Timbre clustering by self–organizing neural networks

Giovanni De Poli, Paolo Prandoni, Paolo Tonella
CSC—DEI University of Padova, Via Gradenigo 6a,
35131 Padova, Italy
tel: +39-49-8287631, fax: 8287699, email: depoli@dei.unipd.it

Introduction

Timbre is that attribute of auditory sensation which allows listeners to rate as different sounds presented in ways altogether similar with respect to intensity, duration, and pitch. The similarity between two sounds can be characterized in physical and mathematical terms only with difficulty because it is a subjective attribute and it depends on a large number of parameters.

J. M. Grey, in his classic work [1], introduces the concept of "timbre space", a means with which he conveyed the vague notion of *similarity* between timbres into the precise notion of a *metric rule* in a three–dimensional space. This space was the result of a multidimensional scaling applied onto a large set of subjective similarity ratings obtained in experimental sessions. A physical interpretation of the reasons for such a spatial distribution was also provided.

In this work we will try to follow the lines of Grey's experiment, but using a neural network as the means to rate timbre differences and to transform them into metric relations. Neural nets have been used already in this field of research [4]; the aim of our work is to simplify timbre multidimensionality, following the lines of Grey's experiment, and to obtain similar results in terms of clusterization and of timbre space. The tools we use are Kohonen self-organizing neural networks (KNN): they show an ability to correctly classify items outside the training set, and they prove highly insensitive to noise. Another reason for their use comes from neurophysiology: the principles of self–organization Kohonen proposes were derived from a model of the cerebral cortex; it is therefore interesting to compare our results with those obtained by Grey starting from subjective judgments.

Grey timbre space

J. Grey's experiments at Stanford University in 1975 were aimed at a thorough investigation in the field of musical timbre. He considered the following synthetic test sounds, obtained from a spectral analysis of recorded true instruments: bassoon (BN), normal cello (S2), E flat clarinet (C1), flute (FL), french horn (FH), english horn (EH), muted cello (S3), oboe (O1, O2), cello sul ponticello (S1), soprano sax (X1, X2, X3), trombone (TM), and trumpet (TP); during the experimental sessions, a group of musically trained listeners provided subjective ratings of the differences between tones. These perceptual data were averaged and arranged in a *similarity matrix*. This matrix was then processed using a multidimensional scaling (MDS) algorithm; the result was the distribution of the timbres in an n–dimensional space; at the same time, the matrix was analyzed using a hierarchical clustering algorithm based on the diameter method, and the result was an independent timbre grouping. The most interesting result was that the clusters thus obtained enclosed timbres located at low distance in the three-dimensional timbre space produced by the MDS algorithm. Grey proposed a physical interpretation for the three dimensions, showing the first dimension to be related to the spectral distribution of energy, the second dimension to the presence of synchronicity in the attack stage through the harmonics, and the third dimension to the presence of high frequency inharmonic noise with low amplitude, during the attack segment.

As timbre, in its definition, is the feature which differentiates sounds under the same conditions of pitch, intensity and duration, Grey first had to equalize the sample sounds with regard to those parameters. This equalization stage featured many psycho-acoustic sessions aimed at the comprehension of the phenomena underlying subjective perception and finally produced a set of sound samples where the timbral issue was the only discriminant.

We used the same data Grey used; they consist of a line–segment approximation of the true evolutions both in frequency and in amplitude of the sound partials as they resulted from a heterodyne analysis of the equalized analog signal.

Self-organizing neural networks

Kohonen's neural networks are inspired by the process that seems to be responsible for the map–like organization of the cerebral cortex; the observable organization of the cortex neurons shows that some zones of the cortex are sensitive to certain stimulations and indifferent to others. The basic mechanism, believed to account for this process of self–organization of the brain, is called the *Hebb principle*; it asserts that if a particular neuron has a considerable reaction to a stimulation, its synapses adapt themselves to the acting stimulus, and a lateral feedback process takes place; an activity bubble is formed in the close neighbourhood of the cell, while cells surrounding the bubble are inhibited. In this way a clusterization process is generated, and the activity bubbles come to be located in different zones of the neural map according to the stimulations to which they are most sensitive.

T. Kohonen formalized this process into a simple numerical algorithm [5]. The arising neural model shows surprising properties of *self organization*: its inner structure modifies to become an n–dimensional projective model of the m–dimensional probability space from which the input samples come. As it is generally $n < m$ (while $n = 1, 2$), the neural map performs a *feature extraction*: along the n axes of the map those input features are mapped which have the largest numerical variance. This also explains the good behavior these models exhibit in the presence of noise : KNN can maximize the amount of information stored because they organize complying to two conflicting requirements: to increase the variance of the outputs of all neurons, with the purpose of recognizing the main features of the inputs; and to introduce a certain degree of redundancy, with the purpose of obtaining correct answers even in presence of noise [6].

3D Kohonen Nets

The first experiments we carried out referred directly to Grey's main result: the three–dimensional timbre space. To obtain results to be compared directly with Grey's, we planned using a 3D neural model, extending Kohonen's equations into the third spatial dimension.

The basic algorithm ruling the self–organization process is the following: at each training step t a new input vector $x(t)$ is presented to the net; the neuron i whose inner values vector m_i is closest to the input vector x is selected as the *best matching neuron*. Different metric rules can underlie this matching criterion; for instance, we adopted the euclidean metric and the "city block distance". Around the best matching neuron a topological neighborhood $N_c(t)$ is defined as the spatial region where the actual learning occours: for all the indices $i \in N_c(t)$

$$m_i(t+1) = m_i(t) + \alpha(t)[x(t) - m_i(t)],$$

while the other neurons are not updated. Both N_c and α, the learning rate, decrease with time: the main structural changes in the net happen at the beginning of the process, when the neighborhood is large, while the remaining steps allow a fine tuning of the neuronal inner values. In our case the topological neighborhood is three–dimensional; its actual shape, cubic or spherical, is not essential, nor is it its shrinking rule which could be linear or exponential.

Literature, however, offers almost no example of three–dimensional forms of this equation. A mathematical analysis of KNN dynamics is extremely difficult; their properties were discovered through experimental simulation and practical applications. For this reason some preliminary experiments were performed to verify this extension.

A first task was to run the classical self–organization test [5, pag. 133] on our new structure: if the input vector x is a random variable with a stationary probability density function $p(x)$, then an ordered image of $p(x)$ will be formed onto the input weights m_i of the processing units. If an uniform distribution over a cubic region is used for the input space, an ordered cubic lattice of points should be obtained as the ultimate structure of the map. Kohonen suggested a minimal number of training steps of 500 times the number of neurons in the net [6, page 1496]; working with this too conservative an estimate. Probably, the more complex lateral interferences in the solid case require allowing a longer phase to structural modifications.

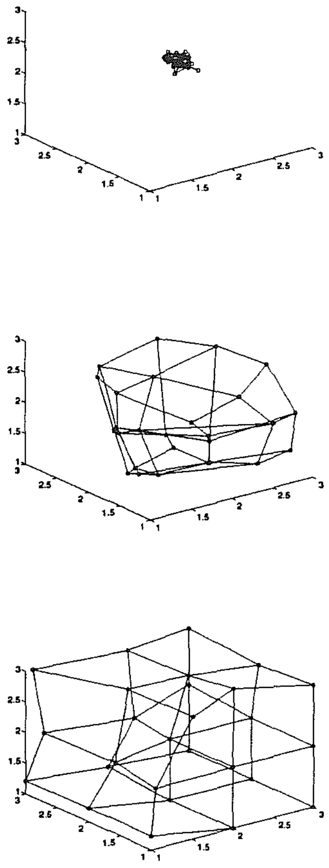

Figure 1: Three stages of evolution in the process of self ordering for a three–dimensional map. (a) startoff, (b) 1000000 iterations; (c) 2000000 iterations.

Figures 1.a, 1.b, and 1.c illustrate two significant steps of this expected evolution and show the validity of the three–dimensional model. In these plots the neuron values m_i are represented as circles in the same coordinate system of the input values, with lines connecting those units which are adjacent in the neuronal array. Figure 1.c shows how adjacent units end up with assuming adjacent values.

3D Clusterization

At this point we presented the net with numerical data obtained directly from those used by Grey in his listening sessions. We used samples of the frequency and amplitude evolutions of the sound signal in their line–segment approximation, so that all of the processing was made by the neural network; we also tried with data outcoming of a pre–processing of sounds, so that the network operated only at the most critical stage, the classification stage [8]. In all cases, the way we used Kohonen networks was somewhat fragile because only few learning samples were available with respect to the number of neurons in the network. A lack of samples causes a great sensitivity in the network final state to the starting random values of the weights. It happens that some of the neurons remain untouched by the learning process and the inner structural organization cannot unfold. It is possible to reduce this sensitivity to the initial conditions computing the mean of the different results obtained in a series of experiences, so that the effect of the initial random weights is canceled by the average [7, page 7]; we studied the convergence properties of the average, and we noted the presence of a final mean configuration with low values of variance, and, accordingly, of the relative error (3% is a typical value).

We obtained the best results using Grey's data directly: in the original line–segment representation all the necessary information to reconstruct the complete heterodine analysis of the timbre is contained and it could thus be used as an input to the neural network. We built an input file containing, for the first 20 harmonics of 12 signals, 10 amplitude envelope samples and 5 frequency envelope samples, and we fed it to a network sized $8 * 8 * 8 = 512$ neurons. After the learning phase, we computed the matrix of relative timbre distances from the spatial location of the best–matching neurons in the map; to this matrix, the same clusterization algorithm used by Grey was applied and the result was:

{(BN FH) [TP (FL S2)] [S1 S3]}
{[(C1 EH TM) O2] X3}

where the brackets split successive levels in the clusterization process. The analogies with Grey's results,

{[(BN TP) FH] [(S2 S3) (FL S1)]}
{C1 (EH X3)}
[O2 TM],

are encouraging; the mismatches are due more to the different times at which grouping occours, than to actual grouping differences.

Timbre Interpolation

The most critical point in the previous experients was the small number of learning samples, which were just the 14 original timbres. Besides averaging results, we tried to increase the number of samples: starting form Grey's original data, and following a line exploited by Grey himself, even though for other purposes [1, pages 75–95], [3], we considered each one of the possible couples of tones and obtained, by an algorithm of linear interpolation of the spectral envelopes, two artificial tones at 1/3 and 2/3 of the distance between the couple extremes. In so doing, we implicitly discarded all information about the frequency evolution of partials, adopting a coding of sounds which Grey calls the *fixed frequency model*. In the end we reached a data set of 200 units.

Clustering algorithms are generally very sensitive to little perturbations in the data points; therefore, even if the timbre space built by the net were not so much different from Grey's, the clustering algorithm could produce a completely mismatching result. Committing the accuracy judgement to a close match between clusterings seems too strict a requirement; however, since Grey does not provide the similarity matrices he used, a comparison between them and our distance matrix, which would be the best criterion, is impossible. In order to obtain a significant index of similarity for the timbre spaces, we exploited the information contained in [1, page 60], that is, the *order* in which clustering occours: 1.(S1, S3), 2.(O1, O2), 3.(BN, TP), 4.(X1, X2), 5.(C1, C2), 6.(X3, EH), 7.(FL, S1), 8.[(BN, TP), FH], 9.[(O1, O2), TM], 10. ...
We define the following *index of disorder*

$$D = \sum_{k=1}^{N} |(d_{k+1} - d_k) - |d_{k+1} - d_k||$$

where d_k is the euclidean distance

between *net* points which, in Grey's space, belong to the k-th cluster. Distances are computed according to the diameter algorithm, that is, the maximum among all possible distances in a group. Clearly $D \geq 0$, and $D = 0$ only if $d_0, \ldots, d_N$ are in ascending order. This latter case does not imply spatial equality, but grants a good similarity. When $D > 0$, distances are in a scrambled order; each inverted group contributes to D with a term proportional to the degree of inversion.

Experiments using the extended data set were run on both two– and three–dimensional networks; convergence now required a huge number of steps and rendered our work a trial to patience. Evaluations of the index of disorder proved consistent in all of the experiences: after a first widely varying phase, due to the large structural changes occourring at the beginning of the self–organization process, the index settled around a value of fifteen (fig. 2), which indicates that few timbres are slightly misplaced. In the analysis of the index behavior, no substantial differences were found between plane and solid networks, suggesting that two dimensional structures manage to locate timbres well enough; since there is a great saving in time, the bulk of the experiments was then conducted on plane nets sized $10 * 5 = 50$ neurons.

Another topological issue came to validate our results: Grey showed how artificial timbres obtained through linear interpolation are acoustically perceived as "half way" between the two timbres from which they originate. Similarly, such an artificial timbre is mapped by the net approximately half way along the line between the two "parents" (fig. 3); this is not obvious if we recall that KNN are *nonlinear projectors*: if linear relations in the input space are preserved, this means that the numerical form into which input samples are coded is well representative of their differences. We refer to this finding as to the inner "coherence" of the neural space.

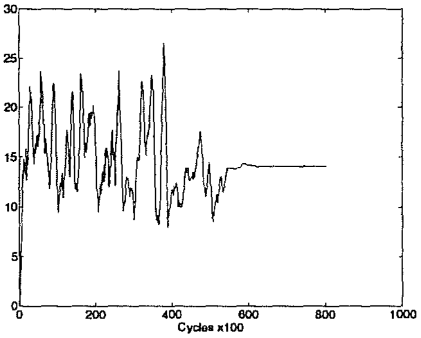

Figure 2: Typical evolution of the index of disorder

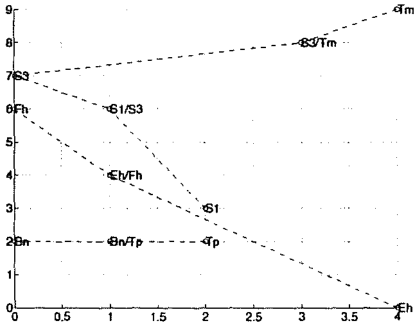

Figure 3: Neural space coherence

Conclusion

KNN are an interesting tool for the classification of a data set belonging to a space with large dimensionality, a task where classical tools for the extraction of high–variance features fail. We obtained maps which, even though different from Grey's timbre space, were not so far away. This suggests that the model underlying the artificial networks principles of self organization resembles, in a way, the features of biological neurons organization. We could ask ourselves, however, which are the legitimate expectations in such experiments. From the wide spatial separation between tones and the inner coherence of the neural space we can infer that the net is capable of efficiently handling a multidimensional feature like timbre; it would be unlikely, however, to have the same results as those obtained from a group of trained listeners. After all, Grey's model was developed in a peculiar environment, and need not be assumed as an absolute target. It would be interesting, among other things, to repeat the tests that led to it with a group of untrained people, or with sound samples of a better quality. In fact, it should be noticed that Grey's synthetic tones are of a low sound quality; future developments will surely profit of a higher quality sampling of the test timbres, and of an adequate signal pre-processing.

With regard to the data reduction techniques, deeper studies are under completion at Padova University; the best results have been obtained using pre–processing based upon Grey's observations, while Charbonneau's methods gave worse final configurations. A development of this work is the substitution, at the initial stage in the process of timbre recognition, of the heterodine analysis with a simulator of the human ear; in this way the operations made on input signals by biological organs and neurons is entirely reproduced by an artificial system.

References

[1] Grey J.M., *An exploration of musical timbre*, Rep. STAN-M-2, Stanford University, 1975.

[2] Grey J.M., *Multidimensional perceptual scaling of musical timbres*, J. Acoust. Soc. Am., 61(5): 1270-1277, 1977.

[3] GordonJ.W., Grey J., *Perception of spectral Modifications on Orchestral Instrument Tones*, Comp. Music J., 2(1): 24-31, 1978.

[4] Feiten B., Frank R., Ungvary T., *Organizations of sounds with neural nets*, Proc. ICMC 91, p. 441-444, 1991.

[5] Kohonen T., *Self-organization and associative memory*, Springer V., Berlin, 1984.

[6] Kohonen T., *The Self-Organizing Map*, Proc. of the IEEE, 78(9): 1464-1480, 1990.

[7] *The Self–Organizing Map Program Package*, Helsinky University of Technology, 1992.

[8] De Poli G., Tonella P., *Self-Organizing Neural Networks and Grey's Timbre Space*, Proc. ICMC 93, 1993.