is very much needed. To support this, a comprehensive survey is required on current advances in this field.

This review paper presents a summary of different audio features and classification techniques used for AMIR task. We focus on musical instrument recognition based on audio signals unless otherwise stated. It is to be noted that the details of other MIR tasks like genre classification, mood classification, artist classification, and music annotation are beyond the scope of this study. The rest of the review article is prepared as follows. We present a concise study on different classes of audio features used for musical classification. Thereafter, we discuss different classifier-feature learning techniques of common choices used for AMIR task. A few unresolved research issues for further study are discussed. In the end, we have concluded our review with a conclusion.

## Overview of Audio Features

Audio features are vital parameters for categorizing a piece of music and musical instruments. There are no rules on what audio features are appropriate for what type of music (monophonic/polyphonic) or instrument kinds. Choosing the optimum audio feature descriptors and classification methods plays a key role in correctly identifying musical instruments.[9] Many different types of audio features have been proposed for the assignment of sound description coming from the speech recognition community and the prior studies on musical instrument sound classification.[10] A complete description of features is described by Peeters.[11] Further, Weihs *et al.*[12] categorized the complete set of audio features taxonomically. From the standpoint of music comprehension, the audio features may be divided into five different classes as follows.

### Timbral Features

Timbral features are utilized to differentiate between sound signals with similar pitch and rhythmic content. The tonal quality of a sound signal associated with varied instruments is captured by timbral features. Timbral features are used predominantly in music classification. To extract the timbral features of a sound signal, the signal is divided into short-time frames which are statistically stationary, by employing windowing function at fixed intervals. In common practice, a Hamming windowing function is used that removes the edge-effects. For each frame, the timbral features are computed with its statistical mean and variance.

### Temporal Features

Extraction of temporal features typically relies on timbral features over the length of timbral feature extraction, which captures the timbre variation over time. The temporal features are created by combining timbre features extracted from several frames. We will be able to construct a richer collection of features for categorization by combining timbre and temporal features in this way. Extraction of temporal features was done in the time domain, requiring less computational effort, and making them easy to put into practice.

### Cepstral Features

Cepstral features are computed from the cepstral analysis of the signal. Cepstral analysis is a nonlinear signal processing approach used in speech and image processing domains. Cepstral features are used to distinguish between dominating instrumental sources. Cepstrum values clustered at the origin provide information about the instrument filter, whereas values distant from the origin provide information about the excitation source.

#### Spectral Features

These features are obtained from the signal transformation in the frequency domain. Generally, an audio signal is segmented into several short-time duration signals followed by a suitable windowing technique to get various local frames. Various signal processing tools can be applied to the local frame signals to obtain different spectral features.

#### Perceptual Features

Perceptual features are generally based on human hearing mechanism. These features are found over local frames of small-time duration varying from 10 ms to 100 ms for music audio samples. Each frame signal is then analysed to obtain different perceptual features.

## Classifier-Feature Learning

In various disciplines, Machine Learning (ML) techniques[13,14] have enabled major breakthroughs in automated data processing and pattern identification capabilities. This includes image processing, speech processing, and musical sound processing. The two major kinds of machine learning methodologies are unsupervised and supervised learning. The third type of learning is known as reinforcement learning; however, it is not included in this paper.

With labelled input and output pairs, the task of supervised learning is to find a correct mapping between input and output. The most extensively used ML category is supervised learning. K-Nearest-Neighbour classifiers (K-NN),[10,15–20] support vector machine (SVM),[7,9,16,21–34] Neural Network (NN) models, also known as Artificial NNs (ANN),[35–50] and Hidden Markov Models (HMM)[51–55] are some examples of supervised learning. Unsupervised learning has no labels; therefore, the purpose is to uncover the very useful structure of data. Anomaly detection, exploratory data analysis, feature learning, and data visualization are just a few available applications. Unsupervised approaches like Principal Components Analysis (PCA),[56] K-means[57] and Gaussian Mixture Models (GMMs)[15,58–64] as well as matrix factorization methods like Non-negative Matrix Factorization (NMF),[68,69] Independent Component Analysis (ICA)[70] and Linear Discriminant Analysis (LDA)[71,72] have been used for decades.

Prior machine learning approaches were constrained to analyse input data in terms of its raw form. As a result, the input to the learning system, normally a classifier, must typically be a hand-crafted portrayal of the feature. Thereby, substantial field expertise and a thorough engineering approach are required.[47] In traditional classification, we are given training and testing data sets, with each example labelled. From the available labelled training data set, a huge array of features are extracted for the given audio samples of individual instruments to capture varied characteristics of the individual class of instruments. Feature descriptors are weighed and evaluated on unlabeled data in a test set. A suitable classifier then classifies the test samples after being trained.[10] Several works[73–76] have used a variant of Martin & Kim[15] classical pattern recognition approach for AMIR.

The generic problem found in the classification of musical instruments is three-fold: (1) data pre-processing, (2) extraction of features, and (3) feature classification.[77] In instrument classification, finding a compact, effective, and robust feature set is the key difficulty. Furthermore, choosing the right classifier is critical for enhancing classification accuracy. A few common choices of classifiers and feature learning for musical recognition tasks are stated below.

#### Classifier Learning

##### *K-Nearest Neighbour (K-NN)*

The K-NN algorithm is the frequently used learning-based algorithm, which preserves the feature vectors of all the training examples. Then, for classifying a new instance, it finds a set of *k* nearest training examples in the feature space and assigns the new example to the class that has more examples in the set. To determine similarity, the Euclidean distance measure has long been used. Despite its ease of implementation, the K-NN technique is sluggish, among others.

Martin and Kim[15] created a hierarchical classification method using a k-NN model applied to a database of 1023 isolated sound samples from 15 orchestral instruments. When no hierarchy was used, they had an 87% family classification success rate and a 61% instrument classification success rate.

The hierarchical technique improved instrument accuracy to 67% but lowered family performance to 79%. Agostini *et al.*[16] discovered a similar result of 66% for 27 instrument classes and 77% for a two-level six-element hierarchy. In this study, the k-NN approach performed poorly when compared to discriminant functions and Support Vector Machines. Eronen & Klapuri[17] correctly identified instrument families (brass, string, and woodwind) with a 94% accuracy rate and individual instruments with an 80% accuracy rate, covering the full ranges of 30 orchestral instruments played with various articulation styles. 44 spectral and temporal features were computed for audio sounds primarily from the MUMS collection. Cross-validation was utilized using Gaussian and k-NN classifiers, with 70 percent/30 percent splits of train and test data. Fujinaga *et al.*[18] reported the recognition rate 50% for 23 orchestral instruments with over 1300 notes from the McGill CD library and 81% for a 3-instrument group. To find the best set of weights for the features, they used a k-NN classifier and a genetic algorithm. Eronen[19] tested a data collection that includes 5286 acoustic and synthetic solo tones from the database with 29 varied western symphonic instruments, 16 of which were included in the test set. Researchers employed a K-NN classifier to extract MFCC, delta MFCC, LPCC, and modulation features. The best results were achieved in solo tone recognition, with 35% for solitary instruments and 77% for families. MFCC delivered the best performance. Kaminskyj & Czaszejko[20] used k-NN classifiers to recognize individual monophonic instrument sounds. Amplitude envelope, constant Q transform, MFCC, and Multi-Dimensional Scaling (MDS) analysis trajectories were used. These features were reduced to a total dimensionality of 710 using Principal Component Analysis (PCA). After that, the k-NN classifiers were trained using various hierarchical schemes. The experiment resulted in 93% recognition accuracy for individual instruments and a 97% for instrument families. Using computerized musical instrument recognition and outlier detection approaches**,** algorithms for automatic elimination of poor instrument samples were proposed by Livshin & Rodet.[10] A K-NN classifier was utilized to test the proposed techniques utilizing 162 feature descriptor values from a database of 20 musical instruments. On a meticulously contaminated sound data set, the introduced Multiclass Interquantile Range (MCIQR) methodology obtained the best evaluation results, deleting 70.1% "bad" samples with a 0.9% false-alarm rate and 90.4% with an 8.8% false-alarm rate.

***Support Vector Machine (SVM)***

SVM is a cutting-edge binary classifier that works on the big margin principle. For the given labelled instances from two classes, SVM finds the optimal separating hyperplane that optimizes the distance between support vectors and the hyperplane. The support vectors are the instances closest to the hyperplane whose labels are most likely to be confused. As a result, the SVM performs better in classification since it focuses on difficult instances.[21]

Marques[22] employed an SVM model on a database of eight solo instruments played by a group of composers. The best classification result was 70% by considering 16 MFCC feature vectors from a sound segment of 0.2 seconds. When she tried to classify longer segments, she found an 83% gain in accuracy. Marques & Moreno[7] developed an SVM classifier based on GMM that can discriminate between eight instruments. The classifier was built with 70% accuracy after testing various parameters such as feature type and classification algorithm. The researchers investigated cepstral, mel-cepstral, and linear prediction coefficients. Agostini *et al.*[16] reported a paper on the content-based classification of musical timbres using SVM, KNN, CDA (Canonical Discriminant Analysis), and QDA (Quadratic Discriminant Analysis). The SVM with RBF (Radial Basis Functions) kernel with 18 descriptors was discovered as the best classifier in recognition of individual instruments within a 46 ms frame, with a success rate of 69.7%, 78.6%, and 80.2% for 27, 20, and 17 instruments, respectively. Essid *et al.*[23] established a new approach to instrument recognition (9 instruments and voice) ranging from solos to quarters considering real music orchestrations. A wide set of acoustic features like temporal features, cepstral features, and perceptual features were extracted and applied to the SVM classifier for recognition and achieved an average accuracy of 53%. Essid *et al.*[24] published a paper on the musical classification that used natural and instrument hierarchical taxonomies. The experiment includes ten instrument classes from various instrument families. The proposed taxonomy comprises cepstral, spectral, perceptual, and MFCC features. Using Gaussian mixture models (GMMs) and SVM, they achieved recognition rates of 87%. To solve the musical instrument classification problem, Deng *et al.*[9] retrieved a vast set of features which includes timbral, perceptual, spectral, along with MPEG-based features. The authors reported an individual classification accuracy of 86.9% and a family classification accuracy of 96.5% for 20 instruments of UIOWA MIS collection using the SVM classifier. They also experimented with a variety of other classifier algorithms, including naive Bayes, k-NN, RBF, and MLP (Multilayer Perceptron). The system performance obtained was most outstanding for MFCC features. Joder *et al.*[25] used a Feature Selection Algorithm (FSA) to extract a wide range of information from solo music of 10 musical instruments. The complete set of features includes temporal, cepstral, spectral, and wavelet features. Binary support vector machine (SVM) classifiers were trained with the produced feature set. They found an average accuracy of 77.82% for one vs. one, 79.08% with PCA, and 73.88% for MAP. Fuhrmann *et al.*[26] proposed a method for automatically categorizing dominating instruments using SVM classifiers trained with features derived from real musical audio data containing 11 pitched and 3 percussive instruments. The pitched and percussive classification tasks reported accuracy of 63% and 78%, respectively. Fuhrmann & Herrera[27] trained the SVM model using tailor-made timbral features based on mean and variance statistics on a data set of 11 modelled instruments. They found precision values up to 0.86 and F-measures larger than 0.65. Ozbek *et al.*[28] released a paper using time-frequency energy localization to classify musical instruments automatically on a database from eight different instruments using a support vector machine (SVM). They got maximum accuracy of around 93% for individual instrument recognition. Wu *et al.*[29] proposed an Expectation-Maximization (EM) technique for polyphonic pitch estimation and instrument identification. The suggested approaches are based on the EM algorithm's estimation of the parameters of each note's spectro-temporal GMM model. The logarithmic transformation and PCA are then used to convert these parameters into a low-dimension timbre feature vector. Finally, using the obtained low dimension timbre features, SVM classifiers were trained to recognize musical instruments with 71% accuracy. By previously dividing the original signal into numerous streams, Bosch *et al.*[30] addressed the identification of dominating musical instruments in polytimbral audio. Concerning the segregation method and the employed SVM model for classification, several strategies are assessed, ranging from low to high complexity. The dataset was from professionally produced recordings, which are notoriously difficult to separate using current source separation algorithms. The instrument recognition accuracy was measured at about 50%. A new cepstrum representation approach, Unified Discrete Cepstrum (UDC), was proposed by Duan *et al.*[31] UDC had the advantage of preventing overfitting by utilizing a natural local adaptive regulator. The authors employed an SVM classification model with UDC and its mel-scale counterpart MUDC (male-scale variant of UDC) to classify 13 different instruments. They reported recognition accuracy for two polyphonic musical notes as 37% and for six polyphonic music notes as 25%. Yu *et al.*[32] employed temporal sum pooling and sparse coding (SC) on cepstrum and used the LIBLINEAR library to train a linear SVM model for classification. The proposed system achieved an F-measure of about 96% in the uni-source dataset and about 69% in the multi-source dataset for classifying 50 instruments. Lin *et al.*[33] developed an audio classification technique employing wavelets and support vector machines (SVMs) to classify 410 audio sounds in 16 different classes. Authors proposed a bottom-up SVM technique over acoustical features along with frequency cepstral coefficients extracted from a public audio database for the audio classification. Experimental results have shown a reduction in classification error from 8.1% to 3.0% and classification accuracy about 100%. Mandel & Ellis[34] employed Support Vector Machines (SVMs) to classify music. They used MFCC as input feature and KL (Kullback-Leibler) divergence-based kernel to measure the song-level similarity. In their experiments, they obtained the classification accuracy of 72.45% and 78.81% for the audio genre and artist identification respectively.

### *Artificial Neural Networks (ANNs)*

ANN is a data processing structure that consists of several interconnected neurons that work together to solve a problem. ANN learns to map between input and output data vectors by adjusting the assigned weights of connecting neurons. Alterations can occur under supervised or unsupervised conditions. A few communicated pioneer works are briefly discussed here.

Kaminsky & Materka[35] used a relatively simple feed forward network with back propagation training algorithm learned to distinguish audio sounds from four different instruments with three input units, five hidden units, and four output units. They achieved an accuracy rate of 97%. Cemgil & Gurgen[36] used a self-organizing three-layer hybrid network to investigate 40 sounds from 10 distinct instrument classes. In experimental results, they obtained a success rate of 97%, 100%, and 94% for three distinct network model architectures. Kostek's team (Kostek & Krolikowski,[37] Kostek & Czyzewski[38]) carried out **s**everal experiments on feedforward neural networks (FNNs) with one hidden layer. Initially, the instruments chosen had similar sounds, but as the test progressed, more categories were introduced latter. For different sets of four classes, accuracy rates of more than 90% were reached. Kostek[39] researched on the classification of 12 distinct instruments played with various articulations. She used multilayer neural networks (NN) trained on FFT (Fast Fourier Transform) based features. It was demonstrated that combining wavelet transform features with MPEG-7 features, the classification accuracy improved to a range of 55% to 98%, with an average of around 70%. Loughran *et al.*[40] presented a classification model to classify musical instruments using MFCC and PCA utilizing multilayer perceptrons. A multi-layered perceptron was trained using principal component analysis. The first 3, 4, and 5 principal components computed from the envelope of the changes in the coefficients were used to train the network. Using four principal components from the first 15 coefficients, they achieved a classification accuracy of 95.88%. Newton & Smith[41] suggested tone descriptors for the classification of musical instruments. In experiments, authors used 2085 isolated musical tones over five instrument groups from the McGill dataset. A neurally inspired tone descriptor was developed using a model of the auditory system's response to sound onset. The neurally-inspired technique had a classification success rate of roughly 75%.

However, deep learning[14] is a method of stacking numerous layers of nonlinear modules for automatically constructing a higher-level representation from raw data. A feedforward network (FNN) with numerous hidden layers of artificial neurons is known as a deep neural network (DNN) or multi-layer perceptron (MLP). When feedback links are introduced to a network, it becomes the recurrent neural network (RNN). RNNs are pretty effective for sequential inputs. RNNs are successful in language modelling[42] and spoken language interpretation.[43,44] A different classifier, convolutional neural network (CNN)[45] is a modified version of the normal neural network model, which finds convolutions on different segmented input signals. Therefore, this model is used to classify audio signals utilizing different timbre features. This is proven in the work of Lee *et al.*[46] with generic audio classification applications. This is utilized in a multilayer CNN extension called a convolutional deep belief network (CDBN). Han *et al.*[47] proposed a CovNet network structure for a vast dataset of 10 instruments and human voice for the recognition of predominant musical instrument task. The CovNet receives mel-spectrogram as an input feature. The network was trained using IRMAS' single-labeled training data, while the multi-labelled test data was used to identify the pre-dominating instrument. The proposed architecture received a micro F1 measure of 0.619 and a macro F1 measure of 0.513. Gururani *et al.*[48] proposed a DNN-based instrument activity detection (IAD) system for detecting the activity of 18 polyphonic instruments. This model was trained using mel-spectrograms obtained from the multitrack datasets. They used one-second audio clips to train the deep neural network and acquired the final prediction score of 80.92 AUC for CNN and 79.22 AUC for CRNN (Convolutional Recurrent Neural Network), per instrument. Gomez *et al.*[49] proposed a hybrid deep neural network-based instrument recognition system which is a combination of fully connected convolutional layers for learning characteristics of spectral-temporal patterns. This is a blend of convolutional and fully connected layers. They investigated the effects of two source separation strategies on instrument recognition for six jazz solo instruments, as well as used transfer learning to fine-tune the trained model. The proposed network shown a micro measure of 0.805 and a macro measure of 0.803. Yu *et al.*[50] used the previously published work of Han *et al.*[47] as a baseline and conducted experiments to compare their results to it. The authors introduced auxiliary classification along with six numbers of additional features extracted from the IRMAS data set of 10 instruments and human voice. They obtained an enhanced accuracy of 0.685 as micro F1 and 0.597 as macro F1 measures. Also, they found an increased accuracy of 10.7% and16.4% in micro and macro measures, respectively, compared to the baseline work of Han *et al.*[47]

***Hidden Markov Models (HMM)***

An HMM model assumes a set of observations produced by another set of hidden states. Thus, in each state, a random measurement from different distribution functions is found to give a joint probability distribution. The model which communicates the highest probability is selected as a likely source for the observation.

Eronen[51] developed a baseline instrument recognizer that employed MFCC and delta cepstrum (AMFCC) coefficients as features and HMMs to describe the feature distributions. A database of isolated notes from 27 Western orchestral instruments, as well as a smaller collection of drumbeats, was used to assess the system. The authors proposed two methods to improve the system's performance. The first was independent component analysis (ICA), and the second was discriminative HMMs training. The highest level of accuracy discovered was 85%. A comparison of six approaches for classifying sports audio signals was presented by Xiong *et al.*[52] They employed Mel-scale Frequency Cepstrum Coefficients (MFCCs) and MPEG audio characteristics for feature extraction. For classification, they used Maximum Likelihood-HMM (ML-HMM) and Entropic-PriorHMM (EP-HMM). The best result obtained with all the combinations was with an accuracy of around 90%. Kitahara *et al.*[53] provided a method for calculating the temporal trajectory of instrument existence probabilities for each F0, visualizing it as an instrogram, a spectrogram-like graphical representation, and applied it to an HMM model. In studies with authentic music, the instrument annotation had an average accuracy of 76.2%, and the instrogram-based similarity measured better approximated actual instrumentation similarity than an MFCC-based one. Eichner *et al.*[54] proposed an HMM model for instrument classification with a database of four instrument types. The recognition are based on solo music pieces played on the instrument under various conditions. They allowed to pass the recordings through a 31-channel mel-scaled filter bank and extracted the first and second-order differences, and then compressed the feature space to 25 dimensions using statistical principal component analysis (PCA). Accuracy of a maximum of 78% was reported in the experiment. Zlatintsi & Maragos[55] proposed the Multi-scale Fractal Dimension (MFD) technique to distinguish different musical instruments. Using PCA analysis, the trials were conducted with 1331 notes from seven different instruments. Authors classified musical instruments using both GMM and HMM models, which were found to be the most promising ones. They reported an error reduction of up to 32% in instrument recognition.

***Gaussian Markov Models (GMM)***

GMMs contemplate the continuous probability density of an observation and model it as a weighted sum of many Gaussian densities. Mean vector, mixture component density, and covariance matrix are taken as the hidden parameters in GMM. The expectation-maximization (EM) or k-means approach is used to estimate the parameter.

Martin *et al.*[15] described a statistical pattern-recognition technique based on Gaussian models with Fisher's multiple-discriminant analysis**.** Perceptually relevant acoustic parameters linked to the physical properties of resonance structure and source excitation were quantified from the output of an auditory model over a full pitch range of 15 orchestral instruments. Approximately 99 percent of the time, the classifiers correctly discriminated transient from continuous tones. Instrument families were identified with an overall success rate of around 90%.The individual instruments with an overall rate of 70% were recognized. Krishna & Sreenivas[58] proposed a classification for solo phrases instead of individual notes. Line Spectral Frequencies (LSF) was taken as features for identifying musical instruments. MFCC and LPCC features were used to evaluate the proposed system. For classification, the K-Nearest Neighbour and Gaussian Mixture Model classifiers were utilized. The best result found for the instrument family level was 95%, and for the individual instrument level was 90% from a data set of 14 instruments. Essid *et al.*[59] trained a GMM classifier with an MFCC feature to distinguish the musical instrument in solo phrases. To denoise the feature data, PCA was used. An overall accuracy of distinguishing five musical instruments was reported at around 67%. Virtanen & Klapuri[60] separated notes using a multi-pitch estimation algorithm and an optional streaming technique that organizes individual notes into sound sources. The most likely note sequence was found using the Viterbi algorithm. The classifiers employed MFCCs (with a 40-channel filter bank) and the first derivatives of MFCCs. The instrument conditional densities of features were modelled using GMM, and the parameters were evaluated using the EM (expectation-maximization) algorithm from the training material. The classification was then done using a Maximum Likelihood classifier. The dataset was artificially created from the RWC dataset. The F1 measure of 59.1 was achieved for 19 distinct pitched instruments with a maximum of six-note polyphony. Burred[61] presented an instrument classification system using Gaussian likelihood for timbre similarity measure with stereo-line source separation as the pre-step. The experimental result was found with an accuracy of 86.7% with a polyphony of two instruments and five classes. They found a better result than the monaural separation, with an accuracy of 79.8%. Heittola *et al.*[62] obtained an accuracy of 59% by using NMF based source filter model along with MFCC and GMM to classify 19 instruments producing six polyphonic notes**.** Diment *et al.*[63] employed a modified group delay (MODGDF) feature, a combination of phase information and MFCC. The authors used a GMM classifier with an EM algorithm and obtained an accuracy of about 71% on a database of 22 instruments. Eronen[64] used GMM classifier to classify 30 orchestral instruments of 7 classes in MUMS database. For the classifier training, researcher selected different features like MFCC, delta MFCC, LPCC, and modulation characteristics. The best result achieved for instrument family recognition was 58%.

Also, some other pioneer classification techniques were reported, which also obtained satisfactory accuracy in the field of musical instrument recognition, such as Brown *et al.*[65] found correct instrument identification accuracy of 79–84% for four classes of instruments with constant-Q coefficients, autocorrelation coefficients, and cepstral coefficients applied to short segments of solo passages from real records. The authors used Bayes decision rules with the K-means algorithm to classify the instruments. Garcia *et al.*[66] proposed a method using individual partials to identify musical instruments. The authors used isolated partials information to find spectral disjointness between the instruments. The data in those features were then used to figure out which instrument was most likely to have that partial. As a result, the sole need for the strategy to work was that each instrument must have at least one isolated partial somewhere in the signal. The experimental result showed an accuracy of 63% using 25 instruments. Vatolkin & Rudolph[67] pointed out the use of different musical features for western and ethnic music from a database of 8 western and 12 ethnic categories. The most suitable features were extracted for classification purposes for each selected category and were used to enhance the accuracy.

#### Feature Learning

Feature Learning is another important issue in music classification. It has a close association with classifier learning. Feature learning aims to improve classification performance by automatically selecting and extracting features. Automatic feature selection and extraction are not the same thing. In automatic feature selection, features are selected directly from many input features following some rules.[78] In feature extraction, an optimized set of features are extracted from the pool of available input features through transformations based on some projection rule and feature mapping.[46,79] The selection and extraction of features can be made in a supervised or unsupervised manner. Using a supervised setting, labelled data is used to enhance the extraction of valuable features that best discriminate between distinct labels.[79] This is accomplished using a variety of metric learning algorithms.[79] Linear Discriminant Analysis (LDA)[71,72] is a key metric learning method for instrument classification that identifies the best dimensional transformation by maximizing the inter-class scatter whereas reducing intra-class scatter. By modelling the fundamental structure of the audio stream, the features are extracted in an unsupervised way without requiring label information.[46] PCA is a common method for unsupervised feature extraction, which reduces the input feature dimension to a lower-dimensional space while maintaining its covariance.[56] Nonnegative matrix factorization (NMF)[68,69] is another method for extracting unsupervised features.

Different researchers have reported a variety of novel features for an automatic musical instrument recognition task in their experimental works. A few of these features are summarized as in Table 1.

### Research Issues

In this section, we discuss three unresolved research issues that deserve further exploration in the future, based on survey of different classifier learning techniques for musical instrument recognition task. These three unresolved research issues are outlined hereunder.

Table 1 — Feature Description

| | |
|---|---|
| TIMBRAL FEATURES | |
| Zero Crossing Rate (ZCR)[3,5,21,67] | Fast Fourier Transform (FFT)[3,7,37–39,61] |
| Spectral Centroid (SC)[3, 5, 15, 17, 21] | Short Time Fourier Transform (STFT)[3, 6, 25, 36, 61] |
| Spectral Roll-Off (SR)[3, 5, 21] | Discrete Wavelet Transform (DWT)[28, 33, 38, 39, 54, 61] |
| Spectral Flux (SF)[3, 5, 21] | Harmonic Centroid (HC)[9, 10, 39, 73, 74] |
| Spectral Bandwidth (SB)[21] | Harmonic Deviation (HD)[9, 73, 74] |
| Spectral Flatness Measure (SFM)[21] | Harmonic Spread (HS)[9, 10, 39, 73, 74] |
| Spectral Crest Factor (SCF)[21] | Harmonic Variation (HV)[9, 10, 73, 74] |
| Amplitude Spectrum Envelope (ASE)[20, 21, 24, 65] | Hamonic Spectral Skewness (HSS)[10, 73, 74] |
| Octave Based Spectral Contrast (OSC)[21] | Harmonic Spectral Kurtosis (HSK)[10, 73, 74] |
| Daubechies Wavelet Coefficient Histogram (DWCH)[3, 5, 21, 25, 28] | Harmonic Spectral Slope (HSSL)[10, 73, 74] |
| Mel-frequency Cepstrum Coefficients (MFCCs)[3, 4, 8, 9, 21, 32, 34, 40, 51, 54, 58, 60, 61] | Harmonic Spectral Decrease (HSD)[10, 73, 74] |
| MFCC Harmonic Partials (MFCC-H)[8] | Harmonic Amplitude (HA)[22, 36, 38] |
| Delta MFCC, Delta-Delta MFCC[21, 51] | Harmonic Roll-Off[10, 73, 74] |
| Fourier Cepstrum Coefficients (FCCs)[22,52] | Harmonic Energy (HE)[10, 11, 16, 29, 37–39, 73–76] |
| Linear Predictive Cepstrum Coefficients (LPCCs)[3, 4, 21, 58, 64, 65, 67] | Low Energy (LE)[3, 5, 67] |
| TEMPORAL FEATURES | |
| Zero Crossing Rate (ZCR)[10, 11, 23–25, 73, 74, 77] | FM (Frequency, Amplitude)[53, 71] |
| Energy Envelope Features (Attack Slope; Log-attack Time; Decrease Slope; Temporal Centroid; Effective Duration; Energy Modulation)[9–11, 18, 26, 27, 29, 38, 39, 73–77] | Mean-Variance (mVar)[26, 27] |
| Rise-time; Attack-time; Decay-time; Sustain-time; Release-time[17, 29, 38, 39, 77] | Auto-Correlation Coefficients (ACS)[10, 23, 24, 65] |
| Group delay features (GDF)[63] | Auto-Regressive Coefficients (ARs)[21, 25, 61] |
| Statistical Moments (SM)[21, 23, 25] | On-Set Duration, Slope[15, 71] |
| AM (Frequency, Amplitude)[15, 17, 19, 21, 23–25, 53, 64, 66, 71] | Tremolo[15, 24] |
| AM (Tremolo, Roughness)[23–25] | (Frequency, Strength, Heuristic Strength) |
| CEPSTRAL FEATURES | |
| Root Mean Square Energy (RMS)[17, 19, 20, 26, 27, 35, 62, 64, 66, 67] | Fractional Fourier Transform (FrFT)[77] |
| Mel-frequency Cepstral Coefficients (MFCCs)[7, 8, 19, 22– 27, 31, 55, 66, 77] | Discrete Fourier Transform (DFT)[8] |
| MFCC Harmonic Partials (MFCC-H)[8] | Fourier Cepstrum Coefficients (FCCs)[33, 77] |
| Delta MFCC; Delta-Delta MFCC[23, 55] | Linear Predictive Cepstrum Coefficients (LPCCs)[7, 17, 19, 22, 26, 27, 64, 66] |
| SPECTRAL FEATURES | |
| Mel-Spectrogram[6, 36, 46–49, 53, 61, 70] | Spectral Variation (SV)[10, 23, 25, 73, 74] |
| Mel-frequency Cepstral Coefficients (MFCCs)[5, 11, 38, 46, 52, 59, 62–65, 67] | Spectral Flatness Measure (SFM)[10, 11, 23, 24, 39, 67] |
| MFCC Harmonic Partials (MFCC-H)[8] | Constant Q-Coefficient ($Q_C$)[20, 23, 24, 61, 65] |
| Delta MFCC; Delta-Delta MFCC[11, 59, 64, 67] | Octave Band Signal Intensities (OBSI)[23–25] |
| Spectral Centroid (SC)[10, 11, 15, 16, 18–20, 23–25, 39, 53, 59, 64–66, 71, 73–77] | Spectral Irregularity[18, 24, 25] |
| Spectral Spread (SS)[10, 11, 39, 73–76] | Spectral Entropy[73] |
| Spectral Roll-Off (SR)[10, 11, 26, 27, 73, 74, 77] | Spectral Harmonicity[10, 16, 26, 27, 29, 67, 75, 76] |
| Spectral Width (SW)[23, 24, 59] | Harmonics (Odd, Even)[10, 11, 15, 26, 27, 37–39, 53, 65, 71, 73–76] |
| Spectral Asymmetry (SA)[19, 23, 24, 59, 66] | Harmonic Energy[10, 11, 16, 29, 37, 38, 73–76] |
| Spectral Skewness (SSW)[10, 11, 18, 24, 25, 59, 73–76] | Tristimulus[10, 11, 18, 26, 27, 37–39, 73, 74] |
| Spectral Kurtosis (SK)[10, 11, 18, 23–25, 67, 73, 74] | Fundamental Frequency ($f_0$)[10, 17, 19, 39, 53, 62, 64, 66, 71, 73–76] |
| Spectral Bandwidth (SB)[16, 29, 67] | Harmonic Deviation (HD)[10, 73, 74] |
| Spectral Contrast, Spectral Brightness[37–39] | Spectral Envelope (SE)[15, 18, 29, 36] |
| MPEG-7 Audio Spectrum Flatness (ASF)[11, 24, 25, 59] | Spectral Energy[8, 65] |
| MPEG-7 Audio Features[39, 52] | Spectral Moments[25–27, 38, 59, 65] |
| Spectral Crest Factor (SCF)[10, 11, 19, 23, 26, 27, 64, 66] | Croma Energy[67] |
| Spectral Slope (SSL)[10, 11, 23, 25, 64, 67, 73, 74] | Short-Time Fourier Transform (STFT)[10, 11, 53, 70, 73–76] |
| Spectral Decrease (SD)[10, 25, 73, 74] | |
| PERCEPTUAL FEATURES | |
| Zero Crossing Rate (ZCR)[9, 50] | Spectral Bandwidth (SB)[9, 16, 29, 33, 50, 67] |
| Mean of ZCR (ZCRM)[9] | Spectral Power[73] |
| Standard Deviation of ZCR (ZCRD)[9] | Spectral Contrast, Spectral Brightness[33, 37–39] |
| Root Mean Square (RMS)[9, 50] | Pitch (Frequency, Variance)[15, 26, 27, 29, 33, 36, 39, 62] |

(*Contd.*)

Table 1 — Feature Description (*Contd.*)

PERCEPTUAL FEATURES

Mean of RMS (RMSM)[9]
Standard Deviation of RMS (RMSD)[9]
Mean of Centroid (Centroid-M)[9, 15, 17, 77]
Standard Deviation of Centroid (Centroid-D)[9, 15, 17, 77]
Mean of Bandwidth (Bandwidth-M)[9]
Standard Deviation of Bandwidth (Bandwidth-D)[9]
Spectral Flux (SF)[9, 22, 66, 77]
Mean of Flux (Flux-M) & Standard Deviation of Flux (Flux-D)[9]
Spectral Centroid (SC)[9, 15, 17, 50]
Spectral Roll-Off (SR)[50]
Mel-Spectrogram[50]
Mel-frequency Cepstral Coefficients (MFCCs)[10, 41, 50, 73–76]
Fourier Cepstrum Coefficients (FCCs)[33]
Delta MFCC; Delta-Delta MFCC[10, 73, 74]
MFCC (Mean & Standard Deviation)[41]
Loudness; Relative Specific Loudness[10, 11, 23, 35, 73–76, 79]
Tempo, Beat[79]
Roughness; Sharpness; Spread[10, 11, 23, 73–76]
Fluctuation Length & Mean Fluctuation Length[10]

### Constraint on Labelled Data

Majority of data sets used to evaluate current music instrument recognition systems are small or mediocre in size. Rather than focusing on efficiency, research has focused on enhancing classification performance. The music industry in today's world is booming. We require more efficient music analysis and classification systems to handle massive data sets. The present methodologies for musical instrument classification face two major issues. In terms of processing time and storage, scalability is the most significant consideration. To streamline feature extraction and address the storage issue, faster pre-processing processes are required to accelerate large-scale classification tasks.

### Learning Similarity Retrieval

A significant problem related to musical instrument classification is finding difficulty in similarity retrieval. The purpose of similarity retrieval is to search a database for similar music. The fact that different similarities are required for different types of music inspire a novel classification technique based on similarity retrieval. This type of classifier learns in an unsupervised fashion. Using exemplar pairings of similar and dissimilar music, we may train a classifier to learn to recover the similarity between the two.

### Usability of Perceptual Features

The way humans perceive and process music in their auditory and neurological systems are highly dependent on perceptual characteristics. As a result, this observation might be used to design a better classifier system. In convolutional neural networks, the perceptual features are processed through multiple hidden layers with numerous nodes that operate as processing units. As a result, training a convolutional neural network entails learning perceptual information and classification rules. However, convolutional neural networks have already been used in musical instrument classification tasks but have not been thoroughly investigated.

## Conclusions

In this review article, we discussed some popular common choices of different state-of-the-art techniques for the recognition of musical instruments. We feel that this study has offered an up-to-date overview of audio features and music classifiers. Humans have a significant ability to recognize musical instrument sound and can make the right decision in a concise time frame. But still, there is a gap between human performance and the automatic musical instrument recognition system performance. As a result, existing AMIR systems still have a lot of scope for development.

## References

1 Shen J, Shepherd J, Cui B & Tan K-L, A novel framework for efficient automated singer identification in large music databases, *ACM Trans Inf Syst*, **27(3)** (2009) 1–31.

2 Liu C-C & Huang C-S, A singer identification technique for content-based classification of MP3 music objects, *Proc Int Conf Inf Knowl Mang* (Chung Hua University) 2004, 506–511.

3 Li T, Ogihara M & Li Q, A comparative study on content-based music genre classification, *Proc Int Res Dev Inf Retrieval* (ACM Toronto, Canada) 2003, 282–289.

4 Zhang T, Automatic singer identification, *Proc Int Conf Multimed Expo* (ICME 2003) (IEEE) 2003, 33–36.

5 Li T & Ogihara M, Music artist style identification by semi supervised learning from both lyrics and content, *Proc Int Conf Multimed* (ACM New York, USA) 2004, 364–367.

6 Seipel F, *Music Instrument Identification using Convolutional Neural Networks*, Master Thesis, Technische Universitat, Berlin, 2018.

7 Marques J & Moreno P J, A study of musical instrument classification using Gaussian mixture models and support vector machines, *Technical Report Series* (Cambridge Research Laboratory) 1999, 1–21.

8 Giannoulis D & Klapuri A, Musical instrument recognition in polyphonic audio using missing feature approach *IEEE/ACM Trans Audio Speech Language Process*, **21(9)** (2013) 1805–1817.

9 Deng J D, Simmermancher C & Cranefield S, A study on feature analysis for musical instrument classification, *IEEE Trans Syst Man Cybern, Part B (Cybern.)*, **38(2)** (2008) 429–438.

10 Livshin A&Rodet X, Purging musical instrument sample databases using automatic musical instrument recognition methods, *IEEE Trans Audio Speech Language Process*, **17(5)** (2009) 1046–1051.

11 Peeters G, A Large set of audio features for sound description (similarity and classification) in the CUIDADO Project, *IRCAM Technol Rep* 2004, 1–25.

12 Weihs C, Ligges U, Morchen F & Mullensiefen D, Classification in music research, *Adv Data Anal Classif*, **1(3)** (2007) 255–291.

13 Jordan M I & Mitchell T M, Machine learning: trends, perspectives, and prospects, *Sci*, **349(6245)** (2015) 255–260.

14 LeCun Y, Bengio Y & Hinton G, Deep learning, *Nature*, **521(7553)** (2015) 436–444.

15 Martin K D & Kim Y E, Musical instrument identification: A pattern-recognition approach, *J Acoust Soc Am*, **14(03)** (1998) 1768–1768.

16 Agostini G, Longari M & Pollastri E, Musical instrument timbres classification with spectral features, EURASIP *J Appl Signal Process*, **1** (2003) 5–14.

17 Eronen A & Klapuri A, Musical instrument recognition using cepstral coefficients and temporal features, *Proc ICASSPIEEE* (Istanbul, Turkey) 2000, 753–756.

18 Fujinaga I & MacMillan K, Realtime recognition of orchestral instruments, *Proc Int Comput Music Conf* (International Computer Music Association, San Francisco) 2000, 141–143.

19 Eronen A, Comparison of features for musical instrument recognition, *Proc IEEE Workshop Apps Signal Proc Audio Acoust* (IEEE) 2001, 19–22.

20 Kaminskyj I & Czaszejko T, Automatic recognition of isolated monophonic musical instrument sounds using *k*NNC, *J Intell Inf Syst*, **24(2/3)** (2005) 199–221.

21 Fu Z, Lu G, Ting K M & Zhang D, A survey of audio-based music classification and annotation, *IEEE Trans Multimed*, **13(2)** (2011) 303–319.

22 Marques J, *An Automatic Annotation System for Audio Data Containing Music*, Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.

23 Essid S, Richard G & David B, Instrument recognition in polyphonic music based on automatic taxonomies, *IEEE Trans Audio Speech Language Process*, **14(1)** (2006) 68–80.

24 Essid S, Richard G & David B, Musical instrument recognition by pairwise classification strategies, *IEEE/ACM Trans Audio Speech Language Process*, **14(4)** (2006) 1401–1412.

25 Joder C, Essid S & Richard G, Temporal integration for audio classification with application to musical instrument classification, *IEEE Trans Audio Speech Language Process*, **17(1)** (2009) 174–186.

26 Fuhrmann F, Haro M & Herrera P, Scalability, generality, and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music, *Proc ISMIR* (International Society for Music Information Retrieval) 2009, 321–326.

27 Fuhrmann F & Herrera P, Polyphonic instrument recognition for exploring semantic similarities in music, *Proc 13th Int Conf Digit Audio Effects* (Graz, Austria) 2010, 1–8.

28 Ozbek M E, Ozkurt N & SavaciF A, Wavelet ridges for musical instrument classification, *J Intell Inf Syst*, **38(1)** (2011) 241–256.

29 Wu J, Vincent E, Raczynski S A, Nishimoto T, Ono N & Sagayama S, Polyphonic pitch estimation & instrument identification by joint modelling of sustained and attack sounds, *IEEE J Sel Top Signal Process*, **5(6)** (2011) 1124–1132.

30 Bosch J J, Janer J, Fuhrmann F & Herrera P, A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals, *Proc Int Soc Music Inf Retrieval Conf* (International Society for Music Information Retrieval) 2012, 559–564.

31 Duan Z, Pardo B & Daudet L, A novel cepstral representation for timbre modelling of sound sources in polyphonic mixtures, *Proc IEEE Int Conf Acoust Speech Signal Process* (IEEE) 2014, 7495–7499.

32 Yu L-F, Su L & Yang Y-H, Sparse cepstral codes and power scale for instrument identification, *Proc IEEE Int Conf Acoust Speech Signal Process* (IEEE) 2014, 7460–7464.

33 Lin C-C, Chen S-H, Truong T-K & Chang Y, Audio classification and categorization based on wavelets and support vector machine, *IEEE Trans Speech Audio Process*, **13(5)** (2005) 644–651.

34 Mandel M & Ellis D, Song-level features and SVMs for music classification, *Proc Int Conf Music Inf Retrieval*, **5** (2005).

35 Kaminskyj I & Materka A, Automatic source identification of monophonic musical instrument sounds, *Proc IEEE Int Conf Neural Netw*, **1** (1995) 189–194.

36 Cemgil A T & Gurgen F, Classification of musical instrument sounds using neural networks, *Proc of SIU97* (Bodrum, Turkey) 1997, 1–10.

37 Kostek B & Krolikowski R, Application of artificial neural networks to the recognition of musical sounds, *Arch Acoust*, **22(1)** (1997) 27–50.

38 Kostek B & Czyzewski A, Representing musical instrument sounds for their automatic classification, *J Audio Eng Soc*, **49(9)** (2001) 768–785.

39 Kostek B, Musical instrument classification and duet analysis employing music information retrieval techniques, *Proc IEEE* (JPROC), **92(4)** (2004) 712–729.

40 Loughran R, Walker J, O'Farrell M & O'Neill M, The use of mel-frequency cepstral coefficients in musical instrument identification, *Proc Int Comput Music Conf (ICMC)* (Belfast, Northern Ireland) 2008, 24–29.

41 Newton M J & Smith L S, A neurally inspired musical instrument classification system based upon the sound Onset, *J Acoust Soc Am*, **131(6)** (2012) 4785–4798.

42 Mikolov T, Karafiat M, Burget L, Cernocky J & Khudanpur S, Recurrent neural network-based language model, *Proc Annu Conf Int Speech Commun Assoc* (INTERSPEECH 2010) **2(3)** (2010) 1045–1048.

43 Mesnil G, He X, Deng L & Bengio Y, Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding, *Proc Annu Conf Int Speech Commun Assoc* 2013, 3771–3775.

44 Yao K, Zweig G, Hwang M-Y, Shi Y & Yu D, Recurrent neural networks for language understanding, *Proc Annu Conf Int Speech Commun Assoc* 2013, 2524–2528.

45 Lecun Y, Bottou L, Bengio Y & Haffner P, Gradient-based learning applied to document recognition, *Proc IEEE*, **86(11)** (1998) 2278–2324.

46 Lee H, Largman Y, Pham P & Ng A Y, Unsupervised feature learning for audio classification using convolutional deep belief networks, *Proc Adv Neural Inf Process Syst*, **22**, 2009.

47 Han Y, Kim J & Lee K, Deep convolutional neural networks for predominant instrument recognition in polyphonic music, *IEEE Trans Audio Speech Language Process*, **25(1)** (2016) 208–221.

48 Gururani S, Summers C & Lerch A, Instrument activity detection in polyphonic music using deep neural networks, *Proc Int Soc Music Inf Retrieval Conf* (ISMIR) (Paris, France) 2018, 569–576.

49 Gomez J S, AbeBer J & Cano E, Jazz solo instrument classification with convolutional neural networks, source separation, and transfer learning, *Int Soc Music Inf Retrieval Conf* (ISMIR) (Paris, France) 2018, 577–584.

50 Yu D, Duan H, Fang J & Zeng B, Predominant instrument recognition based on deep neural network with auxiliary classification, *IEEE/ACM Trans Audio Speech Language Process*, **28** (2020) 852–861.

51 Eronen A, Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs, *Proc 7$^{th}$ Int Symp Signal Process Appl*, **2** (2003) 133–136.

52 Xiong Z, Radhakrishnan R, Divakaran A & Huang T, Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification, *Proc IEEE Int Conf Multimed Expo*, **3** (2003) 397–400.

53 Kitahara T, Goto M, Komatani K, Ogata T & Okuno H G, Musical instrument recognizer "instrogram" and its application to music retrieval based on instrumentation similarity, *Proc IEEE Int Symp Multimed* (IEEE) 2006, 265–274.

54 Eichner M, Wolff M & Hoffmann R, Instrument classification using hidden Markov models, *Proc ISMIR* 2006, 349–350.

55 Zlatintsi A & Maragos P, Multiscale fractal analysis of musical instrument signals with application to recognition, *IEEE Trans Audio Speech Language Process*, **21(4)** (2013) 737–748.

56 Partridge M & Jabri M, Robust principal component analysis, *Proc IEEE Signal Process Soc Workshop* (University of Sydney) 2000, 289–298.

57 MacQueen J, Some methods for classification and analysis of multivariate observations, *Proc Symp Math Statist Probability* (5th Berkeley Symposium) **1(14),** 1967, 281–297.

58 Krishna A G & Sreenivas T V, Music instrument recognition: From isolated notes to solo phrases, *Proc IEEE Int Conf Acoust Speech Signal Process* **4** (2004) IV-265–IV-268.

59 Essid S, Richard G & David B, Efficient musical instrument recognition on solo performance music using basic features, *Proc AES 25th Int Conf* (London, UK) 2004, 89–93.

60 Virtanen T & Klapuri A, Analysis of polyphonic audio using source-filter model and nonnegative matrix factorization, *Adv in Models for Acoust Process (Neural Inf Process Syst Workshop)*, **18** (2006).

61 Burred J J, *From Sparse Models to Timbre Learning: New Methods for Musical Source Separation*, Ph D Thesis, Technical University of Berlin, Berlin, 2008.

62 Heittola T, Klapuri A & Virtanen T, Musical instrument recognition in polyphonic audio using source-filter model for sound separation, *Proc Int Soc Music Inf Retrieval Conf* (Tampere University of Technology) 2009, 327–332.

63 Diment A, Rajan P, Heittola T & Virtanen T, "Modified group delay feature for musical instrument recognition," in *Proc. Int Symp Comput Music Multidiscip Res* (Marseille, France) 2013, 431–438.

64 Eronen A, *Automatic musical instrument recognition*, MS Thesis, Tampere University of Technology, Tampere, Finland, 2001.

65 Brown J C, Houix O & McAdams S, Feature dependence in the automatic identification of musical woodwind instruments, *J Acoust Soc Am*, **109** (2001) 1064–1072.

66 Garcia J, Barbedo A & Tzanetakis G, Musical instrument classification using individual partials, *IEEE Trans Audio Speech Language Process*, **19(1)** (2011) 111–122.

67 Vatolkin I & Rudolph G, Comparison of audio features for recognition of western and ethnic instruments in polyphonic mixtures, *Proc Int Soc Music Inf Retrieval Conf* (Paris, France ) 2018, 554–560.

68 Lee D D & Seung H S, Learning the parts of objects by non-negative matrix factorization, *Nature*, **401** (1999) 788–791.

69 Lee D D & Seung H S, Algorithms for non-negative matrix factorization, *Adv in Neural Inf Process Syst*, **13** (2001) 556–562.

70 Dittmar C & Uhle C, Further steps towards drum transcription of polyphonic music, *Proc Audio Eng Soc* (Audio Engineering Society, Berlin, Germany) 2004, 1–8.

71 Kitahara T, Goto M, Komatani K, Ogata T & Okuno H G, Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps, *EURASIP J Appl Signal Process*, **(1)** (2007) 155–155.

72 West K, *Novel techniques for audio music classification and search*, Ph D Thesis, University of East Anglia, Norwich, U.K., 2008.

73 Livshin A, Peeters G & Rodet X, Studies, and improvements in automatic classification of musical sound samples, *Proc Int Conf Comput Music* (ICMC) (Singapour, Singapore) 2003, 220–227.

74 Livshin A & Rodet X, The importance of cross database evaluation in musical instrument sound classification: A critical approach, *Proc Int Symp Music Inf Retrieval* (ISMIR) 2003, 241–242.

75 Livshin A & Rodet X, Musical instrument identification in continuous recordings, *Proc Int Conf Digital Audio Effects* (DAFX-04) (Naples, Italie) 2004, 222–227.

76 Livshin A & Rodet X, The significance of the non-harmonic "Noise" versus the harmonic series for musical instrument recognition, *Proc Int Symp Music Inf Retrieval* (ISMIR) (NA, France) 2006, 95–100.

77 Bhalke D G, Rama Rao C B & Bormane D S, Automatic musical instrument classification using fractional Fourier transform based MFCC features and counter propagation neural network, *J Intell Inf Syst*, **46** (2016) 425–446.

78 Mierswa I & Morik K, Automatic feature extraction for classifying audio data, *Mach Learn*, **58** (2005) 127–149.

79 Slaney M, Weinberger K & White W, Learning a metric for music similarity, *Proc Int Conf Music Inf Retrieval*, **148** (2008).