Automatically selecting signal descriptors for Sound Classification

Geoffroy Peeters, Xavier Rodet
Ircam - Centre Pompidou
email:peeters@ircam.fr, rod@ircam.fr

Abstract

CUIDADO is a new project (European I.S.T. Project) which aims at providing content-based music applications (Vinet, Herrera, and Pachet 2002). Among these applications is an authoring tool for managing sample databases including search by similarity, search by textual attributes but also a system allowing automatic sound classification based on predefined taxonomies but also allowing user to define its own taxonomies. This last point raises a crucial issue concerning -the design of the classifier but also -the choice of the appropriate signal descriptors in order to perform the classification. This paper concentrates on the design of CUIDADO classifier and on two algorithms for automatically selecting the most appropriate signal descriptors for a given taxonomy: the discriminant analysis and the mutual information.

1 Introduction

Sound classification has raised many interests in the last years (Scheirer and Slaney 1997) (Brown 1998) (Martin and Kim 1998) (Wold, Blum, Keislar, and Wheaton 1999). Most of current sound classification systems rely on the extraction of a set of signal descriptors (such as onset time, spectral centroid,...) which is used latter to perform the classification considering a given taxonomy. This taxonomy is defined by a set of textual attributes defining the properties of the sound such as its source (speech, music, noise, sound effects, instrument name, ...) or its perception (bright, dark, ...) and by a set of parameter's values depending on the model chosen to represent the classes of the taxonomy (multi-dimensional gaussian, gaussian mixture, tree, SVM, ...). The choice of the signal descriptors is specific to each case of classification since the discriminative power of the descriptors depends on the kind of considered sounds (an inharmonicity descriptors is useless to discriminate among only harmonic sounds).

In the case of the CUIDADO classification system, the taxonomy can be user-defined. This involves the system to be able to perform an online-learning including: - choosing among all signal descriptors the ones that are the most relevant for the given taxonomy - estimating from this signal descriptors the parameters of the classes.

2 Signal descriptors

Many different type of signal features have been proposed in the last years in order to describe sound. These come from the speech recognition community (Foote 1994), previous studies on musical sound classification (Scheirer and Slaney 1997) (Brown 1998) (Martin and Kim 1998) (Serra and Bonada 1998) (Wold, Blum, Keislar, and Wheaton 1999) (Jensen and Arnspang 1999) but also from the results of psycho-acoustical studies (Krimphoff, McAdams, and Windsberg 1994) (Peeters, McAdams, and Herrera 2000).

The different choice of features corresponds to different purpose of classification (speech/music/noise, harmonic/percussive sounds, ...). Each set of features is supposed to perform best in its own field. In order to allow covering the wider set of potential taxonomies, in CUIDADO we implemented them all.

2.1 Descriptors taxonomy

The signal descriptors used in our current classification is organized according to the following taxonomy (Herrera, Peeters, and Dubnov 2002). First we distinguish between the time extend validity of the description

Global descriptors: descriptors computed for the whole signal, which meaning is for the whole signal. Example of this are the attack-time of a sound.

Instantaneous descriptors: descriptors computed for each time frame. Example of this are the spectral centroid of a signal which can vary along time. The time vectors of instantaneous descriptors are then processed by a module allowing the modeling of their temporal evolution: mean, standard deviation, derivative, short-term cross-correlation, slope, modulation values.

Inside each class of descriptors, we distinguish descriptors from the kind of signal representation used to extract them.

Temporal descriptors: descriptors (global or instantaneous) computed from the waveform or the signal energy (envelop): Log-Attack Time, Temporal Decrease, Temporal Centroid, Effective Duration, Zero-crossing rate, Cross-correlation

Energy descriptors: descriptors (instantaneous) referring to various energy content of the signal: Global Energy, Harmonic Energy, Noise Energy

Spectral descriptors: descriptors (instantaneous) computed from the Short Time Fourier Transform (STFT) of the signal: Spectral Centroid, Spread, Skewness, Kurtosis, Slope, Decrease, Roll-off point, Variation

Harmonic descriptors: descriptors (instantaneous) computed from the Sinusoidal Harmonic modeling of the signal: Fundamental Frequency, Noisiness, Odd-to-Even Harmonic Ratio, Tristimulus, Deviation, Centroid, Spread, Skewness, Kurtosis, Slope, Decrease, Roll-off point, Variation

Perceptual descriptors: descriptors (instantaneous) computed using a model of the human earring process: MFCC, DMFCC, DDMFCC, Loudness, Specific Loudness, Sharpness, Spread, Roughness

3 Pre-selection of descriptors

Using a wide set of descriptors for the classification may cripple the system since some of them may be irrelevant for the considered class and the estimation of the class parameters may be unreliable. For this reason, a pre-selection of descriptors is necessary. Several techniques has been proposed in order to do that: Principal Component Analysis (Kaminskyj and Materka 1995), Discriminant Analysis(Martin and Kim 1998), Genetic Algorithms (Fujinaga 1998) (sequential backward/forward generation), Neural Networks. Considering the restriction involved by the “online” availability of our system, we considered only computationally attractive techniques: discriminant analysis and mutual information.

Discriminant Analysis (DA) Understanding multidimensional data is the goals of various techniques such as Principal Component Analysis (PCA). The goal of PCA is to perform combination among data such that with a reduced set of orthogonal dimensions most of the initial variance of the data is explained. However, PCA does not allow taking into account data organization such as class belonging. This latter is allowed by the Discriminant Analysis.

Discriminant analysis allows finding combination among variables (in our case the variables are descriptors) in order to maximize discrimination between classes. In the case of the Linear Discriminant Analysis, these combinations are linear. The combinations are represented by a matrix $\underline{\underline{U}}$ which transforms the initial descriptor space D into a new space D' such that only a few axes of D' are necessary to represent class distribution. In the new space, we want the discrimination to be maximum. This criteria can be expressed by choosing $\underline{\underline{U}}$ such that after transformation the ratio of the between-class inertia to the total inertia is maximized.

For a p dimensional descriptors, if we define $\underline{m}$ as the **mean vector** of the descriptors for the whole set of n sounds and $\underline{m}_k$ as the mean vector of the descriptors for the n_k sounds belonging to class k, we can define the **total inertia matrix** $\underline{\underline{T}}$ and the **between-class inertia matrix** $\underline{\underline{B}}$ as

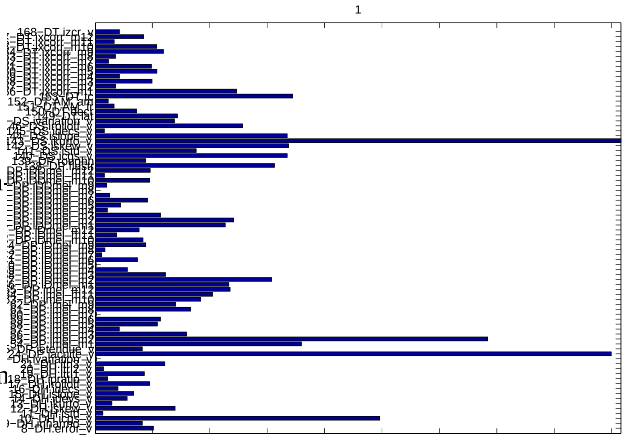

Figure 1: Descriptors selection by DA: weights of the descriptors for the first discriminant axe

$$\underline{\underline{T}} = \frac{1}{n}\sum_{i=1}^{n}(\underline{d}_i - \underline{m})(\underline{d}_i - \underline{m})' \qquad (1)$$

$$\underline{\underline{B}} = \sum_{k=1}^{K}\frac{n_k}{n}(\underline{m}_k - \underline{m})(\underline{m}_k - \underline{m})' \qquad (2)$$

The matrix $\underline{\underline{U}}$ is the one such that after transformation, the ratio between the between-class inertia and the total inertia is maximized. If we note $\underline{u}$ the column vectors of $\underline{\underline{U}}$, this maximization leads to the condition $\underline{\underline{T}}^{-1}\underline{\underline{B}}\underline{u} = \lambda\underline{u}$. The column vectors of $\underline{\underline{U}}$ are then given by the eigen vectors of the matrix $\underline{\underline{T}}^{-1}\underline{\underline{B}}$ associated to the eigen values λ. λ give the discriminative power of each of the new axes.

Descriptors selection with Discriminant Analysis: Each columns of $\underline{\underline{U}}$ represents a combination of the initial descriptors. If the range of each descriptor has been previously normalized, each value in a specific columns gives the weight of each descriptor for a specific dimension and therefore its importance. This is illustred in Figure 1. The selection of the descriptors is based on this weight value: only the descriptors with the biggest weights on each dimensions are retained for the classification.

Mutual Information (MI) Mutual Information is a theory which have been used for features selection as early as 1962. In the context of sound classification, it has been recently used by (Foote 1997) for finding split rules in binary tree construction (binary entropy).

The mutual information between two variables X and Y represents the entropy reduction of X provided by the knowledge of Y. In our case, the mutual information between the class C (qualitative variable) and a specific descriptor D (quantitative variable) is expressed by:

$$I(C,D) = \int\int p(c,d)\log_2\left(\frac{p(c,d)}{p(c)p(d)}\right)\delta c\delta d \qquad (3)$$

The conditional mutual information represents the entropy reduction of C provided by the knowledge of D_1 if we know already the J descriptors D_j. It can be approximated (Battiti 1994) by:

$$I(C, D_1 | D_{j\in[1,J]}) = I(C, D_1) - \alpha \sum_{j=1}^{J} I(C, D_j) \quad (4)$$

where α ranges from 0.5 to 1.

Descriptors selection with Mutual Information: The descriptors are selected according to their mutual information considering a specific set of classes. The first descriptor is the one that leads the largest mutual information given the classes. The following descriptors are the ones with the largest conditional mutual information given the classes and the already selected descriptors.

4 Class modeling

Among the different type of classifier: K-Nearest Neighboring (Fujinaga 1998) (Martin and Kim 1998), Multiple-dimensional classifier, gaussian-mixture, Question-based tree classifier (Jensen and Arnspang 1999), Tree-based vector quantizer classifier (Foote 1997), ... we've chosen a multi-dimensional gaussian model. The choice of the K Nearest Neighboring (KNN) has not been retained since is does not provide an abstraction of the classes and then required the use of the whole database during classification. The choice of a multi-dimensional gaussian mixture model as well as the tree classifiers have not been retained because of their instability and therefore the difficulty to put them in practice in an on-line learning environment.

Learning: For the class k, the parameters of the multi-dimensional function are estimated by the maximum likelihood estimators given the set of pre-selected descriptors of the sounds belonging to the class k. The parameters of the k^{em} class are the mean vector $\underline{\mu}_k$ and the covariance matrix $\underline{\underline{\sum}}_k$.

Evaluation: For a new sound, the descriptor-vector $\underline{d}$ is computed and the probability of the sound to belong to a class k is defined according to Bayes formula $p(k|\underline{d}) = \frac{f(\underline{d}|k)P(k)}{f(\underline{d})}$ where

- $P(k)$ is the "a priori" probability of observing the class (based on the proportion of each classes in the training set and therefore often omitted),
- $f(\underline{d})$ is the distribution of the descriptor-vector $\underline{d}$ which is independent of the classes,
- $f(\underline{d}|k) = A \exp\left(-\frac{1}{2}(\underline{d} - \underline{\mu}_k)' \underline{\underline{\sum}}_k (\underline{d} - \underline{\mu}_k)\right)$ with $A = 1/((2\pi)^{p/2} |\underline{\underline{\sum}}_k|)$ is the conditional probability of observing the descriptor-vector $\underline{d}$ given a class k.

4.1 Descriptor space transformation

The classification can be further improved by the use also of the Discriminant Analysis. This time the Discriminant Analysis is not used for descriptors pre-selection but for space tranformations. The results of the projection of the descriptors on the main discriminant axes is given to the multi-dimensional gaussian model.

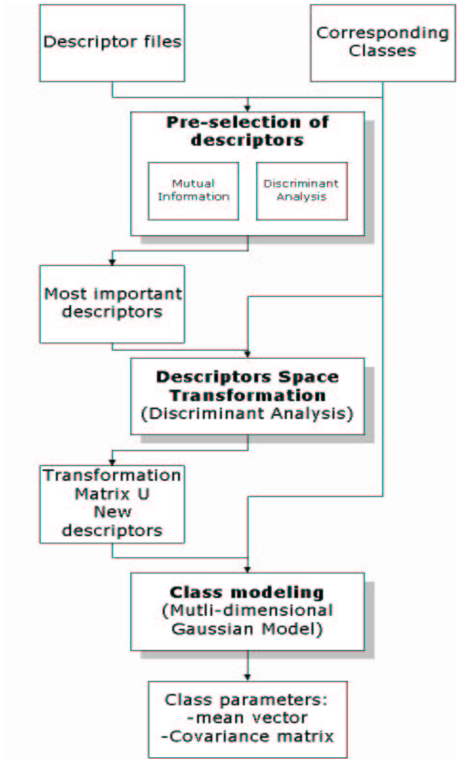
Descriptor files
Corresponding Classes
Pre-selection of descriptors
Mutual Information
Discriminant Analysis
Most important descriptors
Descriptors Space Transformation (Discriminant Analysis)
Transformation Matrix U New descriptors
Class modeling (Mutli-dimensional Gaussian Model)
Class parameters: -mean vector -Covariance matrix

Figure 2: Classification flow-chart

5 Overall design

The overall design of our classification system is depicted in Figure 2. After descriptors pre-selection (by either Discriminant Analysis or Mutual Information) (top part of the figure), a transformation of the space composed of pre-selected descriptors is operated in order to maximize discrimination between classes (middle part of the figure). The result of the projection of the pre-selected descriptors on the main discriminant axes is then given to the class modeling module (bottom part of the figure).

6 Evaluation of the system

6.1 Database used

The evaluatino of the system is performed on a 1400 sounds database composed of extracts from the Ircam Studio OnLine database. The sounds are resampled at 44100 Hz, quantified at 16 bits and mixed in mono. For each considered instrument class, this leads to approximately 100 sounds. Inspired by (Martin and Kim 1998) and (Eronen 2001) taxonomies, we've defined

Score/ Taxonomy	Pizzicato/ sustained	Instrument's family	Instrument's name
All descriptors	96%	89%	86%
Pre-selection by DA	96%	84%	84%
Pre-selection by MI	98%	87%	81%

Table 1: Classification module evaluation

16 instrument classes grouped into 4 instrument families further grouped into pizzicati and sustained instruments (see Figure 3).

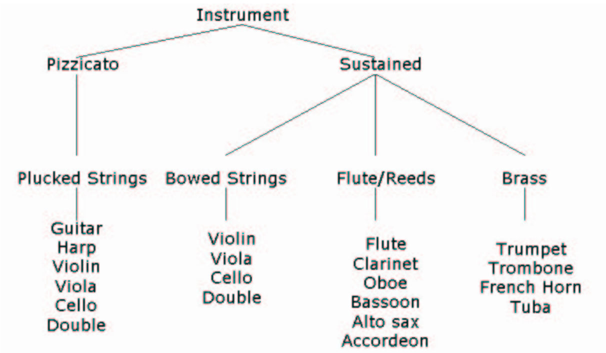

Figure 3: Taxonomy used for instrument classification

6.2 Results

The evaluation is performed using the 1400 sounds of the sound database. The learning is performed by selecting randomly 66% of the sounds of each class (class = pizzicato/ sustained, class=instrument's families or class = instruments). The evaluation is then performed on the remaining 33% of the sounds of each class. The ratings indicated in table 1 correspond to percentage of good classification over the total number of sounds to be classified. Because of the random process, the scores given are mean values over several random sets. For the pre-selection of the descriptors, the selection is performed using all samples of the database. The number of initial descriptors is 81. Pre-selection of descriptors by Discriminant Analysis (DA), (by taking only the descriptors with a value above 20% of the maximum descriptor value on the axe) reduces it to 27 descriptors; using Mutual Information (MI), we kept only the 20 first descriptors. Only the first heigth discriminant axes are considered.

The results, indicated in Table 1, show that the preselection of descriptors using Mutual Information performs better than the one using Discriminant Analysis (higher score with less descriptors). Comparing the results obtained using the whole set of descriptors to the results obtained using only the pre-selected ones shows a slight decrease of performance which is compensated by a 75% reduction of space dimensionality and an equivalent gain of computation-time.

7 Conclusion

In this paper we depicted the current classification system proposed for CUIDADO sample database management application. Considering the possibility given to the user to define its own taxonomies, the system should be able to select automatically which signal features are relevant to perform the classification. We studied the applicability of the Discriminant Analysis and the Mutual Information in order to do that and evaluate them in the context of musical sounds classification. The results shows that among both, the Mutual Information performs best for features selection. Given this open framework, where additional features can be included, further works will concentrate on the evaluation of the system for non-instrumental sounds.

8 Aknowledgement

Part of this work was conducted in the context of the European I.S.T. project CUIDADO.

References

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans on NN 5*(4), 537–550.

Brown, J. (1998). Musical instrument identification using autocorrelation coefficients. In *Proc. Intern. Symposium on Musical Acoustics*, pp. 291–295.

Eronen, A. (2001). *Automatic Musical Instrument Recognition*. Ph. D. thesis, Tampere University of Technology.

Foote, J. (1994). *Decision-Tree Probability Modeling for HMM Speech Recognition*. Ph. D. thesis, Cornell University.

Foote, J. (1997). A similarity measure for automatic audio classification. In *AAAI (Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora)*, USA.

Fujinaga, I. (1998). Machine recognition of timbre using steady-state tone of acoustical musical instruments. In *ICMC*, Ann Arbor, USA, pp. 207–210.

Herrera, P., G. Peeters, and S. Dubnov (2002). Automatic classification of musical instrument sounds. *submitted to JNMR*.

Jensen, K. and K. Arnspang (1999). Binary decision tree classification of musical sounds. In *ICMC*, Bejing, China.

Kaminskyj and Materka (1995). Automatic source identification of monophonic musical instrument sounds. In *IEEE Int. Conf. on Neural Networks*.

Krimphoff, J., S. McAdams, and S. Windsberg (1994). Caractrisation du timbre des sons complexes. ii: Analyse acoustiques et quantification psychophyisique. *Journal de physique 4*, 625–628.

Martin, K. and Y. Kim (1998). 2pmu9. instrument identification: a pattern-recognition approach. In *136th Meet. Ac. Soc. of America*.

Peeters, G., S. McAdams, and P. Herrera (2000). Instrument sound description in the context of mpeg-7. In *ICMC*, Berlin, Germany.

Scheirer, E. and M. Slaney (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *ICASSP*, Munich, Germany.

Serra, X. and J. Bonada (1998). Sound transformations based on sms high level attributes. In *DAFX*, Barcelona, Spain.

Vinet, H., P. Herrera, and F. Pachet (2002). The cuidado project: New applications based on audio and music content description. In *submitted to ICMC*, Goteborg, Sweden.

Wold, E., T. Blum, D. Keislar, and J. Wheaton (1999). Classification, search and retrieval of audio. In B. Furth (Ed.), *CRC Handbook of Multimedia Computing*, pp. 207–226. Boca Raton, FLA: CRC Press.