A REVIEW OF AUTOMATIC MUSICAL INSTRUMENT CLASSIFICATION BASED ON SOUND RECOGNITION SYSTEM

Tee Hao Wong *[1,2], **Rahizar Ramli** †[2], **Ahmad Saifizul** ‡[2], **Ming Foong Soong** ♦[2]
[1]School of Foundation Studies, University College of Technology Sarawak, Sibu, Sarawak, Malaysia
[2] Department of Mechanical Engineering, University of Malaya, 50603, Kuala Lumpur, Malaysia

Résumé

Cet article présente une revue des recherches sur la classification des instruments de musique qui ont utilisé le système d'apprentissage automatique. Les deux principales étapes de la tâche de classification automatique sont discutées, l'extraction des caractéristiques et la classification. La classification des instruments de musique suit le système Hornbostel-Sachs. Dans l'extraction de caractéristiques, les caractéristiques pertinentes couramment utilisées dans la littérature sont répertoriées et organisées dans une taxonomie qui est divisée en fonction du domaine de calcul. Différentes techniques de classification largement utilisées par les chercheurs sont également présentées et passées en revue.

Mots clefs : classification des instruments de musique, apprentissage automatique, extraction de caractéristiques, classification automatique

Abstract

This paper presents a review of research on musical instrument classification which employed the machine learning system. The two main steps in the automatic classification task are discussed: feature extraction and classification. The musical instrument classification follows the Hornbostel-Sachs system. In the feature extraction, the relevant features that are commonly used in the literature are listed and organized in a taxonomy which is divided according to the domain of computation. Different classification techniques that are widely used by the researchers are also presented and reviewed.

Keywords: musical instrument classification, machine learning, feature extraction, automatic classification.

1 Introduction

Two of the various approaches often used in the studies on musical instruments are the acoustical characterization and sound recognition system. Scientists since many years ago started to discover the acoustical characteristics of different types of musical instruments by using various techniques. The initial techniques used are modal analysis and acoustic radiation. Over the years, there are many other new parameters developed and introduced from these fundamental techniques. The common acoustical characterization parameters are mechanical admittance and impedance, sound radiation coefficient, the intensity of the acoustic radiation, anti-vibrational, and transmission parameter to name a few.

Mechanical admittance is defined as the ratio of the velocity, v to the force, F. This characteristic is useful in understanding the body vibration of the musical instrument. The study which used admittance on musical instrument vibration measurements can refer. Reciprocal to the admittance, driving point mechanical impedance on the other hand is defined as the ratio of the applied force, F to the velocity, v produced by the instrument body. Measurement is done by applying the force to the instrument body and the resulting velocity is measured with the accelerometer [1].

Sound radiation characteristics of different musical instruments have been extensively studied as well. Sound radiation coefficient which is defined as the ratio of the material's speed of sound, c to its density, ρ describes how much the sound radiation of the musical instrument body is damped. It can be measured by the vibrational response of the instrument soundboard for a given force [1].

The intensity of the Acoustic Radiation (IAR) parameter is introduced by Tronchin in 2005 on the kettledrums. It is defined as the product between the space-averaged amplitude of the cross-spectrum sound pressure, p, and the velocity, v generated from the surface vibration. As the name suggests, it is a parameter related to acoustic intensity and acoustic radiation [2].

Studies were also carried out in determining the sound characteristics of the woods used in musical instruments. Various woods are tested, analyzed based on the anti-vibration and transmission parameters. The anti-vibration parameter is the reciprocal of the sound radiation ratio produced by the woods. It is the ratio of the longitudinal wave speed, c to the density, ρ of the wood. On the other hand, the transmission parameter is the product of the longitudinal wave speed, c, and the quality factor, Q. The results are then used in the acoustical classification and comparison of the woods used in a different category of musical instruments [3].

* andrew_wth@hotmail.com
† rahizar@um.edu.my
‡ saifizul@um.edu.my
♦ mfsoong@um.edu.my

On the other hand, sound recognition systems started to get more attention due to the growth of digital music. Music information retrieval (MIR) which is the subset of the broader

field of sound recognition, is known to be the field that contributes to the solutions of the musically related task. Sound recognition is a multi-disciplinary field that includes speech recognition, information retrieval, music information retrieval, environmental sound retrieval, etc. Figure 1 below illustrates the general taxonomy of the sound classification scheme introduced by [4]. Under the field of MIR, there are various tasks. For instance, music genre recognition, song identification, mood classification, music annotation, tempo, fingerprinting, etc. One of the tasks is on musical instrument classification.

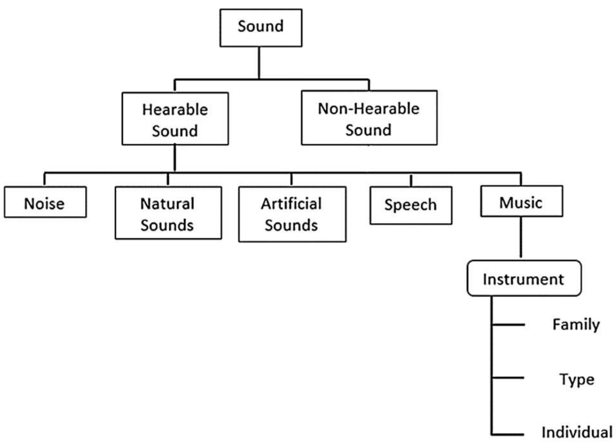

Figure 1: Taxonomy of sound [4].

The application of MIR in the musical instrument classification can help in the identification of the individual musical instrument, its type, and family. It is gaining popularity among researchers, musicians, and acousticians in the efforts of getting a better understanding of the sound produced by musical instruments. As we are currently living in the digital world, where vast amounts of musical databases are made available online. The demands are there for the development of computational tools for the analysis, summarization, classification, and indexing of those musical data [5]. These demands have inspired a growing research attempt in automatic classification of the sound produced by the different types of musical instruments.

This paper aims to review a variety of research efforts on musical instrument classification. Because of the wide variety of applications of music information retrieval as mentioned above, it is difficult to include all relevant works. This paper will only focus on the research and studies done on musical instrument classification. The rest of the paper is organized as follows. Section 2 discussed the musical instrument families and classification followed by sound recognition in Section 3. Section 4 and 5 present the features extraction and classification techniques, respectively. The conclusion is covered in Section 6.

2 Musical Instrument Classification

The globally used musical instrument classification was developed by Curt Sachs and Erich Moritz von Hornbostel in 1914. It is called the Hornbostel-Sachs system (or H-S System). Curt Sachs was a German musicologist and expert on the history of musical instruments. Erich Moritz von Hornbostel was an Austrian musicologist and expert on the history of non-European music [6]. Generally, the H-S system has five top-level classifications, which are shown in Figure 2 below. Initially, there were only four major classes excluding the electrophone. Electrophone class was introduced and added into the system by [7]. The system is then updated in 2011 by the Musical Instrument Museums Online (MIMO) project which aimed to create "a single access point to digital content and information on the collections of musical instruments held in a consortium of European museums". The system consists of several levels below those five major levels which are not shown in Figure 2. A complete list of musical instruments classification which is listed in decimal notation can refer to [8].

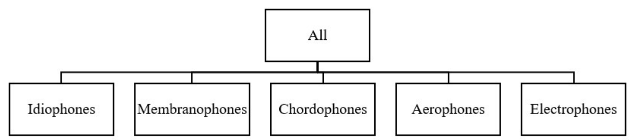

Figure 2: Hornbostel-Sachs musical instrument classification system.

The five top classes in the H-S system are idiophones, membranophones, chordophones, aerophones, and electrophones. They were divided by the primary source of sound from the musical instruments. For idiophones, the sound vibration is coming from the substance of the instrument itself, without requiring stretched membranes or strings. The sound from membranophones instruments is excited by tightly stretched membranes while chordophones instruments are by stretched string between fixed points. As for aerophones, the vibration is mainly from the air itself. Lastly, electrophones are instruments that use materials generating acoustic sound signals, electronically stored data, or electronic circuitry to generate electrical signals that are then transferred to a loudspeaker to produce sound [8].

Hornbostel-Sachs classification system is one of the various systems that have been used throughout history. Several other systems are also widely used. The western classification system is used in the west, dividing instruments into the woodwind family, string family, brass family, and percussion family. The Chinese classification system which is historically proven to be the oldest classifying scheme dates from the third millennium BC. The Chinese classification system groups the instruments according to the materials that are made of. For example, stone, wood, silk, and bamboo [9]. However, these classifying systems are not perfect in classifying all musical instruments. Certain musical instruments can be classified in more than one family. For instance, a piano that has strings, but is struck by hammers. So, it could be classified as a string instrument or percussion instrument according to the western system. The application of any classifying system is dependent on the researcher or musician and their focus or scope of research.

3 Sound Recognition System

As mentioned earlier in section 1, MIR is a subtask of the audio recognition system. The task is dealing with the automatic audio recognition of music signals which at the end will extract the information or characteristics of the music content. Musical instrument class is one of the characteristics that could be obtained by the analysis. The application of an audio recognition system in musical instrument classification is not a new thing as there are numerous attempts done by the researcher on it in recent years. Most of the research done in musical instrument classification have adopted the technique used in speech recognition and speaker identification system. This is because a few features from the speech recognition system can be directly applied to solve the musical instrument classification problem [10].

Generally, the musical instrument classification system consists of three steps, preprocessing, feature extraction, and classification as shown in Figure 3. Most of the research on musical instrument classification emphasized feature extraction which is vital in getting the correct characteristics of the sound processed.

Figure 3: Process of audio recognition system in musical instrument classification.

In the first step, the audio input that is captured by the microphone will go through a windowing process by segmenting the audio into shorter signal chunks. A musical audio signal is usually long and may contain a large number of samples given that the sampling rate is higher than 10 kHz. The audio sample is, therefore, couldn't be analyzed directly and need to go through the pre-processing step. This is because the audio signal is constantly changing. To simplify it, the audio signal is split into a continuous sequence of finite frames of samples. The frames with short scales are then assumed to be not changing much. This process converts the non-stationary audio signal into a stationary signal over a short period [11]. Typically, the segmented frame length is in between 10 to 50 milliseconds and will be overlapping with the adjacent frames for about 25 to 50% [5, 12]. This is to ensure that there are no missing signals during the segmentation process. The frame size, however, is related to the length of the processed sound signal [13].

In the pre-processing step, some research will remove the noise or silence part of the audio input before proceeding to the next step [14]. It can help in reducing the computational complexity of the recognition system. For instance, zero-crossing rate (ZCR) or the energy threshold value is used in the research done to eliminate the unwanted silence part of the audio signal. Other than that, they also applied the pre-emphasis which serves the purpose of compensating the suppressed high-frequency formants during sound production by the musical instruments [5].

The next step of sound recognition is the feature extraction of the audio signal. To classify the audio input into any musical instrument class, it is very crucial to identify the characteristics of the sound produced by each musical instrument. This process is also called parameterization that eventually will build the feature vectors that best represent the musical sound. The built feature vectors contain the most significant characteristics or parameters of the musical sound. This will then be very useful in the classification process. There are various methods to extract the characteristics or features from the audio inputs, which will be discussed in detail in section 4.

The significant parameters of the musical instruments' sound constructed in the feature extraction will be used as a descriptor to represent a similar type of musical instrument or to distinguish between different types of musical instruments. This could be done through the classification process based on various techniques or machine learning algorithms called the classifier. There are many classifiers available currently but the choice of the suitable one depends on the goal of the classification system, the accuracy of the classifier, and avoiding overfitting. In general, the classification algorithm consists of two phases: the training phase and the testing phase. In the training phase, the machine learning algorithm under supervised conditions will build representative acoustic models that best represent the sound class that the system wants to recognize. This is done by taking multiple sound samples of the same musical instrument if the musical instrument type is the goal of the machine learning system. After the algorithm is trained, it will then be tested in the testing phase. The unknown sound samples will be imported into the system for classification. The algorithm will classify the incoming sound signal into different classes based on the information acquired in the previous phase [13].

The effectiveness of the sound classification system is the main concern of the researcher. It is measured by comparing the accuracy of different features or classifiers used in the sound classification system. Until today, researchers are still trying to get the best feature set or classifier that could be used in musical instrument classification. Since 2014, there is an annual competition organized by the MIR community called Music Information Retrieval Evaluation eXchange (MIREX). This event lets the participant test their music classification system in a few categories such as genre, musical instrument, music, mood, and artist classification [11]. Other than that, the MIR community is organizing the meeting through the International Society of Music Information Retrieval Conference (ISMIR) every year since 2014.

4 Feature Extraction

Feature extraction and classifier are important components of the classification system. Feature extraction determines the features to be used for the machine learning system. The problems of classifying the sound samples into different classes based on feature vectors will be addressed. The feature vectors represent the similarities between the sound samples.

The features extracted may be redundant and irrelevant. This will cause a burden for the computation time. Therefore, some of the features will be discarded and only a subset of the features will be used at the end. This process is called feature selection. Both feature extraction and feature selection are very crucial in machine learning. It can ease the computation time by selecting only the useful and relevant features particularly when the dataset is too large [15].

There are several approaches to categorize the features extraction of the audio signal in the machine learning system. Due to the manifold nature of audio features, there is no general taxonomy that could be applied to all fields of research. Hence, it is usually designed according to the research field and purpose of the study. Fu et al. [11] unified the taxonomies of audio features by [16] into a single hierarchical taxonomy. The taxonomy consists of low-level features and mid-level features with the top-level providing the information on the human's perception towards music through the semantic labels. The low-level features in this taxonomy are divided into timbre and temporal features. As for the mid-level features, it contains information on rhythm, pitch, and harmony. The taxonomy is grouped into short-term and long-term features.

Alias et al. [13] extended the taxonomy introduced by [17] in their review on feature extraction techniques on speech, music, and environmentally sound. The taxonomy is classified into physically based and perceptually based approaches. These two approaches are then further divided into different parameterization domains such as time, frequency, wavelet, image, cepstral, etc. This is different from the taxonomy by [17] which listed the parameterization domain on the first level of taxonomy and the physically-based and perceptually based features are put under the frequency domain.

In this paper, taxonomy in Figure 4 will be adopted and the features extraction techniques in the literature for the classification of musical instruments will be reviewed. It is noted that some of the domains may not be relevant in the review of the musical instrument classification therefore it will not be covered in this paper. Only the relevant domain such as time, frequency, cepstral, wavelet domain is covered. The mathematical analysis in detail is beyond the objectives of this study and will not be covered in this paper.

4.1 Time Domain

Also called a temporal domain, the time domain is perhaps the most basic domain for audio signals. It is not complex and easy to extract audio features from. It can be displayed directly from the raw audio signal without further transformation. There are four classes of physical time-domain audio features: zero crossing-based, amplitude-based, power-based, and rhythm-based features.

4.1.1 Zero-Crossing Rate-Based Features

The technique used here is based on the analysis of the rate of change of the sound signal. It is a simple but effective method commonly used in MIR.

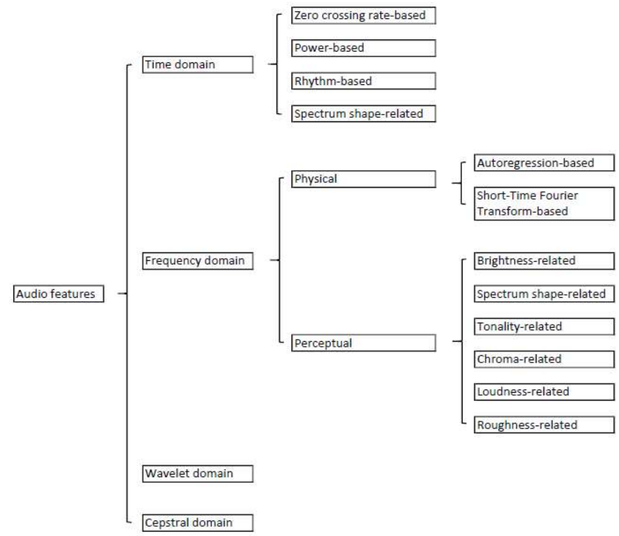

Figure 4: Taxonomy of audio features in musical instrument classification.

Zero-Crossing Rate (ZCR): Known to be one of the easiest features to get from the audio signal. The zero-crossing rate is defined as the number of times the audio signal waveform passed the zero-amplitude level within one second. This feature is widely used in audio classification and machine learning systems. It is measured based on the rate of change of the audio signal and is probably the simplest way for feature extraction. Kedem [18] and El-Maleh et al. [19] mentioned in their papers that the ZCR can provide a rough estimation of the dominant frequency and the spectral centroid in the signal. ZCR is quite popular in the musical instrument classification field.

4.1.2 Power-Based Features

Power-based features are extracted based on the audio signal power. Few relevant features are described below.

Energy: Using the frame-based procedure, the energy feature summarizes the energy distribution of each frame over time. Mitrovic et al. used the term short-time energy to represent this feature [17]. The researchers used this feature for finding the energy distribution in each frame and tried to find the differences between the instruments. Bhalke et al. [5] used time-domain energy as the feature in their musical instrument recognition paper.

Temporal Centroid: Temporal centroid gives the time average over the signal envelope in seconds. It represents the instant moment in time that containing the largest average energy of the signal. The temporal centroid has been used as a time-domain audio feature. It may also be classified as a MPEG-7 feature in the musical instrument classification field [12].

Log Attack Time (LAT): The log attack time characterizes the attack of the sound signal. Musical instruments can produce either instant or smooth transitions of musical sounds. It is computed as the logarithm of the time taken from the start to the first significant local peak [12].

Root Mean Square (RMS): Also named as the volume is the review by [17], RMS is computed by finding the root mean square of the waveform magnitude within the frame [20].

4.1.3 Rhythm-Based Features

Rhythm is a relevant characteristic of musical sound that characterizes the sonic events' structural organization [13]. Feature derived under this taxonomy is discussed here.

Periodicity: Periodicity or tempo is the measure of the rhythmic strength or repetitive structures of audio signals [21]. Periodicity is obtained by applying the autocorrelation function to acquire the mean value of the maximum peaks through all the signal frames.

4.1.4 Spectrum Shape-Related Features

The spectrum shape of the audio signal is another relevant feature that could be employed in the task of musical instrument classification. Spectrum shape-related features are described in the following paragraphs.

Attack, Decay, Sustain, and Release (ADSR) Envelope: The temporal envelope of musical instrument sounds are characterized by attack time, decay time, sustain time, and release time as shown in Figure 5. Attack time is the time taken for the sound signal to rise from zero to the peak. The decay time is the subsequent time to run down the signal level from peak to the sustained level. Sustain time is the main sequence where the signal level remained the same and lastly, the release time represents the time taken for the signal to decay back to the zero levels. ADSR combined up to form a signal envelope that could be extracted as a feature in vector form in the musical instrument classification task [5].

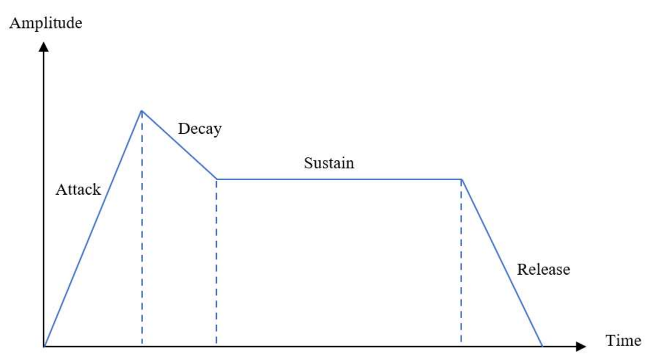

Figure 5: The ADSR envelope

Amplitude Modulation (AM): Amplitude modulation (AM) features are extracted from the audio signal for the peaks which corresponds to the frequency of amplitude modulation. AM has measured over two spectral ranges 4 to 10 Hz and 10 to 40 Hz [22].

Autocorrelation Coefficients (AC): Autocorrelation coefficients (AC) represent the overall shift of the spectrum [23]. Brown reported that AC is useful in musical instrument identifications [24].

Temporal Kurtosis: Temporal kurtosis shows the spikiness of the audio envelope. It is used in measuring the variation of the transients of the audio signal over successive frames [25].

4.2 Physical Frequency Domain

The frequency domain is also named the spectral domain. According to [17], audio features on the spectral domain form the largest set of audio features. They are acquired from autoregression analysis or Short-Time Fourier Transform (STFT). This paper employed the approach by [17] in further dividing the frequency domain into two subsets: physical features and perceptual features. In this section, features extracted in the physical frequency domain for the musical instrument classification task will be discussed first.

4.2.1 Autoregression-based

Autoregression-based features use linear prediction analysis on signal processing. The linear predictor captures the spectral predominance of audio signals [13]. Commonly used autoregression-based features are discussed below.

Linear Prediction Coefficients (LPC): Linear prediction coefficients capture the spectral envelope of the audio signal, such as formant frequencies that could be found in the vocal tract. It has been used extensively in speech recognition applications. The application of LPC in musical instrument classification could be found in the works by [14]. The prediction model used is shown in Figure 6. It consists of the input $u(n)$ which is the periodical sound produced by the musical instrument, $H(z)$ which represents the musical instrument system and the output $o(n)$ represents the music.

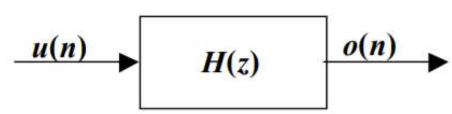

Figure 6: Linear prediction model for musical instrument sound production [14].

Line Spectral Frequencies (LSF): Line Spectral Frequencies are also called Line Spectral Pairs (LSP). It is obtained by finding the root phases of the two polynomials that are decomposed from the LPC [26]. LSF is proved to be more robust when compared to LPC as they provide statistical properties.

4.2.2 Short-Time Fourier Transform-Based Frequency Features

Short-Time Fourier Transform or STFT-based audio features are obtained from the signal spectrogram that is employed by STFT computation. According to [17], there are two ways to yielding the STFT features, either from the spectrogram envelope or from the STFT phase. The application of STFT-based features in musical instrument classification is found to be mostly, if not all, from the spectrogram envelope. These features are widely employed by researchers and are discussed below.

Spectral Flux: Spectral flux (SF) is defined as the 2-norm of the frame to frame spectral amplitude difference vector by [27]. SF measures the changes in the spectrum shape over time. Signals without much variation like noise will show low SF, while the high SF indicates sudden changes that are useful in detecting certain information like the onset of sound.

Spectral Peaks: As defined by [28], spectral peaks are the constellation maps that display the most significant local peaks in the time-frequency signal distortions. The advantage of this feature is that it is highly robust to noise since the significant peak frequencies are usually free from noise disturbance. This feature is used by [28] in the Shazam search engine.

Audio Spectrum Envelope: Audio spectrum envelope (ASE) is defined as the log-spectrum frequency power spectrum that produced a reduced spectrogram of the original audio signal. ASE consists of coefficients that describe the power spectrum density within a series of frequency bands. Categorized as a MPEG-7-based low-level descriptor, it is suitable for automatic musical sound recognition [29].

4.3 Perceptual Frequency Domain

Another division of frequency-based features is the perceptual domain. Perceptual features have a semantic meaning as the human auditory perception. In this section, several perceptual features will be included and discussed.

4.3.1 Brightness-Related Perceptual Frequency Features

The brightness of an audio signal characterizes the frequency spectrum distribution. An audio signal is considered bright when it is dominated by high frequencies. Brightness is also defined as the balancing point of the signal energy [27].

Spectral Centroid: Spectral centroid (SC) is one of the commonly used features. It describes the center of the gravity (centroid) of the spectral energy. It can also be defined as the first moment which is the frequency position of the mean value of the spectrum [30]. Deng et al. in their work on musical instrument classification defined that the SC measures the average frequency weighted by the sum of spectrum amplitude within each frame [12].

Sharpness: Even though it is often treated to be similar to the spectral centroid, sharpness is computed based on the specific loudness instead of the spectrum magnitude. The sharpness of a sound increases as the strength of the high frequencies of the spectrum increases [31].

4.3.2 Spectrum Shape-Related Perceptual Frequency Features

Spectrum shape is considered one of the popular and widely used approaches in MIR. The relevant set of spectrum-shape-related features are listed below.

Bandwidth: Bandwidth is also called a centroid width. It shows the weighted average of the deviations between the spectral components with the spectral centroid [32]. It is the second-order statistic of the spectrum which could distinguish the tonal sounds and noise-like sounds. Bandwidth can be defined from the logarithmic approach or the power spectra [20]. Alternatively, it could also be computed from the entire spectrum or within the spectrum subbands [33]. According to the MPEG-7 standard, bandwidth is defined as the audio spectrum spread (ASS) which is obtained by computing the standard deviation of the signal spectrum.

Spectral roll-off point: Spectral roll-off point is defined as the *N*% percentile of the power spectral distribution. *N* is set at the 95th percentile by [27]. It's a measurement of the skewness of the spectral shape.

Spectral flatness: Spectral flatness measures the flatness of the frequency distribution of the power spectrum. It is calculated by taking the ratio between geometric and the arithmetic mean of a subband in the power spectrum [33]. Spectral flatness can differentiate between noise-like sounds and tonal sounds. Noise-like sounds and tonal sounds are high and low in ratio, respectively. This is beneficial in the musical instrument classification task.

Spectral crest factor: This feature is the contrast of spectral flatness. The spectral crest factor measures the spikiness of the power spectrum. It can be obtained by finding the ratio of the maximum power spectrum and the mean power spectrum of a subband. Opposite to the spectral flatness, noise-like sounds will show a low spectral crest factor while tonal sounds give a higher spectral crest factor. Eronen and Klapuri applied crest factors in their research on musical instrument classification [34].

Entropy: Another measurement of spectral flatness is entropy. It is used in measuring the noisiness of the audio signal. Shannon entropy is usually computed in different subbands [33].

Spectral slope: Spectral slope is a measurement of the inclination of the spectrum shape by applying the linear regression method [35].

Spectral skewness and kurtosis: Spectral skewness is defined as the asymmetry of the spectral distribution around the spectral centroid. Spectral kurtosis, on the other hand, tells the spikiness of the frequency spectrum. The value of spectral kurtosis is high if the spectrum is spikier and low if it is flatter [25].

4.3.3 Tonality-Related Perceptual Frequency Features

The review by [13] put the features under the tonality category differently from the review by [17]. According to [13], tonality features are related to the fundamental frequency which is defined as the lowest frequency of the stationary harmonic sound signal. Tonality describes the structure of the sounds that constitute the fundamental frequency and its partials. Tonality-related features that are widely used in musical instrument classification will be listed and discussed below.

Fundamental Frequency: Denoted as "F naught" or F0, the estimation of fundamental frequency could be done with several approaches, such as spectral methods, autocorrelation methods, or cepstral methods. In the review by [17] and some other literature, the fundamental frequency is denoted as a pitch of the audio signal. Work by [22] extracted fundamental frequency as a feature in instrument recognition.

Harmonicity: Also called partials, harmonics are the integer multiples frequencies of the fundamental frequency. They are often denoted as F1, F2, F3, etc. Harmonicity features can distinguish between periodic and non-periodic sound signals and are commonly employed in recognizing musical instruments. There are two measurements of harmonicity according to the MPEG-7 standard. The first one is the Harmonic ratio which measures the proportion of harmonic components in the power spectrum. The other one measures the upper limit of harmonicity (ULH) which estimates the frequency beyond the spectrum that no longer contains harmonic structure [36].

Inharmonicity: Fundamental and its subsequent harmonics may not always show perfect harmonicity (integer multiples of F0) in the real situation. The actual location of the harmonics may deviate away from its ideal location. This is called inharmonicity and is one of the features extracted in musical instrument timbre classification [37].

MPEG-7 Spectral Timbral Descriptors: Several features are closely related to the harmonic structure of the sound according to the MPEG-7 standard. They are found to be suitable in the discrimination of musical instrument sounds. The features are harmonic centroid, harmonic deviation, harmonic spread, and harmonic variation. The harmonic centroid is the amplitude-weighted average of the harmonic frequencies which is related to the sharpness and brightness. Harmonic deviation measures the deviation of the harmonic peaks from their neighboring harmonic peaks. The harmonic spread is the power-weighted root-mean-square deviation of the harmonic peaks obtained from the harmonic centroid. It is related to the bandwidth of the harmonic frequencies. Lastly, harmonic variation describes the correlation between the two adjacent harmonic peak amplitudes. It represents the harmonic variability of the harmonic structure over time. The application of these features could be found in the work by [12].

Jitter: Jitter determines the deviations of the cycle-to-cycle fundamental frequency. Barbedo and Tzanetakis in their work on the classification of musical instruments describe jitter as the measurement of the stability of the partial over time [38].

4.3.4 Chroma-Related Perceptual Frequency Features

The chroma-related feature is considered as the perceptual feature by [17] and is mainly used in musical information retrieval as it could describe the octave invariance of the sound signal. Chroma is normally ranged to 12 pitch classes, with each class one note of the twelve-tone equal temperament [39]. Two notes with a separation of one or more octaves are said to be having the same chroma. The same chroma means that the notes will produce the same effect on human auditory perception.

Chromagram: Chromagram is computed from a logarithmic Short-Time Fourier Transform to the spectrogram that represents the energy of the 12 pitch classes. It maps all spectral audio information into one octave which results in spectral compression. This could be used in describing the harmonic musical sound signals.

4.3.5 Loudness-Related Perceptual Frequency Features

Loudness is one of the perceptual features that the human auditory system can sense in listening to the sound signal. Loudness-related perceptual features aim to simulate human hearing ability in the audio retrieval system. Peeters et al. defined loudness as the subjective impression of the sound intensity [23].

Loudness: Loudness is computed from the normalized power spectrum of the input frame which subtracts an approximation of the absolute threshold of hearing. It is then filtered by gammatone filter banks and the frequencies across are summed to obtain the power of each auditory filter. These powers which represent the internal excitations will be compressed, scaled, and summed across the filters to extract the loudness estimation [40].

Specific Loudness Sensation: Specific loudness sensation is a measurement of loudness in sone units. Sone units are defined as a perceptual scale for loudness measurement according to [23]. Pampalk et al. computed this feature by merging the spectral masking effect and the Bark-scale frequency analysis [41].

4.3.6 Roughness-Related Perceptual Frequency Features

Roughness is a fundamental hearing sensation that measures the sensory dissonance of the sound signals. According to [42], the amplitude variations which change rapidly will cause unpleasantness and reduce the noise quality, hence deducing that the sound is rough. Computation of roughness can refer to the work by [31, 40]. The application of roughness as a feature in musical instrument classification can see [38].

4.4 Wavelet Features

The application of wavelet is based on the division of the continuous-time signal or given function into different scale components [13]. Wavelet transform can extract the desired time-frequency components of the musical sound signal. The wavelet is decomposed into sub-bands which will be further analysed. The characteristics information of the particular musical sound signal can then be obtained. According to [43], comparing to Fourier transform, wavelet transform has advantages in showing the functions consisting of discontinuities and sharp peaks. It is also good in constructing and deconstructing finite non-stationary signals.

Daubechies Wavelet coefficient histogram features: Proposed by [44] in their study on music genre classification, Daubechies wavelet coefficient histogram is applied by decomposing the audio signal by Daubechies wavelet. Histograms are built from the wavelet coefficients obtained for each subband. The histograms estimate the waveform variation of each subband. Wavelet features are obtained by computing the first three statistical moments and the energy of the coefficients subband.

4.5 Cepstral Features

Introduced by [45] with the concept of "cepstrum", cepstral features represent the smoothed frequency based on the logarithmic magnitude. It was first employed in speech analysis [46] and is now widely used in various fields of audio information retrieval.

4.5.1 Perceptual Filter Bank-Based Features

Perceptual filter banks-based features are computed based on the cepstral domain. The sound signal is first Fourier transformed; the magnitude is then converted into the logarithmic scale. Discrete Cosine Transform will be performed on the previous result to decorrelate the output data.

Mel-frequency cepstral coefficients (MFCCs): Also called MFCC, this feature is very well known in automatic speech recognition and audio content classification. MFCC is designed and computed based on the human auditory model. To extract the MFCC features, the audio signal is framed into short frames and the periodogram estimate for each frame is computed. Mel frequency is then applied to the power spectra before the energy in each filter is summed. All the filterbank energies are then logarithmized and lastly, they are decorrelated by the Discrete Cosine Transform (DCT). Only 8-13 DCT coefficients will be used to represents the spectral shape of the audio signal. The first DCT coefficients represent the spectrum's mean power. The second coefficient represents the spectral centroid. Higher-order coefficients are related to spectral details like pitch [17]. Figure 7 shows the MFCCs obtained from the flute musical instrument.

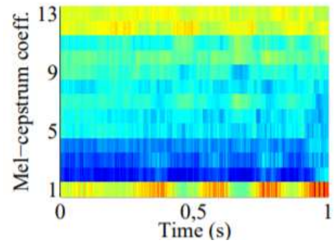

Figure 7: MFCCs for the flute musical instrument [17].

4.5.2 Autoregression-Based Features

Autoregression analysis is often used in signal processing. This technique uses linear prediction analysis that can predict the value of every signal sample by the linear combination of previous values [47].

Complex Cepstrum: According to the [48], complex cepstrum is the inverse Fourier transform of the logarithm of the signal's Fourier transform. Application of the complex cepstrum on the musical instrument recognition can read the work by [49].

Linear Prediction Cepstral Coefficients (LPCC): Linear prediction cepstral coefficients are the alternative for linear prediction coefficients (LPC) discussed earlier above. They are obtained by the inverse Fourier transform of the log magnitude frequency response of the linear prediction spectral envelope [50]. In comparison to LPC, LPCC is more robust in representing the spectral envelope.

5 Classification

After feature extraction and selection, classifiers are used in the machine learning system to classify the isolated musical sounds into the instrument and its family. In this section, several techniques commonly used in automatic musical instrument classification will be discussed. It is worth noting that the accuracy or effectiveness of the classifier is affected by many factors (number of samples, combination with different features, number of samples used in the testing phase and training phase, etc.). Therefore, the classifiers in the following paragraphs will not include the accuracy obtained by each literature reviewed in this paper.

5.1 K-Nearest Neighbours

Also denoted as KNN, this classifier is one of the popular machine learning algorithms. In the training phase, it will store the feature vectors from all the training samples and then use them in classifying the new test samples. By referring to the set of k nearest training samples in the feature set, the new sample will be assigned to the class with the most examples in the set. The system is using the Euclidean distance measurement method. Details of how the classification process goes can refer to the Figure 8 below.

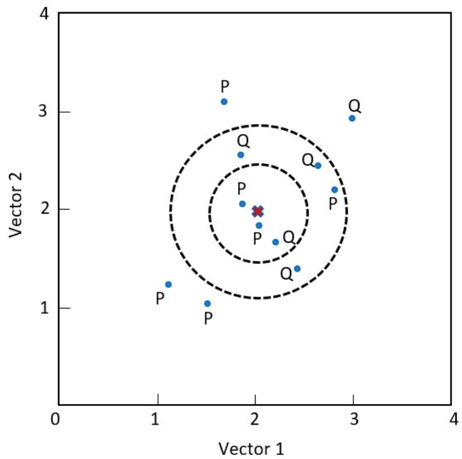

Figure 8: Design of the KNN technique.

From Figure 8 above, the cross is the target of the classification. If k=3 is selected (inner circle), then the cross would be categorized as class P as the three nearest neighbours to that cross are mostly from class P. However, if k=7 is selected (outer circle), the cross would now be categorized as class Q with Q be the majority neighbours.

k-Nearest Neighbours is a simple algorithm that is widely used in the automatic machine learning system, but some downsides are to be considered when implementing this technique. According to [51], this algorithm is lazy and requires stores all the training samples in the memory to generate a decision for the new sample. It is also highly sensitive to the irrelevant features which could dominate the distance metrics. Heavy computational load is another drawback of this algorithm.

5.2 Support Vector Machine

Another popular classifier used is the support vector machine (SVM). It is based on the statistical learning theory developed by [52]. The working principle of SVM is looking for the optimal linear hyperplane which gives the lowest generalization errors when classifying the unknown test sample. The linear hyperplane is mapped so that the margin between the different categories is separated as wide as possible. It serves as the borderline between the categories. The new test samples could be categorized based on which side they fall when they are mapped into space. The hyperplane is a linear line when the features can be separated into 2 dimensions. It will become a 2D plane when it is displayed in three-dimension space. This approach can be used when linearly hyperplane couldn't separate the data in 2-dimensional space and requires higher dimensional space to do so. This is achieved by applying the so-called "kernel trick" as illustrated in Figure 9. Kernel trick transforms the low dimensional input space to a higher dimensional space so that the segregation (hyperplane) could take place.

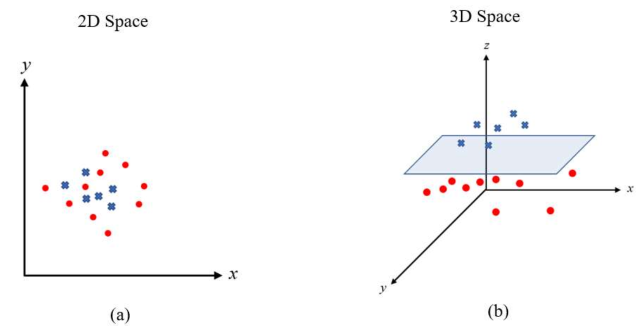

Figure 9: (a) Inseparable data in 2D space (b) Hyperplane separating the data in 3D space with the kernel trick.

Even though SVM is a popular algorithm used by many in their research, SVM does still gives some drawbacks. In the multiclassification task, SVM needs to perform a series of interconnections between the classes. Computation-wise, it is an intensive process to work on. Also, there is a risk of selecting the less optimal kernel function during the process.

5.3 Decision Trees

Decision trees have been pervasively implemented in classification tasks and machine learning systems. This technique attempts to focus on the relevant features and abandons irrelevant ones in the construction of the tree. A decision tree is built top-down that begins with the most informative root node. Usually, two branches will split from the root which represents different descriptor values or attribute. Each node in the tree represents the test of the samples' attributes, and the descendant node represents the result of the test. The complete tree is built by repeating the training process recursively with the training samples. After that, pruning work will be carried out to avoid overfitting. The decision tree is commonly used in supervised learning methods which produce high accuracy, stability, and are easy to interpret.

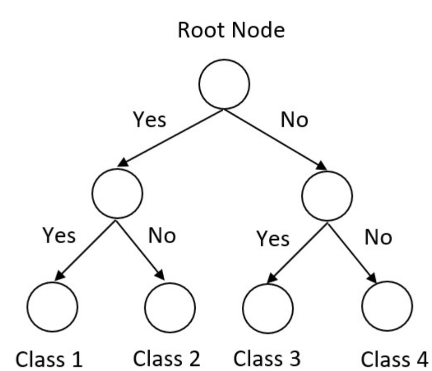

Figure 10: Decision tree in classifying 4 classes of musical instruments.

In the musical instrument classification task, the decision tree can also help in identifying the best feature in discriminating instruments. From the literature, the common decision tree algorithm used is J48 or also known as C4.5. This algorithm is also called a statistical classifier which is developed by [53].

5.4 Naive Bayesian Classifiers

Naïve Bayesian Classifier (NBC) is the classification technique based on the Bayes' Theorem. This technique uses a conditional probability model in the prediction of classes. Naïve Bayes classifier assumes that the classification features are independent, hence it is called "naïve". Like the other classifiers, NBC will be trained by collecting enough training samples. The probabilities of different classes and features will be obtained by counting the frequencies of their occurrence in the training phase. A new sample can then be classified based on conditional probabilities. BNC is one of the easy and fast algorithm one can use in the classification task. It requires less training data and is not computationally intensive. However, this algorithm is known as a bad estimator. It is also too "naïve" by assuming that the features are completely independent in real life. Deng et al., in their works on the feature analysis for musical instrument classification, used the NBC technique as one of the classifiers [12].

5.5 Artificial Neural Networks

Artificial neural networks (ANN) are inspired by biological neural networks. It is constructed based on a large collection of interconnected artificial neurons. These neurons are arranged into layers in which the transmission of signal happens from the input layer to the output layer by the connection called edges. These edges have a weight that tells the strength between the connecting layers. The weight may change during the learning process. With sufficient training samples, the network becomes capable of predicting the outcome from the input. This learning process can be done either supervised or unsupervised. The prediction accuracy of ANN is getting better when more examples are processed. It keeps on learning and refining the weight for every sample processed. Implementation of ANN in musical instrument classification can be found in [10].

5.6 Hidden Markov Models

Abbreviated as HMM, Hidden Markov Model is a statistical Markov model that contains two components. The first is a set of hidden variables that is unobservable directly from the data while the second is another set of variables that are conditional on the first set of hidden variables [54]. HMM is used in predicting a sequence of the hidden variables from a set of observed variables. This allows the model to generate a random measurement in each state from a variety of distributions.

5.7 Gaussian Mixture Models

The Gaussian mixture model is a probabilistic model to representing the subpopulation that is normally distributed within the overall population. Without needing to know which the data point belongs to, this allows GMM to automatically learn the subpopulation. This model can do the clustering of groups of data mixed. This is done by the computation of the three parameters which are the mean, covariance, and the mixing probability of the Gaussian mixture. Due to this, GMM is unsupervised learning. This classifier has been used in speech recognition, image pattern recognition, and musical instrument classification. GMM is one of the popular classifiers used intensively in instrument classification. For instance, refer to [14].

5.8 Discriminant Analysis

Discriminant analysis is a technique used in machine learning to find the linear, quadratic, or logistic functions of the features that characterize or separates samples into two or more predefined classes. Discriminant analysis is related to the multivariate analysis of variance (MANOVA) and regression analysis. This technique could determine the most discriminative features of each class and the most similar or dissimilar classes. Martin and Kim used linear discriminant analysis in their research on musical instrument identification [55].

5.9 Higher-Order Statistics

Higher-order statistics (HOS) is the technique that uses the sample function with cubic power or higher. Conventional techniques (lower-order statistics) are functions with constant, linear, or quadratic functions. Mean and variance is examples of lower-order statistics. HOS in the analysis of musical signals used skewness and kurtosis as the estimation of the shape parameters.

6 Conclusion and Future Work

In this paper, two important steps in the process which are feature extraction and classification are reviewed. The application of the MIR in classifying musical instruments into different families or individual instruments is gaining wide interest from researchers and musicians. Different approaches have been used and they harvested different results. The effort is to obtain the best feature set which contains either individual or the combination of temporal, spectral, cepstral, and other properties of sound in the classification task. Choosing a good classifier is also important in which it can better identify the subtle characteristics of different instruments or families.

From the review in this paper, we have discussed various approaches in the features and classifiers used in classifying monophonic instrumental sounds. While the coverage of this paper is not exhaustive, it is apparent that there is no specific feature or classifier which can be considered as the best in the musical instrument classification task. Most of the works by the literature reviewed in this paper are on the comparison of the accuracy obtained by the different combinations of features and classifiers. The different combination shows different accuracies and also projected both advantages and disadvantages. The selection of the appropriate features or classifier is dependent on the specific task of classification. For instance, the complexity of the learning phase, database size, real-time limitations, etc. However, it can be concluded that fewer features used in the sound recognition system will usually achieve better accuracy and also reduces the computational burden.

It can be noticed that certain literature reviewed is working on traditional musical instruments. These efforts sparked the interest of the authors in working towards the classification of the sound of the local traditional musical instruments. However, our research interest is not in the classification of the individual instrument or families. To our knowledge, an approach yet to be tested is the ability of the sound classification system in identifying the sound quality produced by the traditional musical instrument. The current work by authors or any other future efforts can be focusing on the instruments' sound quality. It is agreeable that "no instruments are 100% alike", hence the quality of each instrument might differ from one another and this is something worth study for. The subtle differences in sound quality might create another tough challenge in the MIR field, but it is worthy to explore for and it is hoped that the new exploration may produce useful knowledge in the future.

Remerciements/Acknowledgments

This paper is extracted from the Ph.D. work of the first author who is currently under the sponsorship of the University College of Technology Sarawak (UCTS). This work is also funded by UCTS under Grant No 4/2018/01.

References

[1] J. Meyer, Acoustics and the performance of music: Manual for acousticians, audio engineers, musicians, architects and musical instrument makers, Springer Science & Business Media2009.

[2] L. Tronchin, Modal analysis and intensity of acoustic radiation of the kettledrum, The Journal of the Acoustical Society of America 117(2) (2005) 926-933.

[3] S. Yoshikawa, Acoustical classification of woods for string instruments, The Journal of the Acoustical Society of America 122(1) (2007) 568-573.

[4] D. Gerhard, Audio signal classification: History and current techniques, Citeseer2003.

[5] D. Bhalke, C.R. Rao, D.S. Bormane, Automatic musical instrument classification using fractional fourier transform based-MFCC features and counter propagation neural network, Journal of Intelligent Information Systems 46(3) (2016) 425-446.

[6] E.M. Von Hornbostel, C. Sachs, Systematik der musikinstrumente. Ein versuch, Zeitschrift für Ethnologie 46(H. 4/5) (1914) 553-590.

[7] F.W. Galpin, A textbook of European musical instruments: their origin, history and character, London: Williams & Norgate1937.

[8] M. Consortium, Revision of the Hornbostel-Sachs classification of musical instruments by the MIMO Consortium, Available in:< http://network. icom. museum/cimcim/resources/classification-of-musical-instruments (2011).

[9] D.E. Hast, J.R. Cowdery, S.A. Scott, Exploring the world of music: an introduction to music from a world music perspective, Kendall Hunt1999.

[10] I. Kaminsky, A. Materka, Automatic source identification of monophonic musical instrument sounds, Proceedings of ICNN'95-International Conference on Neural Networks, IEEE, 1995, pp. 189-194.

[11] Z. Fu, G. Lu, K.M. Ting, D. Zhang, A survey of audio-based music classification and annotation, IEEE transactions on multimedia 13(2) (2010) 303-319.

[12] J.D. Deng, C. Simmermacher, S. 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.

[13] F. Alías, J.C. Socoró, X. Sevillano, A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds, Applied Sciences 6(5) (2016) 143.

[14] J. Marques, P.J. Moreno, A study of musical instrument classification using gaussian mixture models and support vector machines, Cambridge Research Laboratory Technical Report Series CRL 4 (1999) 143.

[15] M.N. Murty, V.S. Devi, Introduction to pattern recognition and machine learning, World Scientific2015.

[16] N. Scaringella, G. Zoia, D. Mlynek, Automatic genre classification of music content: a survey, IEEE Signal Processing Magazine 23(2) (2006) 133-141.

[17] D. Mitrović, M. Zeppelzauer, C. Breiteneder, Features for content-based audio retrieval, Advances in computers, Elsevier2010, pp. 71-150.

[18] B. Kedem, Spectral analysis and discrimination by zero-crossings, Proceedings of the IEEE 74(11) (1986) 1477-1493.

[19] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, Speech/music discrimination for multimedia applications, 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), IEEE, 2000, pp. 2445-2448.

[20] Z. Liu, J. Huang, Y. Wang, T. Chen, Audio feature extraction and analysis for scene classification, Proceedings of First Signal Processing Society Workshop on Multimedia Signal Processing, IEEE, 1997, pp. 343-348.

[21] L. Lu, H. Jiang, H. Zhang, A robust audio classification and segmentation method, Proceedings of the ninth ACM international conference on Multimedia, 2001, pp. 203-211.

[22] A. Eronen, Comparison of features for musical instrument recognition, Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575), IEEE, 2001, pp. 19-22.

[23] G. Peeters, A large set of audio features for sound description (similarity and classification) in the CUIDADO project, CUIDADO IST Project Report 54(0) (2004) 1-25.

[24] J.C. Brown, O. Houix, S. McAdams, Feature dependence in the automatic identification of musical woodwind instruments, The Journal of the Acoustical Society of America 109(3) (2001) 1064-1072.

[25] A. Klapuri, M. Davy, Signal processing methods for music transcription, Springer Science & Business Media2007.

[26] J.P. Campbell, Speaker recognition: A tutorial, Proceedings of the IEEE 85(9) (1997) 1437-1462.

[27] E. Scheirer, M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, 1997 IEEE international conference on acoustics, speech, and signal processing, IEEE, 1997, pp. 1331-1334.

[28] A. Wang, An industrial strength audio search algorithm, Ismir, Washington, DC, 2003, pp. 7-13.

[29] H.-G. Kim, N. Moreau, T. Sikora, Audio classification based on MPEG-7 spectral basis representations, IEEE Transactions on Circuits and Systems for Video Technology 14(5) (2004) 716-725.

[30] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on speech and audio processing 10(5) (2002) 293-302.

[31] E. Zwicker, H. Fastl, Facts and Models, Springer, 1999.

[32] E. Wold, T. Blum, D. Keislar, J. Wheaten, Content-based classification, search, and retrieval of audio, IEEE multimedia 3(3) (1996) 27-36.

[33] A. Ramalingam, S. Krishnan, Gaussian mixture modeling of short-time Fourier transform features for audio fingerprinting, IEEE Transactions on Information Forensics and Security 1(4) (2006) 457-463.

[34] A. Eronen, A. Klapuri, Musical instrument recognition using cepstral coefficients and temporal features, 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), IEEE, 2000, pp. II753-II756 vol. 2.

[35] F. Morchen, A. Ultsch, M. Thies, I. Lohken, Modeling timbre distance with temporal statistics from polyphonic music, IEEE Transactions on Audio, Speech, and Language Processing 14(1) (2005) 81-90.

[36] X. Zhang, W.R. Zbigniew, Analysis of sound features for music timbre recognition, 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07), IEEE, 2007, pp. 3-8.

[37] G. Agostini, M. Longari, E. Pollastri, Musical instrument timbres classification with spectral features, EURASIP Journal on Advances in Signal Processing 2003(1) (2003) 943279.

[38] J.G.A. Barbedo, G. Tzanetakis, Musical instrument classification using individual partials, IEEE Transactions on Audio, Speech, and Language Processing 19(1) (2010) 111-122.

[39] R.N. Shepard, Circularity in judgments of relative pitch, The Journal of the Acoustical Society of America 36(12) (1964) 2346-2353.

[40] M. McKinney, J. Breebaart, Features for audio and music classification, (2003).

[41] E. Pampalk, A. Rauber, D. Merkl, Content-based organization and visualization of music archives, Proceedings of the tenth ACM international conference on Multimedia, 2002, pp. 570-579.

[42] P. Daniel, R. Weber, Psychoacoustical roughness: Implementation of an optimized model, Acta Acustica united with Acustica 83(1) (1997) 113-123.

[43] S. Mallat, A theory for multiresolution signal decomposition: The wavelet representation Fundam, Pap. Wavelet Theory 2 (2009) 494-513.

[44] T. Li, M. Ogihara, Q. Li, A comparative study on content-based music genre classification, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003, pp. 282-289.

[45] B.P. Bogert, The quefrency alanysis of time series for echoes; Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking, Time series analysis (1963) 209-243.

[46] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE transactions on acoustics, speech, and signal processing 28(4) (1980) 357-366.

[47] T.E. Tremain, The government standard linear predictive coding algorithm: LPC-10, Speech Technology (1982) 40-49.

[48] A.v. Oppenheim, R. Schafer, T. Stockham, Nonlinear filtering of multiplied and convolved signals, IEEE transactions on audio and electroacoustics 16(3) (1968) 437-466.

[49] J.C. Brown, Computer identification of musical instruments using pattern recognition with cepstral coefficients as features, The Journal of the Acoustical Society of America 105(3) (1999) 1933-1941.

[50] Q.-Z. Wu, I.-C. Jou, S.-Y. Lee, On-line signature verification using LPC cepstrum and neural networks, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 27(1) (1997) 148-153.

[51] T.M. Mitchell, Machine learning. 1997, Burr Ridge, IL: McGraw Hill 45(37) (1997) 870-877.

[52] V. Vapnik, Statistical learning theory Wiley, New York 1 (1998) 624.

[53] J. Quinlan, C4. 5: programs for machine learning, Elsevier (2014).

[54] P. Herrera-Boyer, G. Peeters, S. Dubnov, Automatic classification of musical instrument sounds, Journal of New Music Research 32(1) (2003) 3-21.

[55] K.D. Martin, Y.E. Kim, Musical instrument identification: A pattern-recognition approach, The Journal of the Acoustical Society of America 104(3) (1998) 1768-1768.


HBK
HOTTINGER BRÜEL & KJÆR
TECHNOLOGY BY
Brüel & Kjær
NEED TO SIMPLIFY NOISE MEASUREMENT AND ANALYSIS?
JOB DONE.
No matter which profession you are in, you need a sound level meter solution that gets your job done faster, easier and problem-free. The new B&K 2245 gives you absolute confidence and control through user-friendly mobile apps and functionality tailored for your task, including on-site analysis, photo embedding, smart data handling on your PC, and more.
To simplify your job-to-do, visit www.hbkworld.com/2245
HBM
Brüel & Kjær
64.5