AUDITORY MODELLING AND SELF-ORGANIZING NEURAL NETWORKS FOR TIMBRE CLASSIFICATION

Piero Cosi*, Giovanni De Poli** and Giampaolo Lauzzana**

* Centro di Studio per le Ricerche di Fonetica - C.N.R.
P.zza Salvemini, 13 - 35131 Padova, Italy
cosi@csrf00.pd.cnr.it

** Universita' di Padova, Dipartimento di Elettronica ed Informatica,
Via Gradenigo, 35100 Padova, Italy.
depoli@dei.unipd.it

ABSTRACT

A timbre classification system based on auditory processing and Kohonen self organizing neural networks is described. Preliminary results are given on a simple classification experiment involving 12 instruments in both clean and degraded conditions.

INTRODUCTION

Timbre is a sound feature which can be hardly analyzed in physical and in mathematical terms due to its dependency on a great number of parameters. The aim of this work is to reduce timbre multidimensionality in order to obtain a simple and accurate tool for timbre classification starting from sound signals. In his classical work, J. Grey (1977) determined a three-dimensional (3D) space in which different instrument sounds were mapped. This space was produced by applying multidimensional scaling to subjective similarity judgments between the timbres of 16 traditional instruments. The interpretation of the coordinates explained the main factors affecting timbre discrimination. The first dimension can be interpreted as spectral distribution of the energy, the second dimension as the presence of synchronicity in the attack stage through the harmonics and the third is connected with the presence of high frequency inharmonic noise with low amplitude, during the attack segment. This space cannot be used directly to classify timbres. In fact the classification of a new timbre requires the repetition of all the psychoacoustic experiments with listening groups. Wessel (1979) proposed a method to compute the brightness of a sound starting from its spectrum and showed

that brightness is correlated with the principal axis of the timbre space. However no method is presently available to compute the coordinates of a timbre in the other dimensions. The determination of the best parameters to classify timbres is still an open problem. In (DePoli & Tonella,1993; De Poli et al., 1993) we tried to classify timbres with the Grey parameters throughout 3D Kohonen maps. Feiten (1992) pre-processed the spectrum by a simplified ear model before training a Kohonen map for timbre spatialization. Leman(1991, 1992) employed an ear model and Kohonen map in order to realize a good ear-brain combination as ontological foundation to musicology.

Similarly to speech analysis, Fourier analysis in combination with filter-bank techniques or cepstrum analysis have been used for many years in order to reduce timbre representation complexity. Recently, in speech analysis and recognition, the introduction of auditory models (Cooke et al., 1993) which explicitly consider non-linear phenomena occurring in the perception mechanism, has given promising results especially when speech is highly degraded by noise (Hunt & Levebvre, 1988). On the other hand, Neural Networks (**NN**) (Rumelhart & McClelland, 1986) have already proved their classification capability in various pattern recognition tasks. For these reasons, in the timbre classification system being considered, auditory modelling and neural network techniques are combined together. In particular S. Seneff's auditory modelling (Seneff, 1988) was used in the analysis stage, while a bidimensional Kohonen Self Organizing Map (**SOM**) (Kohonen, 1984, 1990) was used in the classification stage.

AUDITORY MODELLING MOTIVATIONS

Every sound classification and recognition task is preceded by an acoustic analysis front-end, aiming to extract significant parameters from the time signal. Normally, this analysis is based on a model of the signal or of the production mechanism. Short-Time Fourier Transform (STFT), Cepstrum, and other related schemes (Rabiner & Shafer, 1978) were all developed strictly considering physical phenomena that characterise the speech waveform and are based on the quasi-periodic model of the signal. On the other hand LPC technique and all its variants were developed directly by modelling the human speech production mechanism. Even the most simple physical models of musical instruments are highly non linear; thus they are not suitable to be used for analysis purpose. In music research and speech recognition the focus is on perceived sound rather than on physical properties of the signal or of the production mechanism. To this purpose, lately, almost all these analysis schemes have been modified by incorporating, at least at a very general stage, various perceptual-related phenomena. Linear prediction on a warped frequency scale STFT-derived auditory models, perceptually based linear predictive analysis, are a few simple examples of how human auditory perceptual behaviour is now taken into account while designing new signal representation algorithms. Furthermore, the most significant example of attempting to improve acoustic front-end

with perceptual related knowledge, is given by the Mel-frequency cepstrum analysis of speech (Davies & Mermelstein, 1980), which transforms the linear frequency domain into a logarithmic one resembling that of human auditory sensation of tone height. In fact, Mel Frequency Cepstrum Coefficients (MFCC) are almost universally used in the speech community to build acoustic front-end for Automatic Speech Recognition (ASR) systems.

All these sound processing schemes make use of the "short-time" analysis framework (Rabiner & Shafer, 1978). Short segments of sounds are isolated and processed as if they were short segments from a sustained sound with fixed properties. In order to better track dynamical changes of sound properties, these short segments which are called *analysis frames*, overlap one another. This framework is based on the underlying assumption that, due to the mechanical characteristics of the generator, the properties of the signal change relatively slowly with time. Even if overlapped analysis windows are used, important fine dynamic characteristics of the signal are discarded. Just for that reason, but without solving completely the problem of correctly taking into account the dynamic properties of speech, "velocity"-type parameters (simple differences among parameters of successive frames) and "acceleration"-type parameters (differences of differences) (Furui, 1986) have been recently included in acoustic front end of almost all commercialized ASR systems. The use of these temporal changes in speech spectral representation (i.e. ΔMFCC, $\Delta\Delta$MFCC) has given rise to one of the greatest improvements in ASR systems.

Moreover, in order to overcome the resolution limitation of the STFT (due to the fact that once the analysis window has been chosen , the time frequency resolution is fixed over the entire time-frequency plane, since the same window is used at all frequencies), a new technique called Wavelet Transform (WT), characterized by the capability of implementing multiresolution analysis, has been recently introduced (Kronland-Martinet & Grossmann, 1991). With this new processing scheme, if the analysis is viewed as a filter bank, the time resolution increases with the central frequency of the analysis filters. In other words, different analysis windows are simultaneously considered in order to more closely simulate the frequency response of the human cochlea. As with the preceding processing schemes, this new auditory-based technique, even if it is surely more adequate than STFT analysis to represent a model of human auditory processing, it is still based on a mathematical framework built around a transformation of the signal, from which it tries directly to extrapolate a more realistic perceptual behaviour .

Cochlear transformations of acoustic signals result in an auditory neural firing pattern significantly different from the spectral pattern obtained from the waveform by using one of the above mentioned techniques. In other words, spectral representations such as the *spectrogram*, a popular time-frequency-energy representation of speech, or either the *wavelet spectrogram*, or *scalogram*, obtained using the above described multiresolution analysis technique are quite different from the true *neurogram*. In recent years, basilar membrane, inner cell and nerve fiber behaviour have been extensively studied by auditory physiologists and neurophysiologists and knowledge about the human auditory pathway has become more accurate. A number of studies have been accomplished and a

considerable amount of data has been gathered in order to characterize the responses of nerve fibers in the eighth nerve of the mammalian auditory system using tone, tone complexes and synthetic speech stimuli. Phonetic features probably correspond in a rather straightforward manner to the neural discharge pattern with which speech is coded by the auditory nerve.

Various auditory models which try to *physiologically* reproduce the human auditory system have been developed in the past (Greenberg, 1988), and, even if they must be considered only as an approximation of physical reality, they appear to be a suitable system for identifying those aspects of the acoustic signal that are relevant for automatic speech analysis and recognition. Furthermore, with these models of auditory processing, perceptual properties can be re-discovered starting not from the sound pressure wave, but from a more internal representation which is intended to represent the true information available at the eighth acoustic nerve of the human auditory system.

Advanced Auditory Modelling (AM) techniques not only include "perception-based" criteria instead of "production-based" ones, but also overcome "short-term" analysis limitations, because they implicitly retain dynamic and non-linear sound characteristics. For example, the dynamics of the response to non-steady-state signals, as also "forward masking" phenomena, which occur when the response to a particular sound is diminished as a consequence of a preceding, usually considerably more intense signal, are important aspects captured by efficient auditory models (Seneff, 1988). Various evidences can be found in the literature (Zue et al., 1989; Cosi et al., 1990) suggesting the use of AM techniques, instead of the more classical ones, in building speech analysis and recognition systems. Especially when speech is greatly corrupted by noise (Cosi, 1993), the effective power of AM techniques seems much more evident than that of classical digital signal processing schemes.

AUDITORY PROCESSING

The computational scheme proposed in this paper for modelling the human auditory system, apart from small differences regarding the filter bank designing strategy, refers essentially to the joint *Synchrony/Mean-Rate* (S/M-R) model of *Auditory Speech Processing* (ASP), recently proposed by S. Seneff (1988), resulting from her important studies on this matter (Seneff, 1984, 1985, 1986). The overall system structure, whose block diagram is illustrated in Fig. 1, includes three stages: the first two deal with peripheral transformations occurring in the early stages of the hearing process while the third one attempts to extract information relevant to perception. The first two blocks represent the *periphery* of the auditory system. They are designed using knowledge of the rather well known responses of the corresponding human auditory stages (Kiang et al., 1965; Sinex & Geisler, 1983). The third unit attempts to apply an effective processing strategy for the extraction of important speech properties like an efficient representation for locating transitions between phonemes useful for speech segmentation, or spectral lines related to formants useful for phoneme identification.

The signal, band-limited and sampled at 16 kHz, is first pre-filtered through a set of four complex zero pairs to eliminate the very high and very low frequency components. The signal is then analyzed by the first block, a *40-channel critical-band linear filter bank*. Fig 2 shows the block diagram of the filter bank which was implemented as a cascade of complex high frequency zero pairs with taps after each zero pair to individual tuned resonators. Filter resonators consist of a double complex pole pair corresponding to the filter center frequency (CF) and a double complex zero pair at half its CF. Although a larger number of channels would provide superior spatial resolution of the cochlear output, the amount of computation time required would be increased significantly. The bandwidth of the channels is approximately 0.5 Bark, which corresponds to the width of one critical band, that is, a unit of frequency resolution and energy integration derived from psychophysical experiments (Zwicker & Terhardt, 1990). Filters, whose transfer functions are illustrated in Fig. 3, were designed in order to optimally fit physiological data like those observed by N.Y.S. Kiang et al. (1965). Frequencies and bandwidths for zeros and poles of each filter were designed almost automatically by an interactive technique developed by S. Seneff and described in her Thesis (Seneff, 1985). As for the mathematical implementation of the 40-channel critical-band filter bank, it is described on the top of Fig. 4, where serial (FIR) and parallel (IIR) branches are illustrated in detail.

< insert Fig. 1 >

< insert Fig. 2 >

< insert Fig. 3 >

< insert Fig. 4 >

The second stage of the model is called the *hair cell synapse model* (see Fig. 1). It is non-linear and is intended to capture prominent features of the transformation from basilar membrane vibration, represented by the outputs of the filter bank, to probabilistic response properties of auditory nerve fibers. The outputs of this stage, in accordance with S. Seneff (1988), represent the probability of firing as a function of time for a set of similar fibers acting as a group. Four different neural mechanisms are modelled in this non-linear stage. A *half-wave rectifier* is applied to the signal in order to simulate the high level distinct directional sensitivity present in the inner hair cell current response. This rectifier is the first component of this stage and is implemented by the use of a saturating non linearity. The instantaneous discharge rate of auditory-nerve fibers is often significantly higher during the first part of acoustic stimulation and decreases thereafter, until it reaches a steady-state level. The *short-term adaptation* module, which controls the dynamics of this response to non steady-state signals which is due to the neurotransmitter release in the synaptic region between the inner hair cell and its connected nerve fibers, is simulated by the so called "membrane

model", which was conceived following the work by R.S. Goldor (1985). This model influences the evolution of the neurotransmitter concentration inside the cell membrane. The third unit implements the observed *gradual loss of synchrony* in the nerve fiber behaviour as the stimulus frequency is increased, and it is implemented by a simple low-pass filter. The last unit is called *Rapid Adaptation* and implements the very rapid initial decay in discharge rate of auditory nerve-fibers occurring immediately after acoustic stimulation onset, followed by the slower decay, due to short-term adaptation, to a steady state level. This module performs "Automatic Gain Control" and is essentially inspired by the refractory property of auditory nerve fibers (Swami & Swami, 1983). The final output of this stage is affected by the ordering of the four different components due to their non-linear behaviour. Consequently, as underlined by S. Seneff (1988), each module is positioned by considering its hypothesized corresponding auditory apparatus (see Fig. 1). As for the mathematical implementation of the four modules of the hair-cell synapse model, this is illustrated in the central block of Fig. 4. Fig. 5 describes the result of the application of the model to a simple 1000 Hz sinusoid. Left and right plots refer respectively to the global 60 ms stimulus and to its corresponding first 10 ms window in different positions along the model.

The third and last stage of the model, mathematically described on the bottom of Fig. 4, is formed by the union of two parallel blocks: the *Envelope Detector* (ED), implemented by a simple low-pass filter, which, in accordance with S. Seneff (1988), by smoothing and down sampling the second stage outputs, appears to be an excellent representation for locating transitions between phonemes, thus providing an adequate basis for phonetic segmentation, and the *Synchrony Detector* (SD), whose block diagram as applied to each channel is shown in Figure 6, which implements the known "phase locking" property of the nerve fibers. This block enhances spectral peaks due to vocal tract resonances. In fact, auditory nerve fibers tend to fire in a "phase-locked" way responding to low frequency periodic stimuli, which means that the intervals between nerve fibers tend to be integral multiples of the stimulus period. Consequently, if there is a "dominant periodicity" (a prominent peak in the frequency domain) in the signal, with the so called *Generalized Synchrony Detector* (GSD) processing technique (Seneff, 1984, 1985), only those channels whose central frequencies are closest to that periodicity will have a more prominent response.

< insert Fig. 5 >

< insert Fig. 6 >

In Fig. 7, an example of the output of the model, as applied to a clean BClarinet sound is illustrated for the envelope (a) and the synchrony (b) detector module respectively. The use of the GSD parameters (Fig. 7b) allowed to produce spectra with a limited number of well defined spectral lines and this represents a good use of sound knowledge according to which harmonics are sound parameters with low variance. Due to the high level of overlapping of filter responses, envelope

parameters (Fig. 7a) seem less important for classification purposes but maintain their usefulness in capturing very rapid changes in the signal. Thus they should be more significant considering transient sounds instead of sustained ones.

< insert Fig. 7a & 7b >

In order to prove the robustness of auditory parameters, the same BClarinet sound with gaussian random noise superimposed at a level of 5 dB S/N ratio was analyzed. It is evident, from a comparison between Figures 8a and 8b that the harmonic structure is well preserved by the GSD parameters, even if the sound is greatly corrupted by quite a relevant noise. Figure 9 shows, in time domain, the great difference of a portion of the signal in the clean and noisy conditions.

< insert Fig. 8a & 8b >

< insert Fig. 9 >

The computation time of the joint S/M-R model of ASP is about 100 times real-time on a SUN SparcStation. The system structure is suitable for parallelization with special purpose architectures and accelerator chips. At the present time the model has been also implemented on a floating-point Digital Signal Processor and the obtained computation time is about 10 times real-time (Cosi et al., 1991).

SELF ORGANIZING MAP

Due to its topology-preserving feature and also to its pattern-matching capability a bidimensional Kohonen SOM was chosen for the classification stage. The topology-preserving feature of SOMs let a multidimensional space, in which a particular stationary probability function p(x) is defined, be represented as a two-dimensional or even three-dimensional image, by using minimum variance similarity criteria between space vectors thus giving rise to excitatory or inhibitory interactions between the different nodes of the map. In other words in the SOM, the number of nodes or neurons is substantially lower than the number of vectors used for training the map. In fact, each node represents a cluster of the input space, in the sense that each vector of that cluster excites always the same neuron. On the other hand, considering the pattern-matching ability of the SOM, the ratio between nodes and vectors is reversed, thus leading to a better classification capability and, at the same time, to a higher level of 'energy continuity' among excited neurons. In other words, the map is able to generalize similarity criteria even to vectors not utilized during the training phase. The net topology can be chosen following certain criteria originally proposed by Kohonen. A rectangular

topology seems to be the best, considering the orientation of the vectors or weights of the network versus the vectors to be classified during the learning phase. As for the weight initialization phase, an algorithm extracting the first two dominant eigenvectors within the space of vectors being considered was used and, successively, all the weights were initialized with a random combination of such vectors. This should lead to start the learning phase in a very effective position in order to better reach the convergence of the algorithm.

EXPERIMENT

A limited set of sound samples played by classical musical instruments, representative of the timbre range of a typical orchestra, was utilized for the classification experiment. The CD Library of McGill University was utilized in order to extract the following target sound samples:

Timbre	**CodeName**	**McGill reference**
Alto Trombone	**(ATrbne)**	**CD #2 McGill**
B Clarinet (B flat)	**(BClrto)**	**CD #2 McGill**
Tumpet (B)	**(BTrump)**	**CD #2 McGill**
Flute C (no vibrato)	**(CFluteNV)**	**CD #9 McGill**
Violoncello (no vibrato)	**(CelloNV)**	**CD #9 McGill**
France Horn	**(FHorn)**	**CD #2 McGill**
Oboe	**(Oboe)**	**CD #2 McGill**
Pipe Organ	**(POrgan)**	**CD #10 McGill**
Piano HamburgSteinway	**(Piano)**	**CD #9 McGill**
Tenor Sax	**(TSax)**	**CD #3 McGill**
Violin (no vibrato)	**(ViolinNV)**	**CD #9 McGill**
Viola (no vibrato)	**(VlaNV)**	**CD #9 McGill**

Following the original Grey experiment, timbres were selected with pitch C4 (C fourth octave) corresponding to 261.6 Hz. This note belongs to the pitch range of all considered instruments. The signal was sampled with 16 bits at a sampling frequency of 32 KHz and successively undersampled at 16 KHz by software. The reason of this undersampling was due to the fact that the Seneff auditory processing, and in particular the filter-bank stage, was developed for speech signal sampled at 16 KHz. In the experiment being described this corresponds to use up to 30 harmonics for each sound.

The introduced quality loss is not dramatic in our classification task because the significant harmonics are sufficiently represented. According to Grey's psychoacustic finding, temporal characteristics of the attack seem to retain most of the information for timbre discrimination. Thus only 300 ms for each sound, corresponding to the attack and a small sustained phase, were selected. As illustrated in Fig. 2, GSD parameters retain relevant spectral information while envelope parameters seem less suitable for a classification task and consequently they were not used to train and test the network. In a first experiment each filter-bank channel was sampled every 5 ms giving rise to 2400 data for each item (300 ms / 5 ms x 40 channels). The SOM was designed considering a value less than 30% for the ratio of weight_vector dimension and the total number of nodes of the network. A rectangular structure was chosen with 36x24 nodes leading to a network of 864 nodes. Due to the high complexity of such a structure, classification results could be considered satisfactory only after a very long training phase. For such a reason, a second experiment was set up in which, while maintaining the same network complexity, input data were strongly simplified. Following neurophysiological findings, neurons, after a pulse was fired, have a latency period of 1-3 ms in which no more pulses can be fired. Thus the firing sequences of a single neuron can be synchronised to the input signal if and only if the input sound frequency is less than 1 KHz. With higher frequencies, while single neurons cannot retain their siynchronization property, on the contrary, groups of adjacent neurons can still be synchronized with the input sound waveform if a mean firing rate is considered for that group. However, over 4-5 KHz, the synchronization property is completely lost even in a mean sense. In the Seneff auditory model this phenomenon is captured by the "synchrony reduction" module, as indicated in Fig. 1, which is essentially implemented by a low pass-filter. Following these considerations only the first 20 channels of the auditory model were considered spanning a frequency band ranging from 130 to 1300 Hz. Instead of using all 300 ms sampled every 5 ms for each input timbre, only 3 vectors corresponding to the minimum, the mean and the maximum value of the corresponding GSD parameters were considered, leading to a total of 60 elements for each item. The discrimination power of the SOM in this experiment was considerably reduced, and, for example, very dynamic timbres such as Piano and Sax, were often confused. A third experiment was finally designed in which, instead of considering only 3 vectors relating to minimum, mean and maximum GSD parameters, as in the second experiment, a total of 6 vectors obtained for each item by sampling, independently of GSD parameters, every 50 ms the complete 300 ms duration, were used to train and test the network. The same frequency range of the previous experiment (130-1300 Hz, 20 channels) and the same SOM structure were considered. As already underlined, the network weights were initialized with a linear combination of the first two dominant autovectors of the input pattern space. Learning was divided in two phases: a first general training phase and a fine-adjustment phase. The aim of the first phase is to substantially order the map of neurons while the aim of the second one is to continuously improve and refine the discrimination power of the network.

The following formulas rule the learning algorithm:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \mathbf{hc}_i(t) \times \left| \mathbf{x}(t) - \mathbf{m}_i(t) \right| \qquad (1)$$

$$\mathbf{hc}_i(t) = \begin{cases} \alpha(t) & \text{with } i \in Nc(t) \\ 0 & \text{with } i \notin Nc(t) \end{cases} \qquad (2)$$

where **m** in (1) represents the map weights learning function at step t+1, **x** represents the input vector and **hc** represents the excitation function which remains constant at step t for the whole excitation near space set defined by **Nc**(t). The only limitation given for **a** is that it shall be continuously decaying in time in order to grant for the learning algorithm to converge and it is selected such that 0 < **a**(t) <1. Typically **a**(t) decays to zero after a preselected number of training set presentations. The exact decay schedule is not critical, but Kohonen has noted that the convergence of the algorithm consists of the following two distinct phases: initial formation of map order and final convergence. The learning phase is usually chosen as a piece wise linear decay with the second phase lasting 10 to 100 times longer than the first phase. The same decaying characteristic applies to the near space set function Nc and in particular a linear decaying function was chosen. In fact, starting from an initial radium of action Nco the near space function linearly reduces, step by step, to 1 which represents the adjacent set of neurons. Considering, as usual, the same alpha for each node of the map, the only adjustable parameters during the learning phase are: Nco, hco and the learning time or the number of iterations.

Following these considerations, a high value of hco, relatively to that assigned in the second phase, and a value for Nco equal to half the diagonal of the map were set in order to obtain a rapid ordering of the vector of weights during the first learning iterations while letting the SOM to simultaneously maintain its high level of generalization capability. Usually this phase does not require a great number of iterations, and, in this experiment, we stopped after 1000 iterations. In the second fine-adjusting phase, the excitation function shall be lower both in intensity and in radium of action in order to obtain a better refining and calibration of the map built in the first phase, thus improving its classification power within the input space vectors without reducing its generalization feature. In particular a low value is chosen for hc0 and half the value of the previous phase is chosen for Nco. The number of iterations for this phase, as indicated above, shall be quite high and the chosen value was 5000.

RESULTS

In the following table the various parameters used while training the network are summarized and, in particular, in the fifth column, the quantization errors per sample in both phases are indicated.

Phase	hco	Nco	N. of iterations	QError/sample
1	**0.3**	**20**	**1000**	**0.199824**
2	**0.08**	**10**	**5000**	**0.000016**

Figure 10 illustrates a typical answer map (BClarinet) of the self organizing network showing the different excitation values (Z axis) of its nodes. The image shown in Fig. 11 illustrates, all twelve overlapped excitation planes, computed over a fixed threshold corresponding to 95% of the global dynamic excitation range. Each instrument was clustered in a single region of the map which contains also the input best-matching pattern. All twelve instruments can be quite well distinguished by the map. Moreover, the map is interestingly organized from a topological point of view. In fact, information is coded by the map in a way satisfying certain human mental criteria of timbre classification in space. For example, neurons corresponding to wind instruments and those corresponding to string ones are contiguous in the map. Moreover, Piano and Pipe Organ are not arranged in any of these two categories. In conclusion, auditory modelling and Kohonen self organizing map lead us to obtain a bidimensional representation of timbre space.

< insert Figure 10 >

< insert Figure 11 >

In order to consider the real classification capability of the obtained SOM, a test experiment was set up with two different aims. The first one was intended to test the recognition capability of the network with samples of the same instruments used during training but coming from a different source, while, the second was intended to verify the generalization capability of the network if used to classify different instruments from those used for the learning phase. Test samples were the following:

Timbre	CodeName	Reference
E Clarinet	**(EClarinet)**	**CD #2 McGill**

Bachian Trumpet	**(BhTpt)**	**CD #2 McGill**
Bass Clarinet	**(BassClrto)**	**CD #2 McGill**
Tenor Trombone	**(TTrbne)**	**CD #2 McGill**
English Horn	**(EHorn)**	**S3 G.E.M. synt.**
Oboe2	**(Oboe2)**	**S3 G.E.M. synt.**
Baritone Sax	**(BSax)**	**CD #3 McGill**

Comparing Fig. 10 and Fig. 12, referring to the answer of the network to BClarinet (learning) and EClarinet (test) and Fig. 13a and Fig. 13b relative to BTrumpet (learning) and Bachian Trumpet (test) it is worth noticing the high level of similarity in the answer maps.

< insert Figure 12 >

< insert Figures 13a & 13b >

Moreover, stimulating the network with a new timbre, sensibly different from those used for learning, produces an activation of a neuronic area located near that excited by the most similar training timbre. This similarity is not only a numerical property of the describing parameters but is also verified by listening to the timbres. In the case of the Tenor Trombone, the best-match answer, shown in Figure 14, was in the middle between the best-match of the BTrumpet and that of the Alto Trombone. By listening to these two instruments, the interpolating answer spontaneously given by the network, among the two learning timbres and the target test one, can be strongly appreciated.

< insert Figure 14 >

In Speech Technology, auditory modelling techniques have already shown their superiority versus more classical ones principally when speech is greatly corrupted by noise (Cosi, 1993). In order to verify the robustness of the proposed auditory recognition set of parameters for timbre classification, the complete system was tested with some instruments previously utilized while training the network corrupted by zero-mean gaussian noise at various SNR levels. In Fig. 15 the answer map of the network is shown for the BClarinet with 0 db SRN noise. This map is very similar to the response of the clean tone shown in Fig. 10. Both in the processing phase and in the parameter synthesis phase a great tolerance to the noise can be observed granting to this kind of analysis a high level of competitivity with respect to other more classical analysis techniques like Fourier Analysis, Cepstrum, Lpc etc.... . In Fig. 16 it is possible to appreciate differences between the parameter representing the clean BClarinet set and the noisy (0 db SRN noise level) one. The SOM further reduces this difference and correctly recognizes the timbre.

< insert Figure 15 >

< insert Figure 16 >

CONCLUSIONS

These results are obviously not extendible to all instruments, principally due to the fact that the present learning data-set was rather limited to be statistically significant. However, they are sufficient to demonstrate the usefulness of the chosen approach in order to find a possible definition of an hypothetical timbrical space. To reach this goal, a more complete learning set of instruments will be taken into consideration in the near future.

REFERENCES

Cooke M., Beet S. & Crawford M. (Eds.) (1993). *Visual Representation of Speech Signals*. Cichester: John Wiley & Sons.

Cosi P. (1993). Auditory Modelling for Speech Analysis and Recognition. In M. Cooke, S. Beet and M. Crawford (Eds.), *Visual Representation of Speech Signals* (pp.205-212). Cichester: John Wiley & Sons.

Cosi P., Bengio Y. & De Mori R. (1990). Phonetically-Based Multi-Layered Neural Networks for Vowel Classification. *Speech Communication, 9(1),* 15-29.

Cosi P., Dellana L., Mian G.A. & Omologo M. (1991). Auditory Model Implementation on a DSP32C-Board. *Proceedings of GRETSI-91,* Juan Les Pins.

Davis S.B. & Mermelstein P. (1980). Comparison of Parametric Representation of Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP 28(4),* 357-366.

De Poli G. & Tonella P. (1993). Self-Organizing Neural Networks and Grey's Timbre Space. *Proceedings of 1993 International Computer Music Conference* (pp. 441-444). Tokyo: Waseda University.

De Poli G., Prandoni P. & Tonella P. (1993). Timbre clustering by self-organizing neural networks. *Proceedings X Colloquium on Musical Informatics*. Milan: University of Milan.

Feiten B., Frank R. & Ungvary T. (1991). Organizations of sounds with neural nets. *Proceedings of 1991 International Computer Music Conference* (pp. 441-444). San Francisco: ICMA.

Furui S. (1986). Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP 34(1),* 52-59.

Goldhor R.S. (1985). *Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features* (RLE Technical Report N. 505). Cambridge: MIT Press.

Greenberg S. (Ed.) (1988). Representation of Speech in the Auditory Periphery. *Journal of Phonetics, Special Issue, 16(1).*

Grey J.M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of Acoustical Society of America, 61(5),* 1270-1277.

Hunt M.J. & Lefebvre C. (1988). Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 215-218). New York, N.Y.: IEEE Press.

Kiang N.Y.S., Watanabe T., Thomas E.C. & Clark L.F. (1965). *Discharge patterns of single fibers in the cat's auditory-nerve fibers.* Cambridge, MA: MIT Press.

Kohonen T. (1984). *Self-organization and associative memory*. Berlin: Springer Verlag.

Kohonen T.. (1990). The Self-Organizing Map. *Proceedings of the IEEE, 78(9),* 1464-1480.

Kronland-Martinet R. & Grossmann A. (1991). Application of time-frequency and time-scale methods (wavelet transform) to the analysis, synthesis and transformation of natural sounds. In G. De Poli, A. Piccialli & C. Roads (Eds.), *Representations of musical signals* (pp. 45-85). Cambridge: MIT Press.

Leman M. (1991). Emergent Properties of Tonality Functions by Self-Organization. *Interface, 19(2-3),* 85-106.

Leman M. (1992). Tone Context by Pattern Integration over Time. In D. Baggi (Ed.), *Computer generated music* (pp. 117-137). Los Alamitos, CA: IEEE Computer Society Press.

Rabiner L.R. & Shafer R.W. (1978). *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall.

Seneff S. (1984). Pitch and spectral estimation of speech based on an auditory synchrony model. *Procedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 36.2.1-36.2.4). New York, N.Y.: IEEE Press.

Seneff S. (1985). *Pitch and spectral analysis of speech based on an auditory synchrony model* (RLE Technical Report No. 504). Cambridge: MIT Press.

Seneff S. (1986). A computational model for the peripheral auditory system: application to speech recognition research. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 37.8.1-37.8.4). New York, N.Y.: IEEE Press.

Seneff S. (1988). A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing. *Journal of Phonetics, 16(1),* 55-76.

Sinex D.G. & Geisler C.D. (1983). Responses of auditory-nerve fibers to consonant-vowel syllables. *Journal of Acoustical Society of America, 73,* 602-615.

Swami D.H. & Swami A. (1983). The transmission of signals by auditory-nerve fiber discharge patterns. *Journal of Acoustical Society of America, 74,* 493-501.

Wessel D. (1979). Timbre Space as Musical Control Structure. *Computer Music Journal*, 3(2), 45-52.

Zue V.W., Glass J., Philips M. & Seneff S. (1989). Acoustic Segmentation and Phonetic Classification in the SUMMIT System. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 389-392). New York, N.Y.: IEEE Press.

Zwicker E. & Terhardt E. (1980). Analytical expression for critical-band rate and critical bandwidth ad a function of frequency. *Journal of Acoustical Society of America, 68(5),* 1523-1525.

CAPTIONS TO THE FIGURES

Fig. 1. Block diagram of the joint Synchrony/Mean-Rate model of Auditory Speech Processing.

Fig. 2. Block diagram of the 40-channel critical-band linear filter bank.

Fig. 3. Frequency responses of the 40-channel critical-band linear filter bank.

Fig. 4. Mathematical framework of the joint Synchrony/Mean-Rate model of Auditory Speech Processing.

Fig. 5. Result of the application of the four modules implementing the hair-cell synapse model to a simple 1000 Hz sinusoid. Left and right plots refer to the global 60 ms stimulus and to its corresponding first 10 ms window, in different positions along the model.

Fig. 6. Block diagram of the Generalized Synchrony Detector (GSD) module.

Fig. 7. Output of the model, as applied to a clean B Clarinet sound: (a) envelope, (b) synchrony.

Fig. 8. Synchrony parameter output of the analysis of the same B Clarinet of Fig 7b, superinposed with a gaussian random noise at a level of 5 db S/N ratio.

Fig. 9 Time domain representation of a portion of the B Clarinet signal in (a) clean and (b) noisy conditions (5 db SNR).

Fig. 10. Answer map of the SOM to the B Clarinet stimulus.

Fig. 11. Resulting map. The regions identify the timbres and are computed over a fixed threshold corresponding to 95% of the global dynamic excitation range.

Fig 12. Answer of the map to the E Clarinet test timbre, to be compared with fig. 10.

Fig. 13. Answer of the map to: (a) B Trumpet (learning timbre); (b) Bachian Trumpet (test timbre).

Fig. 14. Activation map of Alto Trombone test timbre. The best match answer is in the middle between the best-match of B Trumpet and Tenor Trombone.

Fig 15 Answer map of the B Clarinet timbre corrupted with noise at 0 db SNR level.

Fig. 16. Comparison of the parameters representing BClarinet timbre in clean and noisy conditions. Six equally spaced samples from synchrony output of each channel are successively shown along the X axis.