Analysis and Differentiation of Music For Predominant Instrument Recognition Using Neural Network

Dr. M. Sulthan Ibrahim
Assistant Professor and Head, Dept of Computer Science
Government Arts and Science college, Veerapandi-625534, Theni, Tamil Nadu, India.

Abstract- *Music retrieval becomes dominant in Jukeboxes and sound systems. As the interest of listeners is varying fine to time and based on other factors like environment, psychology, etc. Identifying the type of music category helps the sound provider and listener to make their interested choice. This paper presents a software methodology based on spectrum analysis and neural network which classifies the category of music based on known features. Several audio classifications are presented and the results are obtained on audio framing, spectrum analysis, feature extraction and classification on machine learning approach.*

Keywords- Convolutional neural networks, deep learning, instrument recognition, multi-layer neural network, and music information retrieval.

I. INTRODUCTION

MUSIC can be said to be built by the interplay of various instruments. A human can easily identify what instruments are used in music, but it is still a difficult task for a computer to automatically recognize them. This is mainly because music in the real world is mostly polyphonic, which makes the extraction of information from audio highly challenging. Furthermore, instrument sounds in the real world vary in many ways such as for timbre, quality, and playing style, which makes identification of the musical instrument even harder.

In the music information retrieval (MIR) field, it is highly desirable to know what instruments are used in the music. First of all, instrument information per se is important and useful information for users, and it can be included in the audio tags. There is a huge demand for music search owing to the increasing number of music files in digital format. Unlike text search, it is difficult to search for music because input queries are usually in text format. If instrument information is included in the tags, it allows people to search for music with the specific instrument they want. In addition, the obtained instrument information can be used for various audio/music applications. For instance, it can be used for a tailored instrument-specific audio equalization and a music recommendation services. In addition, it can be used to enhance the performance of other MIR tasks. For example, knowing the number and type of the instrument can significantly improve the performance of audio source separation, automatic music transcription, and genre classification. Instrument recognition can be performed in various forms. Hence, the term “instrument recognition” or “instrument identification” might indicate several different research topics. For instance, many of the related works focus on studio-recorded isolated notes. To name a few, Eronen used cepstral coefficients andtemporal features to classify 30 orchestral instruments with several articulation styles and achieved a classification accuracy of 95% for instrument family level and about 81% for individual instruments. Diment et al. used a modified group delay feature that incorporates phase information together with mel-frequency cepstral coefficients (MFCCs) and achieved a classification accuracy of about 71% for 22 instruments.

Sparse coding on cepstrum with temporal sum pooling and achieved an F-measure of about 96% for classifying 50 instruments. They also reported their classification result on a multi-source database, which was about 66%. Line spectral frequencies (LSF) with a Gaussian mixture model (GMM) and achieved an accuracy of about 77% for instrument family and 84% for 14 individual instruments.

More recent works deal with polyphonic sound, which is closer to real-world music compare to monophonic sound. In the case of polyphonic sound, a number of research studies used synthesized polyphonic audio from studio-recorded single tones. A non-negative matrix factorization (NMF)-based source-filter model with MFCCs and GMM for synthesized polyphonic sound and achieved a recognition rate of 59% for six polyphonic notes randomly generated from 19 instruments. Various spectral, temporal, and modulation features with PCA and linear discriminant analysis (LDA) for classification. They reported that, using feature weighting and musical context, recognition rates were about 84% for a duo,

78% for a trio, and 72% for a quartet. The uniform discrete cepstrum (UDC) and mel-scale UDC (MUDC) as a spectral representation with a radial basis function (RBF) kernel support vector machine (SVM) to classify 13 types of Western instruments. The classification accuracy of randomly mixed chords of two and six polyphonic notes, generated using isolated note samples from the RWC musical instrument sound database, was around 37% for two polyphony notes and 25% for six polyphony notes. As shown above, most of the previous works focus on the identification of the instrument sounds in clean solo tones or phrases. More recent researches attempt to solve instrument identification in a polyphonic situation, but artificially produced polyphonic music is still far from professionally produced music. Real-world music has many other factors that affect the recognition performance. For instance, it might have a highly different timbre, depending on the genre and style of the performance. In addition, an audio file might differ in quality to a great extent, depending on the recording and production environments.

In section II, we introduce emerging deep neural network techniques in the MIR field. Next, the system architecture Section includes audio pre-processing, the proposed network architecture with detailed training configuration, and an explanation of various activation functions used for the experiment. Section IV, the evaluation section, contains information about the dataset, testing configuration including aggregation strategy, and our evaluation scheme. Then, we illustrate the performance of the proposed ConvNet in Section V, the Results section, with an analysis of the effects of analysis window size, aggregation strategy, activation function, and identification threshold. We also present an instrument-wise analysis and single predominant instrument identification as well as a qualitative analysis based on the visualization of the ConvNet's intermediate outputs to understand how the network captured the pattern from the input data.

II. LITERATURE SURVEY

1. Zhiyao Duan, et al (2014), proposed a method that novel cepstral representation called the uniform discrete cepstrum (UDC) to represent the timbre of sound sources in a sound mixture. Different from ordinary cepstrum and MFCC which have to be calculated from the full magnitude spectrum of a source after source separation, UDC can be calculated directly from isolated spectral points that are likely to belong to the source in the mixture spectrum (e.g., non-overlapping harmonics of a harmonic source). Existing cepstral representations that have this property are discrete cepstrum and regularized discrete cepstrum, however, compared to the proposed UDC, they are not as effective and are more complex to compute. The key advantage of UDC is that it uses a more natural and locally adaptive regularizer to prevent it from over fitting the isolated spectral points.

2. Tomas Mikolov, et al (2010), proposed a new recurrent neural network based language model (RNNLM) with applications to speech recognition is presented. Results indicate that it is possible to obtain around 50% reduction of perplexity by using mixture of several RNN LMs, compared to a state of the art back off language model. Speech recognition experiments show around 18% reduction of word error rate on the Wall Street Journal task when comparing models trained on the same amount of data, and around 5% on the much harder NIST RT05 task, even when the back off model is trained on much more data than the RNN LM. We provide ample empirical evidence to suggest that connectionist language models are superior to standard n-gram techniques, except their high computational (training) complexity.

3. Yann LeCun, et al(2015), introduced a new method, deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels ofabstraction. These methods have dramatically improved the state-of-the-art in speech rec¬ognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the back propagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolution nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shown light on sequential data such as text and speech.

4. Grégoire Mesnil, et al (2013), implemented several important recurrent-neural-network architectures, including the Elman-type and Jordan-type recurrent networks and their variants. One of the key problems in spoken language understanding (SLU) is the task of slot filling. In light of the recent success of applying deep neural network technologies in domain detection and intent identification, we carried out an in-depth investigation on the use of recurrent neural networks for the more difficult task of slot filling involving sequence discrimination. To make the results easy to reproduce and compare, we implemented these networks on the common Theano neural network toolkit, and evaluated them on the ATIS benchmark. We also compared our results to a conditional random fields (CRF) baseline. Our results

show that on this task, both types of recurrent networks outperform the CRF baseline substantially, and a bi-directional Jordan-type network that takes into account both past and future dependencies among slots works best, outperforming a CRF-based baseline by 14% in relative error reduction.

5. Tetsuro Kitahara, et al (2007), designed a new solution to the problem of feature variations caused by the overlapping of sounds in instrument identification in polyphonic music. When multiple instruments simultaneously play partials (harmonic components) of their sounds overlap and interfere, which makes the acoustic features different from those of monophonic sounds? To cope with this, we weight features based on how much they are affected by overlapping. First, we quantitatively evaluate the influence of overlapping on each feature as the ratio of the within-class variance to the between-class variance in the distribution of training data obtained from polyphonic sounds. Then, we generate feature axes using a weighted mixture that minimizes the influence via linear discriminant analysis. In addition, we improve instrument identification using musical context. Experimental results showed that the recognition rates using both feature weighting and musical context were 84.1% for duo, 77.6% for trio, and 72.3% for quartet; those without using either were 53.4, 49.6, and 46.5%, respectively.

6. Alex Krizhevsky, et al (2012), implemented a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce over-fitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

7. Jan Schlüter and Sebastian Böck (2014) proposed a methodology for Musical onset detection. It is one of the most elementary tasks in music analysis, but still only solved imperfectly for polyphonic music signals. Interpreted as a computer vision problem in spectrograms, Convolutional Neural Networks (CNNs) seem to be an ideal fit. On a dataset of about 100 minutes of music with 26k annotated onsets, we show that CNNs outperform the previous state-of-the-art while requiring less manual pre-processing. Investigating their inner workings, we find two key advantages over hand-designed methods: Using separate detectors for percussive and harmonic onsets, and combining results from many minor variations of the same scheme. The results suggest that even for well-understood signal processing tasks, machine learning can be superior to knowledge engineering.

III. BACKGROUND METHODOLOGY

The ability of traditional machine learning approaches was limited in terms of processing input data in their raw form. Hence, usually the input for the learning system, typically a classifier, has to be a hand-crafted feature representation, which requires extensive domain knowledge and a careful engineering process. There are many variants and modified architectures of deep learning, depending on the target task. Especially, recurrent neural networks and ConvNets have recently shown remarkable results for various multimedia information retrieval tasks.

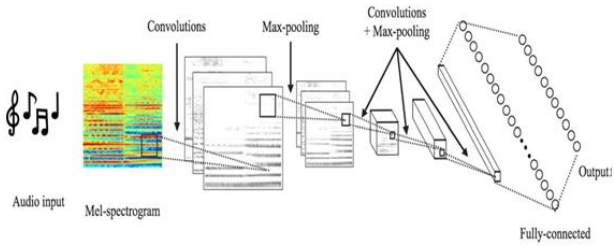

Fig.1 Schematic of the Existing ConvNet

A. Audio Pre-processing

The convolutional neural network is one of the representation learning methods that allow a machine to be fed with raw data and to automatically discover the representations needed for classification or detection. Although it aims to learn a feature representation "automatically", appropriate pre-processing of input data is a crucial factor generating a good feature representation.

In the first pre-processing step, normally a stereo input audio is converted to mono by taking the mean of the left and right channels, and then it is down sampled to 22,050 Hz from the original 44,100 Hz of sampling frequency. This allows us to use frequencies up to 11,025 Hz, the Nyquist

frequency, which is sufficient to cover most of the harmonics generated by musical instruments while removing noises possibly included in the frequencies above this range. Secondly, all audios are normalized by dividing the time-domain signal with its maximum value, and then it is converted to a time-frequency representation using short-time Fourier transform (STFT) with 1024 samples for the window size (approx.46 ms) and 512 samples of the hop size (approx. 23 ms). Next, the linear frequency scale-obtained spectrogram is converted to a mel-scale. We use 128 for the number of mel frequency bins, following the representation learning papers on music annotation by Nam et al. and Hamel et al., which is a reasonable setting that sufficiently preserves the harmonic characteristics of the music while greatly reducing the dimensionality of the input data. Finally, the magnitude of the obtained mel-frequency spectrogram is compressed with a natural logarithm.

From the figure 1, shows that the schematic of the existing ConvNet containing 4 times repeated double convolution layers followed by max-pooling. The last max-pooling layer performs global max-pooling, and then it is fed to a fully connected layer followed by 11 sigmoid outputs.

B. Network Architecture

ConvNets can be seen as a combination of feature extractor and the classifier. The existing ConvNet architecture is inspired by a popular AlexNet and VGGNet structure, which are very deep architecture using repeated several convolution layers followed by max-pooling, as shown in Fig. 1. This method of using smaller receptive window size and smaller stride for ConvNet is becoming highly common especially in the computer vision field such as in the study from Zeiler and Fergus and Sermanet et al. which has shown superior performance in ILSVRC-2013.

C. Training Configuration

The training was done by optimizing the categorical dataset. The training was regularized using dropout with a rate of 0.25 after each max-pooling layer. Dropout is a technique that prevents the over-fitting of units to the training data by randomly dropping some units from the neural network during the training phase. Dropout rate after a fully connected layer was set to 0.5 because a fully connected layer easily suffers from over-fitting. The initialization of the network weights is another important issue as it can lead to an unstable learning process, especially for a very deep network. We used a uniform distribution with zero biases for both convolution and fully connected layers following Glorot and Bengio.

D. Activation Function

The activation function is followed by each convolutional layer and fully connected layer. Several activation functions are used in the Literature. The traditional way to model the activation of a neuron is by using a hyperbolic tangent (tanh) or sigmoid function. However, non-saturating nonlinearities such as the rectified linear unit (ReLU) allow much faster learning than these saturating nonlinearities, particularly for models that are trained on large datasets. A number of recent works have shown that the performance of ReLU is better than that of sigmoid and tanh activation.

IV. PROPOSED SYSTEM

In this proposed method the input audio signal is framed into several window oh different size(where size units into seconds).Fourier transform is used.where the time signal is converted into frequency signal.In the proposed method the audio signal is framed.From the frame the spectrum information is obtained using the fourier transform. Corresponding to Jazz,pop and rock is given to the neural network algorithm.

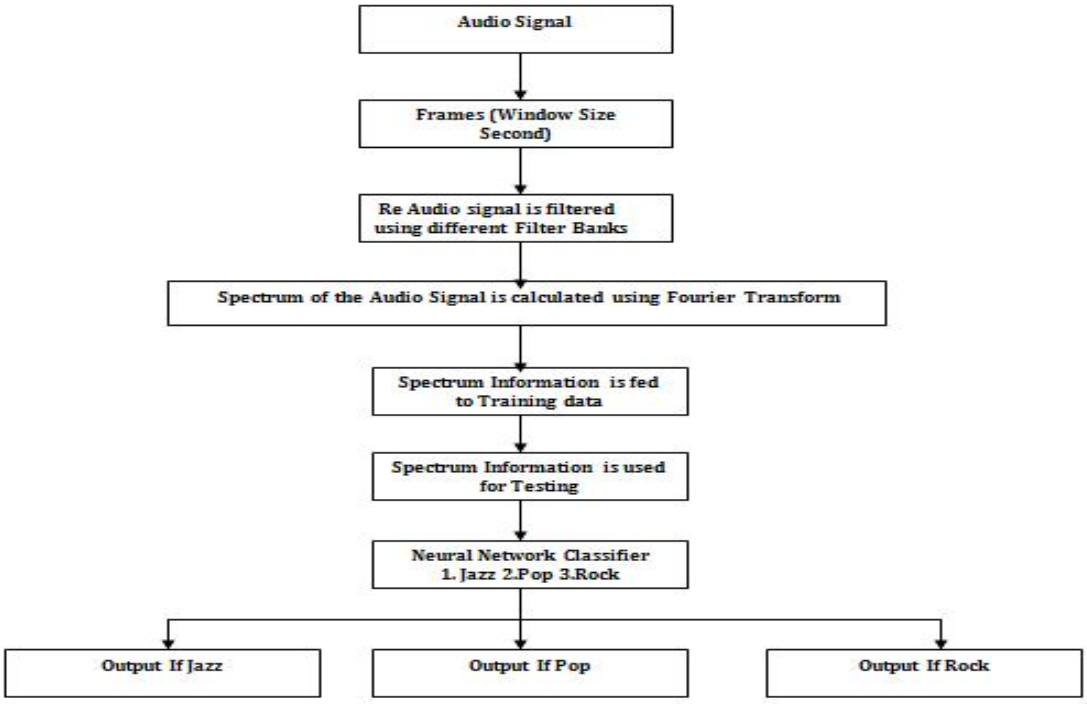

Fig.2 Flow chart of the proposed method

Sound is normally categorize into low ,medium, and high frequency components. ex: drum is 120Hz(low).Flute is 900Hz(high) .Keyboard has all frequencies (low,high,medium).Thus work develops a software to categorize the above sounds based on engineering concepts.(ex.winAMP).But in winAMP we use only to play song and hear the song in different tone effect. This paper address the how a sound is choosen based on the listener choose and song type. This software can be used for mining purpose where specific category can be picked from a large collection of sound database.

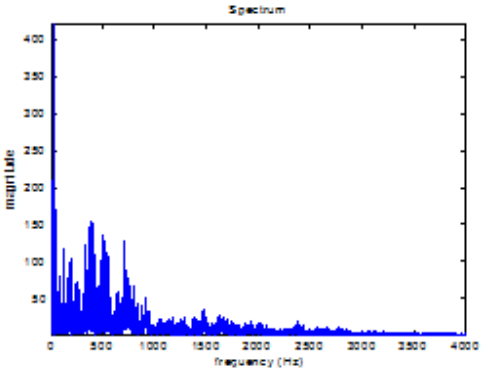

Figure 4. Audio signal spectrum

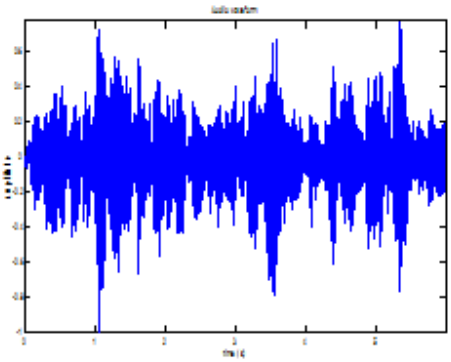

Figure 3. Plotting the spectrum file

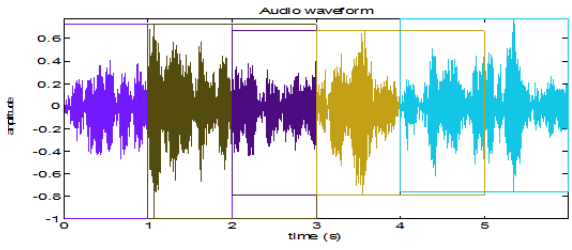

Figure 5. Framing the audio signal

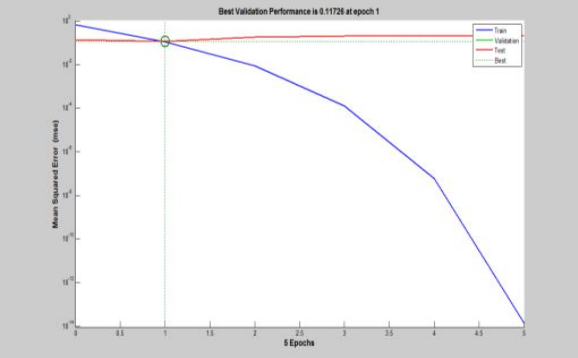

Figure.6 Mean Square Error for 10 neurons

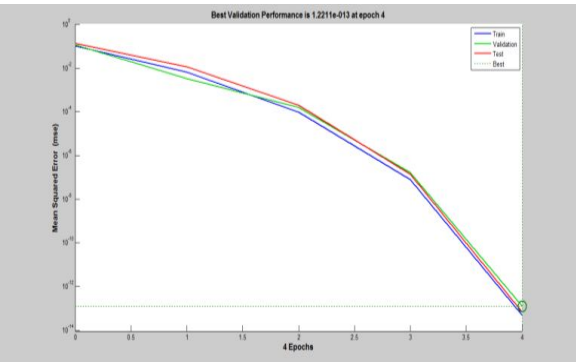

Figure.7 Mean Square Error for 20 Neurons

Table 1: Classification of Audio signal using Neural network

Data Number	Manual Results			Classifier Result
	Jazz	Pop	Rock	
101	✓			100
102		✓		010
103	✓			100
104		✓		010
105	✓			100
106	✓			010
107		✓		100
108			✓	001
109			✓	001

V. CONCLUSION AND FUTURE WORK

The paper presents an analysis and differentiation of music for predominant instrument recognition .The music data is framed and converted to spectrum.The spectrum information is given to neural network which classifies the category .The aim of the work is to develop a musical analyser using software engineering approach .The software developed is In matlab. For the development neural network is proposed. In future the work will be extended to embed the software developed into hardware.

REFERENCES

[1] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in Proc. 2000 IEEE Int. Conf. Acoust., Speech Signal Process., 2000, vol. 2, pp. II753–II756.
[2] A. Diment, P. Rajan, T. Heittola, and T. Virtanen, "Modified group delay feature for musical instrument recognition," in Proc. 10th Int. Symp. Comput. Music Multidiscip. Res., Marseille, France, 2013, pp. 431–438.
[3] L.-F.Yu, L. Su, andY.-H.Yang, "Sparse cepstral codes and power scale for instrument identification," in Proc. 2014 IEEE Int. Conf. Acoust., Speech Signal Process., 2014, pp. 7460–7464.
[4] A. Krishna and T. V. Sreenivas, "Music instrument recognition: From isolated notes to solo phrases," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2004, vol. 4, pp. iv-265–iv-268.
[5] S. Essid, G. Richard, and B. David, "Musical instrument recognition on solo performances," in Proc. 2004 12th Eur. Signal Process. Conf., 2004, pp. 1289–1292.
[6] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in Proc. Int. Soc. Music Inf. Retrieval Conf., 2009, pp. 327–332.
[7] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," EURASIP J. Appl. Signal Process., vol. 2007, no. 1, pp. 155–155, 2007.
[8] Z. Duan, B. Pardo, and L. Daudet, "A novel cepstral representation for timbre modeling of sound sources in polyphonic mixtures," in Proc. 2014 IEEE Int. Conf. Acoust., Speech Signal Process, 2014, pp. 7495–7499.