International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 82 - 85, 2015, ISSN:- 2319–8656

Human Perception and Recognition of Musical Instruments: A Review

Satish Ramling Sankaye
MGM Dr. G.Y. Pathrikar College of CS and IT,
Aurangabad, India

U. S. Tandon
Department of Physics,
College of Natural and Computational Sciences,
Haramaya University,
Dire Dawa, Ethiopia

Abstract: Musical Instrument is the soul of music. Musical Instrument and Player are the two fundamental component of Music. In the past decade the growth of a new research field targeting the Musical Instrument Identification, Retrieval, Classification, Recognition and management of large sets of music is known as Music Information Retrieval. An attempt to review the methods, features and database is done.

Keywords: Musical Instrument; Monophonic; Polyphonic; Classification Data Model;

1. INTRODUCTION

Human perception in musical applications is especially important, since musical sounds are designed merely for human audition. The study of music signal is useful in teaching and evaluation of music. The human vocal apparatus which generates speech also generates music. Therefore, the studies reveal similarities in the spectral and temporal properties of music and speech signals. Hence, many techniques developed to study speech signal are employed to study music signals as well.

To make music, two essential components are needed: the player and the instrument. Hence one of the key aspects in the research of Music has focused on the internal contents of music, the Musical Instrument. In the past decade the growth of a new research field targeting the Musical Instrument Identification, Retrieval, Classification, Recognition and management of large sets of music is known as Music Information Retrieval. Musical instrument Identification is edged on classification of single note (Monophonic), more than one instrument notes at a time (Polyphonic), distinction of instruments in continuous recording or Classification of family/genre. Musical instruments are classified into five families depending on the sound produced as percussion, brass, string, woodwind and keyboard [4], [7].

Table 1: The musical instrument collection

Family	Instruments
Brass	French horn, Trombone, Trumpet, Tuba
Keyboard	Piano, Harmonium
Percussion	Bell, Bongo, Chime, Conga, Cymbal, Dholki, Drum, Gong, Tambourine, Triangle, Timbales, Tympani, Tabla,
String	Guitar, Violin, Sitar, Vichitraveena, Saraswativeena, Rudraveena
Woodwind	Shehnai, Oboe, Saxophone, Flute

The paper is organized as follows: Section 2 describes the different databases studied. The different classification and pattern recognition techniques are discussed in section 3. Finally section 4 furnishes the conclusion.

2. DATABASE

Musical Instrument Identification leads to the aspect of initially recording the sound sample from different sources. It can be recorded directly while playing the instrument by using tape recorder, mobile or any other electronic gadget meant for sound recording in natural environment. Also for the study purpose, the musical instruments are played in an anechoic room at a professional studio. Some of the commonly used databases studied by most of the researchers include:

2.1 Musical Audio Signal Separation (MASS) Dataset

This database was created to help to evaluate of Musical Audio Signal Separation algorithms and statements on a representative set of professionally produced music (i.e. real music recordings). It included several song snips of a few seconds (10s-40s) with the following contents:

- *Tracks (with/without effects):* Stereo Microsoft PCM WAV files (44.1Khz, 24 bits) of every instrumental track including and/or without including effects (plugins enabled or disabled in the project file used for production)
- *Description of the effects*: When available, included a description of the plugins used to modify the tracks without effects.
- *Lyrics:* When available, lyrics are included.

The dataset was compiled by M. Vinyes (MTG former member). Bearlin and Sargon have released the tracks of their songs and Sergi Vila at Garatge Productions and Juan Pedro Barroso at Kcleta Studios [14]. It is available online *www.sargonmetal.com.*

2.2 University of Iowa musical instrument samples

The Musical Instrument Samples Database has been divided into two categories: pre-2012 and post-2012 files. The pre-2012 files are the original sound files that are present on website http://theremin.music.uiowa.edu/MIS.html [17] since the end of May 2014. These sound files were recorded in the Anechoic Chamber at the Wendell Johnson Speech and Hearing Center as early as 1997. This category consists of

mono files for woodwinds, brass, and string instruments at a 16-bit, 44.1 kHz format with Neumann KM 84 Microphones. It also contains stereo files for the most recent recordings of string instruments done between December 2011 and May 2012, and percussion instruments done between March and June 2013 at a 24-bit, 44.1 kHz format with 3 Earthworks QTC40 microphones in a Decca Tree configuration.

The post-2012 files are experimental sound files. They are edited sound files extracted from the University of Iowa Electronic Music Studios website from the pre-2012 category. Each instrument from the string, woodwind, brass, and percussion families, excluding the guitar, piano, and Soundmines folder, has been edited as of July 24, 2014 for public and research use. All files from the string, woodwind, brass, and percussion families have been converted to a 24-bit, 44.1 kHz stereo format. Whenever possible, mid-side processing was applied to these files to widen the stereo field. These files were created in Studio 1 of the University Electronic Music Studios in the Becker Communication Studies

2.3 McGill university master samples

The first release of McGill University Master Samples [MUMS] (Opolko & Wapnick, 1987) featured 3 CDs of recorded, high quality instrument samples. Recently, the library has been expanded to 3 DVDs (Opolko & Wapnick, 2006) and contains samples of most standard classical, some non-standard classical, and many popular music instruments. There are 6546 sound samples in the library, divided between string (2204), keyboard (1595), woodwind (1197), percussion (1087, out of which 743 are non-pitched), and brass (463) families. In principle, each note of each instrument has been recorded separately (44.1 kHz, 24-bit), and most instruments feature several articulation styles. Typically there are 29 samples per instrument, which means that the whole pitch range of the available instruments is not consistently covered. The coverage is nevertheless impressive.

This library is one of the most often used sources of instrument samples within instrument recognition and classification research, sound synthesis and manipulation studies. The library has also been the source for an edited database (SHARC) of steady-state instrument spectra.

2.4 Real World Computing Music Database:

RWC [13] Music Database comprises of four original component Databases. Popular Music Database (100 pieces), Royalty-Free Music Database (15 pieces), Classical Music Database (50 pieces), and Jazz Music Database (50 pieces). Recently two more Database component viz Music Genre Database (100 pieces) and Musical Instrument Sound Database (50 instruments) are added.

The Database of Musical Instrument Sound covers 50 musical instruments and provides, in principle, three variations for each instrument. In all about 150 performances of different instruments are present. To provide a wide variety of sounds, following approach has been taken.

- *Variations (3 instrument manufacturers, 3 musicians):* Each variation featured, in principle, an instrument from a different manufacturer played by a different musician.
- *Playing style (instrument dependent):* Within the range possible for each instrument, many playing styles have been recorded.
- *Pitch (total range):* For each playing style, the musician played individual sounds at half-tone intervals over the entire range of tones that could be produced by that instrument.
- *Dynamics (3 dynamic levels):* Recording was also done for each playing style at three levels of dynamics (forte, mezzo, piano) spanning the total range of the instrument.

The sounds of these 50 instruments were recorded at 16 bit / 44.1 kHz and stored in 3544 monaural sound files having a total size of about 29.1 GBytes and a total playback time (including mute intervals) of about 91.6 hours [13].

3. CLASSIFCATION DATA MODEL

Various Features of Musical Signal have been studied, which are classified as Temporal, Spectral, Time-Domain, and Frequency Domain. Note onset detection and localization is also useful for a number of analysis and indexing techniques for musical signals [2]. Attack, Decay, Sustain and Release [2], [8] are other important features of sound waveform's energy distribution. After studying these feature set the different model are being implemented on the feature set for the Identification of the musical instrument or classifying the excerpt as a member of particular family. The various commonly studied models are discussed below:

3.1 Support Vector Machines

Support Vector Machine (SVM) [15] is a supervised learning method that belongs to a family of linear classifiers used for classification and regression. However, SVM is closely related to neural networks. It is based on some relatively simple ideas but constructs models that are complex enough and it can lead to high performances in real world applications.

The basic idea behind Support Vector Machines is that it can be thought of as a linear method in a high-dimensional feature space nonlinearly related to input space. Therefore in practice it does not involve any computation in the high-dimensional space. All necessary computations are performed directly in input space by the use of kernels. Therefore the complex algorithms for nonlinear pattern recognition, regression, or feature extraction can be used pretending that the simple linear algorithms are used.

The key to the success of SVM is the kernel function which maps the data from the original space into a high dimensional (possibly infinite dimensional) feature space. By constructing a linear boundary in the feature space, the SVM produces non-linear boundaries in the original space. When the kernel function is linear, the resulting SVM is a maximum-margin hyperplane. Given a training sample, a maximum-margin hyperplane splits a given training sample in such a way that the distance from the closest cases (support vectors) to the hyperplane is maximized. Typically, the number of support vectors is much less than the number of the training sample. Nonlinear kernel functions such as the polynomial kernel and the Gaussian (radial basis function) kernel are also commonly used in SVM. One of the most important advantages for the SVM is that it guarantees generalization to some extent. The decision rules reflect the regularities of the training data rather than the incapabilities of the learning machine. Because of the many nice properties of SVM, it has been widely applied to virtually every research field.

3.2 Hidden Markov Model

A hidden Markov model (HMM) [15] is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. An HMM can be considered as the simplest dynamic Bayesian network. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics.

The main characteristic of Hidden Markov Model is that it utilizes the stochastic information from the musical frame to recognize the pattern. In a hidden Markov model, the state is not directly visible, but the dependence of output on the state is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

Hidden Markov Models are widely used as general-purpose speech recognition and musical instrument as well as music identification systems. The basic reason why HMMs are used in music/speech recognition is that a music/speech signal could be viewed as a piecewise stationary signal or a short-time stationary signal.

Another reason why HMMs are popular is that they can be trained automatically and they are simple and computationally feasible to use.

3.3 Gaussian Mixture Model

A Gaussian Mixture Model (GMM) was used as classification tool [10], [15]. GMMs belong to the class of pattern recognition systems. They model the probability density function of observed variables using a multivariate Gaussian mixture density. Given a series of inputs, it refines the weights of each distribution through expectation-maximization algorithms.

In order to construct the models for the music recognition system, they calculated the features for all samples of the database and store the features for each class separately. Then, a Gaussian Mixture Model (GMM), θ_i, for each class is built (i.e., with i = 1...C, where C denotes the number of different classes), using a standard Expectation Maximization (EM) algorithm. EM algorithm is initialized by a deterministic procedure based on the Gaussian means algorithm. A new song is classified into a new category by computing the likelihood of its features given in the classification models, θ_i, with i = 1...C. Summing up these likelihood values, the song is assigned to the class that has the maximum summation value [11].

3.4 Probabilistic latent component analysis (PLCA)

Probabilistic Latent Component Analysis (PLCA) or Non-negative Matrix Factorization (NMF) is efficient frameworks for decomposing the mixed signal into individual contributing components [1]. In NMF approach, the features representing each instrument are the spectral dictionaries which are used to decompose the polyphonic spectra into the source instruments. PLCA interprets this task probabilistically by assuming the spectrum to be generated from an underlying probability density function (pdf), and estimates the joint distribution of observed spectra and a set of underlying latent variables.

Probabilistic Latent Component Analysis [1] is based on modelling the normalized magnitude of the observed spectrum V (f, t) as the probability distribution Pt(f) at time frame index t and frequency bin index f. Pt(f) is factorized into many latent components as

$$P_t(f) = \sum_{p,s,z,a} P_t(f|p,a)P_t(p)P_t(s|p) \times P_t(z|p,s)P(a|s,z).$$

Here, p, s, z, a are the discrete latent variables with Np, Ns, Nz, Na values respectively. At each time t, we know the F0 values indexed by p. We have to identify the underlying source playing at the p^{th} F0. Each source s has dictionaries of envelopes indexed by z. $P_t(f|p, a)$ is the fixed spectrum formed using the source-filter model as

$$P_t(f|p,a) = \frac{e_t(f|p)h(f|a)}{\sum_f e_t(f|p)h(f|a)}$$

Here, $e_t(f|p)$ consists of harmonic peaks at integral multiples of the p^{th} F0 at time t. h(f|a) is the transfer function of the a^{th} filter of a triangular mel-filter bank consisting of 20 filters uniformly distributed on the Mel-frequency scale as in [1].

3.5 Linear Discrimination Analysis (LDA) classifier:

Linear Discriminant Analysis (LDA, also known as Fisher Discriminant Analysis (FDA). LDA [6] has been widely used in face recognition, mobile robotics, object recognition and musical Instrument Classification.

In LDA, they computed a vector which best discriminates between the two classes. Linear Discriminant Analysis (LDA), searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). More formally, given a number of independent features relative to which the data is described, LDA creates a linear combination of these which yields the largest mean differences between the desired classes. Mathematically speaking, for all the samples of all classes, we define two measures:

1) one is called within-class scatter matrix, as given by

$$S_w = \sum_{j\ 1}^{c} \sum_{i\ 1}^{N_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T,$$

Where x_{ji} is the i^{th} sample of class j, μ_j is the mean of class j, c is the number of classes, and N_j the number of samples in class j; and

2) the other is called between-class scatter matrix

$$S_b = \sum_{j\ 1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T,$$

where μ represents the mean of all classes.

The goal is to maximize the between-class measure while minimizing the within-class measure. One way to do this is to maximize the ratio det $[S_b]$ / det $[S_w]$.

3.6 Supervised non-negative matrix factorization:

Supervised Non-Negative Matrix Factorization (S-NMF) method is one the new approach developed for the Identification/Classification of the Musical Instrument [16]. In this approach, a non-negative n × m matrix V (is considered as the features consisting of n vectors of dimension m). The non-negative n × r matrix W (basis matrix) and non-negative r × m matrix H (encoding matrix) in order to approximate the matrix V as:

$$V \approx W.H$$

Where, r is chosen such that (n + m) r < nm. To find an approximate factorization in above equation, Kullback-Leibler divergence between V and W.H is used frequently, and the optimization problem can be solved by the iterative multiplicative rules. But, the basis vectors defined by the columns of matrix W are not orthogonal. Thus, QR decomposition was utilized on W, that is W = QR, where Qn X r is an orthogonal matrix and Rr X r is an upper triangular matrix. At this time,

$$V \approx Q.H'$$

V can be written as a linear combination between an orthogonal basis and a new encoding matrix, where Q contains the orthogonal basis and H' ≈ R.H becomes the new encoding matrix. This method, however, cost a mass of computation for updating W and H iteratively and QR decomposition.

3.7 Classification Methods:

In addition to above classification techniques, some of the most important and common method for identification of Musical Instrument are also studied. DTW algorithm is powerful for measuring similarities between two series which may vary in time or speed [3]. CWT [9] too is wavelet-based feature for discrimination of various musical instrument signals. A semi-supervised learning [5] technique is also suitable for musical instrument recognition. Linear Discriminant Analysis + K-Nearest Neighbors [12] combined method has also been effectively used classification for performing automatic Musical Instrument Recognition.

4. ACKNOWLEDGMENTS

We extend our sincere thanks to Dr. S.C. Mehrotra for his helpful guidance in preparing this review.

5. REFERENCES

[1] Arora, Vipul, and Laxmidhar Behera. "Instrument identification using PLCA over stretched manifolds", Twentieth National Conference on Communications (NCC), IEEE, 2014.

[2] Bello, Juan Pablo, et al. "A tutorial on onset detection in music signals", IEEE Transactions on Speech and Audio Processing, Vol. 13.5 (2005): 1035-1047.

[3] Bhalke, D. G., C.B. Rama Rao, and D. S. Bormane. "Dynamic time warping technique for musical instrument recognition for isolated notes", International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), IEEE, 2011.

[4] Deng, Jeremiah D., Christian Simmermacher, and Stephen Cranefield. "A study on feature analysis for musical instrument classification", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 38.2 (2008): 429-438.

[5] Diment, Aleksandr, Toni Heittola, and Tuomas Virtanen. "Semi-supervised learning for musical instrument recognition", Proceedings of the 21st European Signal Processing Conference (EUSIPCO). IEEE, 2013.

[6] Eichhoff, Markus, and Claus Weihs. "Musical instrument recognition by high-level features" Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Springer Berlin Heidelberg, 2012. 373-381.

[7] Eronen, Antti, and Anssi Klapuri. "Musical instrument recognition using cepstral coefficients and temporal features", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'00. Vol. 2. IEEE, 2000.

[8] Fanelli, Anna Maria, et al. "Content-based recognition of musical instruments", Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, IEEE, 2004.

[9] Foomany, Farbod Hosseyndoust, and Karthikeyan Umapathy. "Classification of music instruments using wavelet-based time-scale features", IEEE International Conference on Multimedia and Expo Workshops (ICMEW). IEEE, 2013.

[10] Hall, Glenn Eric, Hall Hassan, and Mohammed Bahoura. "Hierarchical parametrisation and classification for musical instrument recognition", 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA). IEEE, 2012.

[11] Holzapfel, André, and Yannis Stylianou. "A statistical approach to musical genre classification using non-negative matrix factorization", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 2. IEEE, 2007.

[12] Livshin, Arie, and Xavier Rodet. "Purging Musical Instrument Sample Databases Using Automatic Musical Instrument Recognition Methods", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17.5 (2009): 1046-1051.

[13] M.Goto, et al. RWC Music Database: Music Genre Database and Musical Instrument Sound Database.

[14] M.Vinyes, MTG Mass database, http://www.mtg.upf.edu/static/mass/resources.

[15] Perfecto Herrera-Boyer, Geoffroy Peeters, Shlomo Dubnov, "Automatic Classification of Musical Instrument Sounds", 2002.

[16] Rui, Rui, and Changchun Bao. "A novel supervised learning algorithm for musical instrument classification", 35th International Conference on Telecommunications and Signal Processing (TSP), IEEE, 2012.

[17] University of Iowa Musical Instrument Sample Database, http://theremin.music.uiowa.edu/index.html.