Instrument Sound Description in the Context of MPEG-7

Geoffroy Peeters
Geoffroy.Peeters@ircam.fr
IRCAM
(Analysis/Synthesis Team)
1, pl. Igor Stravinsky
F-75004 Paris - France
http://www.ircam.fr

Stephen.McAdams
Stephen.McAdams@ircam.fr
IRCAM / CNRS
(Musical Perception and Cognition)
1, pl. Igor Stravinsky
F-75004 Paris - France
http://www.ircam.fr

Perfecto Herrera
perfecto.herrera@iua.upf.es
IUA / UPF
31, Rambla
E-08002 Barcelona - Spain
http://www.iua.upf.es

ABSTRACT

We review a proposal made in the framework of MPEG-7 for the description of instrument sounds based on perceptual features. Results derived from experiments dealing with the perception of musical timbre allow us to approximate the main underlying perceptual dimensions from signal descriptors. A combination of these signal descriptors allows a comparison between sounds in terms of an estimated perceived timbral (dis)similarity. Since the proposed model is based on experiments using mainly synthesized sounds, a validation of the model was performed on a database composed of natural musical sounds.

INTRODUCTION

The increasing amount of multimedia data over networks and the need for quick and reliable searches, exchanges and distant manipulations of multimedia databases, require an efficient and universally shared description of these media. MPEG-7 [MPEG-7, 2000] assesses this problem by defining a new ISO standard, to be approved in 2001, for the description of multimedia data.

In the sound field, the description of the media consists of a description of the format (encoding format, sampling rate, etc), meta information (author name, copyright owners, etc), semantic information (type of musical event, name of the instrument recorded), but also of a description of the audio content itself. This description of the sound itself should allow a specification of each sound in the database independently of its taxonomy. Although many variables can be proposed in order to achieve this [Wold et al., 1999] [Martin and Kim, 1998], relying on human perception of sounds allows us to define the most important ones.

In this paper, we present a proposal made in the framework of MPEG-7 for the description of sounds based on perceptual features.

Derived from the results of IRCAM's Musical Perception and Cognition and Analysis/Synthesis teams, we define for each class of sounds a set of variables allowing a description of their most essential perceptual features.

We then discuss the representation of this description and its evaluation in the context of a real database. This evaluation was made in a collaboration between IRCAM and IUA/UPF.

1 INSTRUMENT SOUND DESCRIPTION BASED ON PERCEPTUAL FEATURES

Perception of sounds has been studied systematically since Helmholtz. It is now well accepted that sounds can be described in terms of their pitch, loudness, subjective duration, and something called "timbre". "Timbre" refers to the features that allow one to distinguish two sounds that are equal in pitch, loudness, and subjective duration. The underlying perceptual mechanisms are rather complex. They involve taking into account several perceptual dimensions at the same time in a possibly complex way. "Timbre" is thus a multi-dimensional feature which includes among others, spectral envelope, temporal envelope, and variations of each of them. In order to understand better what the "timbre" feature refers to, numerous experiments ([Plomp, 1970, 1976], [Wedin & Goude, 1972], [Wessel, 1979], [Miller & Carterette, 1975], [Grey, 1977], [Krumhansl, 1989], [McAdams et al., 1995], [Lakatos, 2000]) have been performed.

For the purposes of the description of sounds based on perceptual features in MPEG-7, we rely on three of these experiments: 1) Krumhansl's [Krumhansl, 1989] and 2) McAdams, Winsberg, Donnadieu, De Soete & Krimphoff's [McAdams et al., 1995] experiments, which used the 21 FM-synthetic sounds from [Wessel et al., 1987], mainly sustained harmonic sounds, and 3) Lakatos [Lakatos, 2000] (re-analyzed by [McAdams and Winsberg, 2000]), who used 36 sounds from the McGill University sound library, both harmonic (18) and percussive (18) sounds, although we used only the results of the percussive part.

1.1 Signal Descriptors for the perceptual features

In all of these experiments, people were asked for a (dis)similarity judgment on pairs of sounds. These judgments were then used, through a Multidimensional Scaling (MDS) analysis, to represent the stimuli into a low-dimensional space revealing the underlying attributes used by listeners when making the judgments. People often refer to this low-dimensional representation as a "Timbre Space" see Figure 1.

For each of these experiments, people have tried to *qualify* the dimensions of these timbre spaces, the perceptual axes, in terms of "brightness", "attack", etc. Only recently [Grey and Gordon, 1978] [Krimphoff et al., 1994] [Misdariis et al., 1998] have attempts been made to *quantitatively* describe these perceptual axes, i.e. relate the perceptual axes to variables derived directly from the signal: signal descriptors.

1.1.1 *Quantitative description of the perceptual axes*

In the quantitative description of the perceptual axes, the *position* of each sound in the timbre space is explained using signal descriptor values.

For each axis, the descriptor(s) that best explain(s) the variance of the position of the sounds on this axis is(are) chosen. A mathematical relation between the position on the axis and the value of this(these) descriptor(s) is obtained by a linear regression (multiple-regression) method. The position of a sound in the Timbre Space can then be approximated by a linear combination of the descriptor values of each axis. Assuming the orthogonality of the dimensions of the space, perceptual (dis)similarity can be approximated by use of an Euclidean distance model.

1.1.2 *Quantitative description of the perceptual distance*

"Timbre" as it is currently understood, is a relative feature. In this sense what we are interested in is more the description of the relative positions of the sounds than their absolute position in the Timbre Space.

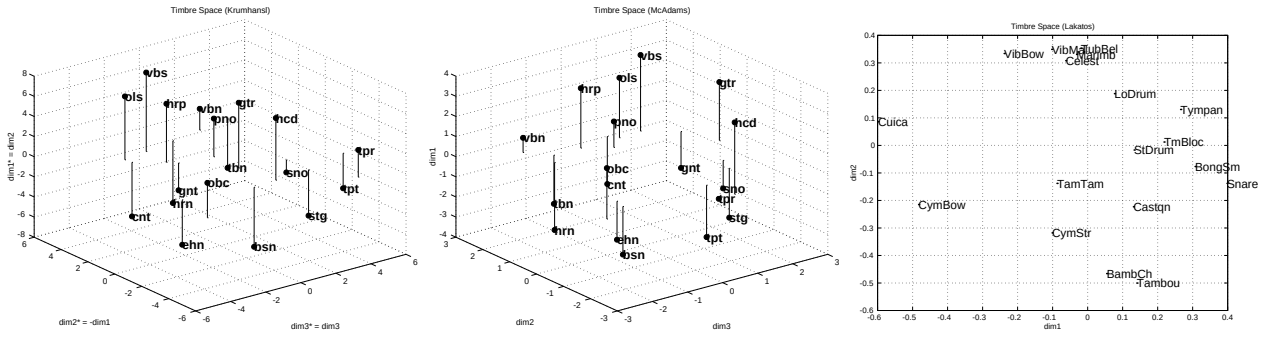

Figure 1: Krumhansl, McAdams et al. and Lakatos (re-analyzed by [McAdams and Winsberg, 2000]) Timbre Spaces

In the quantitative description of the perceptual distance, the perceived *distance* between sounds is explained by a linear combination of descriptor values. For this, the following system of equations is solved using a least-squares resolution: $\underline{A} = \underline{\underline{B}}\,\underline{X}$ where

- $\underline{A}$ is the vector composed of the perceived distance between all possible pairs of sounds (for I sounds, the length of $\underline{A}$ is I(I-1)/2),
- $\underline{\underline{B}}$ is a matrix with elements $b_{i,j}$ being the square of the difference of the value of the descriptor number j for the pair of sounds i,
- $\underline{X}$ is the unknown vector of descriptor coefficients (the length of $\underline{X}$ is equal to the number of descriptors used).

For each choice of a set of descriptors composing $\underline{X}$, an approximation of the vector of perceived distance $\underline{\hat{A}} = \underline{\underline{B}}\,\underline{X}$ is compared to $\underline{A}$. The best set of descriptors $\underline{X}$, i.e. the one with the largest correlation or the smaller modeling error, is chosen.

1.1.3 *Selection of the Audio Descriptors*

For the selection of the signal descriptors that best explain the perceptual features, both *position* and *distance* methods were used. The first one was used in order to check on the independence of the descriptors and to justify the approximation of the perceived distance by a Euclidean model. Then the second one was used in order to estimate the descriptor coefficients.

During the selection, independence of the descriptor to the variation of the analysis parameters (size and shape of the analysis window, hop size) and robustness to the presence of noise in the signal were taken into account. Finally, among the various formulations of each descriptor (mean of the instantaneous values, rms value, maximum value, etc), preference was given to the formulations allowing derivability of descriptor values from one temporal scale to another temporal scale ($x_{T_1+T_2}(t) = \text{mean}[x_{T_1}(t), x_{T_2}(t)]$).

Harmonic Timbre Spaces (see Table):
From the Krumhansl and McAdams et al. experiments, we found results similar to those of Krimphoff et al. and Misdariis et al., except for the third dimension of the McAdams et al. space which, due to the introduction of a new descriptor - the Harmonic Spectral Spread - is now explained at 75% (r=0.87).

After rotation ([dim2, -dim1, dim3]) and normalization (0.43) of the Krumhansl space, in order to make it comparable with the McAdams et al. space, the first dimension is best explained in both cases by Log-Attack-Time, second dimension in both cases by Harmonic Spectral Centroid. Harmonic Spectral Deviation best explains the third dimension of the Krumhansl space, while a combination of Harmonic Spectral Spread and Harmonic Spectral Variation best explain the third dimension of the McAdams et al. space. Orthogonality between third dimensions of Krumhansl and Mc Adams et al. space was therefore considered .

Percussive Timbre Space (see Table):
For the Lakatos experiment (percussive part) the first dimension is best explained by a combination of the Log-Attack-Time and the Temporal Centroid, while the second dimension is best explained by the Spectral Centroid.

1.1.4 *Distance measure*

Using these signal descriptors, each instrument sound can be described in terms of perceptual features and compared to other sounds according to an approximation of the perceived (dis)similarity using the following Euclidean distances:

- Harmonic Timbre Space:

$$d^2(i,j) = (\Delta\text{lat})^2 x_1 + (\Delta\text{hsc})^2 x_2 + (\Delta\text{hsd})^2 x_3 + (\Delta\text{hss}\, x_4 + \Delta\text{hsv}\, x_5)^2 \quad (1)$$

- Percussive Timbre Space:

$$d^2(i,j) = (\Delta\text{lat}\, x'_1 + \Delta\text{tc}\, x'_2)^2 + (\Delta\text{sc})^2\, x'_3 \quad (2)$$

where Δlat, Δhsc, ... are the differences of values for the same descriptor computed on sound i and sound j; $\underline{X}$ and $\underline{X'}$ are the vector of weighting coefficients given in 1.1.2.

1.1.5 *Extraction of the Signal' Descriptors*

Extraction of the signal descriptors is depicted in Figure 2. For the Harmonic Timbre Space, it relies on the previous estimation of an energy function of the signal, the estimation of the fundamental frequency f0 and the detection of the harmonic components of the signal. For the Percussive Timbre Space, it relies on the previous estimation of an energy function of the signal and the estimation of the power spectrum of the signal. In practice, the global spectral envelope was computed as an average over the adjacent harmonic peaks, the Log-Attack Time was computed as the logarithm of the time taken by the energy to go from a threshold of 2% of its maximum amplitude, to 80% of its maximum amplitude.

For validating the proposal, two pilot programs were created. One based on IRCAM's Studio OnLine tools ([Doval and Rodet, 1993], [Depalle et al., 1993]) and one on an extension of the SMS software ([Serra, 1997]).

2 DATA REPRESENTATIONS

In MPEG-7, relationships between descriptors, linking between descriptors and the media, structuring of the document, are done through one or several "Description Scheme(s)". As part of MPEG-7 descriptors, Timbre features take place as a set of variables attached to temporal entities called segments, which are part of the Description Scheme. Due to the nature of the experiments this proposal is based on (monophonic steady sounds

Krumhansl and McAdams et al. experiment signal descriptors

lat: Log-attack time	Defined as the logarithm (decimal base) of the duration from the time when the signal starts to the minimum between [the time when it reaches its maximum value, the time when it reaches its sustained part]
hsc: Harmonic spectral centroid	Defined as the average over the sound duration of the amplitude weighted mean (linear scale) of the harmonic peaks of the spectrum
hss: Harmonic spectral spread	Defined as the average over the sound duration of the amplitude weighted standard deviation (linear scale) of the harmonic peaks of the spectrum, normalized by the hsc
hsv: Harmonic spectral variation	Defined as the average over the sound duration of the one's complement of the normalized correlation between the amplitude (linear scale) of the harmonic peaks of the spectrum of two adjacent frames
hsd: Harmonic spectral deviation	Defined as the average over the sound duration of the deviation of the amplitude (linear scale) of the harmonic peaks of the spectrum from a global spectral envelope

Lakatos experiment (percussive part) signal descriptors

lat: Log-Attack Time	Defined as the logarithm (decimal base) of the duration from the time when the signal starts to the time when it reaches its maximum value
tc: Temporal centroid	Defined as the energy weighted mean of the time of the signal
sc: Spectral centroid	Defined as the amplitude weighted mean of the power spectrum components

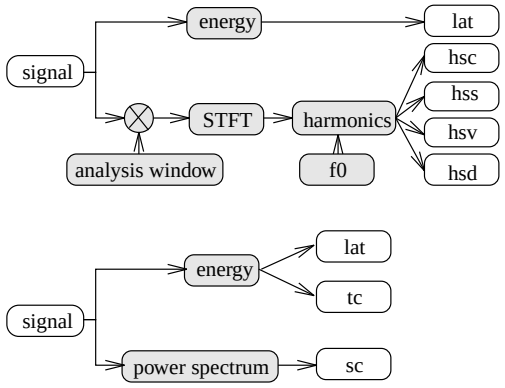

Figure 2: Extraction of the signal descriptors for the Harmonic [TOP] and Percussive [BOTTOM] Timbre Spaces

for the Krumhansl and McAdams et al. experiments, non-mixed, non-sustained sounds for the Lakatos experiment), a confidence coefficient has been added to the set of Timbre descriptors. This coefficient gives us the relevance of the description for the specific segment (blind encoding without any a priori knowledge of the media content may occur). The reliability coefficient takes into account the measure of the periodicity of the signal and the quality of the peak estimation.

3 VALIDATION OF THE DESCRIPTION ON A REAL DATABASE

Given that the main purpose of the descriptors is the facilitation of search-by-similarity operations in databases with musical sounds, the validation of those descriptors was done through a simulated scenario that was conceptually very close to the real one we envision for them (instead of performing another truly psychoacoustical experiment that used (dis)similarity judgments from subjects). Another real-world feature we included is the usage of a large database of non-synthetic sounds. The scenario simulated two different results that had been obtained with two different procedures in a similarity-based search task. The simulated results were supposed to have been generated using the same target sound for both procedures. One of the procedures consisted of a simple random selection of sounds from the database, whereas the other used a given number of the above-discussed descriptors for selecting the sounds. One set of sounds was presented grouped in one half of the screen, and the other set was presented in the other half, as shown in Figure 3. Subjects, who were not aware about the two different selection procedures, had to decide which set of sounds could be considered as being more similar to the target. We expected to find that the sets retrieved with the help of certain combinations of descriptors would be clearly preferred over randomly selected sounds.

Responses, gathered along the continuous grey scale in the figure, were measures of the preference for one set or for the other in terms of similarity to the target. Subjects listened to the sounds by clicking on the note icons, and were allowed to listen to them in any order and as often as they wished. The scores given by the subjects were coded to indicate the degree of preference for the set using the proposed descriptors. Subjects were members of the IUA or IRCAM. The number of participants varied between 18 and 25, depending on the experiment. The sounds used in the experiment were selected using a space-sampling algorithm that ensures a uniform sampling of the feature-space defined by the sound collection. The origin of the sounds was diverse: the Studio OnLine selection specifically compiled for MPEG7 experiments, some public ftp servers, and some copyright-free collections. The number of screens to be judged ranged from 25 to 27, depending on the experiment. Screen allocation of sets was randomized, and so was the order of presentation of screens. In the experimental conditions where several descriptors were involved, we used the weighting scheme presented in section 1.1.4.

In the first two experiments of the series, we studied the sustained sounds. Subjects had to compare a random set versus a set selected using one descriptor in the first one whereas in the second one they compared a random set versus a set selected using the five proposed descriptors. Combining data from both experiments an ANOVA using the number of descriptors as categorical independent variable showed a significant effect on the choice score [$F(1,116)=46.188$, $p \leq 0.001$). Figure 4 [LEFT] shows that when we used the five descriptors, a better selection of sounds could be achieved. Although using only one descriptor globally yields indifferent preferences, there are a couple of them (lat and hsc) that generate sets with higher preference scores.

In the third experiment, devoted to percussive sounds, an ANOVA performed with the number of descriptors as independent variable showed a significant effect on the mean scores ($F(2,51)=65.493$, $p<0.001$). Post-hoc pairwise comparison tests revealed that all differences between the possible combinations are significant. The most relevant result for us is that subjects clearly considered as the most similar to the target those sets of sounds that had been selected using all three descriptors. Figure 4 [RIGHT] shows a summary plot of the results.

From this series of experiments we can reasonably conclude that combinations of the proposed descriptors can be reliably used for performing searches by similarity in databases consisting of musical sounds. Some of the descriptors by themselves yield acceptable although non-optimal retrievals, but a carefully devised weighting scheme can provide results accepted as valid for a large number of people.

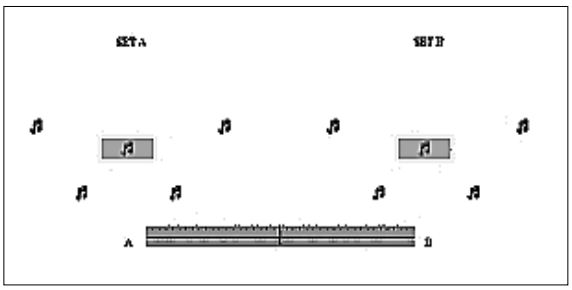

Figure 3: A screen dump of the interface for the experiment

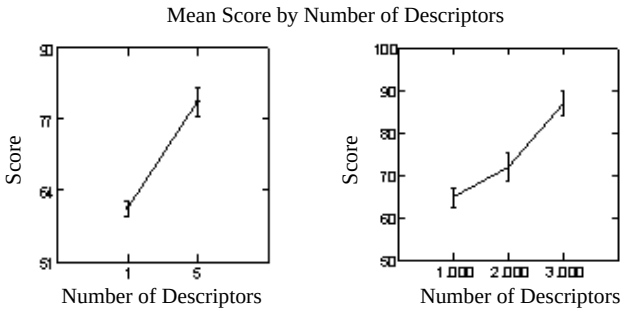

Figure 4: Mean score for the retrieval of sustained sounds [LEFT] percussive [RIGHT] sounds

4 APPLICATIONS

Among the main applications of instrument sound description based on perceptual features are authoring tools for sound designers, musician or database management, retrieval tools for producers or sound design software. Sample databases available today are becoming larger and larger, so much than the usual taxonomical description is becoming poor in comparison to the variety of sounds. These databases would benefit from such a description based on perceptual features (e.g. the Studio OnLine database [Studio-On-Line, 2000] with 160Go, already benefits from such a perceptual description [Misdariis et al., 1998] see Figure 5).

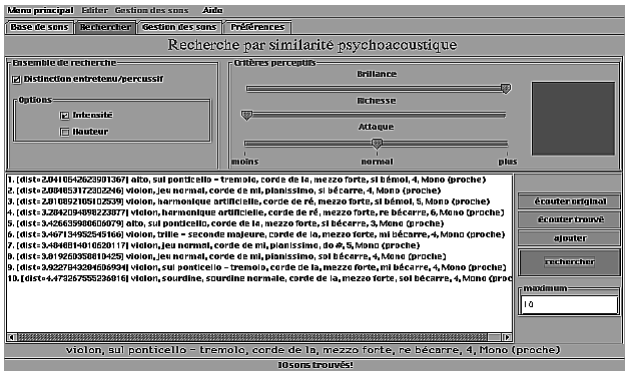

Figure 5: A screen dump of the interface of the search by timbre similarity in the Studio OnLine database

CONCLUSION

For the description of instrument sounds in the framework of MPEG-7, we propose a set of perceptual features derived from experimental studies. Two families of sounds have been considered: sustained harmonic sounds and non- sustained/percussive sounds. This proposal has been validated using a database composed of natural instrument sounds. The validation allows concluding in the adequacy of the proposal in the context of a real database management. Although other variables could be added to the proposed set in order to improve the description, the methodology used here is to rely on experimental studies. Future work will therefore include descriptors derived from other experimental studies dealing with other classes of sounds.

ACKNOWLEDGEMENT

Part of this work was conducted in the context of the European Esprit project 28793 CUIDAD [CUIDAD, 2000].

REFERENCES

[CUIDAD, 2000] CUIDAD (2000). Cuidad working group homepage. http:// www.ircam.fr/ cuidad.

[Depalle et al., 1993] Depalle, P., Garcia, G., and Rodet, X. (1993). Tracking of Partials for Additive Sound Synthesis using Hidden Markov Models. In *Proc. Int. Conf. on Audio, Speech and Signal Proc.*

[Doval and Rodet, 1993] Doval, B. and Rodet, X. (1993). Fundamental Frequency Estimation and Tracking Using Maximum Likelihood Harmonic Matching and HMMs. In *IEEE*.

[Grey and Gordon, 1978] Grey, J. M. and Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am.*, 63(5):1493–1500.

[Krimphoff et al., 1994] Krimphoff, J., McAdams, S., and Winsberg, S. (1994). Caracterisation du timbre des sons complexes. II. Analyse acoustiques et quantification psychophysioque. *Journal de Physique.*

[Krumhansl, 1989] Krumhansl, C. L. (1989). *Structure and perception of electroacoustic sound and music*, chapter Why is musical timbre so hard to understand ?, pages 43–53. S. Nielzen and O. Olsson, Elsevier, Amsterdam (Expcerpta Medica 846) edition.

[Lakatos, 2000] Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics.* in press.

[Martin and Kim, 1998] Martin, K. and Kim, Y. (1998). 2pMU9. Musical Instrument Identification: A pattern-recognition approach. In *Proc. of 136th meeting of ASA*.

[McAdams and Winsberg, 2000] McAdams, S. and Winsberg, S. (2000). A meta-analysis of timbre space. I: Multidimensional scaling of group data with common dimensions, specificities, and latent subject classes. in preparation.

[McAdams et al., 1995] McAdams, S., Winsberg, S., Donnadieu, S., DeSoete, G., and Krimphoff, J. (1995). Perceptual Scaling of synthesized musical timbres: common dimensions, specificities and latent subject classes. *Psychological Research*.

[Misdariis et al., 1998] Misdariis, N., Smith, B., Pressnitzer, D., Susini, P., and McAdams, S. (1998). Validation and Multidimensional Distance Model for Perceptual Dissimilarities among Musical Timbres. In *Proc. of Joint meeting of the 16th congress on ICA, 135th meeting of ASA*.

[MPEG-7, 2000] MPEG-7 (2000). Overview of the MPEG-7 Standard. http:// www.cselt.it/ mpeg/ standards/ mpeg-7/ mpeg7.html.

[Serra, 1997] Serra, X. (1997). *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise. Swets and Zeitlinger Publishers, Poli G. D. ,Picialli A., Pope S. T. and Roads C. edition.

[Studio-On-Line, 2000] Studio-On-Line (2000). http:// www.ircam.fr/ studio-online, http:// sol.ircam.fr.

[Wold et al., 1999] Wold, E., Blum, T., Keislar, D., and Wheaton, J. (1999). *Handbook of Multimedia Computing*, pages 207–227. Boca Raton, FLA: CRC Press, In Furht, B. edition.