Learning optimal descriptors for audio class discrimination *

Bertrand Delezoide, Xavier Rodet
IRCAM - Centre Pompidou
1 Place Igor Stravinsky
Paris, France
bertrand.delezoide@ircam.fr, xavier.rodet@ircam.fr

ABSTRACT

Nowadays, an important and growing quantity of audio information is available by means of public or private databases and TV/radio broadcasts. So researches in audio indexation aim to fulfil the need of (semi-)automatic tools for audio content description. This description involves classifying audio signal into a predefined scene type, and indexing and summarizing the document for efficient retrieval and browsing.

This paper suggests an optimal system for discriminating audio signal on a specific and simple taxononomy: speech, music, and a mixture of these two signal types. Different existing techniques for signal description, class modeling, and classification were studied and implemented. The contributions of this paper include testing existing techniques and combining them to form a complete system. Tests with a wide variety of audio sequences prove the efficiency of this classification system.

1. INTRODUCTION

Nowadays, an important and growing quantity of audio and video information is available by means of public or private databases and TV/radio broadcasts. Nevertheless, the wealth of information arises the problem of an adapted access.

Thus, there is a need for (semi-)automatic tools for description of the contents, which would make it possible to ensure consultation and a convivial and effective handling of the data. Problem arise regarding the extraction of representative information from the contents of the documents, and of organization of their representation in a structure adapted to the processes of request. Indeed, what one calls audio or video content includes extremely varied information that can be interpreted at various levels, giving as many profiles of requests.

This step falls under a current mobility illustrated by the installation of the Iso-mpeg7 standard, which contrary to the previous ones, is not devoted to the transmission but to the description of the contents of the audio and video documents.

This task is assured by a sound (or video) classification system. Most of current sound classification systems rely on the extraction of a set of audio signal descriptors. This set is used later to perform a classification considering a given taxonomy. This taxonomy is defined by properties of the sound such as (speech, music, noise...) and by a set of characteristic values depending on the model chosen to represent the classes belonging. The choice of the signal descriptors is specific to each taxonomy, since the discriminative power of the descriptors depends on the kind of taxonomy chosen to classify sounds. THere is a large scale of possible taxonomies, and the choice of relevant descriptors is not that obvious: For a simple classification problem such as speech/music discrimination, a large amount of descriptors were presented in the literature. It is thus interesting to build a system that is able to perform the choice, among all signal descriptors, of the ones that are the most relevant for the given taxonomy, and then to estimate the parameters of the classes from the signal descriptors of a classified learning database.

This paper suggests an optimal system for discriminating speech, music, and a mixture of these two signal types. The structure of this article is as follows: first of all, simple existing features are listed. In the following section some descriptors selection algorithm are presented (mutual information, discriminant analysis, normalized cuts...). In continuation, models used to represent the classes of a taxonomy are studied (k-nearest neighbors, multi-dimensional gaussian, gaussian mixture...). In the last part, these existing techniques are tested on a simple classification problem: speech, music, and mixture of these two signal types. The best algorithms are combined to form a complete optimal system. Lastly, we expose our conclusions as well as the prospects that appear in the field of audio indexing and multi-media indexing.

2. FEATURES EXTRACTION

2.1 Introduction

One of the keys to the success of any audio content analysis algorithm is the type of features employed for the analysis. These features must be able to discriminate among different target scene classes. Many features have been proposed for this purpose. Some of them are designed for specific tasks, while others are more general and can be useful for a variety of applications. In this section we review the features that we use for our description task.

There are many features that can be used to characterize audio signals. Usually audio features are extracted at two levels: short-term frame level and long term clip level. Here a **frame** is defined as a group of neighboring samples lasting about 10 to 40 ms, within we can assume that the audio signal is stationary and short-term features such as energy and Fourier transform coefficient can be extracted.For the features to reveal the semantic meaning of an audio signal, analysis over a much longer period is necessary. Here we call such an interval an **audio clip**. A clip consists of a sequence of frames, and a clip-level feature usually characterizes how frame-level features change over a clip. The clip boundaries may be the result of audio segmentation such that the frame features within each clip are similar. Alternatively, fixed length clips, usually between 1 and 5 seconds, may be used. Both frames and clip may overlap, and the overlapping lengths depend on the underlying application.

2.2 Frame level features

Most of the frame-level features are inherited from traditional speech signal processing. Generally they can be separated into two categories: time-domain features, computed directly from the audio waveforms, and frequency-domain features, derived from time/frequency representations.

Time domain features:

- The most widely used and easy to compute frame features are **energy** and **volume contour**[1]. They are reliable indicators for silence detection, which may help to segment an audio sequence, or determine a modulation in the signal.
- Besides the volume, the **zero crossing rate (ZCR)**[2] is another widely used temporal feature.
- **Fundamental frequency (F_0)**[3] of an audio waveform is an important parameter in the analysis and synthesis of speech and music. Only voiced speech and harmonic music have well defined fundamental frequencies. But we can still use F_0 as a low-level feature to characterize the fundamental frequency of any audio waveform.
- The **Spectral "Flux"**[4] (or Delta Spectrum Magnitude) is The 2-norm of the frame-to-frame energy difference vector.

Frequency domain features:

- The simplest feature in this category is the **short time energy**. When looking to model signal's energy characteristics more accurately, one can use **sub band short-time energies**[5].
- There are quite a few features that are designed for characterizing the information complexity of audio signals, including **bandwidth**[1] and **entropy**[6].
- **Spectral centroid**[4]: The "balancing point" of the spectral power distribution.
- The 95th percentile of the power spectral distribution or **spectral rolloff point**[4].
- The division of subbands based on human auditory system is not unique. A widely used subband division is the **Melscale subband**[7].
- For automatic speech recognition, many phoneme level features have been developed. **Mel-Frequency Cepstral Coefficients (MFCC)**[7] is one of them.
- The **Discreet Wavelet Transform (DWT)**[8] is a technique for analyzing signal developed as an alternative to the Short Time Fourier Transform (STFT).

2.3 Clip level features

As described before, frame-level features are designed to capture the short-term characteristics of an audio signal. To extract semantic content, one needs to observe the temporal variation of frame features on a longer time scale. This consideration leads to the development of various clip level features, which characterize how frame level features change over a clip.

The extracted frame level features provide a compact representation of an audio clip. In order to further reduce the dimensionality of the extracted feature vectors, statistics over the set of features are used.

The following features are used in our system:

- The mean on a clip of each frame level feature.
- The standard deviation on a clip of each frame level feature.
- Ratios between the subband mean values are computed for subband short-time energies, MFCC and DWT,
- The variation of the ZCR is more discriminative than its exact value, so we use **high zero-crossing rate ratio (HZCRR)**[2].
- **Low short time energy ratio (LSTER)** is defined as the ratio of the number of frames whose short time energy is less than 0.5 times the short time energy average[2].
- From the result of silence detection, we calculate the **Silence Ratio**[1], which is the ratio of the whole silence intervals to the length of the entire clip.
- To measure the variation of an audio clip's amplitude, the **volume dynamic range (VDR)**[1] was defined.

- The signal contour of a speech waveform typically peaks at $4Hz$. To discriminate speech from music, an efficient feature is the **four hertz modulation (4ME)**[4].

3. PRE-SELECTION OF DESCRIPTORS

Our system aims at classifying audio clips by finding the class belonging of those clips according to their 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 preselection of descriptors is necessary. Several techniques have been proposed: Principal Component Analysis[9], Genetic Algorithms[10][11], Neural Networks[12]. This preselection is done in our case by three different algorithms. Peeters[13] presented the first two and the third one is an image segmentation algorithm. Let p be the dimension of the clip descriptor at this step of the system: $p = 93$. A set of classes C, labeled as $c_k \in C$, is given *a priori*, as well as the data vectors d_i (for the i^{th} audio clip), of dimension p.

3.1 Discriminant Analysis

Understanding multidimensional data is the goals of various techniques such as Principal Component Analysis (PCA). PCA aims at performing combinations among the descriptors such that, with a reduced set of orthogonal dimensions, most of the initial variance of the data is explained. However, PCA does not take into account data organization such as class belonging. This is done by the **Linear Discriminant Analysis (LDA)**[14].

Linear Discriminant Analysis seeks a linear combination among variables (descriptors in our case) in order to maximize discrimination between classes according to an inertia criterion. This criterion can be expressed by choosing the combination vector u, so that after transformation, the ratio of the between-class inertia to the total inertia is maximized. For a p dimensional descriptor, if we define m as the mean vector of the descriptors for the whole set of n sounds and m_k as the mean vector of the descriptor vector d_i for the n_k sounds belonging to the class c_k, we can define the total inertia matrix T and the between-class inertia matrix B as :

$$T = \frac{1}{n}\sum_{i=1}^{n}(d_i - m)(d_i - m)' \quad (1)$$

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

We compute the generalized eigenvalues and right eigenvectors of the matrix pair (B, T) that satisfy $BU = TU\lambda$. The columns of the matrix U are the generalized eigenvectors of the matrix pair (B, T). λ, the eigenvalues, gives the discriminative power of the new axes.

Each columns of U represents a linear combination of the initial descriptors. As the range of each descriptor has been previously normalized, each value in a specific column gives the weight of each descriptor for a specific dimension and hence its importance. For our discrimination the weights of the descriptors for the first discriminant axis(first column of U) is used: the one with the biggest eigenvalue. Only the 15 descriptors with the biggest weights are retained for the next operation.

3.2 Mutual Information

Mutual information (MI) is a theory that has been used for feature selection as early as 1962. Battiti[15] has used it, in the context of feature selection for pattern recognition. The mutual information between two variables X and Y represents the entropy reductions of X provided the knowledge of Y. In our case, the mutual information between the class c_k and a specific descriptor d_i is expressed by:

$$MI(c_k, d_i) = \int\int p(c_k, d_i).log_2\frac{p(c_k, d_i)}{p(c_k).p(d_i)} \quad (3)$$

We define the mutual information between the set of classes C and a descriptor d_i as:

$$MI(C, d_i) = \sum_{c_k \in C} MI(c_k, d_i) \quad (4)$$

The descriptors are selected according to their mutual information considering a specific set of classes C. We keep the 15 descriptors with the highest mutual information.

3.3 Normalized Cut

The **normalized Cut** is a theory that has been used for image segmentation [16]. The normalized criterion measures both the total dissimilarity between the different classes and the the total similarity within the classes. c_{l1}, c_{l2} are two classes:

$$Ncut(c_{k1}, c_{k2}) = \frac{cut(c_{k1}, c_{k2})}{assoc(c_{k1}, C)} + \frac{cut(c_{k1}, c_{k2})}{assoc(c_{k2}, C)} \quad (5)$$

where $cut(c_{k1}, c_{k2}) = \sum_{p \in c_{k1}, q \in c_{k2}} w(p, q)$ is the total connection from clips in c_{k1} to clips in c_{k2}, $assoc(c_{k1}, C) = \sum_{p \in c_{k1}, q \in C} w(p, q)$ is the total connection from clips in c_{k1} to all clips in the graph.And

$$w(p, q) = exp(-\|d_i^p - d_i^q\|^2) \quad (6)$$

where d_i^p is value of the descriptor d_i for the clip p, is the similarity measure between the clip p and q.

With this definition of dissociation between classes, the cut that partitions out small isolated points have large Ncut value. Hence by maximizing the Ncut of a class we eliminate descriptors that offer large distribution for a class. We define the normalized cut between the set of classes C and a descriptor d_i as:

$$Ncut(C, d_i) = \sum_{c_k, c_l \in C, k \neq l} Ncut(c_k, c_l) \quad (7)$$

The descriptors are selected according to their normalized cut considering the specific set of classes. We keep the 15 descriptors with the largest normalized cut.

3.4 Choice of the best descriptor

The system has to find the best linear combination of descriptors from the 15 best descriptors.

For that, for each type of descriptors preselection (LDA, MI, Ncut), we test all the combinations of its 15 best descriptors.

We define D_i^q the qth combination of descriptors for the clip i:

$$D_i^q = \sum_{k=1..15} a_k d_i^k \text{ where } a_k = 1 \text{ or } 0. \quad (8)$$

For all the combinations of descriptors D_i^q, LDA, MI and NCut are computed. The combinations of descriptors that provide the largest MI, LDA and NCut are selected for the next step (Table 1). It makes 2^{15} combinations to test for each type of pre-selection, which represents a huge amount of calculation. But it seemed to us the best way to eliminate features that are important in the class dissociation and choose the best combination of feature. For LDA or MI, some heuristic approaches have been suggested in order to reduce the computing time. These techniques are based on selecting features in a stepwise mode. For MI each new feature selected has the highest individual MI and the lowest possible joint mutual information[17]. For LDA, **stepwise linear discriminant analysis**(swLDA) has been introduced by R.Jennrich[18] in the late seventies. It should be emphasized that both techniques are sub-optimal as they are not guaranteed to provide consistently the best solution to the feature selection problem. The system finds for each preselection 4 or 5 features whose combination gives the best class dissociation. Let us remark that, for all of the preselection, the best combination uses some descriptors that were ranked between 10 and 15. It is therefore impossible to rely on the first-ranking to have a good preselection, which raises some questions about the efficiency of stepwise selection models. The arbitrary choice of 15 features is not optimal. But absolutely no existing techniques offer the possibility to eliminating features (a non-linear operation) in an optimal way.

4. DESCRIPTOR SPACE TRANSFORMATION

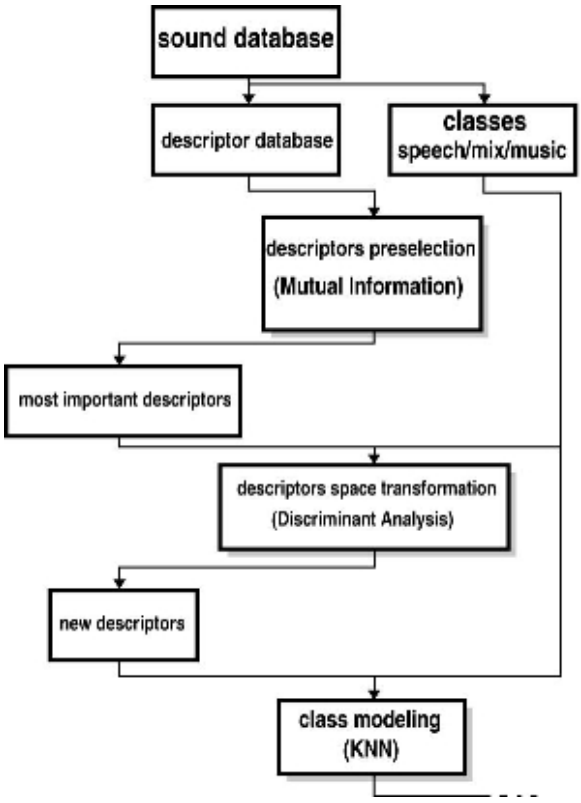
Peeters[13] proposed to improve the classification by the use of Discriminant Analysis. This time the Discriminant Analysis is not used for descriptor preselection but for space transformation. The input of the class modeling part is, then, the results of the projection of the descriptors on the main discriminant axes.

5. CLASS MODELING

There are different types of classifiers, such as: K-nearest neighbors, multi-dimensional-gaussian-model-based classifier, question based tree classifier, tree based vector quantifier. In our system, two classifiers have been tested, a K nearest neighbors system (KNN) and a multi dimensional gaussian model based classifier.

A simple approach to classification is the **K-nearest neighbors (KNN)** method[19].An efficient branch and bound search algorithm has been suggested by W.D'Haes[20] for computation of the KNN in a multidimensional vector space. In a preprocessing step, the set of feature vectors is decomposed hierarchically using hyperplanes determined by principal component analysis (PCA) . During the search of the nearest neighbors, the tree that represents this decomposition is traversed in depth first order avoiding nodes that cannot contain nearest neighbors. This algorithm showed a good behavior for low dimensionality data sets (<=5), which is our case. However KNN does not provide an abstraction of the classes and thus requires the use of the whole database during classification.

Figure 1: overall design of the system



Another approach to classification is the **Multi-dimensional Gaussian Mixture** model classifier [21]. A.Antoniadis[22] developed different methods, here the Kernel Density Estimation have been used. This solution gives excellent results even with coarse settings. For each class, $p(d_i, c_k)$ the conditional probability of observing the descriptor vector d_i given a class c_k is estimated. For a new sound, the descriptor-vector d_i is computed and the probability of the sound to belong to class c_k is defined according to Bayes rules $p(c_k|d_i) = \frac{p(d_i,c_k)}{p(d_i).p(c_k)}$. The system evaluates the probability of the i^{th} clip, given the descriptor vector d_i, to belong to the class c_k. The class label of d_i is then the one with the largest probability.

6. SYSTEM DESIGN

The overall design of our classification method is depicted in Figure 1. After descriptor preselection, the transformation of the space composed of pre-selected descriptors is operated. Then the feature space is given to the class modeling module.

7. EVALUATION OF THE SYSTEM

The following section present an extensive comparison of all the features, the features preselection algorithms and possible classifiers introduced before.

7.1 Database used

Table 1: The best combination of features for the four models

Ncut	MI	LDA	swLDA
MFCC.13.1	MFCC.13.1	MFCC.3.0	MFCC.13.1
spf.1.1	spf.1.1	specfea.4.0	4ME.1.1
4ME.1.1	4ME.1.1	DWT.1.0	hf.1.0
MFCC.12.0	MFCC.6.0	DWT.3.0	
hf.1.0		DWT.9.0	

Table 2: Performance of the two classifiers on the MUSIC database, percentage of music clips classified as music by the speech/music/mixture discriminator

	Ncut	MI	LDA	swLDA
KNN	90	**94,4**	92.7	92
MGauss	80.0	85.5	75	87

The learning database contains 300 labeled sounds composed of 5 seconds excerpts from movies, TV programs and radio programs. The sounds are resampled at $22050Hz$, quantified on 16 bits and mixed in mono. The database is divided into 3 classes: speech, music and mixture music/speech. Indeed, these are the basic set of classes needed in audio structure parsing. A strong attempt was made to collect a dataset that represented as much of the breadth of signals as possible. Thus, the speech class is composed of both male and female speakers, speaking in English, French, German and Japanese , "in studio" and telephonic. The music class is composed of jazz, pop, rap, techno, classical, non-western music, etc, both with and without vocals, plus environmental sounds from movie and TV shows. The third class is the mixture music/speech, it is composed of sounds from movie, TV shows and radio broadcastings. A few special sound clips are uneasy to classify. Clips coming from some contemporary music pieces, certain rap songs etc... may be classified as mixture music/speech. A few noisy speech clips were classified as mixture music/speech. We never tried to avoid this kind of music excerpts.

The evaluation database contains 150 sounds database composed of 5 seconds excerpts from movies, TV programs and radio programs.

7.2 Results from the descriptors preselection

Table 1 presents the results of features preselection for the four models: swLDA(stepwise LDA), LDA, Ncut and MI. Each of the four algorithms finds for 4 or 5 features, the combination of which gives the best class dissociation. Each selected feature is labeled with its family (cepsdes, DWT...),and the index of the feature in his family (13,4...) and 0 or 1. 1 when the feature is a standard deviation of the frame level feature and 0 when it's the mean. The most important selected features are:

Table 3: Performance of the two classifiers on the MIXTURE database, percentage of mixture clips classified as mixture by the speech/music/mixture discriminator

	Ncut	MI	LDA	swLDA
KNN	80	**92**	88	87
MGauss	70	77	75	72

Table 4: Performance of the two classifiers on the SPEECH database, percentage of speech clips classified as speech by the speech/music/mixture discriminator

	Ncut	MI	LDA	swLDA
KNN	97	**97,1**	88,2	85
MGauss	61,8	82,4	55,6	52

- cepsdes.13.1: standard deviation of the ratio between the last and last but one MFCCs.
- spf.1.1: standard deviation of the spectral flux
- 4ME.1.1: standard deviation of the four hertz modulation
- hf.1.0: mean of the percentage of "low-energy" frames

These results put stress on the fact that not all the features are necessary to perform accurate classification. So, the system improves its performance by using only some of the features. Some of the preselected features for speech/music discrimination have already proved their efficiency as Scheirer[4] calls them his "best features". In this case, the starting feature space is specially oriented for our specific classification. In the future system, more features will be implemented so that all kinds of classification can be considered, as the global system remaining unchanged.

7.3 Comparison of the feature preselection algorithms

The feature preselection results appear in tables 2, 3 and 4. Four preselection algorithms are tested: normalized cut (Ncut), mutual information (MI), linear discriminant analysis (LDA) and stepwiseLDA (swLDA). The Mutual Information is apparently the best feature preselection algorithm. Indeed, associated with both KNN and Gaussian Mixture classifiers, the **Mutual Information (MI)** gives better results than the other algorithms in the framework of speech/music/mixture taxonomy.

Table 5: Performance of the speech/music/mixture discriminator, percentage of good classification for speech, music and mixture clips.

	Speech	Music	Mixture
MI,KNN	94,4	97,1	92

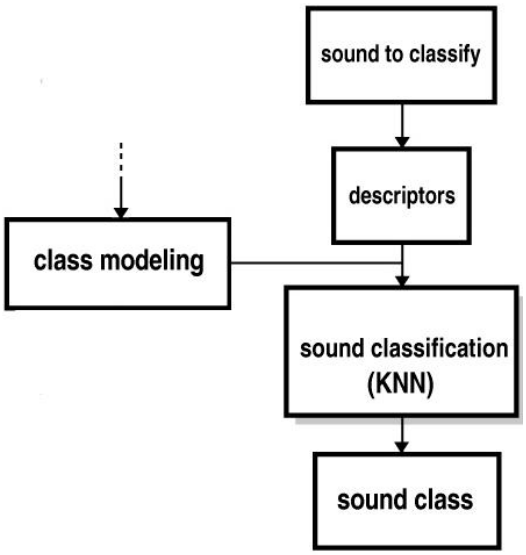
7.4 Comparison of the classifiers

The classification results are apparent upon examination of the table 2, 3 and 4. The **k-nearest neighbors classifier (KNN)** gives better results than multi-dimensional gaussian mixture model classifier (MGauss) in the framework of speech/music/mixture taxonomy. Thus, it seems important to improve the KNN model, notably to provide an abstraction of the classes, so that the system will not compute the whole database to perform a classification. The K-nearest neighboring proposed by W.D'Haes[20] opens the way to an abstraction of the classes in the KNN, by cutting the space into region separated by multiple hyperplanes. Support vector machine (SVM) algorithm may give good results too, but is not implemented yet.

7.5 Results of the optimal system

Table 5 contains the performance of the optimal system that associated **mutual information** model (feature selection) and **k-nearest neighbors** classifier (class modeling).The system is tested on a usual problem of three class discrimination: the discrimination speech/music/mixture of these two signal types.

Figure 2: design of the sound classifier



Tables 2, 3 ,4 and 5 show that it is generally more difficult to classify music than to classify speech. This is not unexpected, as the class of music data, in the world and in our data set, is much broader than the class of speech samples. It seems that there are generally fewer differences between speech samples than between music samples. The results of the discrimination for the third class (the mix of music and speech) are not as good as the others. It is partly due to the size of the mixture database. Indeed, our mixture database coutains less audio clips than the other learning databases, and hence the classification is less effective. Also, the class mixture seems to be less dissociated from the other classes, than speech or music.

The multidimensional classifier that has been built (figure 1) provides a good and robust discrimination between speech music and mixture signals in digital audio.In our future system, after a relevant discrimination, the audio clips belonging to the mixture class could be analyzed with a sound source separation algorithm[24]. So that music and speech could be separated in two different lines. This approach has been treated by [23] for popular music retrieval.

8. CONCLUSION AND FUTURE WORK

This article has questioned the applicability of Discriminant Analysis, Normalized cut, Mutual information, and stepwise Discriminant Analysis in order to perform a feature selection task. We have also studied the performance of two classifiers, KNN and Mgauss. These algorithms were tested in the context of a simple but important class discrimination: speech/music/mixture discrimination. The results show that Mutual Information performs better for feature selection task. And that KNN performs better for class modeling. The best system is presented in Figure 2 and will be integrated later in a more general multimedia analysis system.

This article highlights several important points for the development of a classification system. Firstly, the importance of the choice of the descriptors to be used. Secondly the importance of the choice of a tool for classification. Both feature selection and signal classification techniques (MI and KNN) were used successfully within the framework of a discrimination speech/music/mixture. G.Peeters[13] tested the applicability of these two algorithms on a sound source classification problem. It thus seems interesting to test this type of system for other semantic taxonomies in sound classification but also in video classification. Considering the possibility given to the user to define their own classes, the system should be able to select automatically which signal feature vector is relevant to perform the classification. It can thus be applied to other types of classification and data. In particular, to video scenes classification, by using the joint description of audio and video media. Indeed, the last years advances toward audio and video access at a semantic level, makes it possible to have an intersesting description of both medias, in the framework of class segmentation problem.

9. REFERENCES

[1] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using both audio and visual clues," 2000.

[2] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. ICASSP '96*, Atlanta, GA, 1996.

[3] M. Goto, "A real-time music scene description system: Detecting melody and bass lines in audio signals," 1999.

[4] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 1331–1334.

[5] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," 1998.

[6] Yong Rui, Anoop Gupta, and Alex Acero, "Automatically extracting highlights for TV baseball programs," in *ACM Multimedia*, 2000, pp. 105–115.

[7] L. Rabiner and B. Juang, *Fundamental of speech recognition*, Englewood Cliffs, NJ, 1993.

[8] Tzanetakis, G.Essl, and G.Cook, "Automatic musical genre classification of audio signals," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, Bloomington, Indiana, 2001.

[9] N. Kambhatla and Todd K. Leen, "Dimension reduction by local principal component analys," in *Neural Computation*, 1997, pp. vol.9, no.7, pp. 1493–1516.

[10] D. W. Aha, *Lazy Learning*, chapter 11, pp. 1–5, Kluwer Academic Publishers, Dordrecht, June 1997.

[11] Jihoon Yang and Vasant Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems*, vol. 13, pp. 44–49, 1998.

[12] P.Leray and P.Gallinari, *Analysis of Knowledge Representation on Neural Network Models Representation on Neural Network Models*, chapter Vo.26, p. No.1, 1999.

[13] X. Rodet G. Peeters, "Automatically selecting signal descriptors for sound classification," in *ICMC 2002*. Goteborg (Sweden), September 2002.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press,, 1990.

[15] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Neural Networks,*, vol. 5, no. 4, pp. 537–550, July 1994.

[16] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[17] G. Tourassi et al., "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Medical Physics*, vol. 28, no. 12, December 2001.

[18] R.J. Jennrich, "Stepwise discriminant analysis," in Statistical Metods for Digital Computer, edited by K.Einslein, K. Ralston and H.S. Wilf, (Wiley, New York, 1991).

[19] T. Cover and P. Hart, "Nearest neighbor pattern classification," in *IEEE Trans. Information Theory IT-13*, Munich, Germany, 1967, pp. 21–27.

[20] W. D'Haes, D. Van Dyck, and X. Rodet, "An efficient branch and bound search algorithm for computing k nearest neighbors in a multidimensional vector space," 2001.

[21] R. Duda and P. Hart, *Pattern Classification and scene Analysis*, New York: Wiley, 1973.

[22] A. Antoniadis, "Univariate density estimation," www-lmc.imag.fr/lmc-sms/Anestis.Antoniadis/dess/DESS-Densite.pdf.

[23] Y.Zhuang Y.Feng and Y.Pan, "Popular music retrieval by independent component analysis," in *ISMIR 2002*. 2002, IRCAM - Centre Pompidou.

[24] E. Vincent, X. Rodet, A. Röbel, C. Févotte, R. Gribonval, L. Benaroya, and F. Bimbot, "A tentative typology of audio source separation tasks," in *Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA 2003), Nara, Japan*, 2003.