

Neural Networks and Pitch Detection in Random Acoustic Signals

Nicola Cufaro Petroni^{*}

Mario Guarino[†]

Guido Pasquariello[‡]

1. INTRODUCTION

Neural networks have been used in the last few years to analyze several aspects of musical phenomena. To quote just a few : tonal expectancies [1], tonal semantics [2], temporal evolution [3], composition [4] and so on. In the field of the detection of pitch a research has been started recently [5] on the track of previous studies [6]. It has been shown there that, by means of a backpropagation neural net, it is possible to build a model that is able to distinguish between noises and tones endowed with a recognizable pitch in a random acoustic environment. In particular the implications have been studied of the ability of detecting the presence of a pitch in an acoustic signal when this ability is considered as a primary fact. In this first work a backpropagation neural net has been trained to distinguish among signals with and without pitch. The idea of the pitch is a psychological one meaning the perception of the height of an acoustic signal: whereas in a musical tone we can detect a precise pitch, in a noise we can not. However, even if there exist in nature signals with or without a well recognizable pitch, there are others such that this characteristic is much more ambiguous: they constitute the more interesting challenge to these models.

It is well known [7] that in a first approximation the pitch is essentially connected to the nature of the spectrum, namely of the Fourier Transform of the signal and in particular to its modulus (phases will be neglected). In fact all the information connected to the pitch is encoded in this Fourier Transform, but this must also be considered highly redundant information. Hence we will introduce a few hypotheses in order to eliminate this redundancy, and we will try to check them on the basis of our simulations. In particular the axis of the frequencies has been discretized in a way that allows the individuation of the elements of our language: the tones of the diatonic scale. As a consequence the modulus of the Fourier Transform has been reduced to templates of weights attributed only to the frequencies of this well tempered series of tones. Since the phases are neglected, the number assigned to each frequency interval is a positive real number correspondent to the modulus of the Fourier Transform. Moreover, by means of the octave equivalence, the discretized Fourier Transform has been reduced to templates of just 12 numbers (one for every pitch class) representing the inputs of the neural net, that starting from a set of these templates learns to classify signals with and without pitch. In this case the answer of the net was encoded in just one output number ranging from zero to one.

Recent developments along this line of research [8] have enlarged the scope of the model which is able now not only to distinguish between signals with and without a pitch, but also to characterize the pitch in a precise way. In fact the net presents now 12 inputs and 12 outputs (one for each different pitch

^{*} Dipartimento di Fisica dell'Università di Bari e GNCP-CNR Unità di Bari; via Amendola 173, 70126 BARI (Italy); E-mail CUFARO@BARI.INFN.IT

[†] ICE-CNR; via De Marini 6, 16100 GENOVA (Italy); E-mail GUARINO@MART5.ICE.GE.CNR.IT.

[‡] IESI-CNR; via Amendola 166/5, 70126 BARI (Italy); E-mail PASQUARIELLO@IESI.BA.CNR.IT.

class). The answer of the network with just one of the outputs equal to 1 (the others being equal to zero) represents signals with a pitch belonging to a particular pitch-class; on the other hand the noises are characterized by the absence of a prominent pitch, namely by an answer with 12 equally zero outputs. Moreover the analysis of the signals made by the human ear has been more completely taken into account: in the first paper [5] harmonic tones with an elevated number of partials (31 partials each) were adopted for the simulation, since every line of the spectrum could be attributed to just one frequency interval. However this number was excessive, made our tones too much noisy and neglected the fact that the human ear analysis is based just on a smaller group of prominent harmonics. On the other hand even a pure tone with an unique component of frequency excites in the human ear a wide (non point-like) zone of the basilar membrane. Hence we have chosen to represent harmonic tones not by purely discrete spectra, but by means of the (continuous) excitation levels of the basilar membrane drawn from known empirical data. The reduction to 12 numbers is then performed on this excitation function. The results of both the simulations confirm that, notwithstanding the ambiguities introduced by the reduction process (in general we can not recover the original object starting from the reduced templates of 12 numbers), the compression of the signal here proposed does not cancel out the essential information about the pitch.

2. DESCRIPTION OF THE MODEL

Our model is composed of an adaptative listener embedded in a random acoustic environment supplying simulated external stimuli. A very simple model for a real listener is simulated by means of a backpropagation neural network (BNN) [9] which will undergo a supervised training in the random acoustic environment. In the following we will use BNN whose neuron activations can take any real value in the interval [0,1]. The behaviour of the net as a computing system is determined by the matrix w_{ij} of the synapses and by the form of the transfer function f_j . If these elements are fixed, the net implements a particular function: for every input vector $x = \{x_i\}_{i=1,\dots,n}$ we obtain an unique vector $y' = \{y'_j\}_{j=1,\dots,m}$ as output. However, the decisive feature of a BNN is that it is an adaptative computing device: the outputs $y'(k)$ ($k=1,\dots,N$) of the network corresponding to a set of N input data $x(k)$ ($k=1,\dots,N$) are compared with a set of required output $y(k)$ supplied from the outside and the network parameters are progressively modified so that the outputs will come nearer and nearer to the required outputs, where *near* means near in mean square. The net used in this model is a multi layer perceptron (MLP) network, a particular type of BNN. There is one input layer with 12 neurons, one hidden layer with a variable number of neurons (10,15,20,25) and there is one output layer with 12 neurons. Every neuron in every layer is fully connected with the neurons of both the previous and the following layer. In this work we used one of the classic algorithms for supervised training: the gradient descent algorithm.

In order to build the second element of the model, the random acoustic environment, we will first of all discretize the frequency axis. We will limit ourselves only to a sequence of discrete values of the frequency, namely : $f_k = 2^{k/12} f_0$; $k = 0, \pm 1, \pm 2, \dots$ which are a well tempered semitone apart so that they constitute a sort of (extended) well tempered keyboard. The fundamental objects representing our acoustic signals will be a complex function of the frequency f , namely the Fourier transform $\phi(f)$ of our signal $x(t)$ considered as a function of the time t . If our signal has a continuous spectrum, namely if its Fourier Transform is a continuous function, we produce a discrete spectrum by assigning an average weight only to the points f_k , or more precisely to every (1 semitone) interval around f_k . In order to avoid formal complications we will also suppose that all our signals are limited in band so that we will not consider all the infinite sequence of the f_k : more precisely, if for example $f_0 = 622.3$ Hz (which correspond to the E_5) and we take $k = -60, -59, \dots, 0, \dots, 59$, we will have 120 well tempered frequencies running from about 20 Hz through about 20kHz.

When our signal is periodic the Fourier transform reduces itself to a Fourier series and it is described by means of a discrete harmonic spectrum. If $f^{(1)}$ is the fundamental frequency of this

spectrum, the sequence of the harmonic partials is $f^{(n)} = nf^{(1)}$, where $n \geq 1$. In general these frequencies will not exactly coincide with one of our well-tempered f_k . In the simulations presented here we consider only 6 partials in the spectrum, $f^{(1)}, f^{(2)}, \dots, f^{(6)}$. Then we have to generate the amplitude for every frequency, and we will do this using a Fermi distribution : if $F_n(x)$ is the Fermi distribution function for the n-th partial, it will be enough to generate x_1, \dots, x_m , random real numbers uniformly distributed in $[0,1]$, and then to calculate the amplitudes as

$$c_1 = F_1^{-1}(x_1) \quad \dots \quad c_m = F_m^{-1}(x_m)$$

$$F_n^{-1}(y) = a_n - \frac{\log[(1 + e^{\lambda_n a_n})(1 - y) - 1]}{\lambda_n} ; \quad a_n = \frac{1}{n^p} ; \quad \lambda_n = n^p \quad (p > 0)$$

and p is just a parameter that we can change in order to obtain different slopes in the Fermi distributions. To take into account the analysis of the sound made from the human ear, we have chosen to represent harmonic tones by means of the (continuous) excitation level of the basilar membrane drawn from known empirical data [7] in the so called *critical band rate scale*. This scale is based on the fact that our hearing system analyses a broad spectrum into parts that roughly correspond to critical bands. The unit of this scale is the Bark and the (non linear) relation between the critical band rate z (in Bark), and frequency f (in Herz), is important to understand many characteristics of the human ear. In many cases the following empirical expression [7] is useful to describe the dependence of critical band rate on frequency over the whole auditory frequency range : $z = 13 \arctg(f / f_1) + 3.5 \arctg(f / f_2)^2$ where $f_1 \approx 1.32$ kHz and $f_2 \approx 7.5$ kHz. Moreover the excitation intensity, as a function of the frequency, is:

$$E_n = A_n 10^{2.7|z(f) - z(f_n)|}$$

and since in our model we consider 6 different partials we have to take the function $E(f) = E_1(f) + \dots + E_6(f)$. which can be considered as a good approximation of the energy distribution along the basilar membrane when we perceive a sound with 6 partials.

3. SIMULATION

The experiment is based on acoustic data completely simulated on an Alfa VAX. The BNN software has 12 neurons input layer, 12 neurons output layer, and one hidden layer. In the training phase we have randomly generated training sets of 1400 sample signals with the following structure: 700 harmonic tones (100 with $p = 0$; 150 with $p = 0.5$; 150 with $p = 1$; 150 with $p = 1.5$; 150 with $p=2$) and 700 noises (300 white noises; 200 noises in a 3 octave band; 200 noises in a 1 octave band) where p is the parameter in the Fermi random variable which determines the rate of decay of the higher partials: higher value of p mean faster decay. Remark that the fundamental frequency of every harmonic tone is chosen at random. Every sample is composed of 24 numbers: 12 inputs and 12 target outputs. If we generate a sound, the target is a vector where only one value is 1 and the others are 0; if we generate a noise, the target is a vector with 12 zeros. To test the performance of the trained BNN we used different sample sets: a set of 4000 samples for every class of signals: harmonic tones with $p = 0$, $p = 1$, $p = 2$; harmonic tones with $p = 1$ either without the first harmonic, or without the first and the second harmonic; white noises, noises in a 3 octave band, noises in a 1 octave band. Harmonic tones lacking the first partials are included in the performance sets in order to check the generalization power of the model: they are typical signals connected with the empirical phenomena of the residue pitch [7]. These different kinds of test samples

have been used in 4 different BNN architectures depending on the number of the neurons in the hidden layer : 25, 20, 15, 10. Moreover the simulation with 25 hidden neurons has been performed twice in order to test the stability of the results. To evaluate the results in the test phase we compare the required (y_i) and the actual (y'_i) outputs and we calculate

$$\frac{1}{2} \sum_i |y'_i - y_i| \quad ; \quad i = 1, \dots, 12$$

which represent a measure of the error for every sample. If the required and actual outputs coincide the error is zero and the BNN has correctly performed the classification; if the net identifies a wrong pitch class the error is 1; if the net mistakenly consider a tone as a noise (or vice-versa) the error is 0.5. And, of course, all the intermediate values (and even errors greater than 1) can occur. Hence to evaluate the performance we have divided the interval $[0,1]$ in 10 equal sub-intervals and we have measured the relative frequency of errors falling in every sub-interval (errors greater than 1 are considered as members

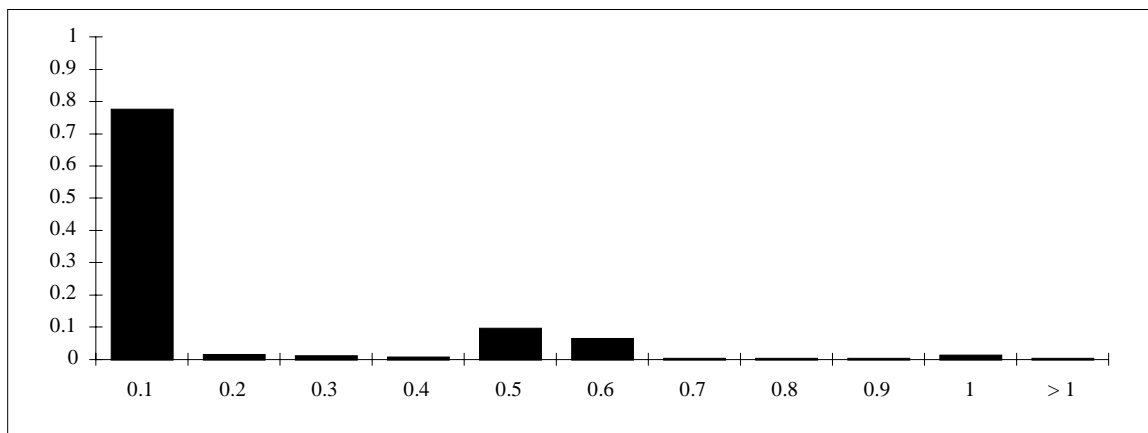


Fig. 1: Histogram of the errors for harmonic tones with $p = 0$

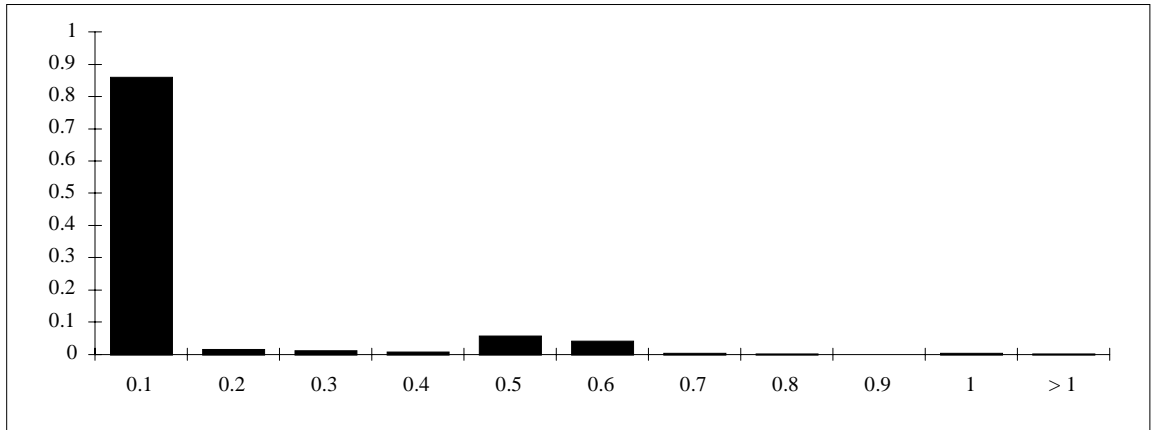


Fig. 2: Histogram of the errors for harmonic tones with $p = 1$

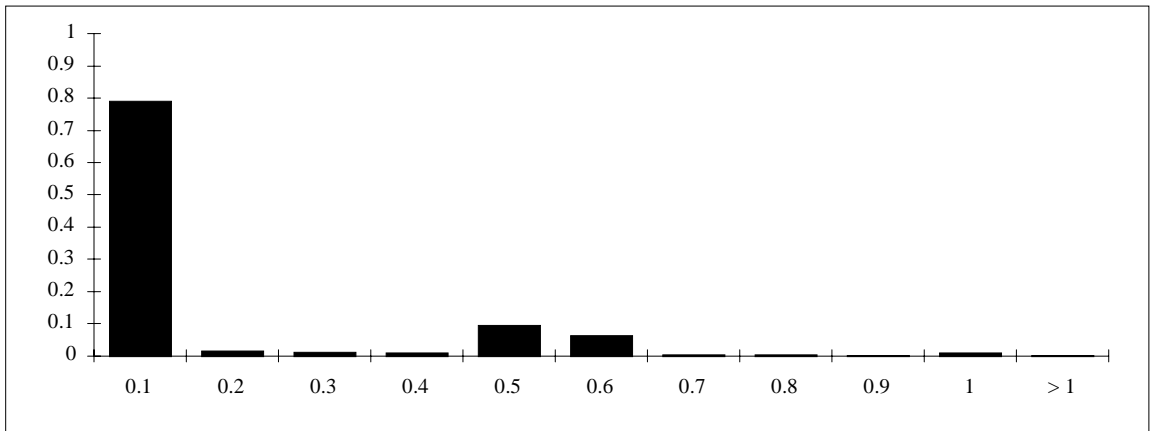


Fig.3: Histogram of the errors for harmonic tones with $p = 1$ and without the first harmonic

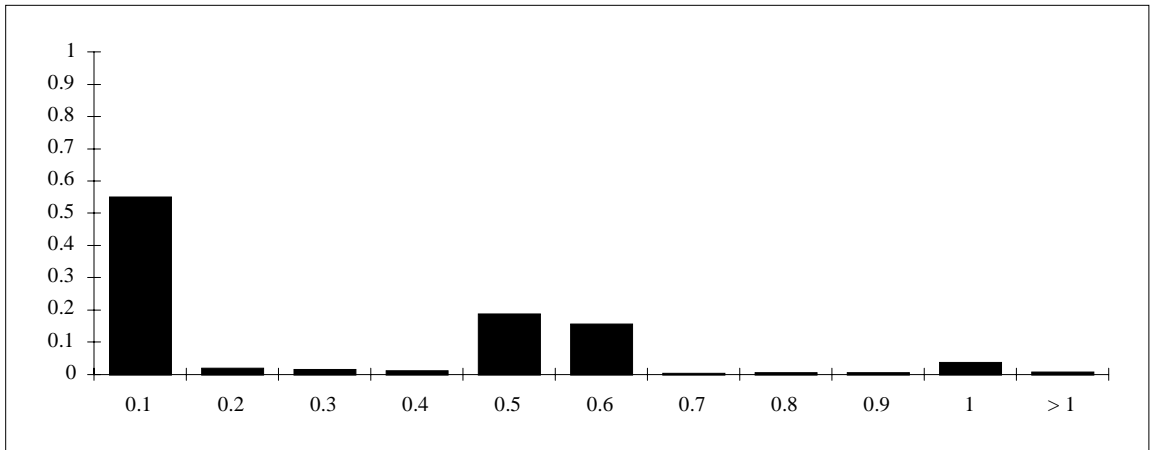


Fig. 4: Histogram of the errors for harmonic tones with $p = 1$ and without the first two harmonics

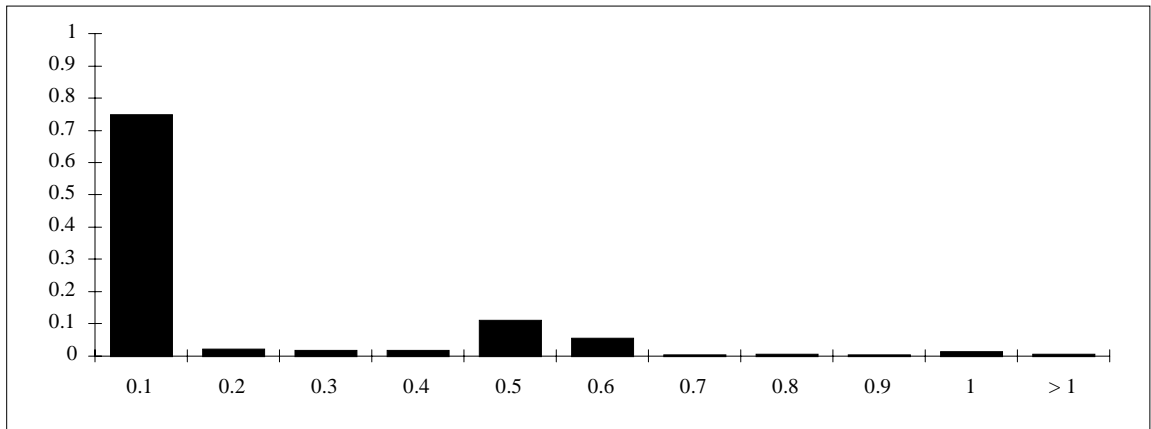


Fig. 5: Histogram of the errors for white noises

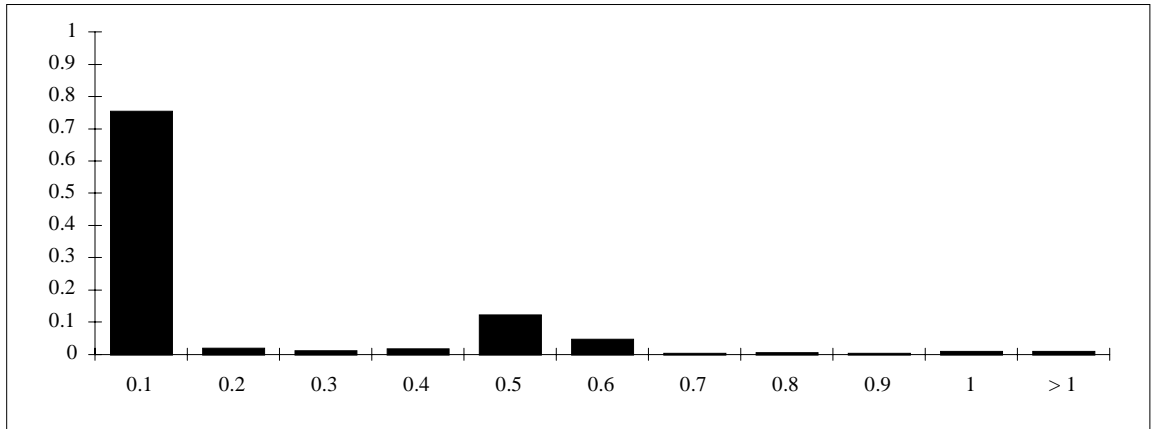


Fig. 6: Histogram of the errors for noises in a 1 octave band

of an 11-th class). In the Fig. 1 to 6 are collected the results (classified by type of sample) where the BNN architecture presents 10 neurons in the hidden layer. Histograms for different architecture are quite similar.

4. DISCUSSION AND CONCLUSIONS

The performances are, generally speaking, satisfactory since the essential information about the pitch of a signal is preserved also in its reduced form. Of course the quality of the detection is better for signals of a type included in the training sets (Fig.s 1, 2, 5, 6), but it is acceptable even for “unheard before” signals (Fig.s 3,4) indicating that the residue pitches can still be classified on the basis of previous different experiences on complete harmonic tones. In fact the results of these simulations show the same qualitative behaviour presented in the case of the complete signals, but for the fact that the detection of the residue pitch is less clear than the perception of the complete pitch, as also happens in the real world. The performances also depends on the value of p : smaller values of p are connected to tones richer in high harmonics and hence more noisy. As for the dependence on the architecture of the BNN it must be remarked that the difference in the global behaviours are dim, even if some bent to be less prone to overtraining is shown by architectures poorer in hidden neurons. Hence the first conclusion is that our simple model of the real acoustic world makes sense since at least it allows one to give on account of some phenomena (the residue pitch detection) by means of simpler facts (the ability to detect the pitch of complete signals).

However this only means that the road is open to investigate more thoroughly the subject in several directions. First of all an obvious modification will be to try to test the model on real-life signals in order to check if our simulated models can in fact be considered as unbiased as we supposed they were. A second line of research could be to explore more closely the behaviour of the network, its stability to randomness in the signals, the need for some regularization in its performances [5,10] and even the possibilities open by different types of Neural Networks (Kohonen maps and unsupervised training) [9]. Furthermore it would be particularly interesting to check if more peculiarly musical concepts as dissonance and consonance can be based on the ability to detect a pitch. In fact, even if it is clear to

everyone that the tonal system is not based (or at least *not only based*) on the presence of detectable pitches, we can nevertheless say that the reference to a tonic is one of the building blocks of the classical tonality. Finally our simulations could be extended to investigate the BNN performances for anharmonic tones, trying to evidenciate possible differences with respect to harmonic ones, and to analyze acoustic signals evolving in time: a direction already indicated by other researchers [3].

REFERENCES

- [1] J.Bharucha: *Music cognition and perceptual facilitation*; Music Perception, 5 (1987) 1. J. Bharucha: *La cognition tonale, l'intelligence artificielle, et les réseaux neuronaux*, proceedings of the symposium sur la Musique et les Sciences Cognitives, Paris, March 1988; P. Mardaga (Liège, 1989). J.Bharucha: *Neural Nets and perceptual learning of tonal expectancies*; proceedings of the conference on Music perception and Cognition, Kyoto, october 1989.
- [2] M.Leman: *Symbolic and Subsymbolic information processing in models of musical communication and cognition*; Interface, 18 (1989) 141. M. Leman: *Emergent properties of tonality function by self-organization*; Interface, 19 (1990) 85. M..Leman: *Schema-based tone center recognition of musical signals*"; J.N.M.R. 23 (1994) 169
- [3] R.D.Gjerdingen: *Learning syntattically sigificant temporal patterns of chords*; Neural Networks, 5 (1992) 551. R.D'Autilia and F.Guerra: *Qualitative aspects of signal processing through Dynamical Neural Networks*, in *Representation of Musical signals*, G.De Poli, A.Piccialli and C.Roads Eds.; M.I.T. Press (Cambridge, 1991)
- [4] S.Dini, A.Gaggiolo, C.Martini and M.Morando: *Man machine interaction in the study of melodic structures*; proceedings in the conference on Cognitive Musicologies; Jyväskylä (Finland) august 1993.
- [5] N.Cufaro Petroni, F.Degrassi, G.Pasquariello: *Detection of pitch in random acoustic signals by Neural Networks*; J.N.M.R. 23 (1994) 369.
- [6] J.Bharucha: *A self organizing neural net model of pitch extraction*; preprint, Dartmouth College (1993). B.Laden : *A parallel learning model of musical pitch perception*; J.New Music Res. 23 (1994) 133
- [7] E.Zwicker and H.Fastl: *Psychoacoustics, Facts and Models*; Springer-Verlag (Berlin, 1990).
- [8] M.Guarino: *Classificazione di segnali acustici mediante rete neurale*; Thesis (Bari, December 1994, in italian); unpublished.
- [9] J.Hertz, A.Krogh and R.G.Palmer: *Introduction to the theory of Neural Computation*; Addison-Wesley (Redwood USA, 1991)
- [10] T.Poggio and F.Girosi: *A theory of networks for approximation and learning*; A.I.Memo 1140 (MIT, 1989)