MUSICAL INSTRUMENT RECOGNITION WITH WAVELET ENVELOPES

PACS: 43.60.Lq

Hacihabiboglu, Huseyin[1,2]; Canagarajah C. Nishan[2]

[1]Sonic Arts Research Centre (SARC)
School of Computer Science
Queen's University Belfast
Belfast, BT7 1NN, UK
Tel: ++44 (0) 28 9027 4761
Fax: ++44 (0) 28 9027 4828
E-mail: h.hacihabiboglu@qub.ac.uk

[2]Digital Music Research Group (DMR)
Dept. of Electrical and Electronic Engineering
University of Bristol
Bristol, BS8 1UB, UK
Tel: ++44 (0) 117 954 5198
Fax: ++44 (0) 117 954 5206
E-mail: n.canagarajah@bris.ac.uk

ABSTRACT

Automatic recognition of instrument type from raw audio data containing monophonic music is a fundamental problem for audio content analysis. There are many methods for the solution of this problem, which use common spectro-temporal properties like cepstral coefficients or spectral envelopes. A new method for instrument recognition utilising short-time amplitude envelopes of wavelet coefficients as feature vectors is presented. The classification engine is a distinctively small multilayer perceptron (MLP) network. A correct classification rate which is comparable to previously reported correct classification rates is attained for a set of three instruments containing flute, clarinet and trumpet.

INTRODUCTION

Timbre was shown to be a multidimensional property of sound [1] describable by both temporal and spectral characteristics. It was shown that the spectral envelopes [2], the cepstral coefficients [3], the spectro-temporal statistics [4], the sub-band energies [5] of the sound signal can be used for instrument recognition and the wavelet coefficients can be used for audio indexing and retrieval [6][7].

Audio content analysis and automatic recognition of musical instruments are frequently carried out using the statistical classification of the special features extracted from a sound signal. The statistical analyses of the feature sets derived from these features reveal information on the type of the instrument being played. Most of the previous research on instrument recognition focuses on *sterile* conditions where, computer synthesized sounds rather than sounds from real-life conditions are employed. This selection of data provides a *clean* set of features for the instruments to be classified. However, this approach is not usually robust enough to classify real sounds.

The feature vectors used in this paper employ the short-time amplitude envelopes of the wavelet coefficients rather than the raw wavelet coefficients or the sub-band energies. Since the discrete wavelet transform (DWT) is a shift variant transform, the wavelet transform of the delayed version of the same sound would give different coefficients. Hence using the raw wavelet coefficients is not desirable. Using the sub-band energy ratios seems to result in an acceptable recognition rate, but loses the temporal dimension of timbre which is essential.

In this work, the three instruments chosen for classification are the E-flat clarinet, the flute and the C trumpet. The reason for such a selection is that, the trumpet and the flute were shown to be undistinguishable by family with statistical analysis (i.e. multidimensional scaling), and the clarinet and the flute are in the same family of instruments (i.e. woodwinds) which are inherently hard to recognize as individual instruments.

Formation of the wavelet envelopes and their relation to the musical instrument sounds will be discussed in the first section. The properties of the *multi-layer perceptron* (MLP) neural network used in instrument classification will be discussed next. Results of a classification task for the three instruments will be provided.

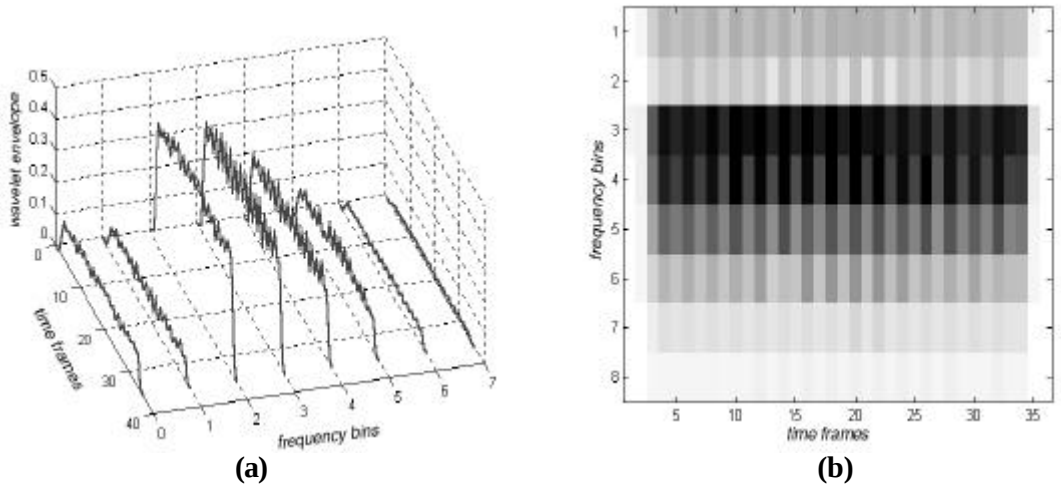
WAVELET ENVELOPES AS AUDIO FEATURES

Discrete wavelet transform (DWT) uses a filterbank structure to obtain a half-band low-pass version and a half-band high-pass version of the signal. The filterbank contains well-defined low-pass and high-pass filters and a subsampling operation after each filter. The low pass version is used as the input to a similar filterbank to get a ¼-band low-pass version, ¼-band high-pass version and a half-band high-pass version. The process is iterated N times to get a N-level DWT. The frequency resolution increases for each iteration while the time resolution decreases [8]. Although application of the wavelet transforms was investigated in detail for many signal-processing applications, research on instrument recognition with wavelets is limited.

The feature vectors used in this research consisted of the *wavelet envelopes*. which were formed using the ratio of the RMS amplitude envelopes of the wavelet coefficients of the leaf nodes of a dyadic wavelet tree to the RMS amplitude of the original signal.

$$F_{i,j} = \sqrt{\sum_{n_j} c_i(n)^2 \Big/ \sum s(n)^2}$$

where $c_i(n)$ is the wavelet coefficients of node i and $s(n)$ is the signal to be analysed and $F_{i,j}$ is the i^{th} element of the feature vector representing the j^{th} frame of discrete wavelet transform. Frame length is chosen as 1024 samples, which corresponds to 46.4 ms. The wavelet used in the derivation of the feature set is the Symmlet-17 wavelet. As the different instruments have different amplitude envelopes, it was necessary to normalise signals for a proper DWT decomposition.



(a) **(b)**

Fig. 1 (a) 3D and (b) 2D, representations of the wavelet envelope for flute playing G4.

Properties of the subjective qualities of these sounds can also be observed from the feature sets (see Fig. 2). Flute sound shows more frequency modulation and the frequency content is spread through frequency bins. Trumpet sound has higher frequency content in the attack portion, but has less frequency modulation than flute. Clarinet sound has steady characteristics and signal energy is probably concentrated in the fist few harmonics of sound. Other than these, similarity of signal amplitude envelope and wavelet envelopes can be examined from the 3D representations of the feature vector sets.

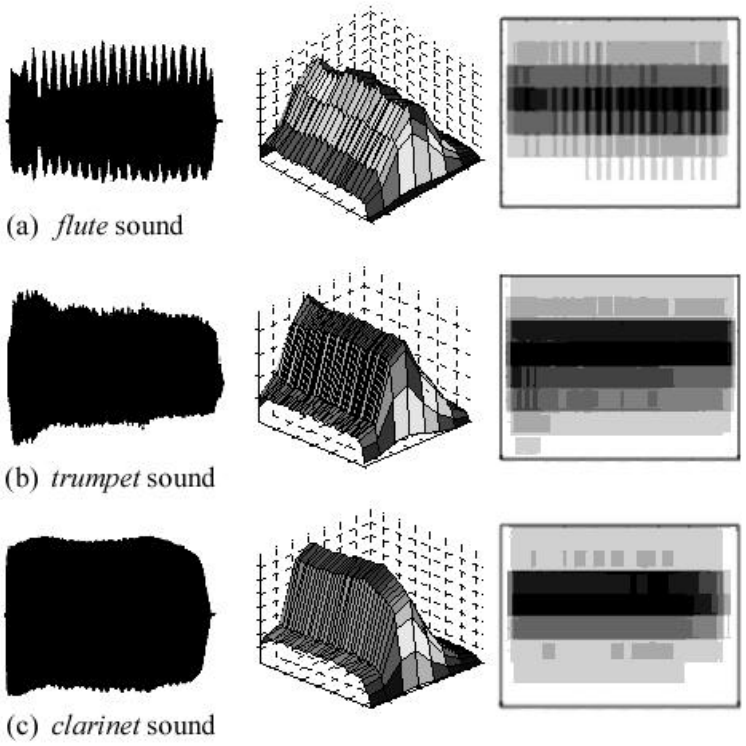


Fig. 2 The waveforms and wavelet envelopes in 3D and 2D representation of (a) flute, (b) C trumpet, and (c) E-flat Clarinet (all playing the note A4).

MLP NEURAL NETWORK CLASSIFIER

The MLP used in the automatic instrument recognition task contains one hidden layer, with 8 neurons and an output layer with 3 neurons (see Fig. 3). Feature vectors for recognition are formed using 8 element vectors derived from the wavelet envelopes of audio signals columnwise.

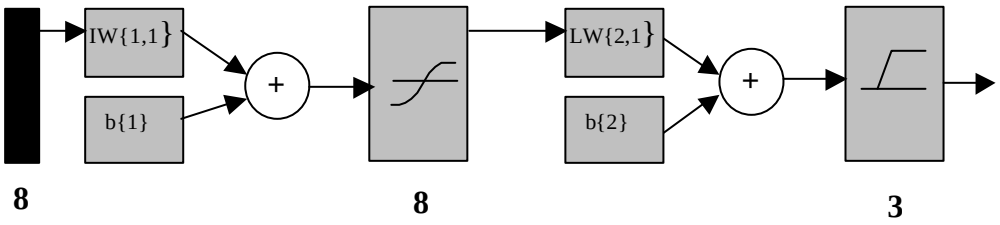


Fig. 3 Structure of the MLP used for instrument recognition. IW, LW and b are input weight, layer weight and bias respectively.

The MLP has 88 connections, which is significantly low compared to the other neural network structures previously proposed for similar purposes. Through extensive simulations, the activation function for neurons in the hidden layer was selected to be *tangent sigmoid* and *saturating linear activation function* was selected for the output layer.

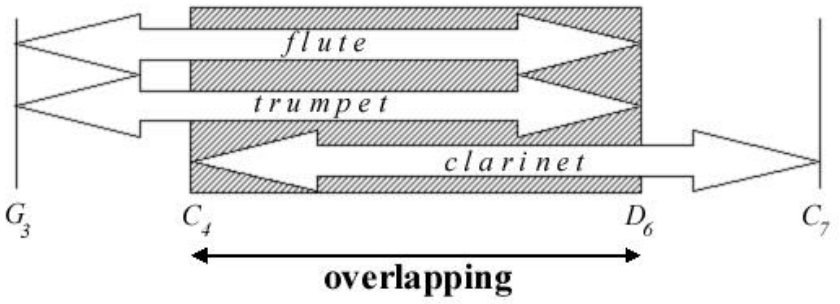


Fig. 4 Relative playing ranges of the instruments.

The MLP was trained with *Levenberg-Marquardt* (LM) method of backpropagation [9] with a training data set consisting of feature vectors extracted from isolated instrument sounds. These sounds contained full playing ranges of the flute, the C trumpet and the E-flat clarinet. The training data set was obtained from the McGill MUMS CD Vol.1 [10]. The flute data set contained all sounds from C_4 to C_7. The clarinet and trumpet data sets contained all sounds from G_3 to D_6. This selection of instruments gave an overlapping range of notes. (See Fig. 4). All acoustical artefacts related to recording were assumed to be eliminated beforehand as these sample sounds were recorded in a recording studio. This provided a reliable data set for training the MLP. Table 1 contains the statistical properties of training set.

Instrument	Number of Sounds	mean length (s)	STD of length (s)
Flute	37 (C_4-C_7)	4.24	0.46
Trumpet	32 (G_3-D_6)	6.70	1.22
Clarinet	32 (G_3-D_6)	4.59	0.80

Table 1 Statistical properties of the sounds in the training set.

Since the training set was highly redundant, training was made using the sequential mode rather than the batch mode. Targets were determined in such a way that only one output should be '1' at a time. The goal for mean square error was set to 0.02 to prevent the neural network from overfitting the training set. The MLP attained the goal in 53 epochs. This number is quite low when compared to number of epochs needed to train a similar network using gradient descent methods. Training with gradient-descent methods requires number of epochs in the order of thousands for generalization.

RESULTS AND DISCUSSIONS

Special attention was given to selection of different types of recordings of each instrument while choosing the test data set. Test set contained recordings of professional and amateur players, highly improvised jazz pieces and synthesised sounds of instruments, sounds with much background noise and professional recordings. This selection is believed to make the results more reliable and more suitable for real-life conditions.

Each feature vector represents only a 46.4 ms long portion of a sound. Therefore, a long sound is represented by a large number of feature vectors while a short sound is represented by a small number of feature vectors. Table 2 gives details about the durations of the sounds used in the test phase for the MLP neural network.

Instrument	Number of Sounds	Average Length (s)	STD of Length(s)
Flute	19	6.63	5.49
Trumpet	18	11.79	10.63
Clarinet	10	7.38	6.45

Table 2 *Statistical properties of sounds in the test set.*

It was observed that the MLP performs quite robustly for highly improvised sound signals (i.e. in jazz recordings) and non-standard playing techniques (i.e. in amateur performances), which were not included in the training phase. The MLP misclassified 31.8% of flute sounds, 11.1% of trumpet sounds, and 20.0% of clarinet sounds. The overall correct classification rate is 78.72% for recognition of one in three instruments.

CONCLUSIONS AND FUTURE WORK

A new set of audio features for monophonic instrument recognition was proposed in this paper. The correct classification rates achieved by a simple MLP are promising. The simplicity of the proposed system makes it viable for real-time applications.

Reduction of the effect of temporal properties in classification of the instruments is the disadvantage of using an MLP network as a classifier. Temporal properties of the feature set may be exploited much better using time-dependent neural networks such as hidden Markov models (HMMs) or time-delay neural networks (TDNNs). Shift-invariant wavelet basis [11] or matching pursuit [12] may also be suitable for extracting spectro-temporal information from sound data.

BIBLIOGRAPHICAL REFERENCES

[1] J. M. Grey, *"Multidimensional perceptual scaling of Musical Timbres"*, J. Acoust. Soc. Am., Vol. 61, No. 5, 1977.

[2] A. T. Cemgil and F. Gurgen *"Classification of Musical Instrument Sounds Using Neural Networks"*, Proc. of SIU'97, Istanbul, Turkey, 1997.

[3] J. Brown, *"Computer Identification of Musical Instruments using Pattern Recognition with cepstral coefficients as features"*, J. Acoust. Soc. Am. , Vol. 105, No. 3, 1999.

[4] K. Martin, *"Toward Automatic Sound Source Recognition: Identifying Musical Instruments"*, NATO Comp. Hear. Adv. St. Inst., Il Ciocco, Italy, July 1998.

[5] H. Hacihabiboglu and N. Canagarajah, *"Instrument Based Wavelet Packet Trees in Audio Feature Extraction"*, Proc. of International Symposium on Musical Acoustics (ISMA'01), Perugia, Italy, 2001.

[6] R. Subramanya and A. Youssef, *"Wavelet-based Indexing of Audio Data in Audio/multimedia Databases"*, Proc. of the International Workshop on Multimedia Database Management Systems, 1998.

[7] G.Li, and A. A. Khokhar, *"Content-Based Indexing and Retrieval of Audio Data usingWavelets"*, IEEE International Conference on Multimedia and Expo (II) 2000.

[8] M. Vetterli, and J. Kovacevic, *"Wavelets and Subband Coding"*, *Prentice-Hall*, 1995.

[9] M. Hagan and M. B. Menhaj, *"Training feedforward networks with the Marquardt Algorithm"*, IEEE Trans. on Neural Networks, Vol. 5, No. 6, 1994.

[10] McGill University Master Samples
WWW Page: http://www.music.mcgill.ca/resources/mums/html/mums.html

[11] I. Cohen, S. Raz, , and D. Malah, *"Shift Invariant Wavelet Packet Bases"*, Proc. 20th ICAASP (IEEE International Conference on Acoustics, Speech, and Signal Processing), 1995.

[12] S. Mallat, "*A Wavelet Tour of Signal Processing*", Academic Press, 1998.