

NON-SUPERVISED SYSTEM AND METHOD FOR MULTIREOLUTION THEMATIC ANALYSIS AND STRUCTURING OF AUDIO STREAMS

The invention relates to a non-supervised automatic system and
5 method for multi-resolution thematic analysis and structuring of audio streams.
Generally, the invention relates to audio stream content monitoring and
analysis applications.

It applies notably in the fields of monitoring and analysis of
environments. It is used for audio signals that are based on themes or topics,
10 that is to say composed of sound patterns that repeat over time.

The present invention can be used:

- within the framework of multimedia searching: the search being
performed in a thematic space,
- within the framework of automatic summarizing of audio streams: the
15 major themes of a document being in essence a potential summary of
the latter; within the framework of an indexation of audio documents,
multimedia database optimal structuring being based on a thematic
catalog,
- within the framework of language recognition: the phonotactic
20 properties specific to each language being perceived as discriminating
themes or concepts,
- in the field of audio compression: the structuring of audio streams into
themes being advantageously included in a dictionary based shape
compression scheme (conventional scheme in audio compression), the
25 compression of the shapes being optimized on the basis of themes.

The invention also allows the long-term analysis of an acoustic or
sound activity representative of an environment and of the activity taking place
therein.

In the subsequent description the expression "audio stream" and
30 "audio signal" designate one and the same object.

The term "frame" relates to a temporal analysis window. Initially the
frame consists of samples but it is common parlance in the field to also speak

of frame in respect of a vector of acoustic parameters, extracted from a signal frame.

An audio document within the framework of the present description is a series of segments which can be represented on the basis of themes. A segment is therefore a series of consecutive frames. One of the objectives of the present invention is to find the optimal size of the documents as a function of the likelihood of the themes obtained so as to segment an audio stream into relevant themes.

A segment is a set of frames (variable number of frames and potentially a single frame). The segment arises from temporal segmentation. For each segment a representative of the segment is computed. This representative is a vector of acoustic parameters.

An audio is the representative of a class. This word is obtained in our case by a non-supervised clustering algorithm:

On the basis of the acoustic vectors the method will group together by class (clustering) the vectors that are close in the acoustic space. Each class possesses a favored representative or centroid. This representative (centroid) is the audio word.

In applications for recognizing sound phenomena in an audio stream, the technical problems to be solved are notably the following:

- How to detect sets and/or series of sound and/or abnormal characteristic events representative of an environment,
- How to obtain solutions that are robust to background noise and to its variabilities.

In the field of monitoring and analysis of environments, the conventional systems known from the prior art rely mainly on image and video processing technologies.

The thematic analysis of a document (text or audio) is the phase which is aimed at establishing the meaning thereof, using the sense of the unitary elements or words of the document.

The notion of word is extended to the audio field, and this description will speak interchangeably of word or "audio" word. The model obtained by latent semantic analysis better known by the abbreviation LSA, is

a tool which allows for semantic lexical relationships on the basis of the distribution of the cooccurrences of audio words in a predefined context or an environment. This LSA model allows the analysis of documents in a space of representations that are characteristic of the concepts or themes of one or
 5 more documents.

A theme is then characterized by the occurrence of a set of audio words sharing semantic ties.

Probabilistic versions of the LSA model have been proposed. The models known to the person skilled in the art such as the PLSA model (the
 10 abbreviation standing for Probabilistic Latent Semantic Analysis) or the LDA model (the abbreviation standing for Latent Dirichlet Allocation) are based like the LSA model on the analysis of the occurrence of the words so as to deduce therefrom a set of representative concepts by way of latent random variables (Bayesian inference). These models, which perform better than the LSA
 15 model, have been used successfully in text analysis but also in image and video analysis (indexation, classification and scrutiny).

However, despite their performance, these procedures (LSA, PLSA, LDA, etc.) exhibit the following drawbacks: they arise from text analysis and rely on the joint analysis of a document set. In the case of audio streams,
 20 for example, the size of the audio streams is no longer clearly defined and the known procedures of the current state of the art use an arbitrary a priori which consists in choosing a single and fixed document size, the document being regarded as the audio stream.

These models do not therefore make it possible to carry out precise
 25 segmentation without supplementing the analysis with a phase of grouping the documents into identical concepts. In the case of audio and/or video streams, the problem is all the more significant as the very notion of document size is not clearly defined.

In the subsequent description, the following terms will be used

30 P: Number of vectors (or classes) at the lowest level of the dendrogram.

M: Number of "audio" words of the dictionary.

S: segment of fixed or variable size.

c_k : Centroid of index k .

x_n : Vector n of the segment S .

$d(S, c_k)$: Distance between the segment S and the centroid c_k .

Q : Number of classes closest to the segment S used for the fuzzy classification.

5 $wf(S, k)$: Degree of membership of the segment S in the class k .

$wf(S, k = 1, \dots, Q)$: Q fuzzy classification based associated degrees of membership in the segment S .

F : fuzzy index.

N : Number of documents.

10 d_j : Document of index j

D : Set of N documents.

w_i : Word of index i

W : Set of M words.

K : Number of themes or topics

15 z_k : Theme of index k

Z : Set of K themes

$P(w_i, d_j)$: Joint probability of documents and of words

$P(d_j)$: Probability of a document d_j .

$P(w_i|z_k)$: Probability of observing a word w_i knowing the theme z_k .

20 $P(z_k|d_j)$: Probability of observing the theme z_k knowing the document d_j .

$P(w_i|d_j)$: Conditional probability of a word w_i in a document d_j .

$L(P|D)$: Log-likelihood of the parameters of the model P .

$n(w_i, d_j)$: Term of the cooccurrence matrix of size $M \times N$.

$Ns(j)$: Maximum number of segment of the document d_j .

25 $L(W|d_{test})$: Log-likelihood of the PLSA model inferred for the document d_{test} .

$tmin$: Minimum size of a document.

$tmax$: Maximum size of a document.

ps : Segmentation spacing

l_i : Likelihood of document i .

s_i : Start instant of document i

t_i : Size of a document i

f_i : End instant of document i .

0 : Optimal subset of documents without temporal overlap, the sum of whose
5 likelihoods l_i is maximal.

One of the subjects of the present invention relates to a method which relies on an optimal selection in the sense of the likelihood, non-supervised, of the size of the documents, based on an exhaustive search for
10 the temporal supports on which the likelihood of the documents with respect to the concepts/themes is highest. The optimal search is performed over a set of different sizes or durations and over a set of initial and final positions of documents within an audio stream to be analyzed.

The invention relates to a non-supervised method of thematic
15 analysis and structuring of an audio stream, said audio stream comprising a set of documents each consisting of several segments, said audio stream comprising one or more patterns or themes repeated over time, characterized in that it comprises at least the following steps:

- Extracting the acoustic parameters of the audio stream to be analyzed,
20 E1
- A step of low-level segmentation of said audio stream so as to obtain a set of segments S, E2
- A step of fuzzy classification of said segments by using a dictionary of M predefined audio words so as to characterize each segment S by a
25 set of Q audio words and of Q degrees of membership w_{fk} of a word in a class, E4,
- Generating a cooccurrence matrix for each document of said audio stream, an element $n(w_i, d_j)$ of the cooccurrence matrix corresponds to a combination of the degrees of membership of all the words of the
30 document d_j which is close to the class i (in the sense of the fuzzy classification), the size $M*N$ of the matrix corresponds to the number M of words and N to the number of documents, E5a,

- For a finite segmentation set (document size t_i , start of the analysis of the audio stream) and on the basis of the associated cooccurrence matrix, determining, by adaptation of the predefined models, the likelihoods l_i associated with each pair (document size t_i , start of analysis), E5c, the document size belonging to a predefined interval $[t_{min}, t_{max}]$,
- Searching for the optimal segmentation in terms of document size, in the sense of the maximum likelihood for each triplet (size of the document, start of analysis, likelihood) by application of an algorithm of dynamic programming type, E6, under the constraint of selecting documents without temporal overlap.

The method can comprise a learning step adapted for generating the dictionary of the audio words and the statistical models comprising at least the following steps:

- A step of extracting the acoustic parameters of the audio stream to be analyzed, E1,
- A step of low-level segmentation of said audio stream so as to obtain a set of segments S , E2,
- A step of non-supervised construction of a dictionary of audio words, E3,
- A step of fuzzy classification of said segments by using a dictionary of predefined audio words so as to characterize a segment S by a set of Q audio words and Q degrees of membership w_{fk} of a word in a class, E4,
- Generating a cooccurrence matrix, for each document of said audio stream, an element $n(w_i, d_j)$ of the matrix corresponds to a combination of the degrees of membership of all the words of the document d_j which is close to the class i in the sense of the fuzzy classification, the size of the matrix corresponds to the number M of words and to the number of documents $M*N$, so as to model the way in which the audio words appear in a document, E5a,
- Estimating statistical models adapted so as to decompose an audio stream into a mixture of themes defined as a multinomial distribution of

words of a dictionary and representative of the conditional probability of observing a word in an audio stream, E5b.

The low-level segmentation step is, for example, a uniform segmentation with segments of fixed size.

5 The low-level segmentation step can be executed with segments of variable size.

The low-level segmentation step is, for example, a segmentation by dendrogram.

10 The non-supervised construction of the dictionary of audio words can be executed by grouping the segments together by class with the aid of an algorithm of LBG type.

The method according to the invention uses, for example, a statistical model of PLSA type defined as a mixture characterizing a model of joint probability of documents and of words.

15 According to a variant embodiment, the parameters of the PLSA model are estimated by maximizing the log-likelihood given by an algorithm of EM (Expectation Maximization) type.

The dynamic programming algorithm can implement a weighted intervals scheduling algorithm.

20 The invention also relates to a system for non-supervised thematic analysis and structuring of an audio stream, said audio stream comprising a set of documents each consisting of several segments, said audio stream comprising one or more patterns or themes repeated over time, said system being adapted for implementing the method comprising one of the
25 aforementioned characteristics, characterized in that it comprises at least the following elements:

- A database comprising a pre-established dictionary of audio words, and statistical models,
- One or more acoustic sensors linked to an analysis module comprising
30 a first module for extracting the acoustic parameters of said audio stream, said first module being linked to a buffer memory, the output of the buffer memory being linked to a module for fuzzy classification of said segments arising from the buffer memory, a module for

constructing cooccurrence matrices and for adapting the models, a module adapted for determining the likelihood of the models and an optimal segmentation search module executing an algorithm of dynamic programming type applied to the likelihood measurements.

5 The system can comprise a learning chain comprising at least the following elements: one or more acoustic sensors, a device comprising a filter and an analog converter and then a module adapted for extracting acoustic parameters and a buffer memory linked to a module for creating audio words and a classification module and a module for creating statistical models, the
10 set of audio words and statistical models arising from this learning module are stored in a database.

Other characteristics and advantages of the device according to the invention will become more apparent on reading the description which follows
15 of an exemplary embodiment given by way of wholly nonlimiting illustration in conjunction with the figures which represent:

- Figure 1, an exemplary system architecture according to the invention,
- Figure 2, a flowchart of the main steps of the method according to the invention,
- 20 • Figure 3, an illustration of an example of low-level segmentation with a grouping according to N classes,
- Figure 4, a multi-resolution semantic analysis scheme for a low-level segmentation of fixed size,
- Figure 5, a multi-resolution semantic analysis scheme for a low-level
25 segmentation of variable size,
- Figure 6, a Dynamic Programming based Optimal segmentation scheme, and
- Figure 7 an exemplary system according to the invention.

30 In order to better elucidate the manner of operation of the method and of the system according to the invention, the description will be described with reference to the monitoring of a subway station where the audio streams or sound signals all have a common temporal structure. This structure is based, for example, on the arrival and the departure of trains. The themes

most represented in this case are: the arrival of a train, the opening of the doors, the alighting and the boarding of travelers, the closing of the doors and the departure of the train. In this case a sound signal to be analyzed will see at least 4 themes or patterns repeated over time.

5 As a function of the daily timetables, the frequency of the trains varies, consequently the duration of the various events specific to each theme or pattern is variable. For example, a waiting period between two trains may be variable in duration and in content, as a function of crowdedness, for example, but the themes or patterns representing this waiting period will
10 originate from similar statistical models. The likelihood of these models will, in essence, be maximal over these waiting periods. Similar reasoning can be applied in the case of shorter events, such as arrival, departure or the closing/opening of the doors.

 Figure 1 represents an exemplary detection system put in place to
15 analyze the content of sound signals or audio streams in a monitored environment, here a subway platform.

 The system 1 comprises one or more acoustic sensors 2, such as microphones, of sound noise present in a zone to be monitored or for which an analysis of sound events is requested. The data received on an acoustic
20 sensor are transmitted firstly to a device 3 comprising a filter and an analog converter, both known to the person skilled in the art, and then via an input 4 to a module 6 carrying out the extraction of acoustic parameters and a buffer memory 7 of first in first out type better known by the abbreviation "FIFO". The buffer memory 7 is used so as to undertake a low-level temporal
25 segmentation. The output of the buffer memory feeds a learning module. The function of the learning module is notably, during a learning phase, to generate a set of "audio words" (creation module 8 and classification module 9) used during the subsequent analysis of an audio stream or sound signal, and of the statistical models (module 10). The set of audio words and of
30 statistical models arising from this learning module are stored in a database 11. This database 11 is, for example, initialized during a learning phase of the system and will be able to be updated in the course of the operation of the system according to the invention.

The same type of system can be used during the phase of analyzing a sound signal or audio stream. One will speak of test phase in this case. The signal is detected by the acoustic sensor or sensors 20, and then will be processed in the same manner as in the case of learning by means of
 5 a filter 21 and of a module 22 carrying out the extraction of acoustic parameters. The processor is adapted for executing the steps which will be detailed subsequently of the method according to the invention. The output of the buffer memory 23 can be linked to an audio words classification module 24, and then thereafter to another device for analyzing or processing the data
 10 after analysis of the audio stream (structuring of an audio stream). On the basis of the audio words obtained by classification, the module 25 estimates the likelihood of the statistical models previously learnt and stored in database 11. The module 26 produces a set of results 27 allowing the analysis and the structuring of the audio stream arising from the sensor 20. This analysis
 15 phase can also be dubbed the test phase.

To summarize, the method according to the invention will implement at least the following steps in a learning and test phase:

- the non-supervised creation of a dictionary of audio words by a grouping or "clustering" technique known to the person skilled in the art
 20 (K-means, EM (Expectation-Maximization) or others).
- the computation of PLSA or LDA models or of equivalent models in this technical field for a set of sizes of documents varying from a minimum bound to a maximum bound, generating a cooccurrence matrix modeling the appearance of the words in each of the documents.
 25 Models are also computed by varying the start-of-analysis delay for one and the same size of documents.

Step 1, E1: extraction of the acoustic parameters (learning and test phase).

The first step corresponds to the extraction of the acoustic
 30 parameters from the audio stream. This first step makes it possible to extract from the audio stream acoustic parameters which will be used subsequently in the method to analyze the sound signal. These parameters make it possible notably to represent the audio signal in an acoustic space of reduced

dimension. The acoustic parameters which will be used are in general spectral, temporal or spectro-temporal parameters related to the spectral and temporal shape of the signal. These parameters are called vectors of acoustic parameters or vectors of parameters. It is thus possible to use a modeling of the spectral envelope of the noise picked up by the acoustic sensor or microphone, such as the cepstral parameters or cepstral vectors. The audio stream in this case will be modeled by a sequence of cepstral vectors. These vectors of parameters are extracted in a very conventional manner known to the person skilled in the art on a constant temporal support fixed at the start of the analysis. For example, it is possible to work with a time window (or frame) of a few tens of milliseconds (ms). Moreover and so as to ensure better continuity of the vectors, an overlap between consecutive frames of 50% (5 ms) is customarily used. Before extraction of the parameters each frame is weighted with a weighting window (Hamming, Hanning, etc.).

Step 2, E2: low-level temporal segmentation (learning and test phase).

The second step corresponds to a low-level temporal segmentation. The first segmentation level makes it possible notably to select the fineness of the analysis desired. It can be conducted in two different ways:

- Uniform segmentation with segments of fixed size,
- Automatic segmentation with segments of variable size.

In the case of uniform segmentation, the size of the segments is predefined and it arises from a compromise between computation time and fineness of the final segmentation by theme or pattern. The size of the segments produced is exactly a fixed number of frames produced in step 1.

In the case of segmentation by dendrogram, the size of the segments is defined as a variable number of frames. A segment is therefore a concatenation of consecutive frames. In this case the size of the segments is in accord with the local statistics of the audio signal to be segmented.

The second way of proceeding is based on a low-level automatic temporal segmentation. The number of frames per segment is variable and is determined by dendrogram. Analysis of a frame results in a vector of parameters which is stored in the buffer memory 7. The size of this buffer memory (number of vectors stored) is equal to the number of elements

(vectors in the present case of application) of the dendrogram. This size which represents the duration of the segment is of the order of a few hundred ms, or indeed a few seconds for highly stationary background noise. The size of the buffer memory is chosen, for example, so as to incorporate at the minimum an
 5 audio event considered a priori to be elementary.

The dendrogram is used here to obtain in an automatic manner a segmentation of the audio signal. The principle consists in grouping together in a "bottom-up" approach the elements input to the dendrogram. This procedure makes it possible to obtain a segmentation for all the various
 10 possible levels, stated otherwise for a number of segments ranging from the initial number of elements or vectors to a single segment.

In the case of the bottom up, one starts from the finest slicing (ie frame-wise slicing) and the frames are grouped together to form segments. Therefore the maximum number of segments is the maximum number of frames in the
 15 buffer memory (elements = frames)

On initializing the method, each element is the representative of its class. If P is the number of elements (vectors) input to the dendrogram, there are therefore P classes at the lowest level. Thereafter, the number of segments is decremented so as to go to the higher level after grouping of the
 20 two closest classes according to a distance criterion defined as a function of the parameters used. The criterion is for example chosen as a function of the type of acoustic parameters which were used to characterize the audio or sound background. Several types of possible groupings exist according to the distance that it is sought to minimize for the selection of the classes to be
 25 grouped together. The 4 main procedures known to the person skilled in the art are as follows:

- minimum distance between the vectors of the classes or "single linkage",
- maximum distance between the vectors of the classes or "complete linkage",
- 30 • mean distance between the vectors of the classes or "average linkage",
- distance between the mean vectors (centroids) or "centroid linkage".

In Figure 4 is represented an example of groupings according to the N classes for a bottom-up approach; the vertical axis corresponding to the

vectors, the horizontal axis showing diagrammatically the buffer memory of the dendrogram. On completion of this grouping, the method makes it possible to obtain 3, and then 2, and then a vector represented by a single letter R for the grouping.

5 The automatic segmentation method must finally select in an automatic manner a segmentation level which will be considered optimal according to a criterion to be defined.

10 A first criterion consists in applying a threshold S_{max} on the distance between the last two classes grouped together (the further one rises through the levels of the dendrogram, the larger the distance between the classes to be grouped together). This therefore entails minimizing the number of possible segments while minimizing the grouping distance considered.

Other criteria can be used such as for example:

- 15 • the maximum Max over the segments of one and the same level of the (Max over the vectors of one and the same segment (distance between the vector and the mean vector of the class associated with the segment)),
- 20 • the maximum Max over the segments of one and the same level of the (Average over the vectors of one and the same segment (distance between the vector and the mean vector of the class associated with the segment)),
- 25 • the minimum Min over the segments of one and the same level of the (Min over the vectors of one and the same segment (normalized inter-correlation between the vector and the mean vector of the class associated with the segment)),
- the minimum Min over the segments of one and the same level of the (Average over the vectors of one and the same segment (normalized inter-correlation between the vector and the mean vector of the class associated with the segment)).

The alternating notations that may be used are for example:

$$30 \quad \begin{aligned} & \underset{k=1, \dots, P-j}{MAX} \left\{ \underset{x \in \text{Segment } k}{MAX} \left\{ distance(x, c_k) \right\} \right\} \\ & \underset{k=1, \dots, P-j}{MAX} \left\{ \underset{x \in \text{Segment } k}{MOYENNE} \left\{ distance(x, c_k) \right\} \right\} \end{aligned}$$

$$\min_{k=1, \dots, P-j} \left\{ \min_{x \in \text{Segment } k} \left\{ \text{correlation}(x, c_k) \right\} \right\}$$

$$\min_{k=1, \dots, P-j} \left\{ \text{MOYENNE}_{x \in \text{Segment } k} \left\{ \text{correlation}(x, c_k) \right\} \right\}$$

where j =level index = 0, ..., $P-1$ (Level 0: P classes, Level $P-1$: 1 class)

5 where k = segment index (class) for level $j = 1, \dots, P-j$

where x is a vector of the segment (class) of index k

where c_k is the mean vector (centroid) of the segment (class) k .

In cases 1) and 2), the distance criterion must be less than a threshold while minimizing the number of segments. In cases 3) and 4), the
10 correlation criterion must be greater than a threshold while minimizing the number of segments.

On completion of step 2, the method has a segment set S of fixed or variable size which characterizes the audio stream.

The following steps 3 and 4 correspond to the "clustering" and
15 classification steps and are based on methods that are standard in shape recognition, for example. Step 3 is only carried out during the learning whereas step 4 is common to the learning and to the test. The aim is to model the acoustic space by computing representatives or "centroids" that are optimal for each class of similar vectors, in the sense of a distance.

20 **Step 3, E3:** construction of the audio word dictionary (learning phase)

The third step corresponds to the construction of a dictionary by groupings of the similar segments or "clustering". This creation corresponds to a non-supervised creation of the dictionary of audio words by a conventional technique of grouping by class or "clustering" of LBG, K-mean or other type.

25 The above-described segmentation is applied to the whole of the audio signal to be analyzed. The segments S thus obtained are grouped together by class by a learning algorithm of the LBG (Line-Buzo-Gray) type, for example. This algorithm makes it possible to associate each segment with a class, itself represented by a mean vector or centroid. A simple criterion of

distance between a segment and the centroid consists in computing the aggregate distance of the centroid from each of the vectors of the segment considered. The distance used is the following:

$$d(S, c_k) = \sum_{n=1}^{N_{vectors}} distance(x_n, c_k)$$

5 = distance between the segment S and the centroid c_k of index k

where x_n is a vector of the segment S, $n = 1, \dots, N_vectors$.

The number of classes may either be fixed a priori, or determined in an automatic manner by using a stopping criterion based on the minimum distance between centroids; it is not necessary to increase the number of
10 centroids if the latter are sufficiently close according to a certain criterion. The determination of the threshold used for this stopping criterion can be based on a quadratic spectral distance. This spectral distance can generally be computed on the basis of the parameters used by computing the associated spectral envelopes. An alternative consists in determining the threshold on the
15 basis of the correlation between the distances used with the parameters and the spectral distances.

In so far as certain normal events may be under-represented, a modified learning version consists in using an algorithm which makes it possible to position centroids uniformly distributed in the parameter space. An
20 exemplary algorithm is explained hereafter:

1. Computation of the mean vector on the basis of the set of learning data.
2. Search for the segment furthest from the mean vector and initialization of the first centroid as being the mean vector of the segment found ($nb_class = 1$).
- 25 3. Iterations making it possible to increment the number of classes ($nb_class = nb_class + 1$): the segment which maximizes the aggregate distance to the centroids identified at the previous iteration is searched for.
4. Stopping criterion: either when the predefined number of classes is reached, or when the distance between the segment found and the
30 centroids identified at the previous iteration is less than a threshold.

The above-described phase makes it possible to initialize a dictionary of "audio words" uniformly distributed in the parameter space or

acoustic space. In a second phase, a step of classifying each segment is carried out. Each segment is associated with the centroid that is closest in the sense of the spectral distance. This classification makes it possible to associate a set of segments with each centroid. The centroid of each class is
 5 thereafter updated by taking the mean vector of parameters corresponding to the vector computed as the average of the parameters of each segment, associated with one and the same class.

This is the procedure conventionally used for updating centroids (LBG, K-means, etc.).

10 On completion of these system initialization steps, a dictionary of M audio words which are the centroids of each class is available and will be able to be used to analyze the audio stream. This dictionary of audio words is stored in the database.

The steps which will now be described correspond to the steps of
 15 modeling (learning) and analysis (testing) of an audio stream or audio signal received on the sensors of the analysis system according to the invention by using the audio words of the dictionary.

Step 4, E4: fuzzy classification of segments of an audio stream (learning and test phase)

20 A fourth step corresponds to the fuzzy classification of each segment of an audio stream analyzed by the system using the dictionary of audio words determined in step 3. A conventionally used procedure is to choose as representative a word of the dictionary (which is itself a representative of a class of the dictionary) according to a distance or proximity
 25 criterion.

In order to improve this conventional procedure, a series of degree of membership in the classes of the dictionary of classes used is employed. In this case a segment will be represented as a weighted linear combination of the distances to the M words of the dictionary (or M classes of the dictionary).
 30 This procedure is better known by the expression "Fuzzy Classification" (or "Fuzzy Clustering"). Each audio segment is associated with a subset (whose size $Q < M$ is fixed in advance, for example $Q=3$) of classes of the dictionary, which are represented by their centroid, on the basis of a weighting vector,

whose degree of membership $wf(S, k)$ of index k associated with the centroid c_k of index k (or with the class of index k) is given by:

$$wf(S, k) = \frac{1}{\sum_{n=1}^Q [d(s, c_k) / d(s, c_n)]^{1/(F-1)}}$$

where F is the fuzzy index (or the parameter which characterizes the degree of fuzziness) greater than 1.

In the given example, the $Q=3$ closest classes of each segment S are chosen.

On completion of this step 4, a set is available consisting of several segments of an audio stream where each segment S of an audio stream is characterized by Q words of the dictionary $Q_{k=1...Q}$ and the Q degrees of membership $wf(S, k = 1, \dots, Q)$. A segment S is indexed with the words of the dictionary. This indexation relates to the quantization of the acoustic space obtained by "clustering".

The optimum is computed in the sense of the global distortion estimated over the set of available segments and the centroids already computed. The process is iterative and based on the minimization of the distortion. After convergence, the optimal centroids are available. The classification is thereafter carried out by choosing, for each segment, the centroid which best represents the segment S of the analyzed audio stream.

Step 5, E5a, E5b, E5c: creation of matrices of cooccurrence of audio words and estimation/adaptation of the model (learning and test phases)

The fifth step corresponds to the creation of the cooccurrence matrices and then to the inference of the statistical thematic models.

The computation of PLSA statistical models is carried out for a set of sizes of documents varying from a minimum bound to a maximum bound (system parameters). These models are also computed with several values of delays (or start of analysis) in the audio stream for one and the same size of audio stream. In the case of low-level segmentation with constant size or uniform segmentation, the delay is expressed as a number of frames. In the case of automatic segmentation, the delay is expressed as a number of segments.

This step is aimed at producing for a finite candidate segmentation set (size t_i of documents d_i , start s_i of the analysis of the document of index i), a

likelihood set arising from several models. The plurality of pairs (size t_i of documents d_i , start s_i of the analysis of the document of index i) therefore allows a multiresolution analysis whose temporal fineness is parametrizable.

The PLSA (Probabilistic Latent Semantic Analysis) model was initially proposed by T. Hofman ('Probabilistic Latent Semantic Analysis', Proceedings of Uncertainty in Artificial Intelligence, UAI'99, Stockholm) and can be considered to be a probabilistic version of latent semantic analysis (LSA). The aim is to decompose an audio stream into a mixture of latent themes (hidden themes or hidden topics) defined as a multinomial distribution of words of a vocabulary (or dictionary).

It is possible to generate audio documents on the basis of a set of audio words and of a PLSA model. The PLSA model is described on the one hand by the probability of observing a word knowing a theme and, on the other hand, by the probability of observing a theme knowing a document.

The PLSA model can be applied to discover themes in audio documents by ascertaining for each audio segment S of the signal to be analyzed the Q the closest words of the dictionary and the Q degrees of membership $wf(S, k)$.

The audio documents are considered to be discrete random variables $d_j \in D = d_1, \dots, d_N$, where N is the number of documents. The words are represented by another discrete random variable $w_i \in W = w_1, \dots, w_M$, where M is the number of audio words. The PLSA model is a statistical model which makes it possible to establish ties between a latent variable (theme or topic) $z_k \in Z = z_1, \dots, z_K$ where K is the number of themes with the occurrence of a word in a document. The PLSA model is then defined as a mixture characterizing a model of joint probability of documents and of words

$$P(w_i, d_j) = P(d_j)P(w_i|d_j) = \sum_{k=1}^K P(w_i|z_k)P(z_k|d_j)$$

The model thus defined is based on the following probabilities:

- the probability of a document $d_j: P(d_j)$
- the probability of observing a word w_i knowing the theme $z_k: P(w_i|z_k)$,

- the probability of observing the theme z_k knowing the document d_j : $P(z_k|d_j)$.

The PLSA model defines the conditional probability of a word in an audio stream $P(w_i|d_j)$ as the mixture of a distribution of words that relates to a theme $P(w_i|z_k)$. The weights of the mixture are given by the distribution of the themes in a document $P(z_k|d_j)$. The parameters of the PLSA model are computed using the principle of maximum likelihood. On the basis of an audio document set D , the log-likelihood of the parameters of the model P is given by:

$$L(P|D) = \sum_{i=1}^M \sum_{j=1}^N n(w_i, d_j) \log P(w_i, d_j)$$

The optimization of this equation is carried out using, for example, the EM (Expectation-Maximization) algorithm on the set of audio documents. As a function of the phase (learning or test), two different uses of the EM algorithm are envisaged:

- Learning phase: complete EM,
- Test phase: reduced EM so as to adapt the existing models to the test data.

Step E5.a: Computation of the cooccurrence matrix (learning and test phases)

This step entails the estimation of the $M \times N$ cooccurrence matrix whose terms $n(w_i, d_j)$ are the coefficients of cooccurrence of the words w_i in the document d_j . These terms are updated for each segment S of the audio documents by using for each of the words of the segment S , the Q audio words and the Q degrees of membership $wf(S, k = 1, \dots, Q)$.

Let $wf(S, k = 1, \dots, Q)$ be the degrees of membership in the Q classes k associated by fuzzy classification of the segment S and let $\{S_j(p = 1, \dots, Ns(j))\}$ be the $Ns(j)$ segments S of the audio document $d_j \in D$.

For each audio document $d_j \in D$ and for each audio word, $w_i \in W$, $n(w_i, d_j)$ is given by:

$$n(w_i, d_j) = \sum_{p=1}^{N_s(j)} wf(S_j(p), i)$$

Step E5.b: Estimation of the PLSA model (learning phase)

This step allows the estimation of the parameters of the PLSA model by maximizing the log-likelihood of the parameters of the model P given by (Complete EM algorithm):

$$L(P|D) = \sum_{i=1}^M \sum_{j=1}^N n(w_i, d_j) \log P(w_i, d_j)$$

On completion of this step, the parameters of the PLSA model are stored in database.

Step E5.c: Adaptation of the PLSA model to the test data (test phase)

This step entails the determination of the parameters of the PLSA model (reduced EM algorithm) followed by the estimation of the log-likelihood of this model inferred for the document d_{test} according to the formula:

$$L(W|d_{\text{test}}) = \sum_{i=1}^M n(w_i, d_{\text{test}}) \log \sum_{k=1}^K P(w_i|z_k) P(z_k|d_{\text{test}})$$

A high likelihood is the indication of the good fit of the PLSA model to the audio data processed.

This step is aimed at producing for a finite candidate segmentation set (document size t_i , start of the analysis of the audio stream s_i), a likelihood set arising from several models. The plurality of pairs (document size, start of the analysis) therefore allows a multiresolution analysis whose temporal fineness is parametrizable. In a more detailed manner, this step is structured as two distinct steps:

- for each pair (t_i, s_i) and over the entire duration of the signal, a cooccurrence matrix is estimated. The aim of this matrix is to model the way in which the audio words appear in each document. The conventional way is simply to count for each document the number of occurrences of the words. This procedure presupposes that one has a

very good definition of the words. In the case of the time-varying audio signals this definition is generally very difficult to obtain. So as to improve this procedure the method will use as occurrence coefficient a combination of degree of membership in the classes of the audio dictionary associated with each word of each document (step 5.a). These coefficients are deduced directly from the classification step (step 4).

- For each pair (document size, start of the analysis) over the entire duration of the signal and on the basis of a cooccurrence matrix (estimated as described in 5.a), a thematic statistical model is estimated (learning phase). In the same manner, the likelihood of this model for each of the documents is also computed (test phase).

The significant point of steps 5.a, 5.b and 5.c resides in the size of the documents. Indeed, the size of the documents is fixed right from the creation of the cooccurrence matrix of size $M \times N$ (where N is the number of audio streams and M is the number of audio words). The number of audio words is fixed and arises from steps 3 and 4. The number of audio documents for a given signal is related to the size of the audio stream itself.

Being given, for one and the same audio signal, the possibility of choosing a variable number of audio streams, it is then possible to choose documents of variable size. If moreover, still for one and the same audio signal, the method specifies different start-of-analysis instants, it is possible to obtain a multi-resolution scheme for analysis and inference of PLSA models (Figure 4 in the case of a low-level segmentation with fixed size frame and Figure 5 in the case of a low-level segmentation with variable size). Therefore for one and the same signal and for a plurality of pairs (document size, start of analysis), it is possible to obtain a plurality of likelihoods associated with a document size and a temporal position in the audio stream to be analyzed.

On completion of this step, a set of likelihoods l_i associated with a set of document size t_i whose minimum size t_{min} and maximum size t_{max} are parameters of the system, and with a set of temporal positions s_i (instant of temporal analysis of the signal), is produced by following a multi-resolution analysis scheme.

Figure 4 presents the multi-resolution scheme proposed on the basis of a low-level segmentation with frame of fixed size (20 ms). All the PLSA analyses are carried out on documents of size which are multiples of the size of the initial frame. The delay and therefore the temporal position of each analysis is likewise equal to the size of a frame of 20 ms.

Figure 5 presents the multi-resolution scheme proposed on the basis of a low-level segmentation with frame of variable size which is a multiple of the size of the acoustic vectors extraction window (20 ms). All the PLSA analyses are performed on documents of size expressed as a number of segments. The delay and therefore the temporal position of each analysis is likewise equal to the size of the current segment (and therefore a multiple of 20 ms).

The two proposed schemes guarantee perfect synchronism of all the analyses, whatever the size of the documents. That is to say and more precisely, whatever the size of a document, it is possible to find a document of different size which starts immediately after (Figure 4 and Figure 5). The synchronism of the multiresolution analysis is therefore ensured. The object of the following step is to find which path is optimal in terms of maximum likelihood of the PLSA models under the constraint that the documents do not overlap temporally.

Step E6: Optimal segmentation (test phase)

The sixth step corresponds to the search for the optimal segmentation in the sense of the maximum likelihood under the constraint of selecting documents without temporal overlap.

The choice of the optimal structuring (size and position of each of the documents) is then obtained on the basis of an algorithm of Dynamic Programming type, applied to the measurements of the likelihood associated with each document assumption. Each document assumption is characterized by a triplet (duration of the document, start of the analysis and likelihood). This step is described in Figure 6. The final result is the production of an optimal segmentation in the sense of the likelihood of the estimated themes. This segmentation is robust to variations in local statistics and makes it possible to have a simpler interpretation level for a monitoring operator.

In the example given to illustrate the invention, the problem is a problem of searching for compatible intervals under the maximum weight constraint. Indeed, each document possesses a size t_i , a start instant s_i , an end instant f_i and a likelihood or weight l_i . The notion of compatibility is defined by the fact that the series of intervals is increasing over time and that no interval overlaps time-wise. This is a sub-problem of Dynamic Programming also called Scheduling of Weighted Intervals (or Weighted Interval Scheduling).

In weighted interval scheduling problems, the objective is to find the compatible subset or subsets of intervals (or documents in our system) without temporal overlap and the sum of whose weights (or likelihood in our system) is the highest. Disjoint segments can also be accepted in the optimal path.

Consider a set D of documents $d_i \in D = d_1, \dots, d_N$ which all possess an associated likelihood l_i . Document $i \in D$ is characterized by a start instant s_i and an end instant f_i . The method will seek an optimal subset O of documents without temporal overlap, the sum of whose likelihoods is maximal. The above-stated problem can be written in the following manner:

$$O = \operatorname{argmax}_{O \subseteq D; \forall i, j \in O \text{ with } f_i \leq s_j \text{ or } f_j \leq s_i} \sum_{i \in O} l_i$$

For $1 \leq i \leq n$, let us put $\text{value}(i)$ as the maximum of the sum of the likelihoods of the documents which do not overlap and having at least i as index. The optimal solution at an instant $j \geq i$ will arise from the previous optimal solution to which will be added the highest weight of a candidate document j to be determined from among the list of documents which is compatible with the optimal solution at the instant j . Under this assumption and according to Bellman's theorem known to the person skilled in the art, we then obtain:

$$\text{value}(i) = \max(l_i + \text{value}(j), \text{value}(i + 1))$$

with j which is equal to the index of the first document which begins after document i finishes.

To implement this theorem, the method uses the following algorithm to find the sum of the likelihoods corresponding to the optimum scheduling:

- Sort the documents by increasing order of start date
- For $1 \leq i \leq n$, find the smallest values of $j \geq i$ such that document j does not overlap document i .
- For $i = n$ to 1, compute $\text{value}(i)$ and then store it with the corresponding path (i.e. the optimal series of documents)

The sum of the optimal scheduling is equal to $\text{value}(1)$.

On completion of this step, the optimal scheduling of the documents is estimated. This optimal route or path, in a multiresolution trellis of the likelihood values relating to the relevance of the PLSA models, allows us to obtain an optimal segmentation in terms of document size while preserving ties with the themes. This is the final objective of the present system.

Figure 7 shows diagrammatically an exemplary architecture comprising several devices making it possible to record audio sounds or audio streams such as acoustic sensors C_i . The set thus formed is linked, for example, to a computer P_i comprising a module for recognizing abnormal events, and a database containing models for recognizing the abnormal events. Each computer P_i is linked to a central monitoring facility or room comprising, for example, several monitoring screens E_i . The links allowing the transfer of the data from one device to another are, for example, wired links, or wireless links, of Bluetooth type or else the system forms part of a wireless local area network or WLAN.

The system and the method according to the invention propose an analysis/structuring on the basis of a thematic interpretation that is richer than the conventional interpretations based on local detection of specific and/or abnormal unitary events. The process is non-supervised and therefore it does not require any a priori knowledge of the themes and concepts.

The invention makes it possible notably to detect sets and/or series of characteristic sound events that are representative of an environment, to offer solutions that are robust to background noise and to its variabilities.

5 It requires only very few parameters, it suffices to specify the size of the dictionary of audio words, the minimum and maximum size is documents and the temporal resolution of the analysis (delays in the stream) of the PLSA/LDA modeling, for example, in particular no threshold is required.

10 Another advantage is the automation of the processing chain, thereby facilitating its adaptation to a new scenario or a new environment, as well as its adaptation over time.

CLAIMS

1 – A non-supervised method of thematic analysis and structuring of an audio
5 stream, said audio stream comprising a set of documents each consisting of
several segments, said audio stream comprising one or more patterns or
themes repeated over time, characterized in that it comprises at least the
following steps:

• Extracting the acoustic parameters of the audio stream to be analyzed,
10 E1

• A step of low-level segmentation of said audio stream so as to obtain a
set of segments S, E2

• A step of fuzzy classification of said segments by using a dictionary of
M predefined audio words, so as to characterize each segment S with
15 a set of Q audio words and of Q degrees of membership w_{fk} of a word
in a class, E4

• Generating a cooccurrence matrix for each document of said audio
stream, an element $n(w_i, d_j)$ of the cooccurrence matrix corresponds to
a combination of the degrees of membership of all the words of the
20 document d_j which is close to the class i in the sense of the fuzzy
classification, the size $M*N$ of the matrix corresponds to the number M
of words and N to the number of documents, E5a

• For a finite segmentation set, document size t_i , start of the analysis of
the audio stream, and on the basis of the associated cooccurrence
25 matrix, determining, by adaptation of predefined models, the likelihoods
 l_i associated with each pair (document size t_i , start of analysis), E5c,
the document size belonging to a predefined interval $[t_{min}, t_{max}]$,

• Seeking the optimal segmentation in terms of document size, in the
sense of the maximum likelihood for each triplet (size of the document,
30 start of analysis, likelihood) by application of an algorithm of dynamic
programming type, E6, under the constraint of selecting documents
without temporal overlap.

2 – The method as claimed in claim 1, characterized in that it comprises a learning step adapted for generating the dictionary of the audio words and the statistical models comprising at least the following steps:

- 5 • Extracting the acoustic parameters of the audio stream to be analyzed, E1
- A step of low-level segmentation of said audio stream so as to obtain a set of segments S, E2
- A step of non-supervised construction of a dictionary of audio words, E3
- 10 • A step of fuzzy classification of said segments by using a dictionary of predefined audio words, so as to characterize a segment S with a set of Q audio words and Q degrees of membership w_{fk} of a word in a class, E4,
- Generating a cooccurrence matrix for each document of said audio stream, an element $n(w_i, d_j)$ of the matrix corresponds to a combination
15 of the degrees of membership of all the words of the document d_j close to the one class i , in the sense of the fuzzy classification, the size of the matrix corresponds to the number M of words and to the number of documents $M*N$, so as to model the way in which the audio words appear in a document, E5a
- 20 • Estimating statistical models able to decompose an audio stream into a mixture of themes defined as a multinomial distribution of words of a dictionary which are representative of the conditional probability of observing a word in an audio stream, E5b.

25 3 – The method as claimed in claim 1 or 2, characterized in that the low-level segmentation step is a uniform segmentation with segments of fixed size.

 4 – The method as claimed in claim 1 or 2, characterized in that the low-level segmentation step is executed with segments of variable size.

30

 5 – The method as claimed in claim 4, characterized in that the low-level segmentation step is a segmentation by dendrogram.

6 – The method as claimed in claim 2, characterized in that the non-supervised construction of the dictionary of audio words is executed by grouping the segments together by class with the aid of an algorithm of LBG type.

7 – The method as claimed in claim 2, characterized in that use is made of a statistical model of PLSA type defined as a mixture characterizing a model of joint probability of documents and of words.

8 – The method as claimed in claim 2 or claim 7, characterized in that the parameters of the PLSA model are estimated by maximizing the log-likelihood given by an algorithm of EM (Expectation Maximization) type.

9 – The method as claimed in claim 1, characterized in that the dynamic programming algorithm implements a weighted intervals scheduling algorithm.

10 – A non-supervised system for thematic analysis and structuring of an audio stream, said audio stream comprising a set of documents each consisting of several segments, said audio stream comprising one or more patterns or themes repeated over time, said system being adapted for implementing the method as claimed in one of claims 1 to 9, characterized in that it comprises at least the following elements:

- A database (11) comprising a pre-established dictionary of audio words, and statistical models,
- one or more acoustic sensors (2, 20) linked to an analysis module comprising a first module (22) for extracting the acoustic parameters of said audio stream, said first module being linked to a buffer memory (23), the output of the buffer memory being linked to a module for fuzzy classification (24) of said segments arising from the buffer memory (23), a module for constructing cooccurrence matrices and for adapting the models (25), a module (26) adapted for determining the likelihood of the models and an optimal segmentation search module executing

an algorithm of dynamic programming type applied to the likelihood measurements.

- 11 – The system as claimed in claim 10, characterized in that it comprises a
5 learning chain comprising at least the following elements:
one or more acoustic sensors(2), a device (3) comprising a filter and an
analog converter and then a module (6) adapted for extracting acoustic
parameters and a buffer memory (7) linked to a module (8) for creating audio
words and a classification module (9) and a module (10) for creating statistical
10 models, the set of audio words and statistical models arising from this learning
module are stored in a database (11).