Sonological models for timbre characterization

Giovanni De Poli
CSC-DEI, University of Padova, Italy
Giovanni.DePoli@dei.unipd.it

Paolo Prandoni
LCAV/DE, EPFL, Lausanne, CH
prandoni@de.epfl.ch

Abstract

In the research on timbre, two important variables have to be assigned from the onset: the instruments used to analyze and to model the physical sound, and the techniques employed to provide an efficient and manageable representation of the data. The experimental methodology which results from these choices defines a *sonological model*; several different psychoacoustical and analytical tools have been employed to this aim in the past. In this paper we will present a series of experiments conducted at the CSC-University of Padova, which attempted to define an experimental framework for the development of algorithmically-defined timbre spaces. Fundamental to our line of research has been the use of analysis methods borrowed from the speech processing community and of data-representation techniques such as neural networks and statistical tools. The results are very interesting and show several analogies to the classical timbre spaces defined in the literature; this has proved very important in order to explore the qualities of musical timbre in a purely analytical way which does not rely on subjective listeners' ratings.

1 Introduction

Unlike other features of musical sounds, such as pitch or loudness, timbre cannot be linked directly to one physical dimension; its perception is the outcome of the presence and of the absence of many different properties of the sound, the perceptual weight of which is still in many ways unclear. The study of the role of all these concurring factors is no new issue in the psychoacoustical research; the difficulties are however countless inasmuch as listeners are asked to produce unambiguous responses on matters for which language provides an extraordinarily rich set of blurred definitions. For this reason, classic studies by Grey and others ([15], [16]), employed the verbally simpler notion of 'similarity rating' to build a timbre space. The spaces thus obtained, albeit different, have shed some light on the relationships between sensation and cause; they have not led, however, to a clear method to define a set of coordinates into which an arbitrary sound can easily be mapped.

In speech analysis, on the other hand, the problem of unique classification of sounds has been variously addressed; to this end, diverse signal processing techniques were devised to perform efficient data reduction while preserving the appropriate amount of information in order to define an almost univocal mapping between waveform and uttered phoneme.

This work tries to apply some common techniques of sound analysis to the acoustic signal to extract a set of significant parameter useful for the description and the classification of musical timbres. Self Organizing Neural Maps (SOM) and Principal Component Analysis (PCA) are the tools which will be employed to construct low-dimensional spaces with the information thus obtained.

2 Acoustic analysis

The acoustic analysis of a sound is the first step in trying to extract meaningful parameters from the signal (seen as a sequence of samples) while at the same time trying to reduce the large variance inherent in the data. The analysis methods rely on a mathematical (and thus computational) modelization of the signal, and on the study of the parameters describing the model itself. The model chosen to describe the signal must be tailored to the characteristics under study, and must provide a simplified representation of the data. In particular, the number of parameters provided by the model must be sufficiently small, and the variance of the parameters, seen as functions of time, must be significantly reduced with respect to the variance of the original data. This is a fundamental condition in order to be able to represent the signal by a reduced set of points in a data space. It is thus apparent that a fundamental point in acoustic analysis is a data-reduction process; this process must o course be designed keeping in mind the goals for which the analysis is carried out, and the type of sounds to which the analysis is applied.

2.1 Models in acoustic analysis of timbre

The various methods which are commonly employed in timbre research can be broadly divided into three classes: models of the signal, models which take into account the mechanisms underlying the sound-producing mechanisms, and models which take into account the properties of auditory perception. Often, models are employed which try to combine in a more or less simplified way these three different characteristics. Sometimes the model can also encompass an analysis-by-synthesis technique in which a sound similar to the original is re-synthesized from the parameters. This offers the possibility of an immediate perceptual evaluation of the meaningfulness of the data extracted by the model.

The classical models for the representation of sounds used in the field of timbre research usually employ a time-frequency representation of the signal. They are based on the Short-Time Fourier Transform (STFT) in its various formulations. Examples of this include the modelization of the signal in terms of slowly time-varying sinusoids, the spectrogram, the averaged spectral shape, the spectral centroid, and so on.

Models which rely on the knowledge of the sound-producing mechanisms are much less common, maybe except for the classical source plus linear filter modelizations which are widely employed in speech analysis and of which LPC (with all its variants) is most well-known. One can observe that the sound produced by an ensemble of musical instruments is marked by highly nonlinear characteristics, which complicate the development of suitable models. The recent development of sound synthesis based on physical models is an example of how hard it is to estimate the parameters of the model from the sound (see for example the Karplus-Strong algorithm [34]). Knowledge about the acoustical properties of sound generation can however be indirectly taken into account in interpreting the data provided by models which are based on other analytical principles.

The third kind of models, which rely on the characteristics of acoustic perception, have been initially developed by the speech processing community. Initial analysis schemes were all built using "production-based" processing schemes. In other words, Short-Time Fourier Transform (STFT), Cepstrum, and other related speech processing schemes were all developed strictly considering the physical phenomena which characterize the waveform obtained by the electrical transduction of the sound pressure wave. Moreover LPC technique and all its variants were developed directly by modeling the production mechanisms of human speech.

In the last few years, almost all these analysis schemes have been modified by incorporating, at least at a very general stage, various perceptually related phenomena. Linear prediction on a warped frequency scale [31], STFT-derived auditory models [1], perceptually based linear predictive analysis of speech [18] are few simple examples of how the perceptual behavior of the human auditory system is now taken into account while designing new algorithms for signal

representation. The most significant example in attempting to improve acoustic analysis of speech by perceptual related knowledge is given by the Mel-frequency cepstrum analysis of speech [6], which transforms the linear frequency domain into a logarithmic one resembling that of human auditory sensation of tone height (see Section 2.2). 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. Short segments of sound are isolated and processed as if they were short segments from a sustained sound with fixed properties. In order to better track the dynamic changes in sound properties, these short segments, which are called analysis frames, overlap one another. This framework is based on the underlying assumption that the properties of the sound signal change relatively slowly with time.

Even if overlapped analysis windows are used, important fine dynamic characteristics of the sound signal are still discarded. For this reason, although without completely solving the problem of correctly taking into account the dynamic properties of sound, velocity-type parameters (simple differences among parameters of successive frames) and acceleration-type parameters (differences of differences) [14] have been recently included in acoustic front end of almost all ASR systems which are now in the market. The use of these temporal changes in sound spectral representation (i.e. DMFCC, DDMFCC) gave rise to one of the greatest improvements in ASR systems. Currently these innovations in the representation of speech signals have not yet entered the field of timbre analysis; it is very well possible that they could bring new insights and results also to the study and to the classification of non-verbal signals such as musical sounds

One of the fundamental limitations of an STFT-type analysis is 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. As a further improvement, in order to overcome this limitation, the Wavelet Transform, characterized by the capability of implementing multiresolution analysis, has recently been introduced [26, 10]. With this 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 simulate more closely the frequency response of the human cochlea. As with the preceding processing schemes, this new auditory-based technique, even if it is probably more adequate than STFT analysis to represent a model of human auditory sound processing, it is still based on a mathematical model of the signal, from which it tries directly to extrapolate a more realistic perceptual behavior .

Cochlear transformations of sound signals result in an auditory neural firing pattern which is significantly different from the spectral pattern obtained from the sound waveform by using one of the techniques mentioned above. In other words, sound spectral representations such as the spectrogram, a popular time-frequency- energy representation of speech, or either the wavelet spectrogram or the scalogram, obtained using the previous multiresolution analysis techniques, are quite different from the true neurogram. In recent years, basilar membrane, inner cell and nerve fiber behavior 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 clusters and synthetic sound stimuli. Timbre features probably correspond in a rather straightforward manner to the neural discharge pattern with which sound is encoded by the auditory nerve.

Various auditory models which try to physiologically reproduce the human auditory system have been developed in the past, and, even if they must be considered as only an approximation of physical reality, they appear to be a suitable system for identifying those aspects of the sound signal that are relevant for automatic sound 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 Modeling (AM) techniques not only follow "perception-based" criteria instead of "production-based" ones, but also overcome "short-term" analysis limitations, because they implicitly retain dynamic and nonlinear sound characteristics. For example, the dynamics of the response to non-steady-state signals, as well as "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 all important aspects which are captured by efficient auditory models. Large evidence can be found in the literature reccommending the use of AM techniques, instead of more classical ones, in building sound analysis and recognition systems. Especially when sound is greatly corrupted by noise, the effective power of AM techniques seems much more evident than that of classical digital signal processing schemes.

The use of auditory models is gaining widespread acceptance not only in the field of speech recognition but also in the timbre research. Recent studies featuring auditory models ([4], [32]) and simplified auditory-based analysis ([11], [13]), gave overall encouraging results.

2.2 Mel Frequency Cepstral Coefficients

In the research on the characterization of musical sounds which was carried out at CSC-University of Padua, the parametric representation known as *Mel Frequency Cepstral Coefficients*, or MFCC, was extensively employed. Although the MFCC are widely used in speech recognition, they are quite new on the scene of musical sound analysis, and therefore a somewhat detailed description of this technique is presented in this section.

The MFCC were first introduced by Davis and Mermelstein [6] in a study comparing different techniques for the coding of monosyllabic words. Out of their 'natural' vocal context, the MFCC are here tentatively used to characterize musical sounds. The coefficients c_i are defined as:

$$c_i = \sum_{k=1}^{N} Y_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{N}\right], i = 1, 2, \ldots, M,$$

where $Y_1 \ldots Y_N$ are the log-energy outputs of a mel-spaced filterbank. Usually a simplified approximation of the mel scale

$$\text{mel}(f) = \begin{cases} f & \text{if } f \leq 1 \text{ kHz} \\ 2595 \log_{10}\left(1 + \frac{f}{700}\right) & \text{if } f > 1 \text{ kHz} \end{cases}$$

is adopted, in which the filterbank is constituted by triangular filters with constant bandwidth up to 1 KHz and with constant-Q above. The resulting global frequency response of the filterbank is flat. Due to the importance of the higher frequencies in music perception as opposed to speech comprehension, in our case the filterbank is made up by $N = 27$ filters and spreads up to 2700 mel $\approx$ 8 kHz as shown in fig. 1. The coefficients are computed for every time frame using a 32 ms Hamming window with a 4 ms time-shift.

As usual in cepstral methods, to obtain a smoothed representation of the spectral envelope, only the first few K coefficients are retained. In the course of tour experiments the first six coefficients appeared to provide enough information; this is in accordance with the similar result holding in the field of speech recognition. This further simplification leads to an overall 95% data reduction ratio. In this representation the coefficient, c_0, is the mean value of the filterbank log-energy output. Its value is directly dependent on the absolute energy of the input signal, which is not of interest when comparing spectral envelopes. Discarding the coefficient c_0 is an efficient way to obtain an energy normalized representation of the signal. To summarize, each time frame is then represented by the first K mel-cepstral coefficients. The first coefficient c_0,

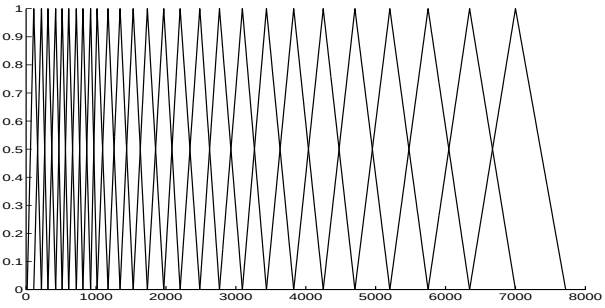

Figure 1: The filterbank.

for the frames corresponding to the attack phase of the signal, can be used to compute the duration of the attack phase, seen as a global parameter of the analyzed sound.

An inverse DCT leads back to the mel-warped spectrum:

$$C(mel) = \sum_k c_k \cos(2\pi k \frac{mel}{B_m})$$

where $B_m = 2700$ mel is the retained bandwidth (in mel) of the signal. With all the coefficients, the reconstruction is exact while, retaining only a subset, a simplified spectral envelope

$$\tilde{C}(mel) = \sum_{k=1}^{K} c_k \cos(2\pi k \frac{mel}{B_m})$$

is obtained. In fig. 2 a single time frame of the input waveform of a clarinet is considered. The spectrum magnitude is shown, together with the filterbank outputs, the mel-cepstrum coefficients c_i and the simplified spectral envelope $\tilde{C}$ reconstructed using the first six coefficients. In fig. 3 the simplified spectral envelope $\tilde{C}$ is plotted against the log of the spectrum magnitude of the signal $|X(f)|$, showing that

$$\tilde{C}(\text{mel}(f)) \approx \log |X(f)|$$

While not a cepstral extraction in the usual sense, the actual effectiveness of the MFCC in speech coding is mainly due to the mel-based filter spacing and to the dynamic range compression in the log filter outputs, properties which both take into account, in a simplified way, the actual processes in the early stages of human hearing. Clearly the spectral envelope is the major factor in the computation of the MFCC, which is advisable in speech analysis where a formantic structure is to be captured. However, almost no musical instrument exhibits a formantic pattern, nor is it sure whether the spectral envelope alone can account for most of the timbre information. The outcome of such a tentative approach is the object of the following sections.

3 Physical timbre space construction

The acoustic analysis yields a data ensemble which measures the acoustic features of the sound. Some parameters are global parameters, and they refer to the sound as a whole; such are for instance the duration of the attack and of the decay phases, or the average power. Other parameters are more strictly time-varying; in a short-time type of analysis their instantaneous value can be associated to the single windowed frames. The description of the sounds is thus constituted of a set of global features plus a set of frame descriptions. The resulting question is then to decide which analysis frames are to be kept and which can be discarded: if all the frames are kept into account, the time-varying characteristics of the signal will dominate; on the other

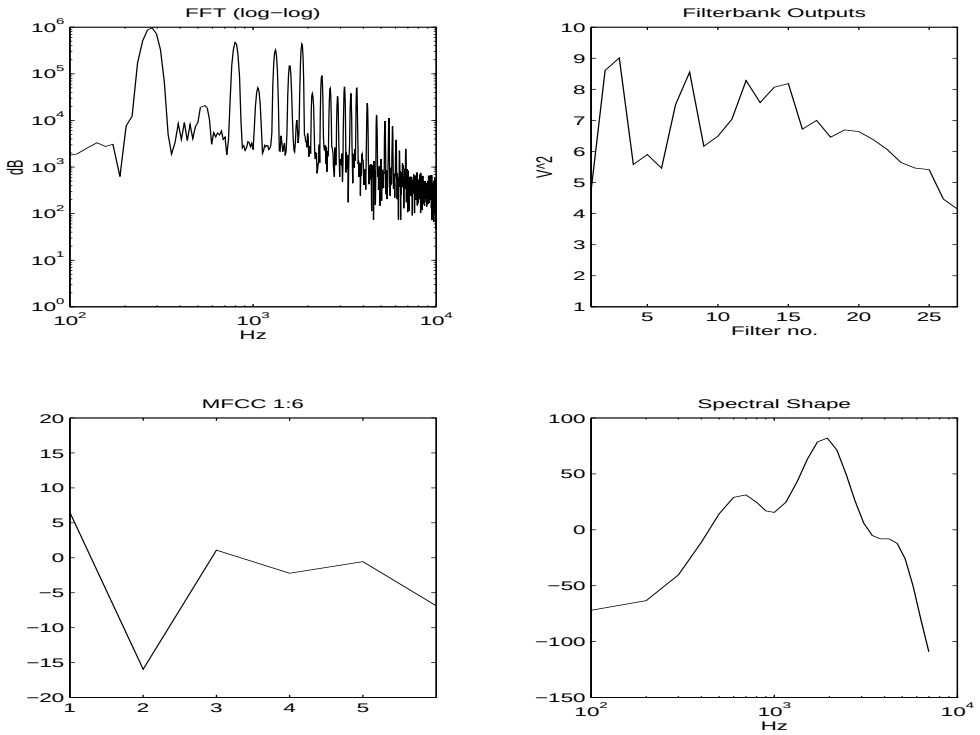

Figure 2: Analysis of a frame of a clarinet.

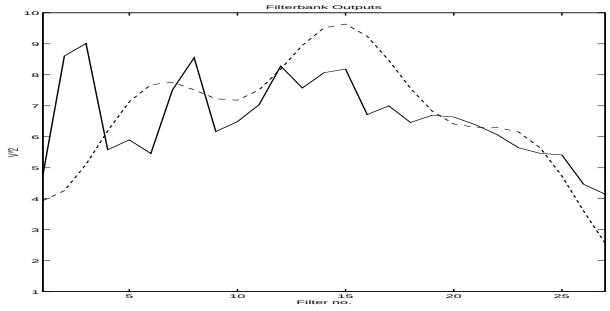

Figure 3: Spectrum magnitude (solid line) compared with simplified spectral envelope (dotted line).

hands, if only a few of the frame-related descriptions are maintained, or, in the limit, only a single, typical frame, the global parameters of the sound will determine the overall representation. The final result of the acoustic analysis is thus a possibly time-varying m-dimensional vector. The aim of the subsequent processing is then to build from these vectors a low-dimensional space in order to represent the signal space with a reduced number of coordinates (2 or 3) which might lead to an easier interpretation of the space itself. This low-dimensional space will be called the *physical timbre space*. The main guideline in the construction of this space is the preservation of the "natural" topology of musical timbres: different sound must be distinguishable and, at the same time, similar sounds must be "close" in some sense. In other words, the concept of acoustic similarity must be coherently mapped onto the concept of distance; the type of metric embedded in this space, which is now used to *measure* the similarity between sounds, has to be carefully chosen keeping in mind its possible interpretations. All this is ultimately compared to the perceptual spaces defined by the psychoacoustic research.

The experiments performed at CSC-University of Padua primarily featured physical timbre spaces obtained from tools such as Kohonen's self-organizing neural maps and Principal Component Analysis. These tools will be described in some detail in the following sections.

3.1 Self Organizing Maps

As explained above, the parametrization algorithm transforms sound signals in sequences of points in a m-dimensional data space. The topological properties of this space are a direct consequence of the ability of the processing technique to extract timbre information from the sound samples. This information is however still too complex to be interpreted directly; in order to simplify and organize the raw data provided by the analysis front-end, a self-organizing neural network can be the instrument of choice.

Kohonen [20] formalized the learning algorithm for self-organizing networks into a simple numerical process whose outcome is the modification of the inner structure of the neural model into a n-dimensional projective model of the m-dimensional probability distribution of the input. With $n < m$ (usually $n = 1, 2$) 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.

Rectangular neural maps are usually employed. A weight vector $\mathbf{w}_i$ with the same dimensionality as the input vectors $\mathbf{x}$ is associated to each neuron, leading to a structure in which all the neurons are ideally connected in parallel to all the input terminals. For each input vector a best matching cell c is found which produces the highest excitation level, and this cell determines the correspondence between the input and a position on the map. The best matching cell is selected by looking for the neuron which minimizes the distance between $\mathbf{x}$ e $\mathbf{w}_i$:

$$c = \arg\min_i \{||\mathbf{x} - \mathbf{w}_i||\}$$

This topological coordinate is considered as the projection of the input vector onto the map. The distance function used in the minimization process can be chosen according to various metric functions; the final topological organization will reflect the properties of the chosen metric.

In the adaptation phase, for each new training input the best-matching cell c is selected and a set of neighboring cells N_c is defined. All the neurons within this set, called the topological neighborhood, are updated by a process of minimization of the distance between the current input and the weight vectors within the set:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \alpha(t)[\mathbf{x}(t) - \mathbf{w}_i(t)], \forall i \in N_c$$

while all the other neurons in the map 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.

Neural maps obtained via this adapting process are spatially organized. There are fundamentally two neural processes which lead to this final result: the spatial concentration of the neural activity around the cell which is best tuned to the current input, and the further fine-tuning of the best-matching cell and of its topological neighbors to the current input.

The network recognizes and enhances the components with the largest variance in the input data. After the self-organization process the weights of the network are tuned to these components. When the input data are partially corrupted by noise, Kohonen neural networks retain their feature extraction capabilities; the performance is still good since the self-organization process follows two different and conflicting requirements: the maximization of the variance of the neural outputs with the purpose of recognizing the most distinctive features in the inputs, and the introduction of redundancy with the purpose of obtaining correct answers even in presence of noise.

The primary result achieved by KNN's is a nonlinear projection of a m-dimensional input space onto a spatial structure of lower dimensionality. The attempt is to obtain a more readable representation of the input data, preserving its main features and discarding unnecessary redundancies and noise. The usefulness of KNN's in this regard is that they require no assumptions on the probability distribution that they are called to model; in this sense they are an extremely convenient and fast exploration tool which allows the researcher to quickly single out, if there are any, the principal features in an otherwise too complex body of experimental data.

3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical tool whose formulation relies on the properties of the *orthogonal linear transforms*. Given a m-dimensional data set, by means of scaling and rotations the principal information is moved onto a reduced set of variables. By carefully choosing the transformation matrix, the new variables prove to be uncorrelated: the information pertaining to each of the variables can thus be analyzed independently. The two major objectives of a PCA analysis are then the *compression of the data*, and the *optimal interpretation of the data*. These goals are achieved if in the original data set the variance in the information is naturally associated to a few principal components, so that the representational loss associated with the dropping of some of the components is relatively small. Geometrically, the PCA defines a set of rotated coordinates in the space in which the new axes coincide with the directions of maximum variance.

PCA, compared to KNN's, offers more precise and measurable results; the set of coordinates provided by PCA algorithms allows the user to obtain quantitative results in the process of fit a model to the data. PCA however is a linear transformation while KNN's can handle more complex input distributions thanks to their locality properties.

4 Experimental results

This section describes some of the results which were obtained in the development of physical timbre space models at CSC, University of Padua. The range of experiments encompassed different acoustic analysis methods, different data reduction techniques, and different sound databases.

An early line of experiments [8, 9] employed a time-frequency representation derived from that used by Grey in his classical paper [15]. We used 2D and 3D SOM's to construct the physical space with encouraging results. In a second experiment [4] we used the joint Synchrony/Mean-Rate (S/M-R) auditory model proposed by Seneff [28] to extract a set of meaningful parameters. A few frames' worth of data for each instrument, in the form of a lumped representation of Generalized Synchrony Detector parameters, was analyzed by means of self organizing map. A 2D map showed a topological organization of timbres which complied with the important criteria

arising in the subjective classification of musical sounds. The combination of an auditory model and of a neural network proved the system extremely robust with respect to additive noise, giving good classification results even in conditions of 0dB SNR. As a later line of research, the data provided by a physical model of the cochlea by Nobili and Mammano was analyzed by means of PCA [24, 25].

4.1 Timbre spaces obtained from the MFCC analysis

As noted above, the Mel Frequency Cepstral Coefficients parametrization, which is widely employed in speech recognition, proved very effective also in the characterization of musical sounds. Different sound databases and several data reduction techniques were taken into account. Musical instruments were considered first; later we considered general ambiance sounds and different types of singing voices. The main results will be presented in this section.

As a general remark, it is important to note how the final topology of the timbre space depends on the inherent metric defined onto the space itself. PCA is for instance an orthogonal transform; since Euclidean distance is preserved under orthogonality, that was the metric we chose to employ in that case. It is however interesting to notice that, for the distance d^2,

$$d^2 = \sum_i (c_i^{(m)} - c_i^{(n)})^2 = \int_0^{2\pi} (C^{(m)}(\omega) - C^{(n)}(\omega))^2 \frac{d\omega}{2\pi}$$

where in a given windowed analysis frame the $c_i^{(m)}$'s are the MFCC coefficients and $C^{(m)}(\omega)$ is the Fourier transform of the frame for the m-th instrument; in the simplified representation in which only a few coefficients are retained, the Euclidean distance between two mel-cepstral vectors corresponds to the Euclidean distance between the simplified spectral envelopes:

$$\hat{d}^2 = \|\mathbf{c}^{(m)} - \mathbf{c}^{(n)}\|_2 \approx \|\tilde{C}^{(m)} - \tilde{C}^{(n)}\|_2.$$

This physical quantity is thus the actual data ruling the topological organization of the physical timbre space obtained from MFCC parametrization; the Euclidean distance was then the general choice for experiments with both self-organizing networks and PCA.

4.1.1 SOM-derived projections

The first data reduction tool we used was a self-organizing map (SOM). Several experiments were carried out with different sets of musical instruments and different sizes for the neural net. Here following is the description of the most complete results. [5]. The actual sound database was drawn from the McGill University Master Samples CD library, considering the first 800 ms of each signal. In the experiment, the mel-cepstral parameter vectors defined above were used as the input to a rectangular net of $15 * 30 = 450$ neurons; the SOM is thus called to perform a projection from a six-dimensional probability space onto a bi-dimensional map. The training database was generated by processing the sound samples of 40 different orchestral instruments (see table 1), all of them playing a C4, approx. 261 Hz, and collecting the *single* six-element vectors arising from the processing of one frame, for all frames. As opposed to a lumped representation of the whole course of the sounds, this approach offers two advantages: it provides a large numerical database for the training phase and it takes into detailed account the inner variability of the musical tones. The training was performed over a random shuffling of the database vectors, in order to eliminate all direct information about their correct order in each tone. After the training, when a sequence of vectors was presented as an input, a sequence of neurons (*neural path*) excited correspondingly. In other words, the processes underlying the neural classification can be viewed as such: the analysis algorithm converts a sound into a sequence of parameter vectors and the neural net converts the sequence of parameter vectors

Label	Instrument	Label	Instrument
alfl	alto flute	lute	lute
bbcl	B♭ clarinet	mar	marimba
basn	bassoon	oboe	oboe
bscl	bass clarinet	obam	oboe *d'amore*
btrp	Bach trumpet	obcl	oboe *classico*
cell	cello	orbp	organ *Baroque*
cepz	cello *pizzicato*	orco	organ *cornet*
clav	harpsichord	orfl	organ *flute*
clst	celesta	ortu	organ *tutti*
corn	cornet	pn	piano
crom	crumhorn	rec	recorder
ctrp	C trumpet	sx	tenor sax
dbbs	double bass	ttb	tenor trombone
dbpz	dbbs *pizzicato*	tuba	tuba
ebcl	E♭ clarinet	va	viola
eh	English horn	vapz	viola *pizzicato*
fh	French horn	vibr	vibraphone
fl	flute	vl	violin
gt	guitar	vlens	violin *ensemble*
harp	harp	vlpz	violin *pizzicato*

Table 1: Labels of instruments.

into a sequence of excited neurons. Acoustic properties of the tones are thus translated into spatial relationships between neural paths.

The experiments started with smaller subsets of instruments. The net was capable of reconstructing the proper sequences of frames and to well distinguish between different instruments. These abilities proved to be irrespective of net size, which affected only the number of neurons associated to one tone and consequently the level of detail reflected in the path. Similarly, the increases in size of the database did not impair the effectiveness of the neural system. In fig. 4 the ultimate results are shown: all the neural paths relative to the 40 instruments of table 1 are represented; while all the time-course of the signals was used in the training phase, in fig. 4 only the steady-state portions are depicted. It can be seen that the paths are generally well separated, with only occasional overlapping for very similar tones (e.g. the percussive sounds of *mar*, *clst*, and *vibr*). The space spanned by the paths themselves is related to the inner variability of the sounds; *pizzicato* strings, for instance, which exhibit clear transitions in timbre during the decay phase, are source to wider neural excitements. Most interesting however is the relative positioning of the paths: in fig. 5 those local groupings are highlighted which correspond to normally defined instrumental families: strings, trumpets, oboes, clarinets, percussion's, and plucked strings. Some anomalies are present, e.g. the violin and the harpsichord, which can however be explained once the structure of the map is taken into account.

A further analysis of the global ordering revealed that the main axis is strictly related to the spectral energy distribution of the steady-state portion of the tone. The background of fig. 5 shows the spectral envelopes associated to the local information of the net; these envelopes are obtained inverse-transforming the six MFCC each neural zone is most sensitive to. A clear horizontal shift is present, from the low-pass prototypes of the left side to the band-pass of the right. This particular spectral information is embedded in the first cepstral coefficient which, possessing the largest numerical variance, is assigned the main axis by the self-organizing algorithm. Since the spectral energy distribution is related to the perceptual quality called

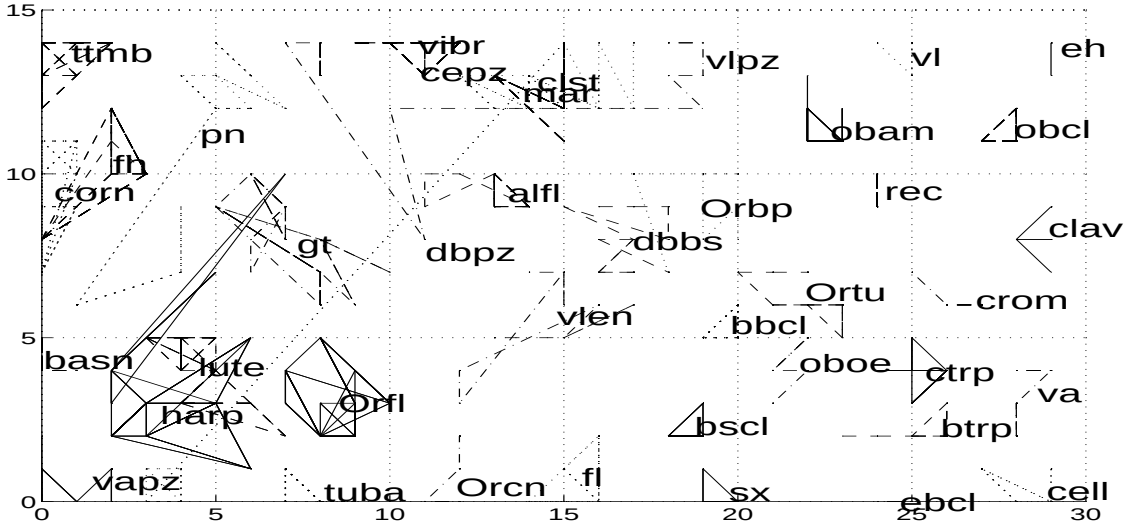

Figure 4: Final neural paths.

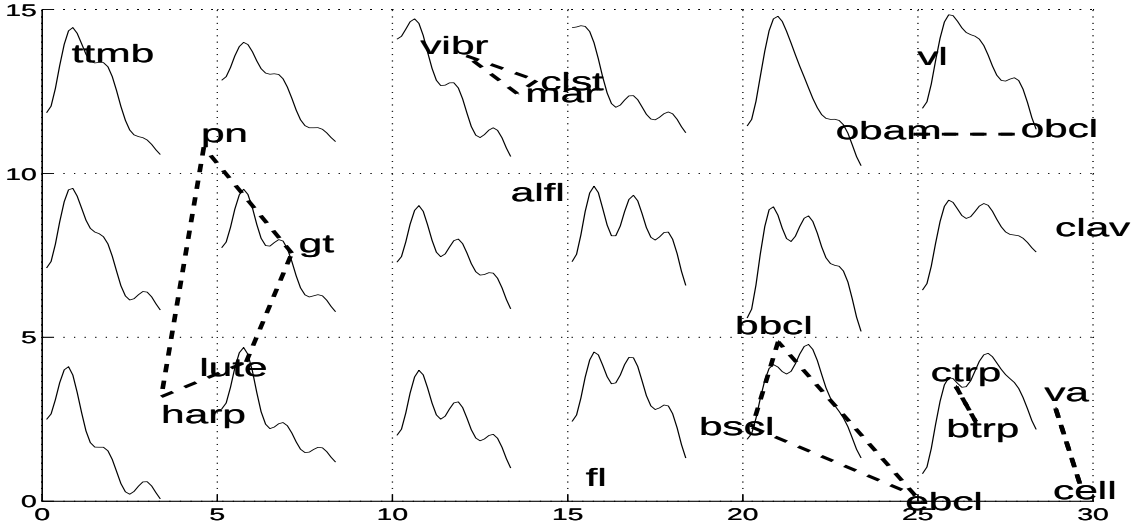

Figure 5: Family clusters and spectral content.

brightness, and since in all the perceptual timbre spaces one of the axis accounts for timbre brilliancy, a comparison between these spaces and the neural model becomes possible, at least with regard to the orderings according to brightness. Table 2 compares the neural net to Grey's and Wessel's [33] timbre spaces; the analogies are clear, especially considering the larger number of instruments with which the net deals.

Net	Grey	Wessel
basn	fh	fh
fh	bscl	fl
fl	basn	basn
bscl	cell	cell
sx	ctrp	bscl
ebcl	fl	sx
ctrp	sx	ctrp
cell	ebcl	ebcl
oboe	eh	oboe
eh	oboe	eh

Table 2: Brightness orderings for instruments.

No global ordering is related to the second axis, which seems to rule the local groupings into which instrumental families are spatially clustered. The grouping anomalies mentioned before can now be explained as the second dimension being overruled by the first. Even though the violin, for example, is generally regarded as akin to the cello, its spectral content is narrower; the converse holds for the harpsichord with regard to the other plucked strings such as the lute or the guitar. In fact, when judging sound similarities, listeners usually take into account articulatory and structural features of the instrumental sources (whether actually present or recognized) which do not pertain to the mere acoustic field, but involve higher-level cognitive processes.

In addition, we must remember that the projection performed by the map is non-linear, and its tendency to fill up the entire available space leads to deformations of the input space; anomalies with respect to the conventional timbre spaces are then to be expected.

4.1.2 PCA-based projections

In order to obtain more quantitative results and to avoid distortion in the timbre space deriving from the use of nonlinear projections, PCA, a linear transformation, was adopted.

Several sets of different musical sounds have been used, all of which led to similar final results. In order to allow a direct comparison between our physical timbre space and the perceptual timbre space, we will describe the analysis [3] of the same set of musical sounds used by Carol Krumhansl in [22], where musical timbre is studied by means of listeners' responses, and by McAdams [21], who provided an acoustic explanation of the dimensions in the perceptual space.

The actual sound samples have been obtained by FM sound synthesis on a Yamaha Tx802 synthesizer (Tab. 4.1.2). Each of the 21 tones has the same $E\flat$ 4 pitch (311.1 Hz), and was sampled at 44.1 kHz. According to McAdams [21], the first and third dimension of the perceptual space are related to the steady-state frequency spectrum of the signal, whereas the second dimension is related to the duration of the attack phase of the sound; in the MFCC coding of the waveform, this parameter is embedded in the variations of the c_0 coefficient over time, as explained in section 2.2. For the moment we will concentrate on the steady-state portion of the signal. The sound signal is sliced into overlapping frames: each frame is converted to a MFCC representation, namely a vector of MFCC coefficients which can be regarded as a point

Horn	Trumpet	Trombone	Harp
Trumpar	Oboleste	Vibraphone	Striano
Sampled Piano	Harpsichord	Tenor Oboe	Oboe
Bassoon	Clarinet	Vibrone	Obochord
Bowed Piano	Guitar	String	Piano
Guitarnet			

Table 3: Musical Instruments (Yamaha Tx802 synthesizer).

in a multidimensional MFCC space. A PC analysis performed on the MFCC-coded instrumental set reveals that the 80% of the variance is concentrated in the first three components. A three-dimensional space is thus able to provide a "correct" topological organization within the limits of the retained information. The resulting physical timbre space is shown in Fig. 6 and Fig. 7. These pictures show that the points corresponding to the analysis of a single instrument are tightly clustered together; as a consequence, the centroid of the cluster can be chosen as the prototypical MFCC coordinate of the instrument. This does not lead to misrepresentations as the tight clusters originated by the different instruments map into well separated regions of the MFCC space.

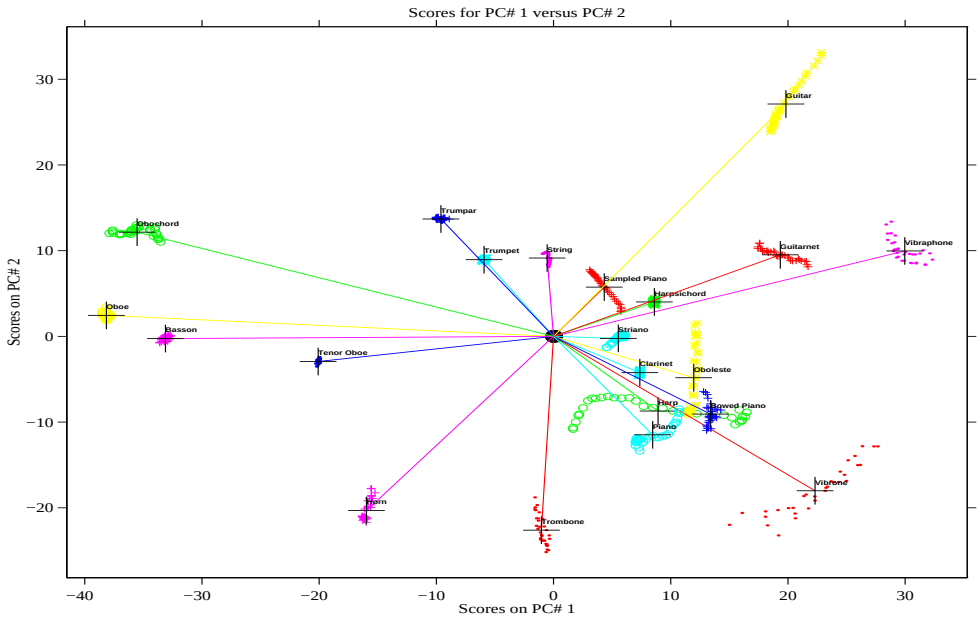

Figure 6: Frame grouping: first and second dimensions.

Several interpretations can be carried out on the space thus obtained. The origin represents the average spectral envelope of all the timbres in the data set; this envelope, plotted in Fig. 8, clearly exhibits the typical lowpass shape which characterizes almost all musical instruments [17]. In fig. 9, 10 and 11 the three direction cosines related to the PCA spatial transformation are shown. Each represents the spectral contribution of the respective spatial dimension of the timbre space.

To better evaluate the global features of the timbre organization thus obtained, it is useful to represent the smoothed spectral envelopes at the extremes of the axes as the sum of the average spectral envelope (the origin) with the eigenvalue-weighted contributions (positive on the right side, negative on the left side) of each of the direction cosines. In this new representation, the

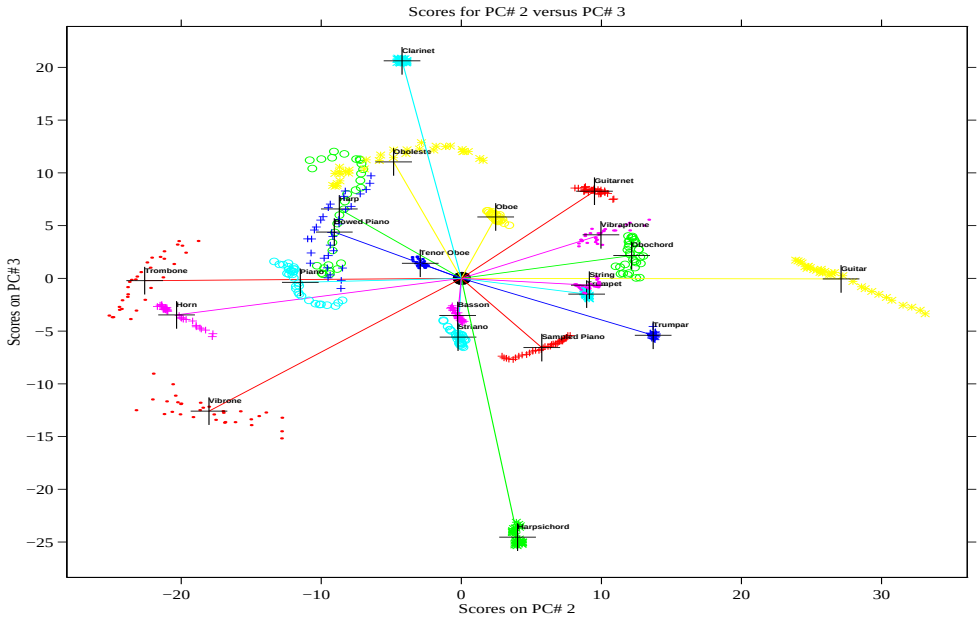

Figure 7: Frame grouping: second and third dimensions.

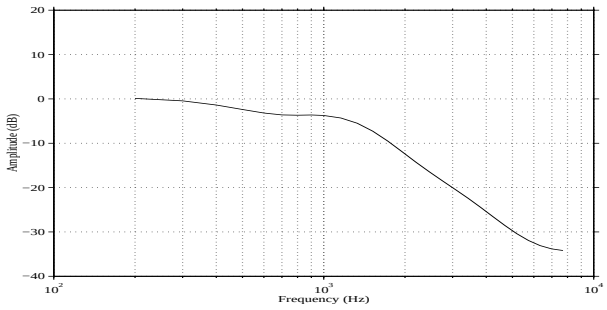

Figure 8: Overall average spectral shape

distribution of the spectral energy along the axes of the new PCA space appears more clearly (fig. 12, 13).

The analysis of the plots shown so far leads to several conclusions. Clearly, the first axis is related to the spectral energy distribution, called *brightness*. The spectral envelope associated to the first principal component (fig. 9) shows a boosting of the low frequencies for positive values of the coordinate, whereas negative values are linked to frequency values above 1.5 kHz. These interpretations are confirmed by the spectral envelopes shown in fig. 12 (horizontal direction). Along this dimension, bright-sounding instruments such as the *oboe*, the *bassoon*, and the *horn* are at a maximum of the geometric distance (and of the perceptual distance too) from the darker-sounding instruments such as the *vibraphone*, the *guitar*, and the *piano*.

The second dimension is correlated to the energy values in a broad frequency range which encompasses the whole band for musical sounds, $0.6 \div 6$ KHz. From the plots of fig. 10 the fundamental characteristics could be assumed to be an apparent spectral irregularity. Fig. 12 (vertical direction) shows that at one end of the axis there is an amplification in the region corresponding to the knee of the spectral envelope, whereas at the other extreme there is a smoothing in the slope discontinuity in the spectral envelope, which changes into an almost monotonic curve. At one end we have *trombone, horn, vibrione*; at the other we have the *guitar*.

The third dimension seems to be associated to subtler aspects of the spectral characteristic;

it is correlated to the energy content of a narrow region of the spectrum centered around 700 Hz. This set of frequencies has an extremely important role in acoustic perception, and is a commonly used parameter in audio equalization. A possible hypothesis would be that this frequency region underlies a differentiation criterion which is similar, albeit finer, to the general distinction between low- and high-pass instruments. At the positive end we have the *clarinet* and at the negative end the *harpsichord*.

If the single instruments are now taken into account, it is interesting to examine the differences in the perception of their sound in relation to their positioning inside the timbre space. Since we have used an Euclidean metric distance, the spatial differences between sounds are related to the differences between spectra; a graphic representation of the space, in which the prototype spectra are displayed in their proper positions, can thus be used to infer judgments on perceptual differences (fig. 14).

The correct ordering of the instruments along the dimension related to *brightness* is an assessment of the interpretation which was given before to the spectral shape contribution due to the first principal component. Brilliant-sounding instruments such as the *oboe*, the *bassoon*, and the *horn* are at a maximum of the geometric distance and of the perceptual distance from the darker-sounding instruments such as the *vibraphone*, the *guitar*, and the *piano*. The ordering proves also *regular*, which supports a notion of coherence in the instrumental ordering along this dimension.

Other experiments with different sound sets confirmed the fact that the first two dimensions extracted by the PCA control the overall spectral shape, the "cutoff" frequency, and the spectral slope; the third dimension is always related to the energy content of a spectral region bounded between 600 and 800 Hz.

PCA was also applied to the set of musical instruments of table 1, previously analyzed by means of neural nets as reported in section 4.1.1. In this case too, the PCA extracts three components with maximum variance (47%, 27%, and 12% respectively). The contribution to the spectral shape determined by the first two coordinates of the resulting space is shown in figure 15; this is to be compared to figure 12. We can clearly notice, in particular along the lower-left to upper-right diagonal, the variation in the cutoff frequency and in spectral slope, the two factor which affect sound brightness. In figure 16 the third direction cosine is also displayed, clearly showing the feature corresponding to the energy content around 600Hz. We are led to conclude that this feature is a differentiating factor in the quality of musical timbre which acts in an independent fashion from the quality called brightness.

4.2 Analysis of different typologies of sounds

4.2.1 Analysis for the timbre characterization of singing voices

The same methods described above, namely mel-cepstral parametrization and PCA data reduction, were applied towards the characterization of the timbre of singing voices. The main results,

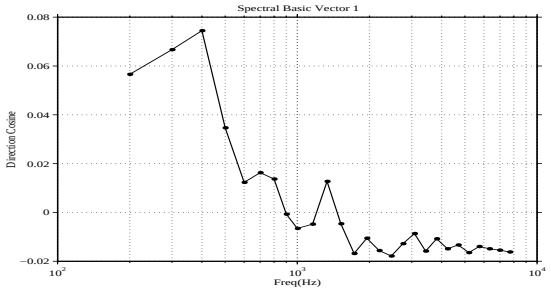

Figure 9: 1st direction cosine of Krumhansl's data set (spectral envelope).

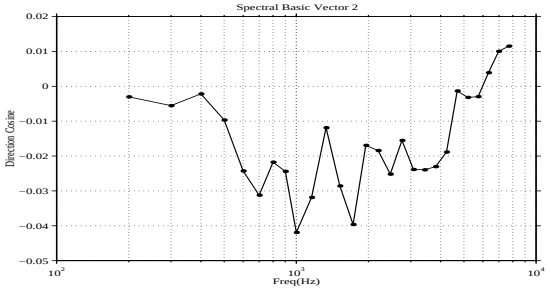

Figure 10: 2nd direction cosine of Krumhansl's data set (spectral envelope)

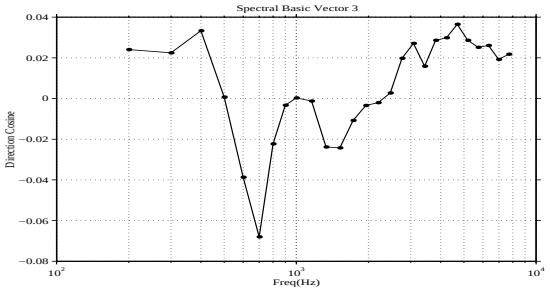

Figure 11: 1st direction cosine of Krumhansl's data set (spectral envelope)

obtained in collaboration with E. Teresi and F. Ferrero, will be presented in the following section. This is a problem which has been addressed very rarely in the literature [2]; we approached it to try to understand which of the acoustic features in vocal sounds are most influenced by the "vocal typology" peculiar to a particular singer. Besides, we wanted to assess the applicability of the parametrization scheme we used for musical instruments to this kind of signals.

A recording session was attended by 14 professional singers (table 4) who were asked to perform the standard full-voice vocalism's for the vowel /a/. The data set was obtained from the single note $D3$ (294Hz), which was sung by all artists for more than 300ms.

n.	singer	symb.
1–2	bass	B
3–5	baritone	b
6–7	tenor	T
8–9	mezzosoprano	M
10–14	soprano	S

Table 4: Numbers and symbols used for singers.

The usual procedure (MFCC parametrization and PCA) provided the three principal components for all of the 616 analysis frames, which represent the (time-varying) coordinates of the sounds in the physical timbre space.

A first examination of the space highlights the degree of separation between the regions corresponding to the individual singers. The differentiation is more present in the plane defined by the first two principal components, and is more pronounced if the singers belong to different vocal typologies.

Figure 17 shows the localization of the points relative to the 14 singers for the first two dimensions. The area pertaining to a single singer is surrounded by an ellipse centered on the center of mass of the input distribution for the singer; its main axis is aligned with the direction of

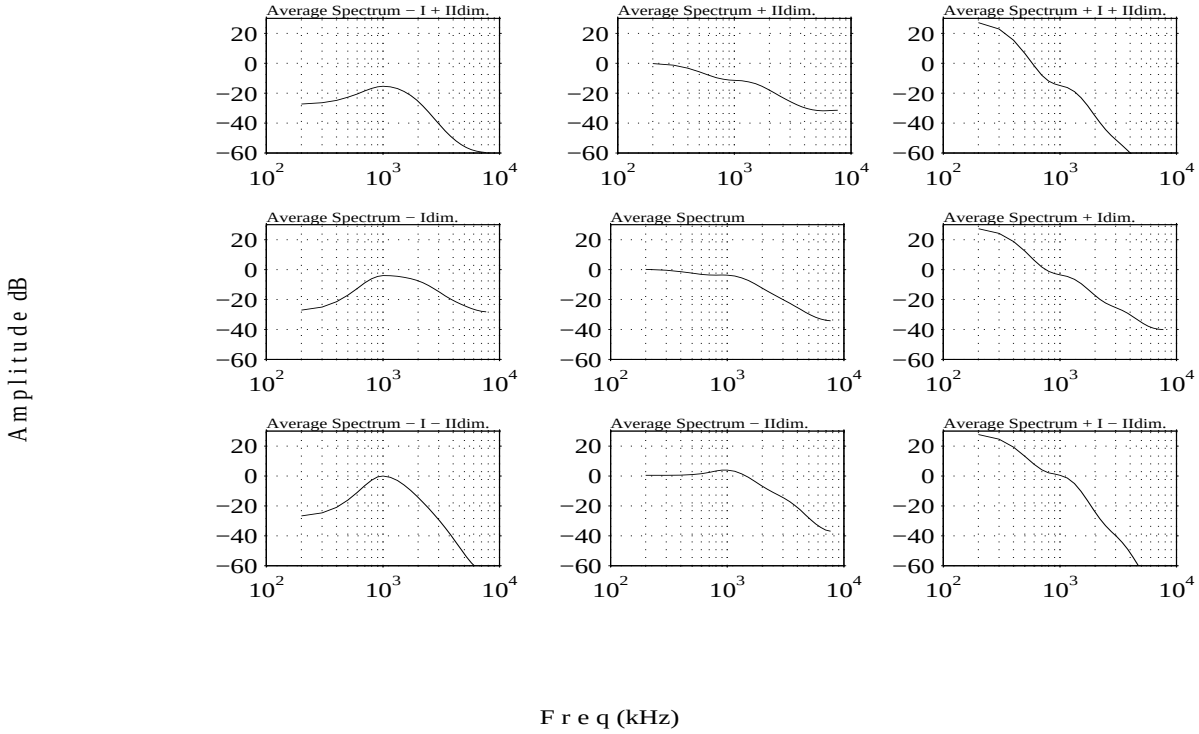

Figure 12: Spectral envelope (1st and 2nd coordinate). Krumhansl data set.

maximum variance for said distribution, with a length equal to two times the standard deviation; the length of the second axis is equal to two times the standard deviation of the data and it is oriented in the direction perpendicular to that of maximum variance. Ellipsoids corresponding to the particular kind of voice (heavier lines) have also been plotted.

The configuration that the points assume in the space show that in order to describe the spectral differences between singers at least two dimensions are needed. The corresponding distribution of the data points approximately defines a square, whose vertices are occupied by the basses, the tenors, the mezzo-sopranos, and the sopranos respectively. The left-hand side of the plot, for the negative values of the first coordinate, corresponds to male voices; for these voices the second coordinate operates the differentiation between basses and tenors, increasing in value. Similarly, female voices belong to the positive-valued half of the first coordinate, in the right-hand side of the plot, and they are differentiated along the second axis.

4.2.2 Recognition of Environmental Sounds by SOM

In order to test the possibility of applying the architecture defined by the conjunction of MFCC analysis and self-organizing maps to a system for monitoring ambiance sounds, a preliminary experiment was carried out by training a 6×12 neural net on a set of non-musical sounds. These eight sounds, typical household sounds like tea-kettles, doorbells, and alarm clocks, were recorded on a pocket tape recorder; the poor signal to noise ratio provided by this setup is a further test of the robustness of the system. The training database was created the usual way by selecting the single cepstral frames of the first eight sounds. The sound labeled *sve2* was put aside for the test phase of the experiment.

The results are very good. Fig. 18 shows how the net managed to clearly identify and separate the different sound sources, despite their inherent complexity; if compared to musical tones, these sounds indeed possess a much more differentiated evolution in time. The node at location $(0, 2)$ represents silence between rings or pulses of the different sounds. The generalizing

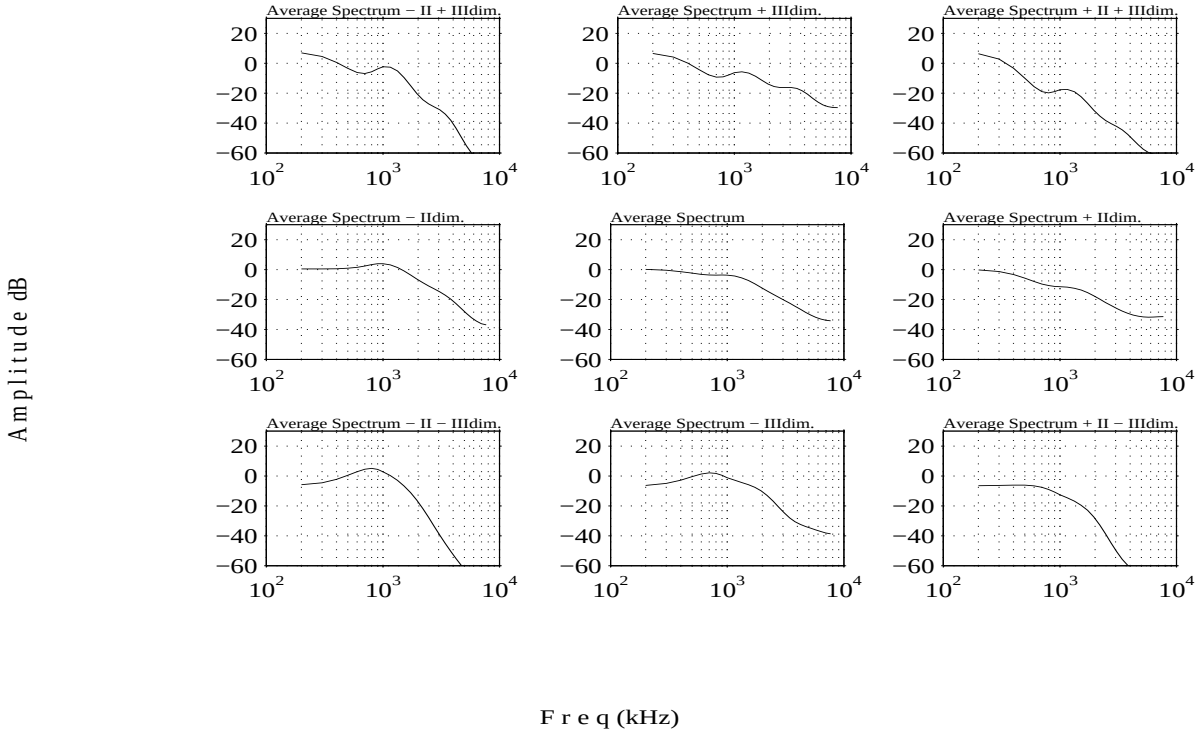

Figure 13: Spectral envelope (2nd and 3rd coordinate). Krumhansl data set.

properties of the system are quite good too; the neural path labeled *sve2*, corresponding to an alarm clock and plotted in the picture with a dash-dot line, is clearly overlapping the path labeled *sve1*, which corresponds to another, similar alarm clock with a somewhat different sound.

These results are a further confirmation of the representational effectiveness of the MFCC's and they suggest the possibility for a particularly efficient way of implementing a self-learning ambiance monitoring system.

5 An application: synthesis from physical timbre space

In this last section an application of the physical timbre space to sound synthesis will be outlined. When the data reduction tool which generates the timbre space is a self-organizing map, since the spectral properties of the analyzed instruments are embedded in the neural lattice in an orderly fashion, a way of converting neural paths into timbre variations is automatically provided. User-defined 'navigations' over the net will be translated into acoustical transitions across timbre prototypes. This possibility arises from the fact that within the network the actual spectral coefficients are stored and organized according to a meaningful distance metric. As we said before, the weight vector $\mathbf{w}_i$ associated to each neuron contains a set of cepstral coefficients so that a prototype spectral envelope is associated to each point in the map. Since the neurons in the net are organized according to similarity rules, an arbitrary trajectory on the map will originate a transition between spectral envelopes with the same properties of timbre similarity. Furthermore, it is possible to fill in the gaps between neurons with interpolated versions of the data set, thus allowing for smooth sound changes along trajectories. A similar concept applies to the space obtained by PCA; since PCA is a linear transform, linear interpolation can be used between points to obtain the intermediate mel-cepstral coordinates.

The experimental results lead to believe that the topological organization provided by the space is related to the actual evaluation of the timbre content in the sound, at least as far as the steady state portion of the sound is concerned. An user-definable palette of sounds is the main

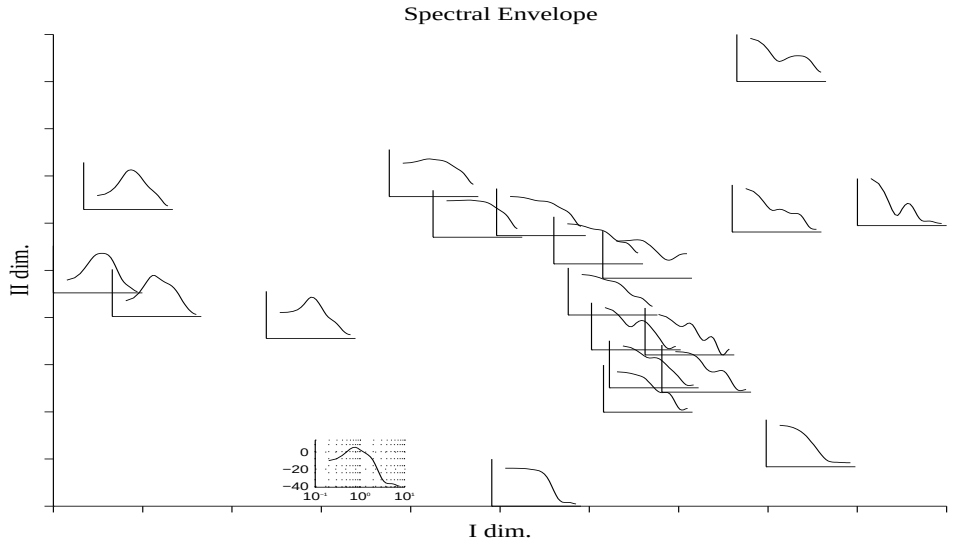

Figure 14: Spectral envelopes (1st and 2nd dimensions)

advantage of this approach, since the spaces derived by SOM or PCA can be seen as a planar mapping of a the inner topology of the particular data set.

A first way to employ the system comes from the ability of the space to represent the relations between sounds in a topological way, thus highlighting both differences and analogies between the elements of the training set. In the synthesis the space can be used to help the user to choose between a predetermined set of sounds according to their different expressivity needs. This would lead to a system of automatic indexing analogous to Feiten's [13]. A different set of analyzed sounds would produce a map with different characteristics; as a consequence, the system would assess itself as a rather flexible tool to represent timbre relations between arbitrary sets of sounds.

A second application towards the synthesis of musical sounds would arise from the MFCC representation itself. Some synthesis parameters can be extracted directly from the MFCC's; the approximate frequency envelope can indeed be obtained as

$$A(f) = \exp[C(\mathrm{mel}(f))].$$

To fully deploy this kind of synthesis it is necessary to go back from the spectral envelopes to the sounds. Of course it is not possible to reconstruct the sound from the spectral envelope only, but the spectral content is indeed a characterizing element of the sound itself which can be used to control the synthesis parameters in many synthesis techniques. From the spectral envelopes, additive synthesis can be carried out as a sum of sinusoids whose instantaneous amplitude is:

$$s(t) = \sum_k A(f_k)\cos(2\pi f_k t + \phi_k)$$

and where the partial frequencies can either be multiples of the fundamental frequency or can be distributed in a less harmonic fashion; they can also vary over time. The amplitude of the partials can be either computed on the fly or retrieved from a table lookup for $A(f)$. When the spectral envelope changes, the table changes too, with the possibility of interpolating the values during the transition.

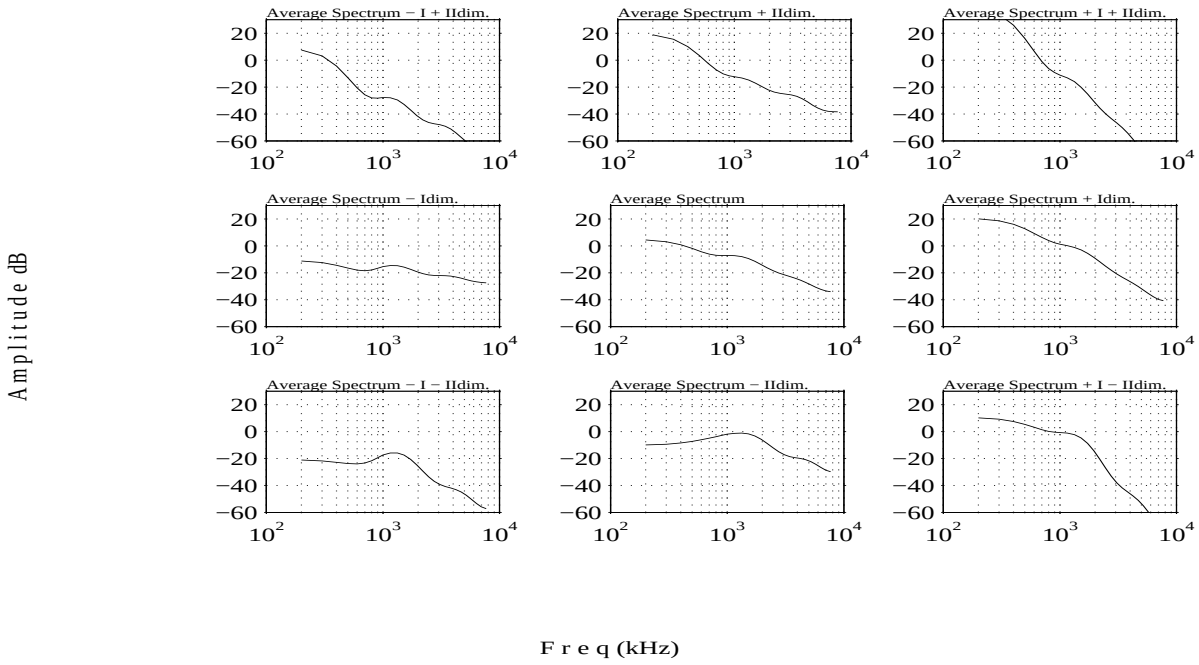

Figure 15: Spectral envelope (1st and 2nd coordinate). McGill data set.

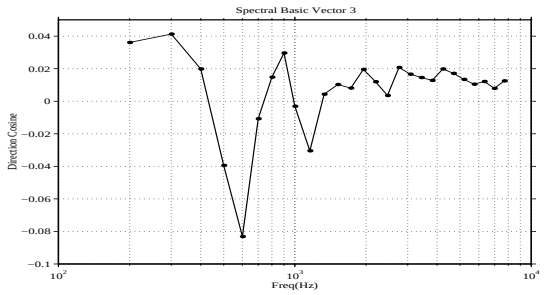

Figure 16: 3rd direction cosine of McGill data set (spectral envelope)

Using the Pitch Synchronous Granular Synthesis [7], a short-time waveform can be computed for each set of MFCC whose spectral magnitude $|G(f)|$ approximates the given spectral envelope, i.e.

$$|G(f)| \approx A(f)$$

When this is computed for all the points in the map, the 'grains' become organized according to the spatial relationships defined by the learning algorithm over the chosen instrumental database. Periodic or quasi-periodic repetition of the grains in time produces sounds with the given spectral envelope.

The spectral envelope can be usefully employed for non-periodic sounds as well [29] by means of inverse FFT of $G(f)$, with random phase values, followed by overlap-and-add of the series of sound segments. Obviously, this stochastic part can be combined with the deterministic part, possibly possessing a different spectral envelope, either by synthesizing the two components separately, or by performing a single inverse FFT for each frame [27].

Since the spectral envelopes embedded in the net are normalized in their amplitudes, the sounds obtained with the methods so far described will have to be scaled by a global amplitude envelope which is not necessarily dependent on their spectral content.

More generally, the spectral envelopes can be employed to control the synthesis by means of filters whose frequency response magnitudes are an approximation of the spectral envelopes

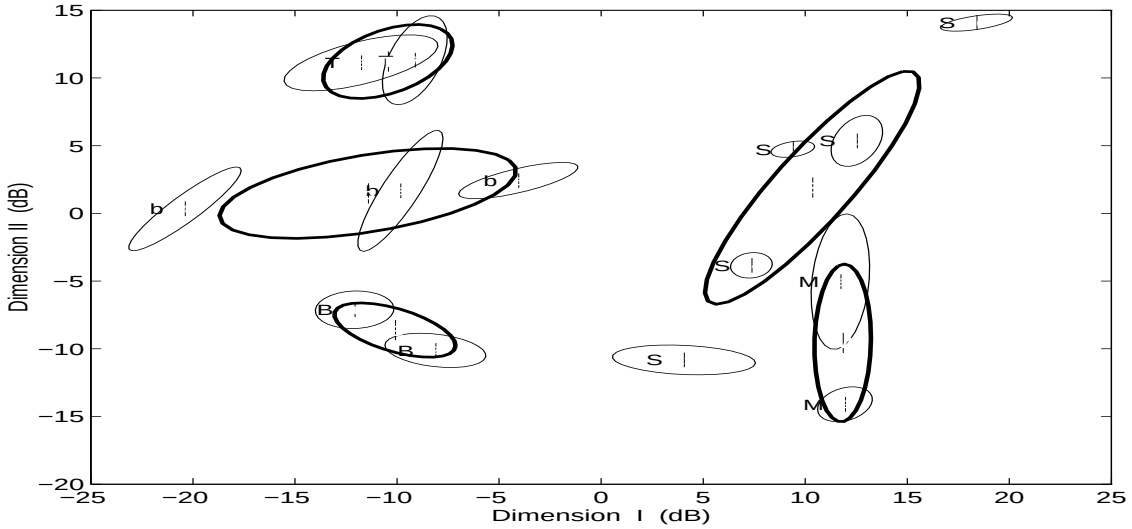

Figure 17: Configuration of points for all the 14 singers (fist two dimensions)

Figure 18: Network trained with ambiance sounds.

themselves. In this case, a different filter is associated to each point in the map. The excitation source for these filters will now play a fundamental role in the resulting sound. A source with a high harmonic content, or with localized energy peaks in different zones of the spectrum will preserve the peculiar quality of the filter. This is a way to control the sound in the so-called cross-synthesis.

6 Conclusions

The results of our research spread in two distinct directions.

On the side of the signal processing tools employed, it appears that the perceptually-based parametrization provided by the Mel-Cepstrum algorithm is well suited to the representation of all perceptually meaningful sounds besides speech; our experiments with natural ambiance sounds have confirmed the fact [5]. The powerful data reduction introduced by the coding technique is well matched to the natural properties of human hearing and retains most of the relevant information.

On the other hand, it has emerged that it is indeed possible to define *algorithmically* a physical timbre space which shows rather surprising analogies with the perceptual spaces previously encountered in the scene of timbre analysis. The properties of timbre which these spaces highlight are fundamentally two: *brightness*, appearing once again as the fundamental axis along which the main differentiation occurs, and a localized energy contribution in the spectrum, called *presence* from an analogy with the mid-band enhancement controls found in audio amplifiers, which seems to account for an independent type of differentiation. While there is surely more than that, one could hypothize that the evolution of classical musical instrument over the centuries has implicitly been ruled by the relevance of these two acoustic qualities. After all, instrumental craftsmen could access and modify these parameters much more easily than, say, the dynamics of the periodic exciter. In other words, the temporal details such as the attack stage, which we left out, are much less amenable to modifications or adjustments, tied as they are to the fundamentals of the instrumental structure; it is not surprising then that temporal clues are the key to *recognizing* an instrument, as they appear to remain constant among the different nuances of tone color. In this light, and on the side of music perception, the results shown here seem to provide good hints for the general debate over the role of temporal details in timbre

perception. While Grey regarded the attack phase as a preeminent factor determining timbre quality, other researchers like Sundberg [30, page 75] maintain the predominance of the features of the steady-state portion when *evaluating* timbre quality, and move the importance of the attack to the act of *recognizing* sounds. The physical timbre spaces we obtained algorithmically seem to endorse this second point of view.

For our analysis to be more complete, however, these temporal factors cannot possibly be disregarded; and, at the same time, we must try to tackle the difficulties of generalizing the results with regard to pitch and to dynamic, These are delicate problems, which have been only hinted at here, and which will be addressed in full in our future research.

References

[1] Blomberg M., R. Carlson,K. Elenius and Bjorn Granstrom, "Auditory Models and Isolated Word Recognition", STL-QPSR Vol. 4, 1983, pp.1-15, 1983.

[2] Bloothooft G. and R. Plomp, "Spectral analysis of sung vowels. III. Characteristics of singers and modes of singing," J. Acoust. Soc. America, vol. 79, no. 4, pp. 852-864, March 1986.

[3] Boatin N., G. De Poli, P. Prandoni, *Timbre characterization with Mel-Cepstrum: a multivariate analysis* Proc of XI Colloquium on Musical Informatics (CIM), pp. 145-148, 1995

[4] Cosi P., G. De Poli and G. Lauzzana, "Auditory modelling and self-organizing neural networks for timbre classification", *Journal of New Music Research*, 23(1): 71-98, 1994.

[5] Cosi P., G. De Poli, P. Prandoni, *Timbre characterization with Mel-Cepstrum and neural nets*, Proc. 1994 ICMC, pp. 42–45, 1994.

[6] Davis, S.B., and Mermelstein, P. *Comparison of Parametric Representations for Monosyllabic Word recognition in Continuously Spoken Sentences*, IEEE Trans. ASSP, vol. 28(4), pp. 357–366, 1980.

[7] De Poli G. and A. Piccialli, "Pitch synchronous granular synthesis". In G. De Poli, A. Piccialli and C. Roads, eds. *Representations of music signals.* Cambridge, Massachusetts: MIT Press, pp. 187-219, 1991.

[8] G. De Poli, P. Prandoni, P. Tonella, "Timbre Clustering by Self–organizing Neural Networks," *Proceedings of the X CIM*, pp. 102-108, Dec. 1993.

[9] De Poli G., P. Tonella, "Self-organizing neural networks and Grey's timbre space", Proc. Int. Computer Music Conf. 1993, p. 260-263, Tokyo 1993.

[10] Evangelista G.P.

[11] Feiten B. Frank, R. and Ungvary, T. (1991). "Organization of sounds with neural nets". *Proceedings of the 1991 International Computer Music Conference.* San Francisco: International Computer Music Association, pp. 441-444.

[12] Feiten B. and S. Gunzel, "Distance measure for the organization of sounds", *Acustica*, vol. 78, pp. 181-184, 1993.

[13] Feiten B. and S. Gunzel. "Automatic Indexing of a Sound Database Using Self-organizing Neural Nets". *Computer Music Journal* 18(3): 53-65, 1994.

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

[15] Grey, J. M., 1975. *An Exploration of Musical Timbre*, Report STAN–M–2, Stanford University.

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

[17] Benade A.H. *Spectral envelopes of Orchestral Instruments*, Journal of the Acoustic Society of America, vol 78, 1985.

[18] Hermansky H., B.A. Hanson and H. Wakita, "Perceptually Based Linear Predictive Analysis of Speech", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-84), paper 13.10, pp.509-512, 1984.

[19] Kohonen, T., *Self-organization and associative memory*, Berlin: Springer Verlag, 1990.

[20] Kohonen, T., "The Self-Organizing Map". *Proceedings of the IEEE*, 78(9): 1464–1480, 1990.

[21] J. Krimphoff, S. McAdams et S.Winsberg, "Cractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique", *Journal de physique IV* Colloque C5, supplément au Journal de Physique III, Volume 4, Mai 1994.

[22] Krumhansl C.L. *Why is musical timbre so hard to understand?*, in S. Nielzen, O. Olsson(ed.), Structure and perception of electroacoustic sound and music, Elsevier, p. 43–53, 1989.

[23] McAdams S., Winsberg S., Donnadieu S., De Soete G., Krimphoff J., "Perceptual scaling of sinthesized musical timbres: Common dimensions, specifities, and latent subject classes", *Psychol. Res.*, vol. 58, pp. 177-192, 1995.

[24] Mannano F., Nobili R., "Biophysics of the coclea I: linear approximation", *Journal of Acoustical Society of America,* Vol. 93, pp. 3320-3332, 1993.

[25] Nobili R., Mannano F., "Biophysics of the coclea II: stationary nonlinear phenomenology", *Journal of Acoustical Society of America,* Vol. 99(3), 1996

[26] Rioul O. and M. Vetterli, "Wavelets and Signal Processing", IEEE Signal Processing Magazine, pp.14-38, October 1991.

[27] Rodet X. and Ph. Depalle, " A new additive synthesis method using inverse Fourier tranform and spectral envelopes". *Proceedings of the 1992 International Computer Music Conference.* San Francisco: International Computer Music Association, pp. 410-411, 1992.

[28] Seneff S., "A joint synchrony/mean-rate model of auditory speech processing", Journal of Phonetics, Special Issue, Vol. 16(1), January 1988, pp. 55-76, 1988.

[29] Serra X. and J. Smith., "Spectral modeling synthesis. A sound analysis/synthesis system based on a deterministic plus stochastic decomposition". *Computer Music Journal* 14(4): 12-24, 1990

[30] Sundberg, J., *The Science of Musical Sounds*, San Diego: Academic Press, 1991.

[31] Strube H.W., "Linear prediction on a warped frequency scale", Journal of Acoustical Society of America, JASA Vol. 68(4), Oct. 1980, pp. 1071-1076, 1976.

[32] Toivianen, P., Kaipainen, M. and J. Louhivuori, "Musical timbre: Similarity ratings correlate with computational feature space distances", *Journal of New Music Research*, 24(3): 282-298, 1995.

[33] Wessel, D., "Timbre Space as a Musical Control Structure", *Computer Music Journal* 3(2): 45–52, 1979.

[34] Evangelista G. and Cavaliere S., "Karplus-Strong Parameter Estimation", *CIM95 Proceedings*, 1995.