TIMBRE RECOGNITION OF SINGLE NOTES USING AN ARTMAP NEURAL NETWORK

D.K. Fragoulis, J.N. Avaritsiotis and C.N. Papaodysseus

National Technical University of Athens
Department of Electrical and Computer Engineering
9 Heroon Polytechniou St., 157 73 Zographou, Athens, Greece

ABSTRACT

In this paper, a model for the perception of musical instrument timbre is presented. The model uses an ARTMAP neural network to distinguish single notes played by five different instruments. The duration of each note is quite sort. The recognition of timbre is based on three acoustic properties: spectral synchrony, slope of the attacks and spectral distribution. Arrays of values of the above properties are used as input patterns. By training the network with a large number of different input patterns a robust pattern recognizer for timbre identification is constructed. The choice of this specific type of neural network model provides the ability for creating timbre categories which can continuously being updated at any point of its operation, while at the same time, knowledge of previously learned categories is retained.

1. INTRODUCTION

Our ability to identify things by listening is based on the information provided by their acoustic properties, and these properties are the result of the production process. We can say generally that acoustic properties belong to and at the same time characterize the source. Moreover, these properties evolve over time. Typically, the changes are slow, continuous, and regular so that it is possible to track a sound over time. The recognition of timbre is a complex procedure based on the exploitation of the information that is hidden inside the acoustic properties. However it is reduced to a classic pattern recognition procedure by evaluating the acoustic properties and assign the values of these properties to timbre patterns. The evaluation of the acoustic properties is succeeded by introducing a qualitative description of them. The performance of the recognition process is enhanced if information derived by more acoustic properties is used. An important observation about sounds is that they can combine together to create mixed sounds.

So we have to distinguish sound properties between local and emergent which are generated by the interaction of the local properties. Since in this work we are examining single notes, we do not consider the problem of sound combination.

2. THE NOTION OF TIMBRE

The term timbre refers to the perceptual qualities of objects and events. We use this term to express what an acoustic event sounds like. Timbre has been thought of as related to one acoustically measurable property such that each note of an instrument or each spoken sound of one voice would be characterized by a single value of that property. The traditional definition of timbre is by exclusion: The quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar. Although this definition may tell us what timbre is not, it does not tell us what timber is.

Due to the interactive nature of sound production, there are many stable and time varying acoustic properties. It is unlikely that any one property uniquely determines the pitch. The sense of timbre comes from the interaction between the properties of the vibration pattern. Therefore the timbre can be perceived in terms of a set of acoustic properties. Thus, the acoustic properties are used to find out what event was most likely to have produced that sound. The connection between these properties and the object is learned by experience.

One of the acoustic properties that listeners use to identify events is spectral shape. It can be characterized by using the relative amplitudes of the partials to derive various statistical measures. The main statistic measures that are used for this purpose follow next: (a) the central tendency, (b) the overall power obtained by summing the squared amplitude of each component, (c) the power spectrum across a set of frequency regions, (d) the variance of the amplitudes. Even if a very simple sound can be specified by its frequency spectrum, this is not the case in complex sounds. If we consider that the spectrum changes from note to note for each musical instrument, then it is clear that timbre identification can not be based on the particular spectral shape of a note.

Another acoustic property is the onset and offset of individual harmonics of a note. For “string” instruments, the differences in the attack and decay of harmonics are due to the variations in the method of excitation. For “wind” instruments, are due to the feedback between the vibration modes of the mouthpiece and sound body. Several qualitative descriptions have been utilized to describe the aforementioned acoustic property. The most common

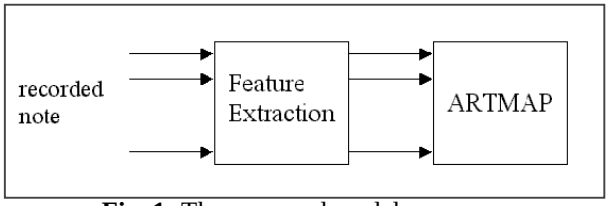

Fig. 1: The proposed model

description is based on the estimation of the degree of synchrony of the attack transients. The information provided by the transients can be used to distinguish among simple-sounds. But for a given instrument the pattern of the transients will vary across notes.

A third acoustic property is noise. Noise can occur when excitation energy is first applied to the source. For example, bowing a violin creates an initial high frequency sound before the bowing stabilizes. Noise can also be continuous. In this case, the breathy sound of a flute created by blowing across the mouthpiece and the hissing of a pipe, can be mentioned as characteristic examples.

Another very important acoustic property concerns the transition between two successive sounds. During the transition, the decay and the attack of the two sounds often overlap. The overlap can hinder the procedure of timbre identification by masking the transient or it can improve that procedure by creating a unique acoustic pattern not heard in discrete sounds.

It is clear from the above that the production of sound yields a large number of acoustic properties that can determine timbre. Some of these properties vary across different notes, duration and intensities. Also the performers may vary the sound by means of the excitation technique, intonation and musical emphasis. As a conclusion we can say that no predominant acoustic property determines timbre. Any single acoustic property can provide some level of timbre identification performance, and combinations of properties usually produce better performance than a single one. In this view, to perform timbre recognition it is necessary first to determine a set of acoustic properties and to introduce a qualitative description of these properties [1] [2]. The choice of the specific properties is case-depended. Then several values of the selected properties can be used for the construction of a template that relates timbre patterns with patterns of acoustic properties. Thus, timbre recognition reduces to a pattern recognition procedure.

Neural networks are used very often in pattern recognition tasks when the existence of robust templates is needed. By training a network with a wide variety of patterns that correspond to notes played by several instruments with different duration, intensity and playing technique, the network can learn to be insensible in small timbre differences caused by these factors while at the same time it can be able to recognize different musical instruments.

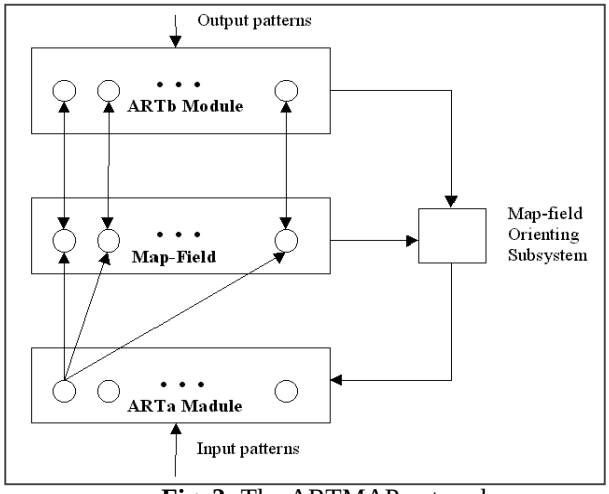

Fig. 2: The ARTMAP network

3. THE PROPOSED MODEL

In this work we propose a model which is able to distinguish notes played by the following musical instruments: piano, guitar, trumpet, saxophone and flute. For convenience several notes played by a specific musician with almost the same duration (1.5 sec) and intensity, have been recorded. The recognition is based on three acoustic properties: (a) the slope of the attack (b) the degree of synchrony of the attack transients (c) the amount of energy in the higher frequencies. In figure 1, a general aspect of the model is illustrated. In the first block the input signal is analyzed and the value of 10 parameters that correspond to the three aforementioned acoustic properties are estimated. The second block is an ARTMAP neural network that takes as input the 10 output values of the feature extraction block and performs the procedure of pattern recognition. The functionality of the ARTMAP network is presented in the below paragraph.

4. THE ART NETWORK

The ART type neural networks are non-supervised, self-organized networks. They differ in topology from most of the other neural networks in having a set of top-down weights as well as a bottom-up one. The main advantage of this type of networks is that are able to learn additional patterns at any time of their operation, while keeping the previous knowledge. An ART1 network can self-organize binary input patterns, while an ART2 can do the same for both binary and analog patterns. An ARTMAP is a supervised neural network that comprises two ART modules, named ART_a and ART_b.

The ARTMAP network used, consists of an ART1 as ART_a and an ART2 as ART_b. Vector $\mathbf{b}_p$ encodes the timbre categories and vector $\mathbf{a}_p$ encodes the information related to the three aforementioned acoustic properties. The basic features of the ARTMAP network are presented below.

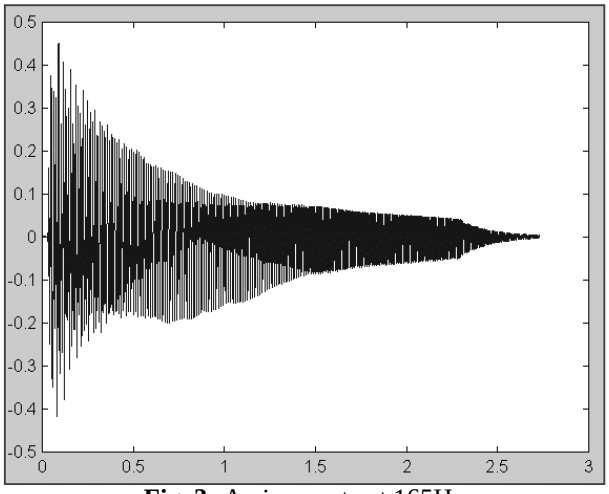

Fig. 3: A piano note at 165Hz

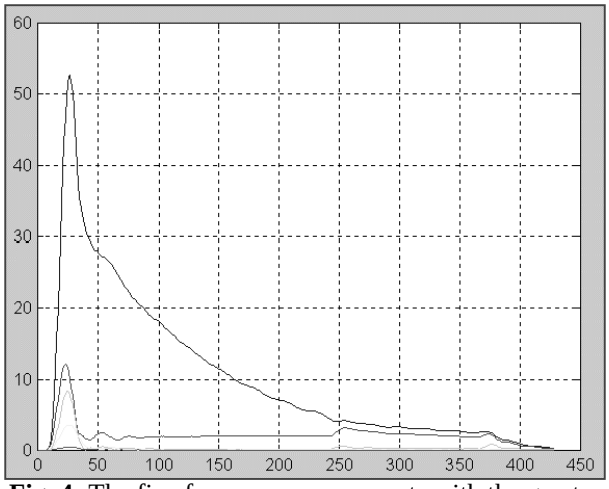

Fig. 4: The five frequency components with the greater amplitude

4.1 The ARTMAP network

As mentioned before, ARTMAP is a self-organized, supervised neural network consisted of two unsupervised ART modules, ART_a and ART_b, and an inter-ART associative memory, called a map filed [5]. ART_a and ART_b are linked by fully connected adaptive connections between the layer of ART_a and the map field, and non–adaptive, bidirectional, one-to-one connections from the map field to the layer of ART_b. The ART_b network self-organizes the desired output pattrens for each input pattern presented to ARTa.

Briefly, a pair of vectors $\mathbf{a}_p$ and $\mathbf{b}_p$ are presented to ART_a and ART_b simultaneously. The ART_a and ART_b networks choose suitable output categories for these vectors. The map field then checks to see if the choice of ART_a can correctly predict the choice of ART_b. If it can, then learning takes place between the map-field node corresponding to the winning F_{b2} node and the F_{a2} pattern. Connections to all other F_{b2} nodes are inhibited. If not, the map field increases the vigilance of ART_a so that ART_a does not choose the same F2 category again but searches on until a suitable F_{a2} category is found. If there are no suitable categories ART_a chooses an uncommitted node, in which case learning can always take place. An ARTMAP network is presented in figure 2.

5. TRAINING OF THE TIMBRE RECOGNITION SYSTEM

In order to derive the values of the three aforementioned acoustic properties for a note, it is necessary to perform calculations both in time and in frequency space. As an example, a note played by guitar at 165 Hz is presented in figure 3. Initially, we filter separately the five frequency components with the greater amplitude, as illustrated in figure 4. Thus we are able to estimate the slope of the attack of each one component in time space. The derived values are related to the first under examination acoustic property. Next, we calculate the time delay of each component relative to the component that appears earlier in time space. These values are describing in a qualitative way the second property. Finally from the shape of the spectrum of the note, which is shown in figure 5, we can calculate the percentage of energy over the 2.5 kHz and derive a value that correspond to the third acoustic property. It is clear from the above that the input layer of the neural network consists of 10 nodes. The output layer which is used to represent the desired number of timbre categories, consists of 4 nodes.

By repeating the above procedure for several sampled notes played by the five musical instruments that are examined, a number of sets of values is derived. These sets of values are used as training patterns for the ARTMAP network. The aim of the training phase is to associate each training pattern with an appropriate output pattern that represents a timbre category. Essentially, the network during this phase learns the input patterns, by adjusting its weights.

The ARTMAP network we have used for timbre recognition, was trained to recognize all the single notes that belong to the basic octave, played by a piano, a guitar, a trumpet, a saxophone and a flute. For convenience the duration of the notes was restricted to be about 1.5 sec. The training samples of the network are notes with fundamental frequency component at: 110Hz, 117Hz, 124Hz, 131Hz, 139Hz, 147Hz, 156Hz, 165Hz, 175Hz, 185Hz, 196Hz, 208Hz, played by piano, guitar, trumpet, saxophone, and flute. Hence the total number of training samples is 60.

6. NETWORK PERFORMANCE

The network was tested using 12 recorded notes with duration close to 1.5 sec. All of these testing notes were different from the notes used in the training phase. The reason was to observe the generalization ability of the network. The ARTMAP network made mistakes in 2 of the 12 test notes. The two failures correspond to guitar notes which were played very softly by the musician. It is quite reasonable to expect that alterations in the way of playing a note, causes important changes to the

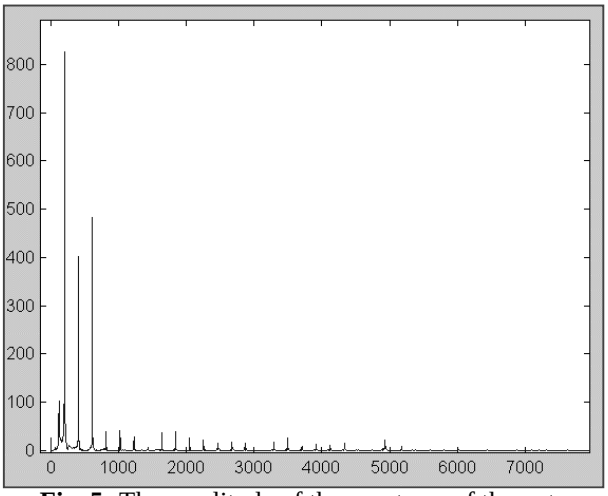

Fig. 5: The amplitude of the spectrum of the note

shape of the spectrum of the note. All the other test samples were successfully identified by the network.

In order to avoid these failures it is necessary to increase the number of training patterns. That way the network would be able to create more categories for each instrument and therefore to succeed a higher performance. The ARTMAP network has the ability to retain knowledge of previously learned patterns as mentioned above. So it can easily be trained with new patterns at any point of its operation and this makes it very convenient for being used in a recognition system.

7. DISCUSSION

This study has shown that the acoustic properties mentioned in the literature as components of musical timbre, provide useful information that can be used for musical instrument recognition.

For our timbre recognition system a supervised network was needed. Using the unsupervised ART networks, we have found that an optimum level of vigilance to provide a satisfactory clustering, could not be estimated. The optimum level of vigilance would be one that separate the patterns from different instruments while clustering together as much as possible patterns which are from the same instrument. Thus, a more advanced network was needed to control the vigilance of the ART networks. ARTMAP was able to succeed that, and to cluster successfully the training patterns.

We should notice that the results in this study could not be compared with human performance on similar tasks, because listeners hear musical phrases rather than a single tone played on a musical instrument.

8. CONCLUSIONS

The ARTMAP network has successfully perform timbre recognition over 5 musical instruments. The mistakes made by the network can be easily explained since the network has been trained with notes played with a specific technique.

If the training set was larger, then the performance of the network would had been higher. It is clear that the capability of the proposed model can be extended to a great variety of playing techniques and to more instruments. This would result in a very robust timbre recognition model for single notes.

9. REFERENCES

[1] Grey, J.M. (1977). Timbre discrimination in music patterns. *Journal of the Acoustical Society of America*, 64, 457-472.

[2] Krumhansl, C.L. (1989) Why is musical timbre so hard to understand? In S. Nielzen and O.Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (pp. 43-53) Amsterdam: Elsevier (Exerpta Medica 846).

[3] C.Harston, A.Maren, R.Pap, Handbook of neural computing applications, (Academic Press) 1990, pp.142-146.

[4] Carpenter, G.A. & Grossberg S.(1987b) ART2:Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26,4919-4930.

[5] Carpenter, G.A. , Grossberg, S. & Reynolds, J.H. (1991a) ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organising neural network. *Neural Networks* ,4, 565-588.