Classification of Musical Instrument Sounds Using Neural Networks

Ali Taylan Cemgil, Fikret Gürgen
Department of Computer Engineering
Boğaziçi University
TR-80815 Istanbul Turkey
{cemgil,gurgen}@boun.edu.tr

December 8, 1997

Abstract

This study introduces the classification of musical instrument sounds by artificial neural networks (ANN). The time varying spectral contents of sounds are estimated based on Short-time Fourier Transform (STFT) and are applied to ANN structures for classification. Recognition results obtained from a multilayer perceptron (MLP), time delay neural network (TDNN) and a hybrid self organizing map radial basis function network (SOM-RBF) are presented and compared.

1 Introduction

1.1 Psychoacoustical Aspects

The human auditory system is a complex and highly nonlinear device. Its physiological structure is well known however the processes that govern the sense of hearing and auditory perception still remain partially understood. Consequentely, to devise a single complete auditory model describing all perceptual phenomena is a difficult task. Nevertheless, several complex models are suggested in the literature summarizing the results of current psychoacoustical experiments [1]. In fact, the state-of-the-art high quality audio coding algorithms [4], [6] and other emerging audio applications [2] make use of these results reformalized in form of several computational models, restated according to the requirements of the application at hand. In this respect, we will focus on musical sound classification capabilities of the human auditory system where a new instance of a sound event is transformed and subsequently classified by subjective measures of human perception.

1.2 Classification of Musical Sounds

In describing musical sound we face four principal parameters: (1) Loudness, (Energy Content, Intensity), (2) Pitch (Frequency difference between the harmonics), (3) Duration (Perceptual duration or physical duration), (4) Tonal Color or Timbre. The first three attributes are one dimensional; each can be described by a single physical measure. On the contrary, a measure in the physical sense does not exist for the timbre. Even a complete definition is hard to give because tonal color is the result of a subjective and non-measurable perceptual process. Grey showed in his important studies the significance of three distinct phenomena for the timbre classification problem, namely (1) the spectral energy distribution; (2) the presence of synchronicity in the transients of the higher

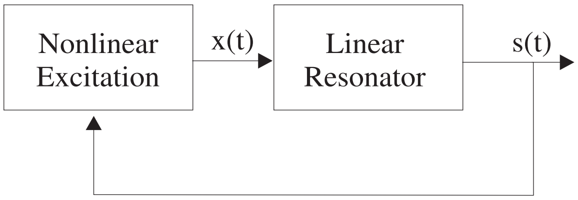

Figure 1: General block diagram of a musical instrument.

harmonics, along with the closely related amount of spectral fluctuation within the tone through time; and (3) the presence of low amplitude, high frequency energy in the initial attack segment [3].

1.3 Timbral Families and Sound Production Mechanisms

Almost all musical instruments can be described in terms of conceptually separate excitation and resonator mechanisms, according to well-known scheme in Fig. 1. A nonlinear element, such as reed or bow, excites a linear, energetically passive, multimode element such as a tube or string. In some cases, the influence of the excitation can be considered as feed-forward (guitar, percussion instruments), but generally the linear element in turn influences the operation of the nonlinear element (woodwinds, reed organ, bowed strings, etc.), as suggested by the lower path in Fig. 1. As a remarkable fact, it is common sense to classify musical sounds according to the traditional sound generation mechanisms. For example, we normally speak of woodwinds, brass, percussions, strings etc. and consider the sound of a violin to be nearer to a cello rather than a trumpet. Generalizing the relationship between the sound production mechanism and the timbre family, one can conclude that the rich frequency content of the excitation is shaped in time by the resonator, and this time varying frequency pattern is taken as an important cue for timbre classification.

2 Experiments

2.1 Data

The sound database is recorded from the standard set of Soundblaster AWE32. We sampled 40 different sounds from 10 timbral families in the chromatic scale from A3 to A4 at 44100kHz using 16 bits/sample. The sounds themselves are sampled versions of real acoustical sources.

2.2 Preprocessing

In building an automated sound classifier, (as it is the case with any pattern recognition schema) the success is highly dependent on the application specific preprocessing, where important features are kept. In sound classification, as an additional burden, significant features are not easy to specify a priori. Therefore we depend on psychoacoustical and physical aspects of musical sound to transform the vast amount raw signal data into a perceptually correct domain for subsequent classification. In the well-known method of discrete STFT analysis:

$$X(n,k) = \sum_{-\infty}^{\infty} x(m)w(n-m)e^{-j2\pi km/N}R_N(k)$$

from the Fourier transform point of view, the window $w[n]$ slides along the signal $x[n]$. $R_N(k)$ is a rectangular window of size N. At the end of the analysis we obtain Fourier transform coefficients defined on a uniform grid over the time frequency plane for each time index n and frequency index

k. Definitely for our purposes only a subset of those coefficients are necessary. The following observations are made: Audio signal $x[n]$ from a musical instrument exhibit quasi-periodicity. We get a frame out of the signal by windowing with some window function $w[n]$. Hence we expect that $|X(\omega)|$ contains a pattern of roughly evenly spaced peaks which are shaped after the spectrum of the analysis window, i.e. given a period of P for $x[n]$, $|X(\omega)|$ ideally consists of a set of spectral peaks at multiples of $2\pi/P$. If P is unknown, it can be estimated from the spectra by the P&G Method [9].

We make the following assumptions to reduce the vast amounts of coefficients on the lattice by thresholding : (1) The local maxima (exceeding a threshold) on the direction of the frequency axis are equally spaced in frequency with about $2\pi/P$. (2) The difference between successive harmonics is approximately equal to the fundamental pitch frequency. (3) The sets of local maxima evolve parallel to time axis. (3) The instrument is monophonic.

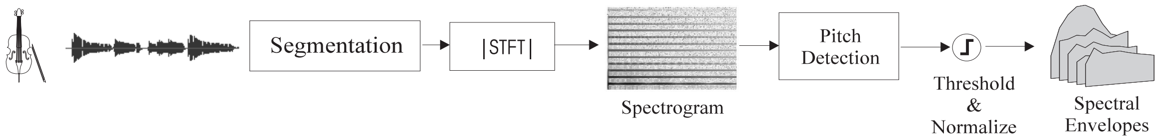

Figure 2: Scheme of the Preprocessor

Along with these heuristics, we extract a set of discrete 'harmonic envelopes' $x_h[n]$ from the STFT modulus, where h and n are harmonic and frame index respectively. In the present work, we let $h = 1, 2, \ldots, 12$ and $n = 1, 2, \ldots, 20$. Sample patterns resulting from this STFT analysis (50% overlapping frames, Hamming window of length 1024 samples), are shown in Fig. 3.

3 Classification

In general, a classification problem can be wieved as a function approximation problem. A classifier in this sense is a mapping $f : X \to C$ where X is the set of input vectors and C is the set of target classes with $|C| = N$. If the classes are discrete and cannot be ordered (e.g. letters in character recognition, phonemes in speech), N different functions $f = [f^1, f^2 ..., f^c, ..., f^N]$ are estimated where each$f^c : X \to Y$ corresponds to a discriminant function and $Y = \{y : y \in [0, 1]\}$. One key issue is the choice of a model (or equivalently a topology) $\Im(\mathbf{w}, \cdot)$ for f^c where $\mathbf{w}$ are the model parameters. In our study, we use three standard well known topologies (1) Multilayer Perception, (2) Time-Delay Neural Network (3) Self-organizing Network.

3.1 MLP

A MLP (Fig 4) has the following functional form:

$$\mathbf{f} = \mathbf{H}[\mathbf{\Phi}_h(\mathbf{W}\mathbf{x})]$$

where $\mathbf{x} = [x_0, x_1, x_2, .., x_p]$, a $p+1$ dimensional input vector with the constant term $x_0 = 1$, $\mathbf{W}$ is a $(h-1) \times (p+1)$ input layer connection weight matrix, $\mathbf{H}$ a $N \times h$ hidden layer connection weight matrix, $\Phi_h(x) = [1, \phi_1, \phi_2, \ldots, \phi_{h-1}]$ is a vector valued nonlinear function where $\phi_i = \exp(x_i)/(1 + \exp(x_i))$ (the sigmoid function).

3.2 TDNN

As an example of a recurrent network, although not recurrent in the same manner as a e.g. Hopfield Network, we considered the time-delay neural network (TDNN) structure. In a sense, TDNN can

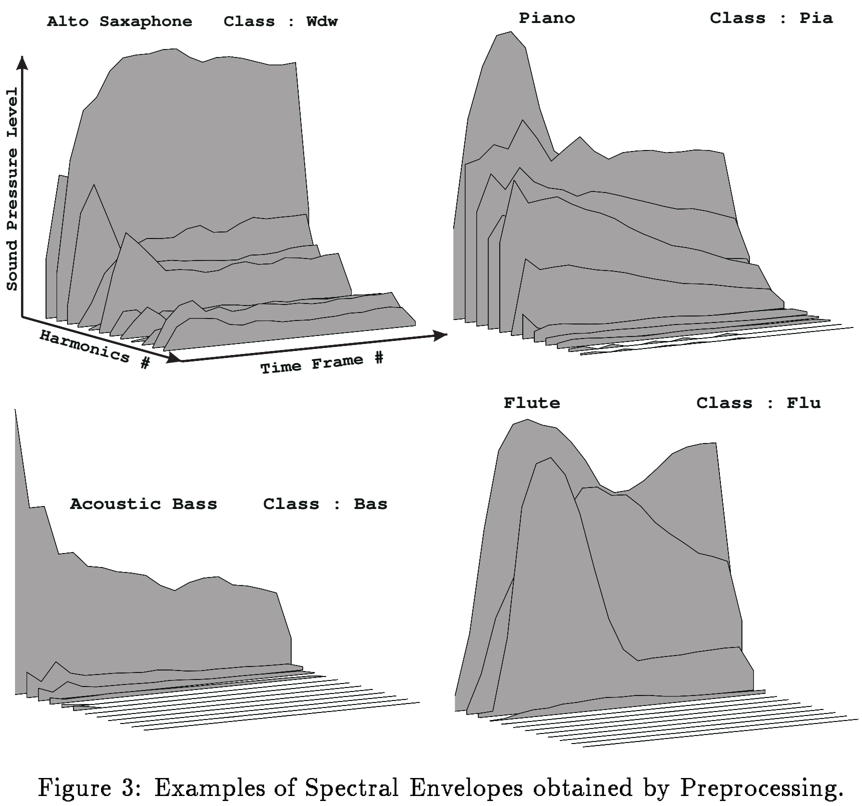

Figure 3: Examples of Spectral Envelopes obtained by Preprocessing.

be considered to be between a static MLP and a recurrent network, in that it is able to pick up the dynamic (time varying) features of the input. A TDNN is basically a MLP structure with temporal input $\mathbf{x}_{K,R,t} = [x_{Rt+1}, x_{Rt+2}, .., x_{Rt+K}]$. The input layer output at time $t = 0, 1, 2, .., (p-K)/R$ are accumulated in the hidden layer as $\mathbf{B} = [b_0, b_1, \ldots, b_{(p-K)/R}]$ where

$$\mathbf{b}_t = \mathbf{\Phi}_{\mathbf{h}}(\mathbf{W}\mathbf{x}_{K,R,t})$$

and the network output $\mathbf{f}$ is computed as

$$\mathbf{f} = \mathbf{HB}$$

where K is the window size and R is the amount of shift between successive windows (Fig 4).

3.3 SOM

The SOM is is the result of a vector quantization (VQ) algorithm which is especially useful for visualization or coding of high-dimensional data [5]. The SOM approximates the input data distribution in a non-parametric and ordered fashion with a low and continuously changing regression surface. We combine the SOM with a RBF network, in which after convergence, the codebook vectors are interpreted as class centers with normal Gaussian units. Subsequently, the outputs are classified by a simple linear perceptron trained with the backpropogation algorithm. The excitation on the resulting map is also very useful for visualizing the timbre clusters Fig 8.

4 Training

Once a model $\Im(\mathbf{w}, \cdot)$ is choosen, the model parameters $\mathbf{w}$ are estimated by minimizing an error measure, which serves as an indicator of the distance between the desired output and the observed output:

$$E = \sum_{i=1}^{N} d(y_i, \Im(\mathbf{w}, x_{\mathbf{i}}))$$

where $d(\cdot)$ is a distance measure (almost always taken as the Euclidian norm for regression and Cross-entropy for classification) and x_i are the ith training data sample from a set of N samples. Since the model $\Im(\mathbf{w}, \cdot)$ is non-linear in general, no close-form solutions to the optimal parameters are known. Therefore, variants of back propagation, a popular steepest descent agorithm, are employed to estimate a sub-optimal solution for $\mathbf{w}$ which is "good enough". In order to avoid overfitting, during training the error on a cross validation set is monitored (in our case this set is the test set), and training is stopped when this error tends to increase.

5 Results

In present work, the ANN structures from the three different topologies shown in Fig. 4 are trained with spectral energy envelopes, where data are divided into a training and test set containing 240 and 160 labeled examples respectively.The first goal is to find the how many frames (unit of time) and harmonics (unit of frequency components) are needed for good generalization. Fully connected MLPs are trained for all combinations of $h = 4..12$ and $f = 4..10$ by the backpropogation algorithm (Fig 5).

TDNN simulations are carried out for all combinations of $h = 3..12$ and $f = 8..12$ for sliding window sizes 3..6 (Fig 6). The best results obtained on the test set are in listed in Table 1.

Two dimensional Feature Maps are trained using the Kohonen's SOM Algorithm for all combinations of $h = 4..12$ and $f = 4..10$ (Fig. 7).

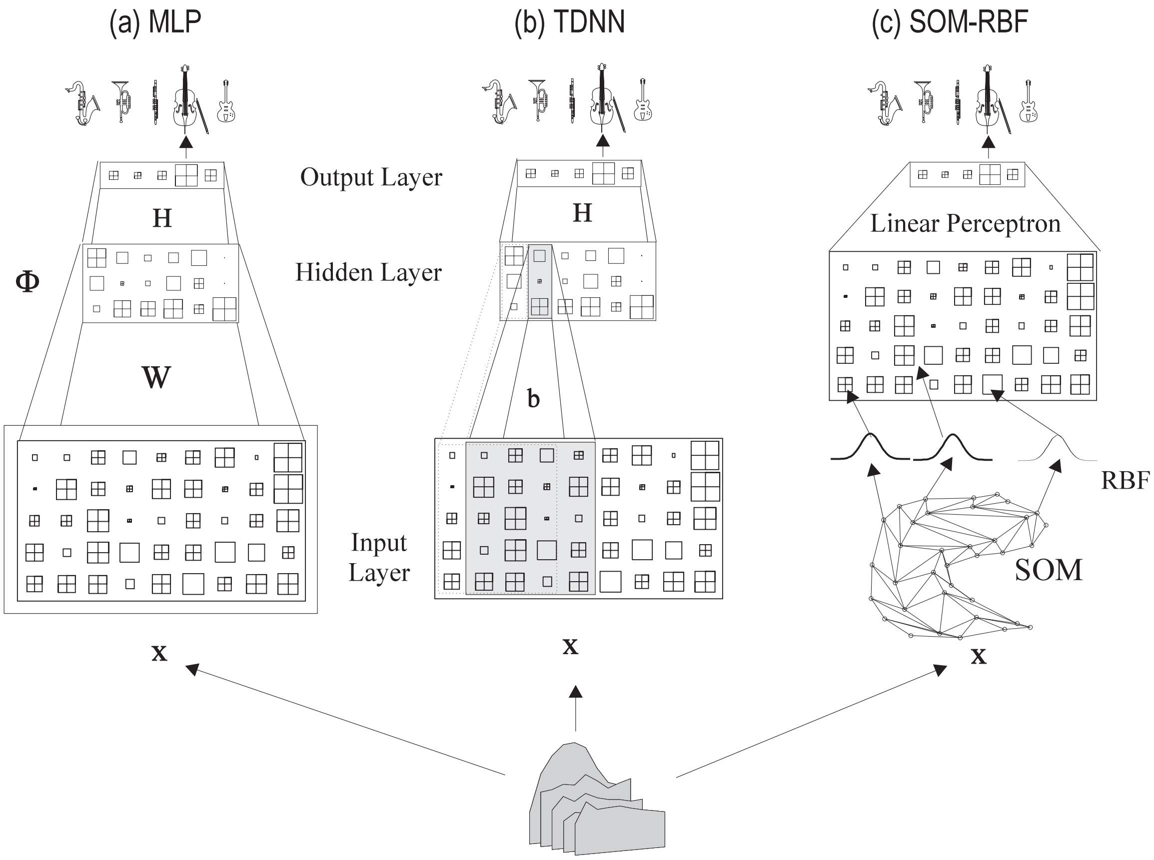

Figure 4: Classifier Structures.

6 Conlusions and Future Research Plans

In the simulation studies carried out on our limited dataset, TDNN model exhibits slightly better performance in terms of success and also with less number of parameters. The SOM-RBF results are promising, although being a few percent below the MLP and TDNN, because they give additional information on how sounds are organized according to timbres rather than just being a black-box as the latter. We expect to get better performance on a larger database by fine tuning the parameters involved.

The result suggest that the success rate improves proportional to the number of harmonics used in the representation and less sensitive to the number of frames, i.e. the spectral content in the attack portion of a sound contains most of the information needed for the sounds characterization. This is also consistent with earlier psychoacoustical results [3], [1].

The timbre model can also be improved in that it is inadequate to represent all aspects of the tonal color such as: (1) Variations of the fundamental pitch period (vibrato), (2) Sound that exhibits inharmonic spectra (percussion instruments, bells, etc.). (3) Very short time characteristics (attack period or pitch onset period), To overcome these shortcomings we plan to do the following improvments: (1) The fundamental pitch contour can also be input, perhaps at a much lower time resolution. (2) A classifier could be built for distinguishing harmonic instruments from percussive instruments. Then different classes would undergo different processings and classifiers.(3) Other transforms than the STFT could be used such as the Wavelet Transform or variants of pseudo-Wigner [8] distribution. Additionally, different portions of the sound can be analysed with different

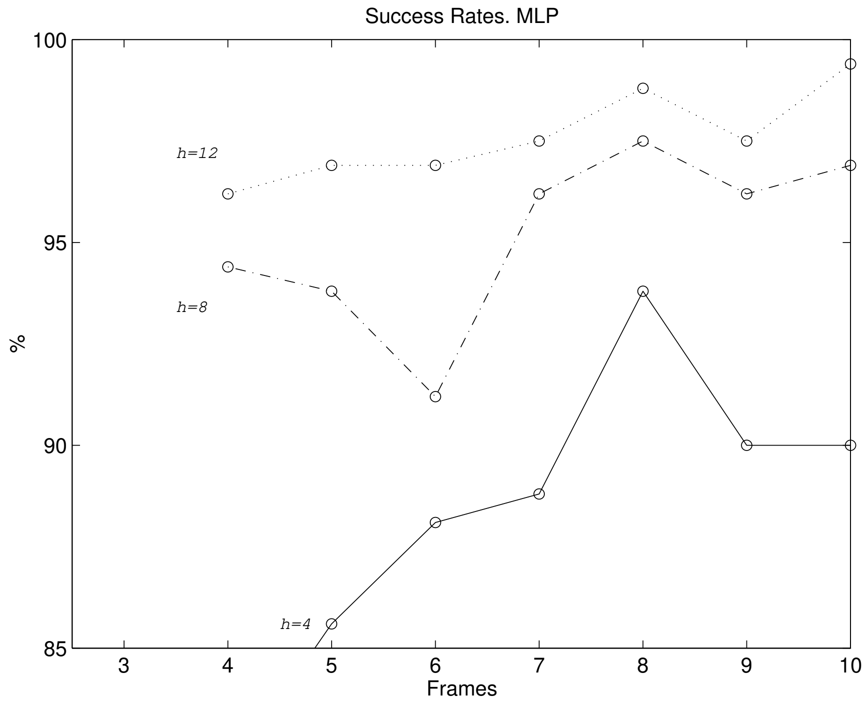

Figure 5: Success Rates : MLP.

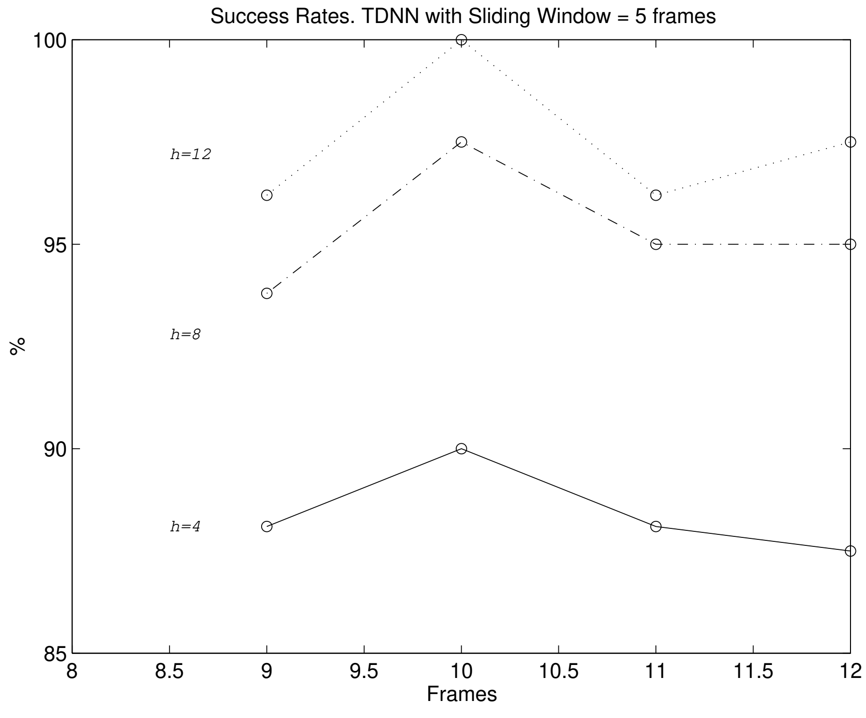

Figure 6: Success Rates : TDNN with Sliding Window Size 5.

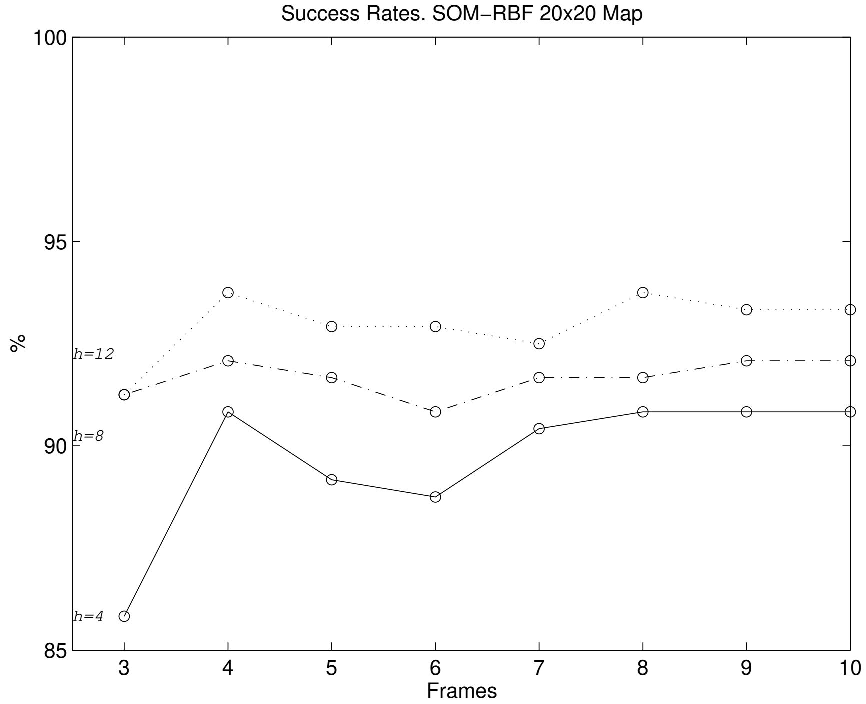

Figure 7: Success Rates : 20x20 SOM-RBF.

methods. e.g. the Attack portion can be analyzed by WT where more time resolution is desired whereas the steady state portion can be well represented by an averaged spectrum over several frames.

The frequency dependent sensitivity and the masking properties (impact of an harmonic to its neighbor harmonics) [7] of the human auditory sysem can be considered in the preprocessing stage for devising a perceptually more meaningful distance measure by weighting the energy envelopes according to their perceptual inportance, rather then using simply the Euclidian metric, which assumes equal variances in each dimension.

Table 1: Comparison of best Results

Network	f	h	Connections	Epochs	Comment	Success%
MLP	10	12	15818	500		97.50
TDNN	10	10	2634	1600	Window = 3	100.00
	10	10	2866	1600	4	97.50
	10	12	3418	1600	5	100.00
	9	9	1830	1600	6	98.80
SOM-RBF	4	12	19600	N/A	2D-20x20	93.75
	4	10	9225	N/A	2D-15x15	94.17
	4	9	3700	N/A	2D-10x10	93.75

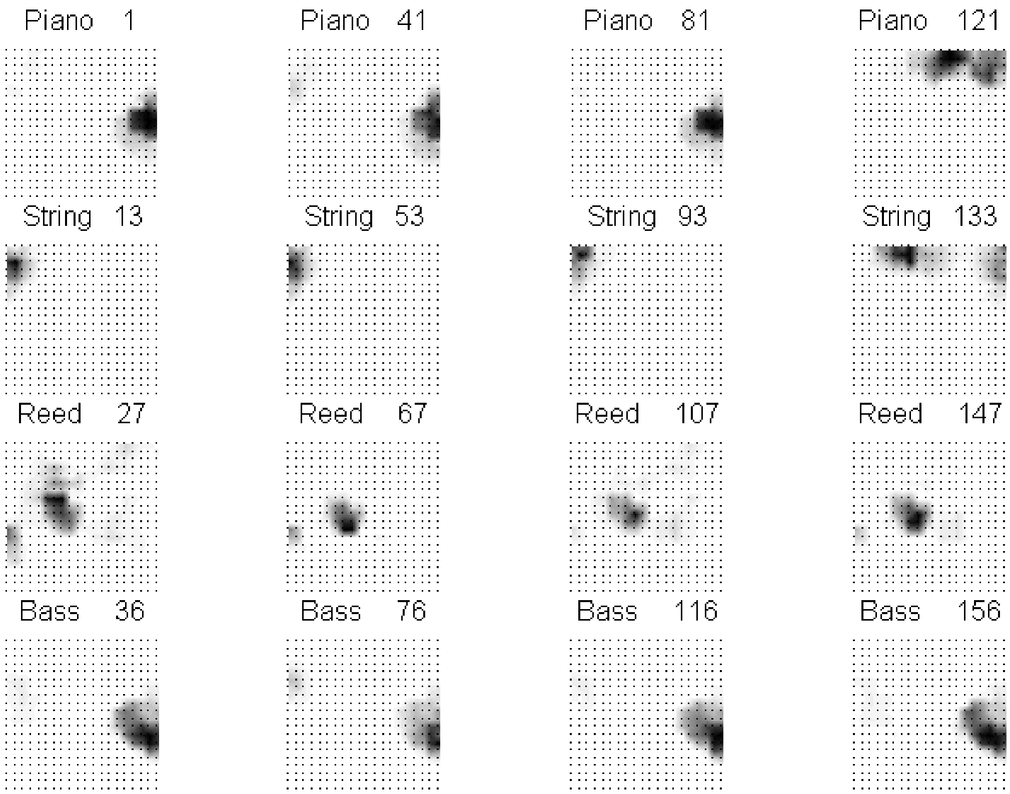

Figure 8: Timbre Clusters.

References

[1] A. Bregman, Auditory Scene Analysis, MIT Press Mass. 1990

[2] B. Feiten, S. Gunzel, Distance Measure for the Organization of Sounds, Acustica, Vol78 pp181-184 1993

[3] J. Grey, Multidimensional perceptual scaling of musical timbres, J. Acoust. Soc. Am. 61 (5), pp1270-1277, May 1977

[4] N. Jayant, J. Johnston and R. Safranek, Signal Compression Based on Models of Human Perception, Proceedings of the IEEE, October 1993, Volume 81, Number 10, Pages 1385-1421

[5] T. Kohonen, The Self Organizing Map, Proc. IEEE, 78 (9) pp 1464-1480 September 1990

[6] P. Noll, Digital Audio Coding for Visual Comminications, Proc. IEEE vol 83 No 6. June 1995 p925-943

[7] J. R. Pierce, The Science of Musical Sound, Scientific American Books, New York 1983

[8] W. Pielemeier, G.Wakefield, A high resolution time-frequency representation for musical instrument signals.,J. Acoust. Soc. Am. 99 (4), pp2382-2397 April 1996

[9] Piszczalski, M. and B. Galler, Predicting Musical Pitch from Component Frequency Ratios, Journal of the Acoustical Society of America vol 66, no 3, pp 710-721, September 1979

[10] B. D. Ripley, Neural Networks and Related Methods for Classification, J. R. Statist. Soc. B(1994) 56 No. 3, pp. 409-456