Timbre Characterization with Mel-Cepstrum and Neural Nets

Piero Cosi
CNR, Padova
cosi@csrf00.csrf.pd.cnr.it

Giovanni De Poli
University of Padova
depoli@dei.unipd.it

Paolo Prandoni
University of Padova
pran@dei.unipd.it

Abstract

In this work the problem of timbre recognition–classification is addressed by combining the properties of a powerful speech-coding technique, the Mel-frequency Cepstral Coefficients, with the feature extraction capabilities of a self-organizing neural network. Acoustic relationships between tones are reflected into spatial relationships onto a neural lattice. Final results are in good agreement with the usual classifications of timbre quality, and offer promising grounds for the construction of a general, analysis-based timbre space.

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 rôle 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 [Grey,1975] and others 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 and a consequent set of coordinates into which an arbitrary waveform 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 an efficient data reduction while preserving the appropriate information. The aim of this work is to test the effectiveness of a speech analysis technique applied to sound signals by building a simple 'artificial listener'. A parametric representation of timbre information defines an underlying multidimensional timbre space; a self-organizing feature map (SOM) is used to discover and portrait the inner relationships of the space. Comparisons to the two- and three-dimensional analogues obtained from psychoacoustical data are the key to interpret the results.

2 Sound analysis

The *Mel Frequency Cepstral Coefficients*, or MFCC, are a parametric representation of acoustic signals widely used in the field of speech recognition. They were first introduced by Davis and Mermelstein [Davis and Mermelstein,1980] 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:

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

$i = 1, 2, \ldots, M$, where $X_1 \ldots X_N$ are the logarithmic energy outputs of a mel-spaced filterbank. Due to the importance of the higher frequencies in music perception as opposed to speech comprehension, in our case the filterbank spreads up to 8 KHz as shown in fig. 1. The coefficients are computed using a 32 ms Hamming window with a 4 ms time-shift; only the first six coefficients are used, providing an overall 95% data reduction ratio. The actual sound database is drawn from the McGill University Master Samples CD

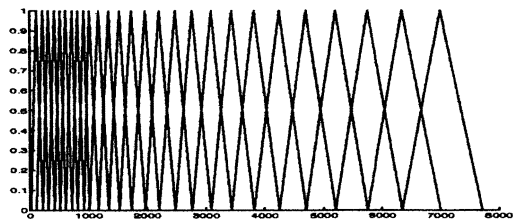

Figure 1: The filterbank.

library considering the first 800 ms of each signal.

While not a cepstral extraction in the usual sense, the actual effectiveness in speech coding is mainly due to the mel-based filter spacing and to the dynamic range compression in the log filter outputs, both of which take into account the actual processes in the early stages of human earing. 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 The neural model

The parametrization algorithm transforms sound signals in sequences of points in a six-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. To explore these properties a self-organizing neural network is used. Kohonen [Kohonen,1990] formalized the learning algorithm for such 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 space from which the input samples come. With $n < m$ (while 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.

In our case the mel-cepstral parameter vectors defined above are used as input to a rectangular net of $15 * 30 = 450$ neurons, so that $n = 2$ and $m = 6$. The training database was built up 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. 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 is performed over a random shuffling of the database vectors, so that there is no direct information upon their correct order in each tone. After the training, when a sequence of vectors is presented as an input, a sequence of neurons (*neural path*) is excited correspondingly. To summarize, the processes characterizing the artificial listener can be viewed as such: the analysis algorithm converts a sound into a sequence of parameter vectors; the neural net converts the sequence of parameter vectors into a sequence of excited neurons. Acoustic properties of tones are thus translates into spatial relationships between neural paths.

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	organo *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.

4 Results

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 the net size, which affects only the number of neurons committed to one tone and thence 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. 2 the ultimate results are shown: all the neural paths relative to the 40 instruments of table 1 are repre-

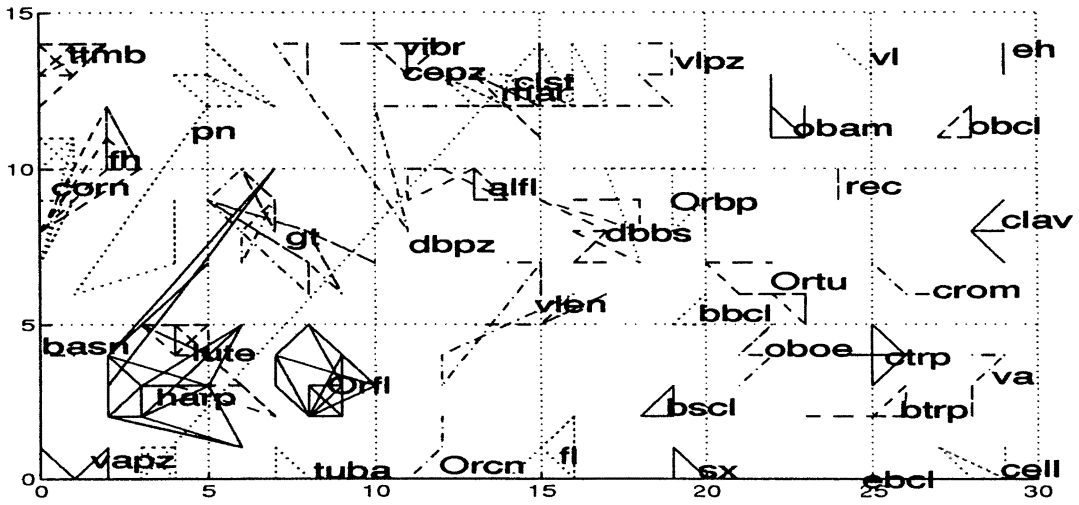

Figure 2: Final neural paths.

sented; while all the time-course of the signals was used in the training phase, in fig. 2 only the steady-state portions are depicted. It can be seen that the paths are generally well separated, with only occasional overlappings 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 sources to wider neural excitements. Most notable, however, is the relative positioning of the paths: in fig. 3 those local groupings are highlighted which correspond to normally defined instrumental families: strings, trumpets, oboes, clarinets, percussions, and plucked strings. Some anomalies are present, e.g. the violin and the harpsichord, which can however be fully explained once the structure of the map is taken into account.

A further analysis of the global ordering reveals that the main axis is strictly related to the spectral energy distribution of the steady-state portion of the tone. The background of fig. 3 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 cofficient 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 *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 brightness orderings. Table 2 compares the neural net to Grey's and Wessel's [Wessel,1979] 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 by which instrumental families are spatially clustered. The grouping anomalies mentioned before can now be explained as an overruling of the second dimension 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

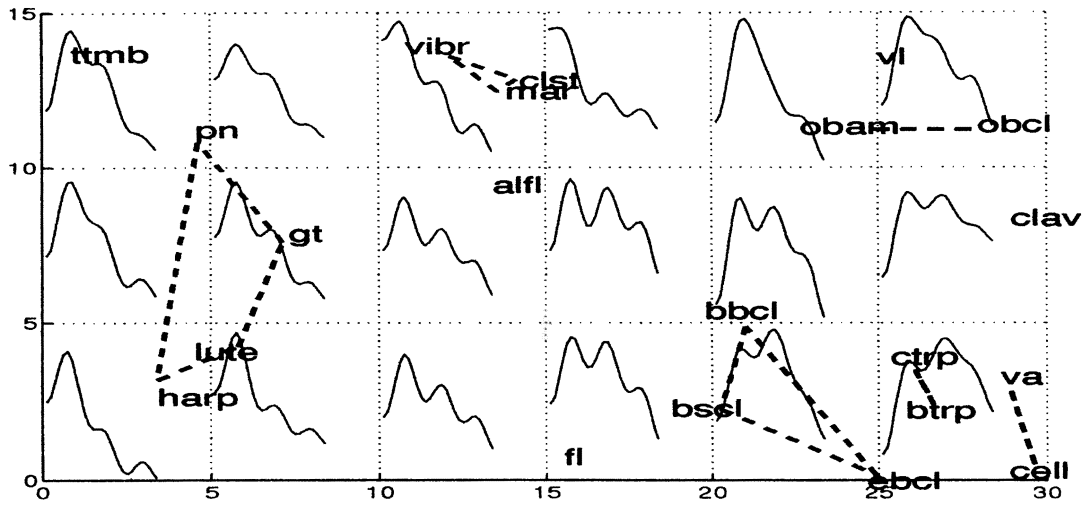

Figure 3: Family clusters and spectral content.

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.

Further studies presently in progress at the University of Padova are confirming the ability of the MFCC to well capture the timbre quality of instrumental tones; this gives the MFCC a broader scope of action in signal coding, and stresses their good match with the properties of human earing. On the side of music perception, the results shown here seem to provide good cues for the debate over the rôle of temporal details in timbre perception. While Grey regarded the attack phase as a preeminent factor determining timbre quality, other researchers like Sundberg [Sundberg,1991, 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 organization provided by the net seems to endorse this second point of view.

Further extensions of this research include the realization of a synthesis control based on the neural timbre space; 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 provided. User-defined 'navigations' over the net will be translated into acoustical transitions across timbre prototypes.

5 Summary

Two techniques were combined in the design of an analysis-based timbre space; the MFCC coding proved effective in extracting the relevant timbre information from sampled sound signals, while a Kohonen's neural network proved effective in organizing the processed data. Analogies with peceptually-based timbre spaces provide encouraging confirmation of the consistency of this combined approach.

References

[Davis and Mermelstein,1980] Davis, S.B., and Mermelstein, P. *Comparison of Parametric Representations for Monosyllabic Word recognition in Continuously Spoken Sentences*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28(4), 1980, pp. 357–366.

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

[Kohonen,1990] Kohonen, T., *The Self-Organizing Map*, Proceedings of the IEEE, Vol. 78, no, 9, 1990, pp. 1464–1480.

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

[Wessel,1979] Wessel, D., *Timbre Space as a Musical Control Structure*, Computer Music Journal, Vol. 3 no. 2, 1979, pp. 45–52.