Recognition of Musical Instruments

Harya Wicaksana, Septian Hartono, & Foo Say Wei

School of Electrical & Electronic Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
Email: {hary0001, sept0001, eswfoo}@ntu.edu.sg

***Abstract*—In this paper an automated method to recognize the musical instruments playing the musical signals is presented. Various features of the musical instruments and musical signals are investigated. The features can broadly be grouped into three categories: temporal, spectral, and cepstral features. A composite neural network structure is proposed as the classifier. The performance of the composite neural network using a set of carefully chosen features is compared with that of the traditional neural network. Experimental results show that the accuracy achieved using composite structure (94%) is significantly higher than that using the traditional structure (88%) when more than four musical instruments are to be distinguished.**

***Keywords*—musical instrument, recognition, neural network**

I. INTRODUCTION

In automatic transcription of musical pieces, the first step is to identify the musical instrument used as the characteristics of the notes are different for musical signals from different musical instruments. [1].

Various algorithms have been explored in detecting the musical instrument from the notes played. De Poli and Prandoni used mel-frequency cepstrum coefficients calculated from isolated tones as an input to a Kohonen self-organizing map, in order to construct timbre spaces [2]. The neural network used was the Kohonen Self-Organizing Map (SOM), which is used more for separation and clustering than for recognition. Langmead [3, 4] also used a self organizing neural network but his focus was more on creating timbre categories.

In this paper, several features of musical signal are explored together with neural network classifiers of different structures to assess the best approaches to identify musical instruments.

The organization of the paper is as follows. In the next section, several features that are used as discriminating variables are described. The structures of the neural networks adopted for the recognition system are discussed in Section III. Results of experiments are summarized in Section IV with concluding remarks presented in Section V.

II. FEATURES

The features extracted from the musical signals may be classified into three different groups: temporal, spectral, and cepstral features. In addition, according to the ways the features are extracted, they may further be classified as primary and secondary features. Primary features are features obtained directly from the musical signal, while secondary features are derived from primary features, which are usually represented by binary or integer values.

A. Temporal Features

The waveforms of a single note played by six different musical instruments are presented in Figure 1. It can be observed that there are significant differences in the shape and duration of the waveforms. As such, features extracted from the waveforms are potential differentiators. Temporal features are features obtained directly from the time-domain musical signals. The following temporal features are explored.

Rise Time: The rise time is taken as the time difference between the time at the end of attack and the backtracked position where the magnitude is 25% of the magnitude at the end of attack.

Decay Time: The decay time is obtained as the time difference between the end of attack and the forward position where the magnitude is 25% of the magnitude at the end of attack.

Sustain: Sustain is a measure of the length of time that the sound of a note lasts. If the decay time is larger than 1.5 seconds, the sustain value is taken as 1; otherwise, it is taken as 0. Note that Sustain belongs to the class of secondary feature.

Vibrato: When the number of peaks in the waveform of a note is more than 5, it is assessed that strong vibration is present. A signal with more than 5 peaks is assigned a Vibrato value 1; otherwise the value of Vibrato is set at 0.

Sharp Rise: The next 2 features indicate the sharpness of the peak at the end-of-attack. The value of sharp rise is evaluated with three frames before the end-of-attack using the following expression:

$$s_r = \lfloor \frac{8\mathrm{MS}[p]}{3\sum_{n=0}^{3}\mathrm{MS}[p-n]} \rfloor$$

where p is the value of the frame at the end-of-attack.

1-4244-0387-1/06/$20.00 ©2006 IEEE

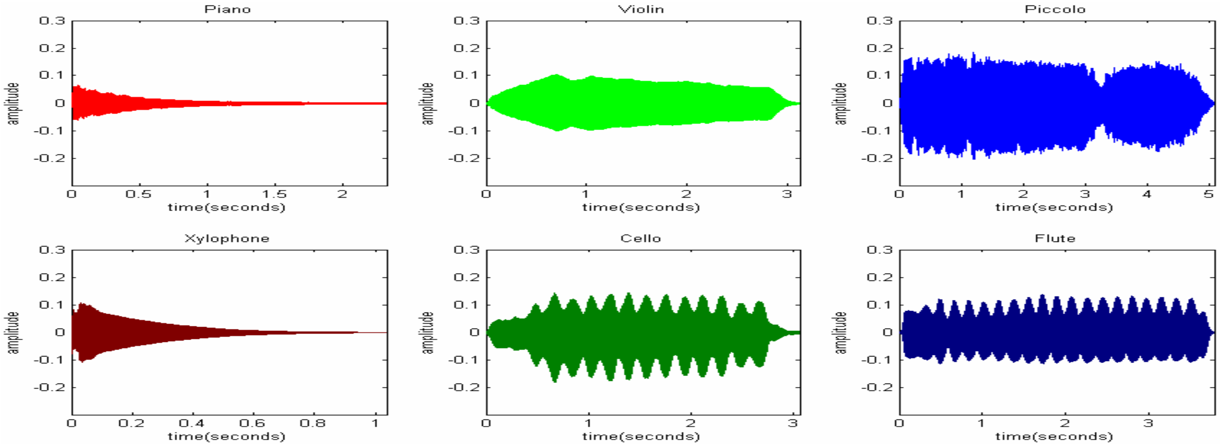

Figure 1. Time-domain signal of 6 musical instruments

Sharp Decay: Likewise, the value of sharp decay is obtained in similar way. However, instead of taking the three frames preceding the frame at the end-of-attack, the three frames after the frame at the end-of-attack are used.

$$s_r = \lfloor \frac{4\mathrm{MS}[p]}{\sum_{n=0}^{3} \mathrm{MS}[p+n]} \rfloor$$

Time from end of attack to time with maximum amplitude: This feature is taken as the time interval between the end of attack of a note and the position of the maximum value, if the maximum value is different from the value at the end of attack.

.Maximum normalized slope: This feature is computed as the maximum value of normalized slope from the beginning of attack to the end of attack. Mathematically, it is given by:

$$\max_{0 \leq n < p} \left(\frac{\mathrm{MS}[n+1] - \mathrm{MS}[n]}{\mathrm{MS}[p]} \right)$$

B. *Spectral Features*

Spectral features are obtained from the samples in the frequency domain of the musical signal. Since the frequency distribution of musical signal is geometrical, constant Q transform [5, 6] is used.

The constant Q transform of a time-domain signal x[n], X[k], is given by:

$$X[k] = \sum_{n=0}^{N[k]-1} W[k,n] x[n] e^{-j2\pi Qn/N[k]}$$

where W[k,n] is a Hamming window and 2πk/N[k] is the digital frequency. In this paper, Q is set at 34(half-note range). The CQT of the note A4 for six types of musical instruments are presented in Figure 2. It can be observed that there are differences in the CQT spectra that can be used for discrimination of musical instruments.

The following spectral features are adopted.

Compressed harmonic ratios: The first 5 harmonics of the note are identified from the CQT. The Compressed Harmonic Ratios (CHR) are computed as the logarithmic values of the ratios of the spectral values of the harmonics to the spectral value of the fundamental frequency.

Normalized spectral centroid: Spectral centroid is obtained by first taking the average of the products of the spectral magnitude and frequency of all spectral components of the CQT starting from one octave before the fundamental frequency f_f to one octave after the fifth harmonic. The average value is then divided by f_f to obtain the normalized spectral centroid.

Odd-even harmonics ratio (Boolean): It is the ratio of the sum of the spectral values of the first and third harmonics to the sum of spectral values of the second and fourth harmonics.

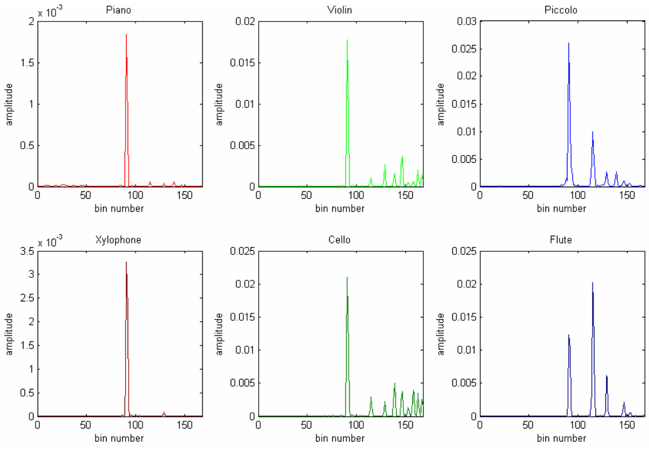

Figure 2. CQT of A4 note of 6 Musical Instruments

C. Cepstral Features

Cepstral coefficients are used widely in musical recognition system and considered as the most important features [7]. In this paper, fifteen cepstral coefficients and the mean of these fifteen cepstral coefficients are also used as a discriminating feature for the recognition system. The cepstral coefficients of a single note played by six different musical instruments are given in Figure 3. It can be seen that some differences exist that can be explored for discriminating the musical instruments.

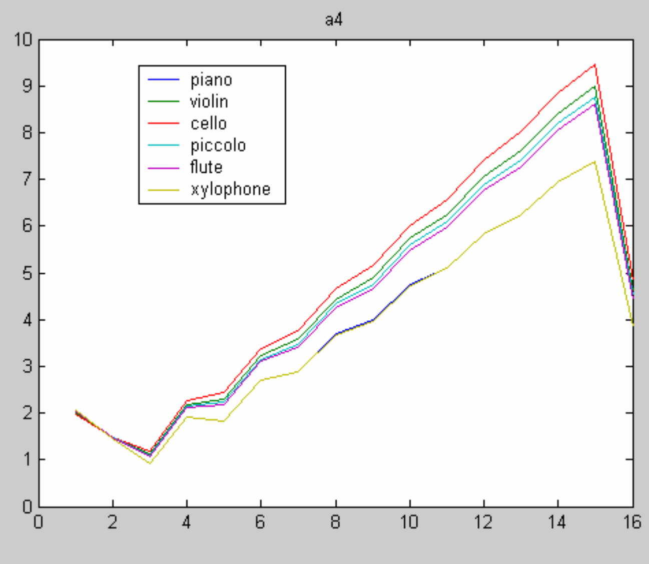

Figure 3. Cepstral Coefficients of 6 Musical Instruments

III. Classifiers

The feed forward neural network model is used in the network design. The network consists of 3 layers. The first layer is the input layer. The number of neurons in the input layer is determined by the number of features selected. For the experiments reported in this paper, 30 input neurons are used. The 2nd layer is the hidden layer. The number of the neurons in the hidden layer is adjusted according to the complexity of the network i.e. the number of neurons in the input and output layer. The 3rd layer is the output layer. The number of neurons in the output layer depends on the number of instruments to be recognized.

Each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiplied by a separate weight value. The weighted inputs are summed, and passed through a limiting function which scales the output to a fixed range of values. Then, based on the values obtained in the output neurons, the type of the instrument can be determined.

The values of the weights between neurons will determine the real distinctiveness or 'intelligence' of the network. Back propagation algorithm is adopted as a method of adjusting the weights between neurons.

A Back propagation network learns by example. Some input samples with the known-correct output for each sample are necessary for training purposes. From the given input samples, the network will produces some output based on the current state of it's synaptic weights (initially, the output will be random). This output is compared to the known-good output, and a mean-squared error signal is calculated. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are calculated to reduce the error signal. This procedure is repeated until the network achieves some pre-determined threshold.

A. Single Network

The traditional structure is the single network system. For this structure, the number of output neurons is equal to the total number of instrument to be discriminated.

B. Composite Network

Instead of using the traditional structure, a composite structure is proposed. For this composite structure, two instruments are distinguished at a time by a simple neural network with two output neurons while the number of input neurons remains the same. For n instruments to be identified, the composite network consists of nC2 networks with 2 output neurons. For example, for 4 instruments: piano (*pia*), violin (*vio*), cello (*cel*), and piccolo(*pic*), six comparisons are required using the *pia-vio, pia-cel, pia-pic, vio-cel, vio-pic,* and *cel-pic* networks, as shown in Figure 4. The output values of each pair-wise comparison network ranges from 0 to 1. The sum of the output values of the same instrument is computed. Instrument which has the highest total score from the nC2 networks is taken as the correct instrument for the input musical data.

By focusing only on two instruments at a time, the training error for each network is lower, Another advantage of using the composite network is in the scalability of the system. If it is required to add another instrument, there is no need to carry out the complete training of the entire system. For example if a new instrument is introduced to the system, only new N networks are required to be trained to differentiate the newcomer from the existing N instruments. To reduce the number of instrument, it can be done by removing the network that contains the unwanted instrument

IV. Experimental Results

Experiments were carried out using samples of notes obtained from MUMS CD from McGill University. The instruments selected for the experiments are piano, violin, cello, piccolo, flute and xylophone. 25 samples of notes in similar range are used for training the networks and the remaining samples available from the database, not necessarily in the same range, are used for testing.

As a comparison of the performance of the single network and the composite network, experiments were carried out for both networks for 4 and 6 musical instruments. It is found that 100% accuracy can be achieved using the training data for testing for both 4 and 6 instruments and for both composite and simple networks.

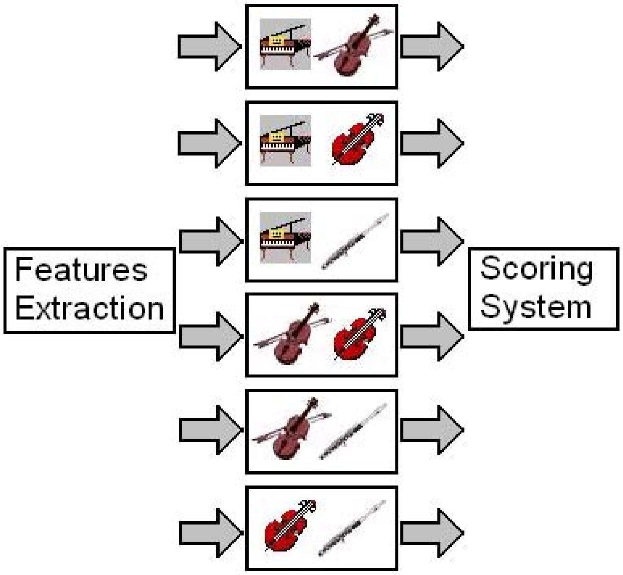

Figure 4. Composite Network System

When testing data were used to assess the performance of the classifiers, less than perfect scores are obtained for both simple and composite networks. The results are summarized in Table I. The performance of the single network is slightly better when only 4 musical instruments are to be distinguished. However, the accuracy achieved by the composite network is significantly better for identifying 6 musical instruments. It is extrapolated that when more instruments are to be identified, composite network will outperform the single network. Apparently, the choice of feature vectors is also an important factor in the performance of the system.

TABLE I. RESULT FROM TEST DATA

Network type	**Number of instruments**	**Correct**	**Incorrect**	**Accuracy**
Composite	4	49	3	94.23%
Single	4	50	2	96.15%
Composite	6	75	5	93.75%
Single	6	70	10	87.50%

V. CONCLUSION

In this paper, a new approach to the recognition of musical instruments is proposed. Experiments were carried out to select a set of critical features of musical signal for the purpose. In addition to the popular cepstral coefficients, other temporal features and spectral features were chosen.

As for the classifier, a composite neural network structure is proposed. For this structure, two instruments are compared at a time and the scores of the pair-wise comparison networks are summed up to select the instrument that is the most likely candidate that produces the musical signal.

Results show that an accuracy of 93.75% can be achieved using the composite network system in discriminating the six musical instruments compared with 87.50% using the single network system.

ACKNOWLEDGMENT

This work was completed when the first and the second author, NTU President Research Scholars, were undertaking the Undergraduate Research on CAmpus (URECA) program in Nanyang Technological University, Singapore.

REFERENCES

[1] G S. Essid, G. Richard, and D. Bertrand. Musical instrument recognition based on class pairwise feature selection. In Proc. of the 5th ISMIR Conf., 2004.

[2] G., De Poli & P. Prandoni, “Sonological Models for timbre characterization”, Journal of New Music Research 26, pp. 170-197, 1997.

[3] C.J. Langmead, “A Theoretical Model for Timbre Perception based on Morphological Representations of Time-varying spectra” (M.A. Thesis), Dartmouth College, 1995a.

[4] C.J. Langmead, “Sound analysis, comparison and modification based on a perceptual model of timbre”, International Computer Music Conference, 1995b.

[5] J.C. Brown, “Calculation of a constant Q spectral transform”, J. Acoust. Soc. Am. Vol. 89 No. 1(Jan 1991).

[6] J.C. Brown & M.S. Puckette, “An efficient algorithm for the calculation of a constant Q transform”, J. Acoust. Soc. Am. Vol. 92 No. 5(Nov 1992).

[7] A. Eronen & A. Klapuri, “Musical instrument recognition using cepstral coefficients and temporal features”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000.