

# An analysis-based timbre space

Paolo Prandoni

31 July 1994

## Abstract

This paper presents a tentative characterization of the timbre quality of musical sounds by means of an algorithmic parametrization of their features. Specifically, the focus is on the stationary part of the signal, which is coded in form of a series of Mel-Frequency Cepstral Coefficients; relationships between different sounds are accounted for in terms of distances in the coefficients space. To assess the effectiveness of this representation, a qualitative/quantitative analysis is performed which follows the lines of the classic psychoacoustic experiment by J. Grey. A two-dimensional timbre space is obtained with good overall classification properties, and whose axes prove in strict relation with parameters which are easily derived from the spectrum of the signals.

## 1 The characterization of sounds

In a perceptually oriented framework, an acoustic sound has a twofold nature: on one side are its many physical features, which comprise information on its time and frequency envelopes, and possibly on the nature of its source; on the other side is its perceptual value, which represents the known general response of human listeners. In this context, to ‘characterize’ means to point out those physical features of the sound which prove to be related to a given set of perceptual values; eventually this leads to a reduced amount of data in which, at best, only the relevant information is preserved.

As an example, in speech analysis the general goal is to relate the signal to the spoken phoneme; in the *acoustic-phonetic* approach to the problem this is accomplished by means of distinct feature detectors, and the subscope of characterizing voiced sounds, for instance, could be roughly described as locating the position of the formants. However, the perceptual values of interest could be others, such as the speaker’s identity or the overall intonation, and in each case the process of characterization would take into account different features of the uttered sound. This process is performed by means of an analysis algorithm in whose design the important factors are the type of perceptual values sought and the a-priori knowledge about the sound source. In the practice, one can discover this strategy in the efforts to portray the known structure of the inner ear in filterbank designs and, at the other side, in the LPC model of the vocal tract as an all-pole, slow-varying filter.

Moving to more general sounds, however, two concurring problems arise: there could be little or no information on the physical constraints imposed by the source to the waveform, and the related perceptual values could be seldom well defined, as a consequence of the non-categorical perceptual processes involved. Musical sounds form a peculiar subset of all perceptually meaningful sounds; from the historical evolution of sound-producing devices, a class of musical instruments has emerged which well samples the space of natural sounds. These instruments span a wide range of acoustical

possibilities and, with their inner variability in sound quality, render the space of musical sounds a continuum in which it is hard to establish a set of perceptual coordinates. Musicians try to deal with this continuum by placing landmarks which refer to the physical nature of the playing instrument: distinction between strings, woodwinds, brasses, percussions, and other families is an example. Further, musical tradition has developed a set of widely accepted adjectives to distinguish between similar sounds which comprise terms such as brilliant, sharp, or hollow, to name a few. It is very difficult, not to say impossible, to find an algorithmic procedure to quantify these qualities. The psychoacoustic researches which tried to quantify in some way the vague notion of timbre have always relied on the consultation of human listeners, with the major drawback that the subjective ratings thus obtained are often affected by the listener's high-level notions of the structural features of the playing instruments rather than being a pure comparison of sounds. Conversely, in this study we tentatively apply a typical speech coding technique known as Mel-Frequency Cepstral Coefficients to a database of instrumental tones. Although no assumption is made on the input waveform, this type of parametrization takes into account the basic phenomena in human hearing and so exploits the first of the two design strategies mentioned above. To evaluate the effectiveness of this approach we followed the lines of a mile-stone psychoacoustic research, John Grey's "An exploration of musical timbre" [5], replacing subjective comparisons between sounds with a metric distance in the space of the parameters.

## 2 Grey's timbre space

Grey analyzed a large database of similarity ratings between sixteen instrumental sounds as was provided by a group of trained musicians, and proposed a three-dimensional timbre space as the final result.

The experiments started with an analysis-synthesis processing of actual instrumental sounds in order to eliminate all 'non-timbre' difference clues in loudness and duration. By presenting pairwise all the instruments in all their combinations and by collecting the correspondent similarity/dissimilarity judgements from the listeners, expressed in a numerical form, a series of data was obtained which was finally averaged in a set of half-matrices of subjective ratings.

These data were processed in two independent ways. The matrices were given as the input to a multidimensional scaling algorithm (MDS) on one hand, and to a hierarchical clustering (HC) algorithm on the other. The MDS analysis was used to discover a metric arrangement of the stimuli in an Euclidean space which could account for the similarity ratings in terms of spatial distances; the clustering algorithm was used to group similar stimuli regardless of a possible spatial structure underlying. The rationale to assess the goodness of the model was the consistency between the two independent analyses: a spatial model was sought in which the clusters grouped stimuli already close one another. This proved to be the case with a three-dimensional solution.

The subsequent step was the interpretation of the physical dimensions related to the three axes; and, while some plausible hypotheses have been laid, they remain somewhat qualitative and not clear-cut, embedding several different causes whose relative weight is unknown. This is clearly acknowledged by Grey himself who calls attention to the high-level influences on the listeners' responses. On the whole, anyway, the first axis is clearly related to the spectral distribution of energy in the sound, while the other two are connected in a mixed fashion to diverse temporal and spectral features of the tones, with the proviso stated above. The major failure this lack of directness calls forth is the impossibility to 'find the coordinates' of an arbitrary sound inside the timbre space; the positioning of new instruments requires that the experiment be undertaken from the start.

### 3 Tools of analysis

The Mel Frequency Cepstral Coefficients (MFCC) were first introduced by Davis and Mermelstein in a study comparing different techniques for the coding of monosyllabic words [3]. They are defined as:

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

where  $X_1 \dots X_N$  are the log-energy outputs of the filterbank showed in fig 1. The bank consists of 27 triangular filters with unit flat overall response, equally wide and equally spaced in a mel scale; such a scale, linear up to 1 KHz and logarithmic above, is defined as follows:

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

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. The coefficients were computed using a 32 ms Hamming window with a 4 ms time-shift. The actual sound files were sampled at 32 KHz, 16 bits from a compact disk; it follows that the algorithm provides a 95% data reduction ratio. The first 800 ms of each signal were considered; since the attack portion does not exceed 100-150 ms, more than 600 ms of steady-state sound are available to the subsequent analysis. While not a cepstral extraction in the usual sense, the 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. Both these features mimic the physiological processes of the inner ear. Further, the automatic energy normalization and the orthogonal basis introduced by the cepstral derivation justifies the use of a simple Euclidean distance in estimating a comparison between coefficients vectors since, from the Parseval's relation, it is:

$$\sum_i (c_i^{(m)} - c_i^{(n)})^2 = \int \log^2 \left| \frac{X_m(\omega)}{X_n(\omega)} \right| \frac{d\omega}{2\pi},$$

in which the  $X(\omega)$ 's and the  $c_i$ 's are the spectrum and the cepstral coefficients of given signals.

The multidimensional scaling algorithm is a standard version of the classic tool developed by Shepard [6, 2].

The clustering algorithm performs a hierarchical clustering of the data according to the diameter method; the grouping strategy always chooses the minimal distance between points, where a minimal distance between an intermediate cluster and an external point is the minimal of all distances for each element of the set.

### 4 Experiments

The major problems in interpreting perceptual data arise from the difficulty of separating the temporal characteristics from the spectral ones; we chose to leave out temporal details altogether to focus as clearly as possible on the frequency features of the sounds. There has been a long debate about the essentiality, when dealing with timbre, of the quick time evolutions of the signal such as the attack phase. It seems however that these time clues are especially necessary when trying to *recognize the sound source*, whereas they are much less important in *evaluating the quality* of a tone.

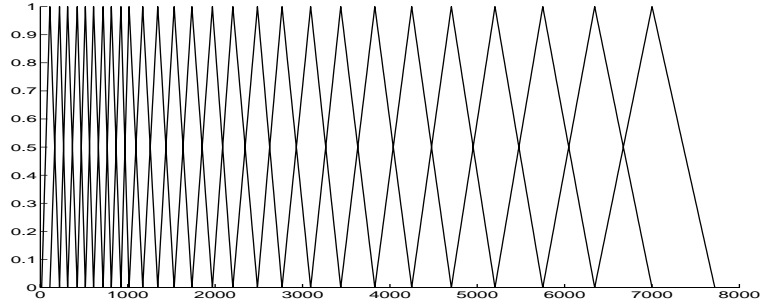


Figure 1: The filterbank used in the computation of the MFCC.

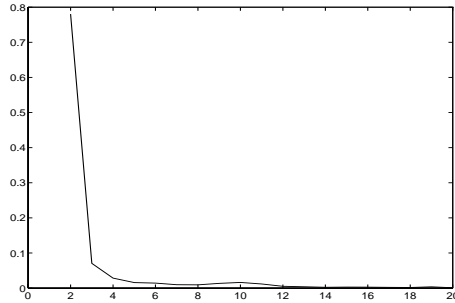


Figure 2: The percentage variation of the rating matrix.

We collected a database of twenty-seven instrumental sounds (table 1), all of which are a C4 (about 268 Hz); this comprises not only the classic sustained instruments, but also percussive and plucked sounds. The audio files, gathered from the CD library McGill University Master Samples, were processed by the MFCC algorithm, which yielded 200 coefficients vectors per file; of these the first 100, in most cases embedding the non-stationary attack stage, were not subsequently employed. The distributions of the 100 points in the coefficients space for each sound proved to form a set of ‘clouds’ more or less tightly but well clustered around their barycenters; the mean radius of these distributions resulted significantly less than the mean distance between barycenters, showing the representations of the instruments to be sufficiently differentiated. From this cursory analysis of the coefficients space, a final representation of sounds seemed adequate in which only the barycenters of the clouds were considered; relations between instruments were portrayed by a matrix of Euclidean distances between barycenters, which from now on will be called the ‘*rating matrix*’.

The number of coefficients needed to obtain a good representation of the data was determined by considering the mean percentage deviation in the rating matrix values against the number of coefficients used. From figure 2 it can be seen that very low values are attained starting from four coefficients; in order to conform to the lines of [3] at no extra charge, we chose to use six coefficients; the mean percentage deviation introduced by a seventh coefficient would be only  $D = 0.96\%$ .

The rating matrix thus computed was finally analyzed by the MDS and the HC algorithms described above. In fig 3 the normalized stress index is represented for an increase of the dimensionality of the solution from one to three. It appears clearly that two dimensions are the proper choice: the final configuration is quickly found in nine iterations, yielding a stress index of  $S = 0.0046$  and a correlation between the two axes of  $6 \cdot 10^{-3}$ .

On the other hand, the instrumental groups produced by the clustering algorithm are shown in

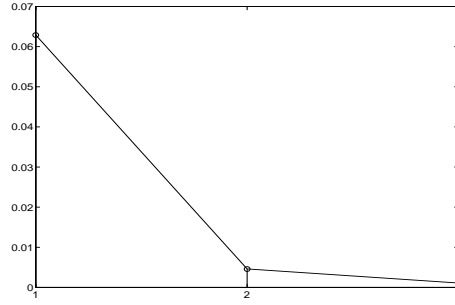


Figure 3: The stress indices for different dimensionalities.

Label	Instrument	Label	Instrument
alf	alto flute	gt	guitar
bbcl	B $\flat$ clarinet	harp	harp
bscl	bass clarinet	lute	lute
btrp	Bach trumpet	mar	marimba
cell	cello	oboe	oboe
clav	harpsichord	obam	oboe <i>d'amore</i>
clst	celesta	obcl	oboe <i>classico</i>
corn	cornet	pn	piano
crom	crumhorn	rec	recorder
ctrp	C trumpet	sx	tenor sax
ebcl	E $\flat$ clarinet	ttb	tenor trombone
eh	English horn	va	viola
fh	French horn	vibr	vibraphone
fl	flute		

Table 1: Instrument labels.

table 2. Qualitatively, that is an extremely good classification which respects almost perfectly the usual family relations between instruments. Group 1 is made up by the two shrillest components of the flute family; group 2 is formed by clarinets; group 3 contains the trumpets; group 4 the strings; group 5 contains the harpsichord alone, which is indeed a ‘hard-to-couple’ instrument; group 6 is formed by the percussive sounds of the vibraphones family; group 7 unites the other trumpet-like instruments; group 8 is occupied by the oboes; group 9 is somewhat meaningless; group 10 covers all the plucked instruments plus the piano. The only two notable misplacements are the oboe in the second group and the English horn in the eighth group; they are however minor ‘errors’ since clarinets, oboes, and horns are not so far off even in Grey’s space.

The important issue is however the perfect consistency of the two concurrent analyses: clusters do not overlap in the two-dimensional timbre space and they embrace localized areas. In figure 4 the timbre space is displayed along with the clusters; as a further step we attempted to provide a physical interpretation of the axes found.

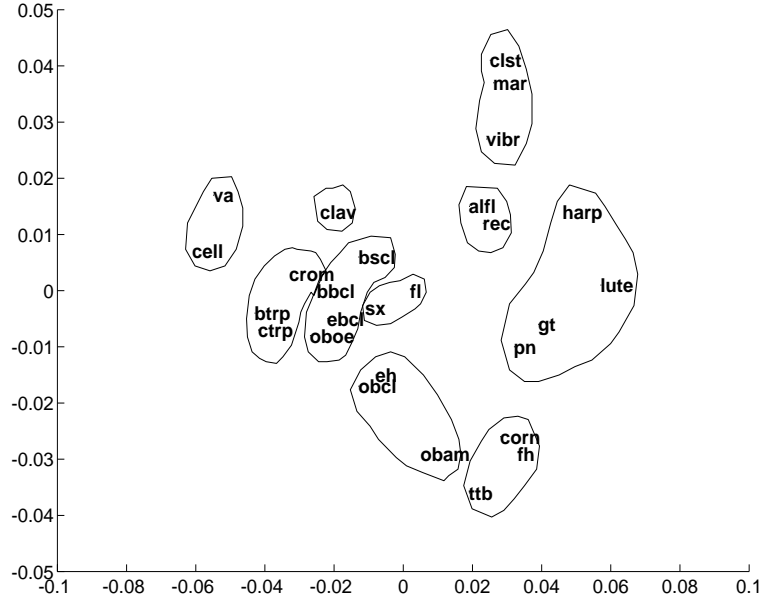


Figure 4: The MDS scaling with the clusters of the HC algorithm.

## 5 Quantitative analysis

In all of their versions, timbre spaces always possess an axis related to the spectral energy distribution of the tone. One of the most important parameters that can be computed from the spectrum of a signal is the barycenter of this distribution, BR, which is a direct measure of the perceptual *brightness* of the sound:

$$\text{BR} = \frac{\int f \cdot X(f) df}{\int X(f) df},$$

#	Instrument labels
1	alfi, rec
2	bbcl, bscl, ebcl, oboe
3	btrp, ctrp, crom
4	cell, va
5	clav
6	clst, mar, vibr
7	fh, corn, ttb
8	obcl, obam, eh
9	fl, sx
10	gt, lute, harp, pn

Table 2: Instrumental clusters.

	$x$	$y$
BR	-0.906	-0.077
PR	-0.160	-0.813

Table 3: Cross-correlations.

which, given the harmonicity of the spectrum, is often computed as:

$$\text{BR} = \frac{\sum_k k a_k}{\sum_k a_k},$$

where the  $a_k$  are the amplitudes of the partials. Indeed, in our timbre space the  $x$ -axis proves to be strictly correlated to the BR index of the sounds employed<sup>1</sup>, with a correlation factor of  $r = 0.906$ ; the BR index is almost completely uncorrelated to the  $y$ -axis ( $r = 0.077$ ). These results agree with and even surpass their past analogues [4].

It is however somewhat more difficult to find a coherent characterization for the vertical axis. One of the best results we obtained was correlating the  $y$ -axis to a PR index (from ‘*presence*’) thus defined:

$$\text{PR} = 10 \log_{10} \int |H(\omega)X(\omega)|^2 \frac{d\omega}{2\pi},$$

where  $H(\omega)$  is a rectangular band-pass filter with unit gain from 700 to 900 Hz. The index is thus a dB measure of the energy content of a localized portion of upper-midband frequency axis; this is in itself a perceptually meaningful region, whose enhancement is a commonly used device in audio equalization to add a different sort of ‘brilliancy’ to the reproduced sound. The correlation factor between PR and the  $y$ -axis is  $r = 0.813$ ; noticeably, in accordance with its being a different sort of tone color, the PR index is little correlated to the  $x$ -axis ( $r = 0.160$ ). Table 3 sums up these results.

The timbre space reconstructed from the computed values of BR and PR is clearly less sharp than the MDS solution; to test its validity, however, the two indices have been computed for a series of piano notes above C4; the location in the space of these notes is clearly clustered around the starting location, as shown in fig. 5. For frequencies above, there happens to be a large shift in the  $y$  coordinate, which is expectable since timbre and pitch are *not* independent.

## 6 Discussion

The results of this study spread in two distinct directions.

On the side of the signal processing tool 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<sup>2</sup>. 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 the spectral features of the steady-state portion of instrumental tones provide two fundamental clues for properly locating timbre in a meaningful

<sup>1</sup>The BR index, and the following PR index, have been computed from the FFT’s of a suitably windowed section of the steady-state portion of the signal

<sup>2</sup>As our experiments with natural environmental sounds have confirmed [1].

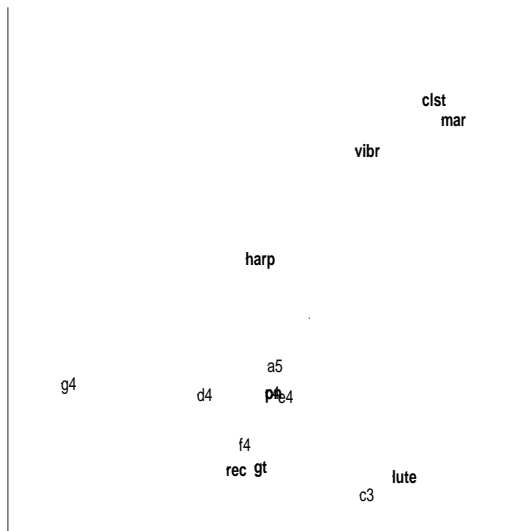


Figure 5: Different piano notes (lighter type) in the reconstructed timbre space.

timbre space. *Brightness* has once again appeared as the fundamental axis along which the main differentiation occurs. The PR index, intended as a peculiarly localized energy contribution in the spectrum and called *presence* from an analogy with the mid-band enhancement controls found in audio amplifiers, seems to account for the vertical differentiation of timbre in our space.

While there is surely more than that, it is however plausible that the evolution of musical instruments has implicitly pursued the adjustment of these two very immediate hearing factors, which are in fact the actual variables allowable to the craftsman. 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.

For our analysis-based timbre space to be complete, however, these clues 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] Cosi P., De Poli G., and Prandoni P., *Timbre Characterization with Mel-Cepstrum and Neural Nets*, ICMC 1994 Proceedings, Aarhus, Denmark, 1994.
- [2] Coxon, A.P.M., *The User's Guide to Multidimensional Scaling*, Heinenmann, 1982.
- [3] 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.

- [4] Ehresman, D., and Wessel, D. *Perception of Timbral Analogies*, IRCAM Technical Report No. 13, 1978.
- [5] Grey, John M. *An Exploration of Musical Timbre*, Report STAN-M-2, Stanford University, 1975.
- [6] Shepard, R.N., *The analysis of proximities: multidimensional scaling with an unknown distance function (I and II)*, Psychometrika 27, 1962.
- [7] Sundberg, Johan, *The Science of Musical Sounds*, Academic Press, San Diego, 1991.