Timbre Characterization with Mel-Cepstrum: a Multivariate Analysis

Nicola Boatin, Giovanni De Poli, Paolo Prandoni
CSC - Department of Electronics and Informatics, University of Padova, Italy

Abstract

In this paper, 21 musical sounds are coded as a series of Mel-Frequency Cepstral Coefficients. By regarding the coefficients as dimensions, a Principal Components Analysis of the data extracts three factors which together account for about 80% of the total variance The interpretation of these factors contributes to a better understanding of the main variables which shape the quality of a sound.

1 Introduction

Timbre is a multidimensional feature which describes the quality of the sound when other factors such as pitch and loudness are kept constant. A classical method has been employed several times in the past to build a timbre space: similarity ratings provided by a group of listeners are organized by means of a Multidimensional Scaling analysis; the empirical interpretation of the dimensions of the resulting timbre space leads to insightful hypotheses on the major features of musical timbres. A correlation procedure usually follows, in which these features are tentatively paired with physical quantities underlying the sound. These studies provided a first general understanding of the major variables which define the musical timbre as a sound feature; on the other hand, a clear methodology has not yet been proposed to extract some "timbral coordinates" from a given, arbitrary sound.

More recent researches attempted to analyze musical timbre starting directly from the acoustic signal. More or less simplified ear models are the tools employed to extract perceptual parameters from the waveform. These parameters are subsequently mapped onto a timbre space, often exploiting the self-organizing capabilities of artificial neural networks.

The representational capabilities of perceptual qualities offered by the Mel-Frequency Cepstral Coefficients (MFCC) are widely employed in speech recognition systems. In [3] we have verified that these capabilities are extremely effective even when applied to the analysis of sound quality. Sounds which appear timbrically close to the listener lead to representations which are topologically close in the MFCC domain. Moreover, the overall topological organization in the coefficients' space of the MFCC representation of sounds appears to be in general agreement with the usual organization of timbre spaces defined by previous psychoacoustical researches.

In this work we try to extend these results by applying a multivariate analysis to several prototypical sets of MFCC-coded musical sounds. The interpretation of the principal components in the resulting space is the starting point for a better understanding of the main variables which shape the quality of a sound.

2 Parametrization

The *Mel Frequency Cepstral Coefficients* (MFCC) were first introduced by Davis e Mermelstein in a comparative study of different speech coding techniques [1]. In the method, filterbank analysis and cepstral analysis are combined: the short-term log-energy output of a mel-spaced filterbank is mapped by a DCT onto the set of the MFCC coefficients. The filterbank is constituted by 27 partially overlapping triangular filters, equally spaced on a mel-frequency scale. The mel scale itself is defined as

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

As opposed to the speech analysis case, the importance of higher frequency components in the perception of musical sounds leads to a version of the filterbank which extends up to 8 KHz. The coefficients are computed using a 23.2 ms Hamming window with a 4 ms time-shift.

An inverse DCT over the coefficients c_i leads back to the spectral envelope, $C(\mu)$, which is expressed in dB and warped along the frequency axis according to the Hz-mel mapping above. As is usual in cepstral techniques, dropping the higher order coefficients (*liftering*) prior to computing the inverse transform leads to a smoothed version of the spectral envelope, $\tilde{C}(\mu)$, which is very often more useful in the subsequent analyses. The first coefficient, c_0, is also dropped. This coefficients is indeed proportional to the energy of the input, so that, by its removal, an amplitude normalized representation of the signal is easily accomplished. This is going to be extremely useful when comparisons between different signals are going to be taken into account.

This cepstral analysis involves a cosine transform, which is an orthogonal transform; the "natural" choice for a distance metric in the MFCC space is thus the Euclidean distance. It can be shown that the Euclidean distance between MFCC sets is equivalent to the distance between (possibly smoothed) spectra in the mel-frequency domain; this seems to be a reasonable measure of the similarities between sounds.

While not a cepstral extraction in the usual sense, the effectiveness of the MFCC 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.

3 Experimental results

Several sets of different musical sounds have been used, all of which led to similar final results. In this paper we will report the analysis of the same set of musical sounds as used by Carol Krumhansl in [2], where musical timbre is studied by means of listeners' responses. The actual sound samples have been obtained by FM sound synthesis on a Yamaha Tx802 synthesizer (Tab. 3). Each of the 21 tones has the same pitch (*Eb* 4, freq. 311.1 Hz), and was sampled at 44.1 kHz.

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 1: Musical Instruments (Yamaha Tx802 synthesizer).

In the present study the quasi steady-state portion of the tone is taken into account. 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 in a multidimensional MFCC space. Experimental evidence shows that the points stemming from the analysis of a single instrument are tightly clustered together. As a consequence, the centroid of the cluster is 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.

The advantages of this representation appear clearly from the analysis of the variance distribution of the data along the different steps of the coding process. Fig. 1 displays the variance (individual and cumulative) of the instrumental data as appears at the output of the filterbank; this is almost uniformly distributed along the entire spectrum, that is, along the axes of the associated space. In other words, in this frequency space there are no predominating axes. In the MFCC space, on the other hand, the variance is primarily distributed across the first coefficients, as appears in fig. 2. By selecting these coefficients as the dominant geometrical coordinates, a MFCC subspace is obtained which adequately represents the data. The first seven coefficients prove to be sufficient to this aim.

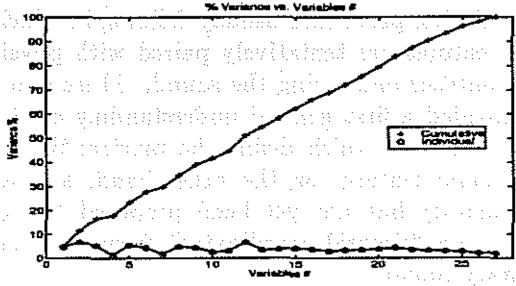

Figure 1: Percent of the total variance associated to the filterbank channels.

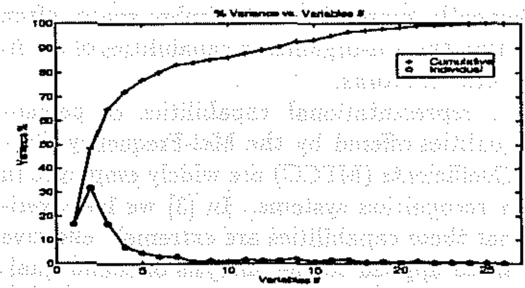

Figure 2: Percent of the total variance associated to the MFCC coefficients

Given the previous results, the next goal is

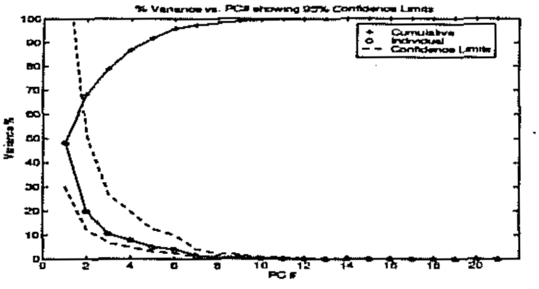

Figure 3: Percent of the total variance associated to the Principal Components

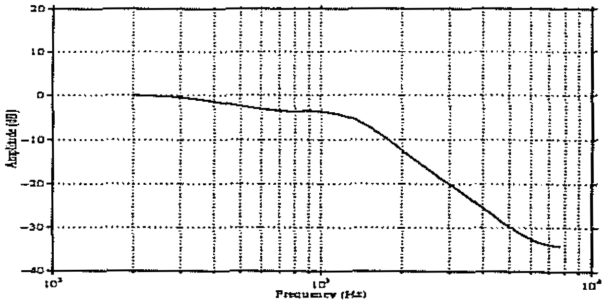

Figure 4: Overall average spectral shape

to find an optimal result in the definition of a minimal subspace in the MFCC space which preserves the relevant information in the signals. The Principal Component Analysis (PCA) is a multivariate statistical tool whose formulation relies on the properties of the *orthogonal linear transforms*. Given a p-dimensional data set, by means of scaling and rotations the principal information is moved onto a reduced set of variables. By carefully chosing the transformation matrix, the new set of variables proves to be uncorrelated: the information pertaining to each of the variables can thus be analyzed in an independent fashion. 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 rotated of coordinates in the space in which the new axes coincide with the directions of maximum variance.

A PC analysis performed on the MFCC-coded instrumental set reveals that the 80% of the variance is concentrated in the first three components 3. A three-dimensional space is thus able to provide a "correct" topological organization within the limits of the retained information.

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. Fig. 4 clearly shows the typical low-pass shape which characterizes almost all musical instruments. In fig. 5, 6 and 7 the three direction cosines related to the PCA spatial transformation are shown. Each one represents the spectral contribution of the correspondent 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 one) of each of the direction cosines. In this new representation, the distribution of the spectral energy along the axes of the new PCA space appears more clearly (fig. 8, 9 and 10).

4 Discussion and results

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. 5) 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. 8. 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 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÷6 KHz. From the plots of fig. 6 it could be assumed that the fundamental characteristics be an apparent spectral irregularity. Fig. 9 shows how

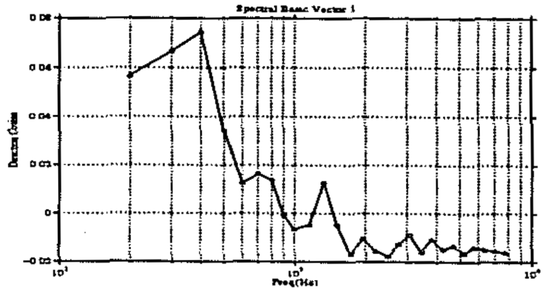

Figure 5: Spectral envelope (1st direction cosine)

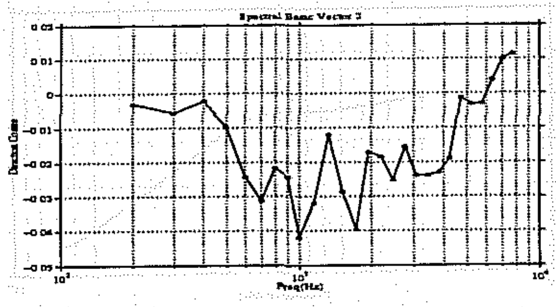

Figure 6: Spectral envelope (2nd direction cosine)

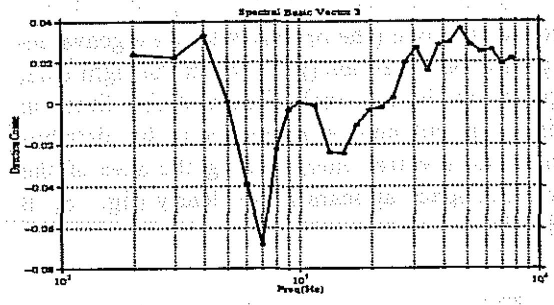

Figure 7: Spectral envelope (3rd direction cosine)

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*.

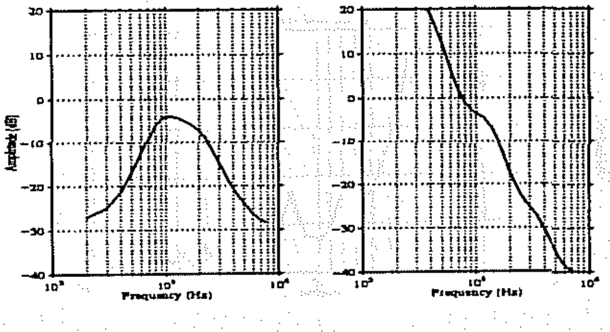

Figure 8: Spectral envelope (1st coordinate)

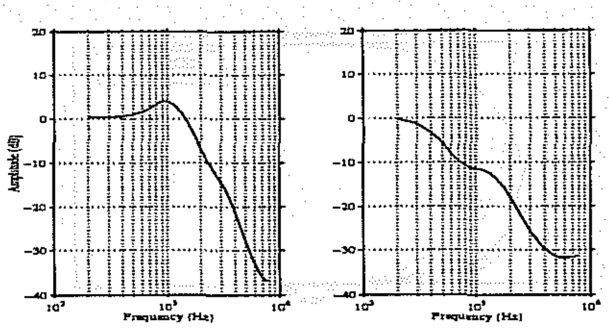

Figure 9: Spectral envelope (2nd coordinate)

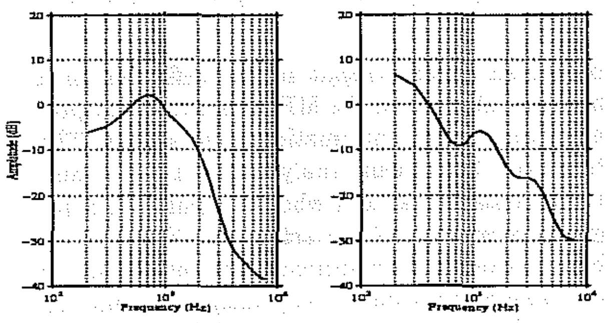

Figure 10: Spectral envelope (3rd coordinate)

5 Conclusions

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 700 and 900 Hz. 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.

References

[1] 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.

[2] 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.

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