(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) **EP 1 941 486 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
23.12.2015 Bulletin 2015/52

(21) Application number: **06809601.5**

(22) Date of filing: **16.10.2006**

(51) Int Cl.:
G10H 1/00 *(2006.01)*

(86) International application number:
PCT/IB2006/053787

(87) International publication number:
WO 2007/046048 (26.04.2007 Gazette 2007/17)

(54) METHOD OF DERIVING A SET OF FEATURES FOR AN AUDIO INPUT SIGNAL

VERFAHREN ZUR ABLEITUNG EINER REIHE VON EIGENSCHAFTEN FÜR EIN AUDIOEINGANGSSIGNAL

PROCEDE PERMETTANT DE DERIVER UN ENSEMBLE DE CARACTERISTIQUES POUR UN SIGNAL D'ENTREE AUDIO

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IS IT LI LT LU LV MC NL PL PT RO SE SI SK TR

(30) Priority: **17.10.2005 EP 05109648**

(43) Date of publication of application:
09.07.2008 Bulletin 2008/28

(73) Proprietor: **Koninklijke Philips N.V.**
5656 AE Eindhoven (NL)

(72) Inventors:
- **BREEBAART, Dirk, J.**
  NL-5656 AA Eindhoven (NL)
- **MCKINNEY, Martin, F.**
  NL-5656 AA Eindhoven (NL)

(74) Representative: **Uittenbogaard, Frank**
Philips
Intellectual Property & Standards
P.O. Box 220
5600 AE Eindhoven (NL)

(56) References cited:
WO-A-88/10540 **WO-A2-98/27543**
US-A- 5 918 223

- **GEORGE TZANETAKIS ET AL: "Musical Genre Classification of Audio Signals" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, US, vol. 10, no. 5, July 2002 (2002-07), XP011079656 ISSN: 1063-6676**
- **HSUAN-HUEI SHIH ET AL: "An HMM-based approach to humming transcription" MULTIMEDIA AND EXPO, 2002. ICME '02. PROCEEDINGS. 2002 IEEE INTERNATIONAL CONFERENCE ON LAUSANNE, SWITZERLAND 26-29 AUG. 2002, PISCATAWAY, NJ, USA,IEEE, US, vol. 1, 26 August 2002 (2002-08-26), pages 337-340, XP010604375 ISBN: 0-7803-7304-9**
- **PETER AHRENDT, ANDERS MENG,JAN LARSEN: "Decision time horizon for music genre classification using short time features" PROCEEDINGS OF EUPSICO, 10 September 2004 (2004-09-10), pages 1293-1296, XP002422658 Retrieved from the Internet: URL:http:// eprints.pascal-network.org/arch ive/ 00000154/01/eusipco04_rev2.pdf> [retrieved on 2006-02-28]**

EP 1 941 486 B1

Description

[0001] This invention relates to a method of deriving a set of features of an audio input signal, and to a system for deriving a set of features of an audio input signal. The invention also relates to a method of and system for classifying an audio input signal, and to a method of and system for comparing audio input signals.

[0002] Storage capabilities for digital content are increasing dramatically. Hard disks with at least one terabyte of storage capacity are expected to be available in the near future. Added to this, the evolution of compression algorithms for multimedia content, such as the MPEG standard, considerably reduces the amount of required storage capacity per audio or video file. The result is that consumers will be able to store many hours of video and audio content on a single hard disk or other storage medium. Video and audio can be recorded from an ever-increasing number of radio and TV stations. A consumer can easily augment his collection by simply downloading video and audio content from the world-wide-web, a facility which is becoming more and more popular. Furthermore, portable music players with large storage capacities are affordable and practical, allowing a user to have access, at any time, to a wide selection of music from which to choose.

[0003] The huge selection of video and audio data available from which to choose is not without problems, however. For example, organization and selection of music from a large music database, with thousands of music tracks, is difficult and time-consuming. The problem can be addressed in part by the inclusion of metadata, which can be understood to be an additional information tag attached in some way to the actual audio data file. Metadata is sometimes provided for an audio file, but this is not always the case. When faced with a time-consuming and irritating retrieval and classification problem, a user might most likely give up, or not bother at all.

[0004] Some attempts have been made in addressing the problem of classification of music signals. For example, WO 01/20609 A2 suggests a classification system in which audio signals, i.e. pieces of music or music tracks, are classified according to certain features or variables such as rhythm complexity, articulation, attack, etc. Each piece of music is assigned weighted values for a number of chosen variables, depending on the extent to which each variable applies to that piece of music. However, such a system has the disadvantage that the level of accuracy in classification or comparison of music tracks similar pieces of music is not particularly high.

[0005] Other known approaches rely on autocorrelation, such as US 5 918 223 A, or cross-correlation, such as Ahrendt et al. "Decision time horizon for music genre classification using short time features", in Proceedings of EUPSICO, 2004.

[0006] Therefore, an object of the present invention is to provide a more robust and accurate way of characterising, classifying or comparing audio signals.

[0007] To this end, the present invention, as claimed in the appended claims, provides a method of deriving a set of features of an audio input signal, particularly for use in classification of the audio input signal and/or comparison of the audio input signal with another audio signal and/or characterization of the audio input signal, which method comprises identifying a number of first-order features of the audio input signal, generating a number of correlation values from at least part of the first-order features, and compiling the set of features for the audio input signal using the correlation values. The step of identifying may comprise, for example, extracting a number of first-order features from the audio input signal or retrieving a number of first-order features from a database.

[0008] The first-order features are certain chosen descriptive characteristics of an audio input signal, and might describe signal bandwidth, zero-crossing rate, signal loudness, signal brightness, signal energy or power spectral value, etc. Other qualities described by first-order features might be spectral roll-off frequency, spectral centroid etc. The first-order features derived from the audio input signal might be chosen to be essentially orthogonal, i.e. they might be chosen to be independent from each other to a certain degree. A sequence of first-order features can be put together into what is generally referred to as a "feature vector", where a certain position in a feature vector is always occupied by the same type of feature.

[0009] The correlation value generated from a selection of the first-order features, and therefore also referred to as a second-order feature, describes the inter-dependence or covariance between these first-order features, and is a powerful descriptor for an audio input signal. It has been shown that often, with the aid of such second-order features, music tracks can accurately be compared, classified or characterised, where first-order features would be insufficient.

[0010] An obvious advantage of the method according to the invention is that a powerful descriptive set of features can easily be derived for any audio input signal, and this set of features can be used, for example, to accurately classify the audio input signal, or to quickly and accurately identify another similar audio signal. For example, a preferred set of features compiled for an audio signal, comprising elements of the first-order and second-order features, does not only describe certain chosen descriptive characteristics, but also describes the interrelationship between these chosen descriptive characteristics.

[0011] An appropriate system for deriving a set of features of an audio input signal comprises a feature identification unit for identifying a number of first-order features of the audio input signal, a correlation value generation unit for generating a number of correlation values from at least part of the first-order features, and a feature set compilation unit for compiling a set of features for the audio input signal using the correlation values. The feature identification unit may

comprise, for example, a feature extraction unit and/or a feature retrieval unit.

[0012] The dependent claims and the subsequent description disclose particularly advantageous embodiments and features of the invention.

[0013] The audio input signal can originate from any suitable source. Most generally, an audio signal might originate from an audio file, which may have any one of a number of formats. Examples of audio file formats are uncompressed, e.g. (WAV), lossless compressed, e.g. Windows Media Audio (WMA), and lossy compressed formats such as MP3 (MPEG-1 Audio Layer 3) file, AAC (Advanced Audio Codec), etc. Equally, the audio input signal can be obtained by digitising an audio signal using any suitable technique, which will be known to a person skilled in the art.

[0014] In the method according to the invention, the first-order features (sometimes also referred to as observations) for the audio input signal might preferably be extracted from one or more sections in a given domain, and generation of a correlation value preferably comprises performing a correlation using pairs of the first-order features of corresponding sections in the appropriate domain. A section can be, for example, a time-frame or segment in the time domain, where a "time-frame" is simply a range of time covering a number of audio input samples. A section can also be a frequency band in the frequency domain, or a time/frequency "tile" in a filter-bank domain. These time/frequency tiles, time-frames and frequency bands are generally of uniform size or duration. A feature associated with a section of the audio signal can hence be expressed as a function of time, as a function of frequency, or as a combination of both, so that correlations can be performed for such features in one or both domains. In the following, the terms "section" and "tile" are used interchangeably.

[0015] In a further preferred embodiment of the invention, generation of a correlation value for first-order features extracted from different, preferably neighbouring, time-frames comprises performing a correlation using first-order features of these time-frames, so that the correlation value describes the interrelationship between these neighbouring features.

[0016] In one preferred embodiment of the invention, a first-order feature is extracted in the time domain for each time-frame of the audio input signal, and a correlation value is generated by performing a cross-correlation between a pair of features over a number of consecutive feature vectors, preferably over the entire range of feature vectors.

[0017] In an alternative preferred embodiment of the invention, a first-order feature is extracted in the frequency domain for each time-frame of the audio input signal, and a correlation value is computed by performing a cross correlation between certain features of the feature vectors of two time-frames over frequency bands of the frequency domain, where the two time-frames are preferably, but not necessarily, neighbouring time-frames. In other words, for each time-frame of a plurality of time-frames, at least two first-order features are extracted for at least two frequency bands, and generation of a correlation value comprises performing a cross-correlation between of the two features over time-frames and frequency band.

[0018] The first-order features of a feature vector, since chosen to be independent or orthogonal from each other, will be features describing different aspects of the audio input signal, and will therefore be expressed in different units. To compare levels of co-variance between different variables of a collection of variables, each variable's mean deviation can be divided by its standard deviation, in a commonly known technique used to calculate the product-moment correlation or cross-correlation between two variables. Therefore, in a particularly preferred embodiment of the invention, a first-order feature used in generating a correlation value is adjusted by subtracting from it the mean or average of all appropriate features. For example, when computing a correlation value for two time-domain first-order features across the entire range of feature vectors, the mean of each of the first-order features is first computed and subtracted from the values of the first-order features before calculating a measure for the variability of a feature, such as mean deviations and standard deviations. Similarly, when computing a correlation value for two frequency-domain features from two neighbouring feature vectors, the mean of the first-order features across each of the two feature vectors is first calculated and subtracted from each first-order feature of the respective feature vector before computing the product-moment correlation or cross-correlation for the two chosen first-order features.

[0019] A number of such correlation values can be calculated, for example a correlation value each for the first & second, first & third, second & third first-order features, and so on. These correlation values, which are values describing the co-variance or interdependency between pairs of features for the audio input signal, might be combined to give a collective set of features for the audio input signal. To increase the information content of the set of features, the set of features preferably also comprises some information directly regarding the first-order features, i.e. appropriate derivatives of the first-order features such as mean or average values for each of the first-order features, taken across the range of the feature vectors. Equally, it may suffice to obtain such second-order features for only a sub-set of the first-order features, such as, for example, the mean value for the first, third and fifth features taken over a chosen range of feature vectors.

[0020] The set of features, in effect an extended feature vector comprising first- and second-order features, obtained using the method according to the invention can be stored independently of the audio signal for which it was derived, or it can be stored together with the audio input signal, for example in the form of metadata.

[0021] A music track or song can then be described accurately by the set of features derived for it according to the

method described above. Such feature sets make it possible to carry out, with a high degree of accuracy, classification and comparison for pieces of music.

[0022] For example, if feature sets or extended feature vectors for a number of audio signals of similar nature, such as those belonging to a single class - e.g. "baroque" - are derived, these feature sets can then be used to build a model for the class "baroque". Such a model might be, for example, a Gaussian multivariate model with each class having its own mean vector and its own covariance matrix in a feature space occupied by extended feature vectors. Any number of groups or classes can be trained. For music audio input signals, such a class might be defined broadly, for example "reggae", "country", "classic", etc. Equally, the models can be more narrow or refined, for example "80s disco", "20s jazz", "finger-style guitar", etc., and are trained with suitably representative collections of audio input signals.

[0023] To ensure optimal classification results, the dimensionality of the model space is kept as low as possible, i.e. by choosing a minimum number of first-order features, while choosing these first-order features to give the best possible discrimination between classes. Known methods of feature ranking and dimensionality reduction can be applied to determine the best first-order features to choose. Once a model for a group or class is trained using a number of audio signals known to belong to that group or class, an "unknown" audio signal can be tested to determine whether it belongs to that class by simply checking whether the set of features for that audio input signal fits the model to within a certain degree of similarity.

[0024] Therefore, a method of classifying an audio input signal into a group preferably comprises deriving a set of features for the input audio signal and determining, on the basis of the set of features, the probability that the audio input signal corresponds to any of a number of groups or classes, where each group or class corresponds to a particular audio class.

[0025] A corresponding classifying system for classifying an audio input signal into one or more groups might comprise a system for deriving a set of features of the audio input signal, and a probability determination unit for determining, on the basis of the set of features of the audio input signal, the probability that the input audio signal falls within any of a number of groups, where each group corresponds to a particular audio class.

[0026] Another application of the method according to the invention might be to compare audio signals, for example, two songs, on the basis of their respective feature sets, in order to determine the level of similarity, if any, between them.

[0027] Such a method of comparison therefore preferably comprises the steps of deriving a first set of features for a first audio input signal and deriving a second set of features for a second audio input signal and then calculating a distance between the first and second sets of features in a feature space according to a defined distance measure, before finally determining the degree of similarity between the first and second audio signals based on the calculated distance. The distance measure used might be, for example, a Euclidean distance between certain points in feature space.

[0028] A corresponding comparison system for comparing audio input signals to determine a degree of similarity between them might comprise a system for deriving a first set of features for a first audio input signal and a system for deriving a second set of features for a second audio input signal, as well as a comparator unit for calculating a distance between the first and second sets of features in a feature space according to a defined distance measure, and for determining the degree of similarity between the audio input signals on the basis of the calculated distance. Evidently, the system for deriving the first set of features and the system for deriving the second set of features might be one and the same system.

[0029] The invention might find application in a variety of audio processing applications. For example, in a preferred embodiment, the classifying system for classifying an audio input signal as described above might be incorporated in an audio processing device. The audio processing device might have access to a music database or collection, organised by class or group, into which the audio input signal is classified. Another type of audio processing device might comprise a music query system for choosing one or more music data files from a particular group or class of music in the database. A user of such a device can therefore easily put together a collection of songs for entertainment purposes, for example for a themed music event. A user availing of a music database where songs have been classified according to genre and decade might specify that a number of songs belonging to a category such as "pop, 1980s" be retrieved from the database. Another useful application of such an audio processing device would be to assemble a collection of songs having a certain mood or rhythm suitable for accompanying an exercise workout, vacation slide-show presentation, etc. A further useful application of this invention might be to search a music database for one or music tracks similar to a known music track.

[0030] The systems according to the invention for deriving feature sets, classifying audio input signals, and comparing input signals can be realised in a straightforward manner as a computer program or programs. All components for deriving feature sets of an input signal such as feature extraction unit, correlation value generation unit, feature set compilation unit, etc. can be realised in the form of computer program modules. Any required software or algorithms might be encoded on a processor of a hardware device, so that an existing hardware device might be adapted to benefit from the features of the invention. Alternatively, the components for deriving feature sets of an audio input signal can equally be realised at least partially using hardware modules, so that the invention can be applied to digital and/or analog audio input signals.

[0031] Other objects and features of the present invention will become apparent from the following detailed descriptions considered in conjunction with the accompanying drawing. It is to be understood, however, that the drawings are designed solely for the purposes of illustration and not as a definition of the limits of the invention.

Fig.1 is an abstract representation of the relationship between time-frames and features extracted from an input audio signal;

Fig. 2a is a schematic block diagram of a system for deriving a set of features from an audio input signal according to a first embodiment of the invention;

Fig. 2b is a schematic block diagram of a system for deriving a set of features from an audio input signal according to a second embodiment of the invention;

Fig. 3 is a schematic block diagram of a system for deriving a set of features from an audio input signal according to a third embodiment of the invention;

Fig. 4 is a schematic block diagram of a system for classifying an audio signal;

Fig. 5 is a schematic block diagram of a system for comparing audio signals.

[0032] In the diagrams, like numbers refer to like objects throughout.

[0033] To simplify understanding of the methods pursuant to the invention and described below, Fig. 1 gives an abstract representation between time-frames t_1, t_2, ..., t_I or sections of an input signal M and the set of features S ultimately derived for that input signal M.

[0034] The input signal for which a set of features is to be derived could originate from any appropriate source, and could be a sampled analog signal, an audio-coded signal such as an MP3 or AAC file, etc. In this diagram, the audio input M is first digitized in a suitable digitising unit 10 which outputs a series of analysis windows from the digitised stream of samples. An analysis window can be of a certain duration, for example, 743ms. A windowing unit 11 further sub-divides an analysis window into a total of I overlapping time-frames t_1, t_2, ..., t_I, so that each time frame t_1, t_2, ..., t_I covers a certain number of the samples of the audio input signal M. Consecutive analysis windows can be chosen so that they overlap by several tiles, which is not shown in the diagram. Alternatively, a single, sufficiently wide analysis window can be used from which to extract the features.

[0035] For each of these time-frames t_1, t_2, ..., t_I, a number of first-order features f_1, f_2, ..., f_f is extracted in a feature extraction unit 12. These first-order features f_1, f_2, ..., f_f might be computed from a time-domain or frequency domain signal representation, and can vary as a function of time and/or frequency, as will be explained in greater detail below. Each group of first-order features f_1, f_2, ..., f_f for a time/frequency tile or time-frame is referred to as a first-order feature vector, so that feature vectors fv_1, fv_2, ..., fv_I are extracted for the tiles t_1, t_2, ..., t_I.

[0036] In a correlation value generation unit 13, correlation values are generated for certain pairs of first-order features f_1, f_2, ..., f_f. The pairs of features may be taken from single feature vectors fv_1, fv_2, ..., fv_I or from across different feature vectors fv_1, fv_2, ..., fv_I. For example, a correlation might be computed for the pair of features ($fv_1[i]$, $fv_2[i]$), taken from different feature vectors, or for the pair of features ($fv_1[j]$, $fv_1[k]$) from the same feature vector.

[0037] In a feature processing block 15, one or more derivatives fm_1, fm_2, ..., fm_f of the first-order features fv_1, fv_2, ..., fv_I, e.g. a mean value, an average value or set of average values can be computed across the first-order feature vectors fv_1, fv_2, ..., fv_I.

[0038] The correlation values generated in the correlation value generation unit 13 are combined in a feature set compilation unit 14 with the derivative(s) fm_1, fm_2, ..., fm_f of the first-order features f_1, f_2, ..., f_f computed in the feature processing block 15 to give a set of features S for the audio input signal M. Such a feature set S can be derived for every analysis window, and used to compute an average feature set for the entire audio input signal M, which might then be stored as metadata in an audio file, together with the audio signal, or in a separate metadata database, as required.

[0039] In Fig. 2a, the steps of deriving a set of features S in the time domain for an audio input signal x(n) are explained in more detail. The audio input signal M is first digitized in a digitization block 10 to give a sampled signal:

$$x[n] = x(\frac{n}{f_s}) \tag{1}$$

[0040] Subsequently, the sampled input signal x[n] is windowed in a windowing block 20 to yield a group of windowed samples $x_i[n]$ of size N and hop-size H for a tile in the time-domain using a window w[n]:

$$x_i[n] = \begin{cases} w[n]x[n+Hi] & \text{for} \quad 0 \le n < N \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

[0041] Each group of samples $x_i[n]$, corresponding to a time-frame t_i in the diagram, is then transformed to the frequency domain, in this case by taking the Fast Fourier Transform (FFT):

$$X_i[k] = \sum_n x_i[n] \exp\{-2\pi jnk / N\} \quad (3)$$

[0042] Subsequently, in a log power calculation unit 21, values for log-domain sub-band power P[b], are computed for a set of frequency sub-bands, using a filter kernel $W_b[k]$ for each frequency sub-band b:

$$P_i[b] = 10 \log_{10}\left(\sum_k X_i[k] X_i^*[k] W_b[k] \right) \quad (4)$$

[0043] Finally, in a coefficient calculation unit 22, the Mel-frequency cepstral coefficients ($MFCC_s$) for each time-frame are obtained by the direct cosine transform (DCT) of each sub-band power value P[b] over B power sub-bands:

$$MFCC_i[m] = \sqrt{\frac{1}{B}} \sum_b P_i[b] \cos\left(\frac{\pi (2b+1) m}{2B} \right) \quad (5)$$

[0044] The windowing unit 20, log power calculation unit 21 and coefficient calculation unit 22 taken together give a feature extraction unit 12. Such a feature extraction unit 12 is used to calculate the features f_1, f_2, ..., f_f for each of a number of analysis windows of the input signal M. The feature extraction unit 12 will generally comprise a number of algorithms realised in software, perhaps combined as a software package. Evidently, a single feature extraction unit 12 can be used to process each analysis window separately, or a number of separate feature extraction units 12 can be implemented so that several analysis windows can be processed simultaneously.

[0045] Once a certain set of time-frames I has been processed as described above, a second-order feature can be computed (over the analysis frame of I sub-frames) that consists of the (normalized) correlation coefficient between certain frame-based features. This takes place in a correlation value generation unit 13. For example, the correlation between the y-th and z-th MFCC coefficient across time is given as follows by equation (6):

$$\rho(y,z) = \frac{\sum_i \left(MFCC_i[y] - \mu_y\right)\left(MFCC_i[z] - \mu_z\right)}{\sqrt{\sum_i \left(MFCC_i[y] - \mu_y\right)\left(MFCC_i[y] - \mu_y\right) \sum_i \left(MFCC_i[z] - \mu_z\right)\left(MFCC_i[z] - \mu_z\right)}}$$

where μ_y and μ_z are the means (across I) of $MFCC_i[y]$ and $MFCC_i[z]$ respectively. Adjustment of each coefficient by subtracting the mean gives a Pearson's correlation coefficient as second-order feature, which is in effect a measure the strength of the linear relationship between two variables, in this case the two coefficients $MFCC_i[y]$ and $MFCC_i[z]$.

[0046] The correlation value $\rho(y,z)$ calculated above can then be used as a contribution to a set of features S. Other elements of the set of features S can be derivatives of the first-order feature vectors fv_1, fv_2, ..., fv_I of a time-frame, calculated in a feature processing block 15, for example mean or average values of the first few features f_1, f_2, ..., f_f of each feature vector fv_1, fv_2, ..., fv_I, taken over the entire range of feature vectors fv_1, fv_2, ..., fv_I.

[0047] Such derivatives of the first-order feature vectors fv_1, fv_2, ..., fv_I are combined with the correlation values in a feature combination unit 14 to give the set of features S as output. The set of features S can be stored with or separately from the audio input signal M in a file, or can be further processed before storing. Thereafter, the set of features S can be used, for instance, to classify the audio input signal M, to compare the audio input signal M with another audio signal, or to characterize the audio input signal M.

[0048] Fig. 2b shows a block diagram of a second embodiment of the invention in which the features are extracted in the frequency domain for a total B of discrete frequency sub-bands. The first few stages, up to and including the computation of the log sub-band power values are effectively the same as those already described above under Fig. 2. In this realisation, however, the values of power for each frequency sub-band are directly used as features, so that a feature vector fv_i, fv_{i+1} in this case comprises the values of power for each frequency sub-band over the range of frequency sub-bands, as given in equation (4). Therefore, the feature extraction unit 12' requires only a windowing unit 20 and log

power calculation unit 21.

[0049] Calculation of a correlation value or second-order feature in this case is carried out in a correlation value generation unit 13' for consecutive pairs of time-frames t_i, t_{i+1}, i.e. over pairs of feature vectors f_i, f_{i+1}. Again, each feature in each feature vector f, f_{i+1} is first adjusted by subtracting from it a mean value μ_{Pi}, μ_{Pi+1}. In this case, for example, μ_{Pi} is calculated by summing all the elements of the feature vector f_i and dividing the sum by the total number of frequency sub-bands, B. The correlation value $\rho(P_i, P_{i+1})$ for a pair of feature vectors f_i, f_{i+1} is computed as follows:

$$\rho(P_i, P_{i+1}) = \frac{\sum_b (P_i[b] - \mu_{Pi})(P_{i+1}[b] - \mu_{Pi+1})}{\sqrt{\sum_b (P_i[b] - \mu_{Pi})(P_i[b] - \mu_{Pi}) \sum_b (P_{i+1}[b] - \mu_{Pi+1})(P_{i+1}[b] - \mu_{Pi+1})}} \quad (7)$$

[0050] The correlation values for feature vector pairs can be combined in a feature combination unit 14', as described under Fig. 2 above, with derivatives of the first-order features calculated in a feature processing block 15' to give as output the set of features S. Again, as already described above, the set of features S can be stored with or separately from the audio input signal in a file, or can be further processed before storing.

[0051] Fig. 3 illustrates a third embodiment of the invention where features extracted from an input signal contain both time-domain and frequency-domain information. Here, the audio input signal x[n] is a sampled signal. Each sample is input to a filter-bank 17 comprising a total of K filters. The output of the filter-bank 17 for an input sample x[n] is, therefore, a sequence of values y[m, k], where $1 \leq k \leq K$. Each k index represents a different frequency band of the filter-bank 17, whereas each m index represents time, i.e. the sampling rate of the filter-bank 17. For every filter-bank output y[m, k], features $f_a[m, k]$, $f_b[m, k]$ are calculated. The feature type $f_a[m, k]$ in this case can be the power spectral value of its input y[m, k], while the feature type $f_b[m, k]$ is the power spectral value calculated for the previous sample. Pairs of these features $f_a[m, k]$, $f_b[m, k]$ can be correlated across the range of frequency sub-bands, i.e. for values of $1 \leq k \leq K$, to give correlation values $\rho(f_a, f_b)$:

$$\rho(f_a, f_b) = \frac{\sum_m \sum_k (f_a[m,k] - \mu_{f_a})(f_b[m,k] - \mu_{f_b})}{\sqrt{\left(\sum_m \sum_k (f_a[m,k] - \mu_{f_a})^2\right)\left(\sum_m \sum_k (f_b[m,k] - \mu_{f_b})^2\right)}} \quad (8)$$

[0052] In Fig. 4, a simplified block diagram of a system 4 for classification of an audio signal M is shown. Here, the audio signal M is retrieved from a storage medium 40, for example a hard-disk, CD, DVD, music database, etc. In a first stage, a set of features S is derived for the audio signal M using a system 1 for feature set derivation. The resulting set of features S is forwarded to a probability determination unit 43. This probability determination unit 43 is also supplied with class feature information 42 from a data source 45, describing the feature positions, in feature space, of the classes to which the audio signal can possibly be assigned.

[0053] In the probability determination unit 43, a distance measurement unit 46 measures, for example, the Euclidean distances in feature space between the features of the set of features S and the features supplied by the class feature information 42. A decision making unit 47 decides, on the basis of the measurements, to which class(es), if any, the set of features S, and therefore the audio signal M, can be assigned.

[0054] In the event of a successful classification, suitable information 44 can be stored in an metadata file 41 associated, by a suitable link 48, with the audio signal M. The information 44, or metadata, might comprise the set of features S of the audio signal M as well as the class to which the audio signal M has been assigned, along with, for instance, a measure of the degree to which this audio signal M belongs to that class.

[0055] Fig. 5 shows a simplified block diagram of a system 5 for comparing audio signals M, M' such as can be retrieved from databases 50, 51. With the aid of two systems 1, 1' for feature set derivation, feature set S and feature set S' are derived for music signal M and music signal M' respectively. Merely for the sake of simplicity, the diagram shows two separate systems 1, 1' for feature set derivation. Naturally, a single such system could be implemented, by simply performing the derivation for one audio signal M and then for the other audio signal M'.

[0056] The feature sets S, S' are input to a comparator unit 52. In this comparator unit 52, the feature sets S, S' are analysed in a distance analysis unit 53 to determine the distances in feature space between the individual features of the feature sets S, S'. The result is forwarded to a decision making unit 54, which uses the result of the distance analysis unit 53 to decide whether or not the two audio signals M, M' are sufficiently similar to be deemed to belong to the same group. The result arrived at by the decision making unit 54 is output as a suitable signal 55, which might be a simple

yes/no type of result, or a more informative judgement as to the similarity, or lack of similarity, between the two audio signals M, M'.

[0057] Although the present invention has been disclosed in the form of preferred embodiments and variations thereon, it will be understood that numerous additional modifications and variations could be made thereto without departing from the scope of the invention. For example, the method for deriving a feature set for a music signal could be used in a audio processing device which characterises music tracks, with possible applications for generation of descriptive metadata for the music tracks. Furthermore, the invention is not limited to using the methods of analysis described, but may apply any suitable analytical method.

[0058] For the sake of clarity, it is also to be understood that the use of "a" or "an" throughout this application does not exclude a plurality, and "comprising" does not exclude other steps or elements. A "unit" or "module" may comprise a number of blocks or devices, as appropriate, unless explicitly described as a single entity.

Claims

1. A method of deriving a set of features (S) of an audio input signal (M), which method comprises:

   - identifying a number of first-order features (f_1, f_2, ..., f_f) of the audio input signal (M);
   - generating a number of correlation values (ρ_1, ρ_2, ..., ρ_l) from at least part of the first-order features (f_1, f_2, ..., f_f);
   - and compiling the set of features (S) for the audio input signal (M) using the correlation values (ρ_1, ρ_2, ..., ρ_l) wherein different first-order features (f_1, f_2, ..., f_f, f_a, f_b) are extracted from one section (t_1, t_2, ..., t_l) in a given domain of the audio input signal (M), and the generation of a correlation value (ρ_1, ρ_2, ..., ρ_l, ρ) comprises performing a correlation using pairs of different ones of the extracted first-order features (f_1, f_2, ..., f_f, f_a, f_b) of the section in this domain.

2. A method of deriving a set of features (S) of an audio input signal (M), which method comprises:

   - dividing the audio signal into a plurality of frequency sub-bands;
   - identifying a number of first-order features (f_1, f_2, ..., f_f) of at least one of said frequency sub-bands of the audio input signal (M);
   - generating a number of correlation values (ρ_1, ρ_2, ..., ρ_l) from at least part of the first-order features (f_1, f_2, ..., f_f);
   - and compiling the set of features (S) for the audio input signal (M) using the correlation values (ρ_1, ρ_2, ..., ρ_l) wherein the first-order features (f_1, f_2, ..., f_f, f_a, f_b) are extracted from different time-frames (t_1, t_2, ..., t_l) of the audio input signal (M), and the generation of a correlation value (ρ_1, ρ_2, ..., ρ_l, ρ) comprises performing a correlation using first-order features (f_1, f_2, ..., f_f, f_a, f_b) of the same frequency sub-band of different time-frames (t_1, t_2, ..., t_l).

3. A method according to claim 2, wherein, for each time-frame (t_1, t_2, ..., t_l) of a plurality of time-frames, a first-order feature vector (fv_1, fv_2, ..., fv_l) is extracted as a function of time , and generation of a correlation value (ρ_1, ρ_2, ..., ρ_l) comprises performing a cross-correlation between certain elements of the feature vectors (fv_1, fv_2, ..., fv_l) over a number of the feature vectors (fv_1, fv_2, ..., fv_l).

4. A method according to claim 2, wherein, for each time-frame (t_1, t_2, ..., t_l) of a plurality of time-frames, a first-order feature vector (fv_1, fv_2, ..., fv_l) is extracted as a function of frequency, and generation of a correlation value (ρ_1, ρ_2, ..., ρ_l) comprises performing a cross-correlation between certain elements of the feature vectors (fv_1, fv_2, ..., fv_l) of two time-frames (t_i, t_{i+1}) over frequency.

5. A method according to any of the preceding claims, wherein a first-order feature (f_1, f_2, ..., f_f) used in generating a correlation value (ρ_1, ρ_2, ..., ρ_l) is adjusted by a mean of corresponding first-order features (f_1, f_2, ..., f_f) prior to generation of the correlation value (ρ_1, ρ_2, ..., ρ_l).

6. A method according to any of the preceding claims, wherein the set of features (S) comprises a number of correlation values (ρ_l, ρ_2, ..., ρ_l) and a derivative of at least a number of the first-order features (f_1, f_2, ..., f_f).

7. A method of classifying an audio input signal (M) into a group and determining, on the basis of the set of features (S) of the audio input signal (M), the probability that the audio input signal (M) falls within any of a number of groups, where each group represents a particular audio class, wherein the set of features (S) has been derived using a method according to any of claims 1 to 6.

8. A method of comparing audio input signals (M, M') to determine a degree of similarity between the audio input signals (M, M'), which method comprises:

- deriving a first set of features (S) for a first audio input signal (M);
- deriving a second set of features (S') for a second audio input signal (M');
- calculating a distance between the first and second sets of features (S, S') in a feature space according to a defined distance measure;
- determining the degree of similarity between the first and second audio signals (M, M') based on the calculated distance,

wherein the first and second set of features (S) have been derived using a method according to any of claims 1 to 6.

9. A system (1) for deriving a set of features (S) of an audio input signal (M), comprising:

- a feature identification unit (12,12') for identifying a number of first-order features (f_1, f_2, ..., f_f) of the audio input signal (M);
- a correlation value generation unit (13,13') for generating a number of correlation values (ρ_1, ρ_2, ..., ρ_l) from at least part of the first-order features (f_1, f_2, ..., f_f);
- and a feature set compilation unit (14,14') for compiling the set of features (S) for the audio input signal (M) using the correlation values (ρ_1, ρ_2, ..., ρ_l) wherein different first-order features (f_1, f_2, ..., f_f, f_a, f_b) are extracted from one section (t_1, t_2, ..., t_l) in a given domain of the audio input signal (M), and the generation of a correlation value (ρ_1, ρ_2, ..., ρ_l, p) comprises performing a correlation using pairs of the first-order features (f_1, f_2, ..., f_f, f_a, f_b) of the section in this domain.

10. A system (1) for deriving a set of features (S) of an audio input signal (M), comprising:

- a feature identification unit (12,12') for identifying a number of first-order features (f_1, f_2, ..., f_f) of the audio input signal (M);
- a correlation value generation unit (13,13') for generating a number of correlation values (ρ_l, ρ_2, ..., ρ_l) from at least part of the first-order features (f_1, f_2, ..., f_f);
- and a feature set compilation unit (14,14') for compiling the set of features (S) for the audio input signal (M) using the correlation values (ρ_1, ρ_2, ..., ρ_l) wherein the first-order features (f_1, f_2, ..., f_f, f_a, f_b) are extracted from different time-frames (t_1, t_2, ..., t_l) of the audio input signal (M), and the generation of a correlation value (ρ_1, ρ_2, ..., p_l, ρ) comprises performing a correlation using first-order features (f_1, f_2, ..., f_f, f_a, f_b) of the same frequency sub-band of different time-frames (t_1, t_2, ..., t_l).

11. A classifying system (4) for classifying an audio input signal (M) into a group, comprising a probability determination unit (43) for determining, on the basis of the set of features (S) of the audio input signal (M), the probability that the input audio signal (M) falls within any of a number of groups, where each group represents a particular audio class, wherein the set of features (S) has been derived using a method according to any of claims 1 to 6.

12. A comparison system (5) for comparing audio input signals (M, M') to determine a degree of similarity between the audio input signals (M, M'), comprising:

- a comparator unit (52) for calculating a distance between a first and second sets of features (S, S') in a feature space according to a defined distance measure, and for determining the degree of similarity between the audio input signals (M, M') on the basis of the calculated distance, wherein the first and second set of features (S) have been derived using a method according to any of claims 1 to 6.

13. An audio processing device comprising a classifying system (4) according to claim 11 and/or a comparison system (5) according to claim 12.

14. A computer program product directly loadable into the memory of a programmable audio processing device comprising software code portions for performing the steps of a method of deriving a set of features (S) according to claims 1 to 6 or for performing the steps of a method of classifying an audio input signal (M) according to claims 7 or for performing the steps of a method of comparing audio input signals (M, M') according to claim 8, when said program is run on the audio processing device.

15. A database comprising a set of features (S) derived of an audio input signal (M), wherein the set of features (S) has been derived using a method according to any of claims 1 to 6.

Patentansprüche

1. Verfahren zur Ableitung einer Reihe von Eigenschaften (S) eines Audioeingangssignals (M), wobei das Verfahren Folgendes umfasst:

- Identifizieren einer Anzahl von Eigenschaften erster Ordnung (f_1, f_2, ..., f_f) des Audioeingangssignals (M);
- Erzeugen einer Anzahl von Korrelationswerten (ρ_1, ρ_2, ..., ρ_l) von wenigstens einem Teil der Eigenschaften erster Ordnung (f_1, f_2, ..., f_f);
- und Erstellen der Reihe von Eigenschaften (S) für das Audioeingangssignal (M) unter Verwendung der Korrelationswerte (ρ_1, ρ_2, ..., ρ_l), wobei unterschiedliche Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) aus einem Abschnitt (t_1, t_2, ..., t_l) in einem gegebenen Bereich des Audioeingangssignals (M) extrahiert werden, und die Erzeugung eines Korrelationswertes (ρ_1, ρ_2, ..., ρ_l, ρ) das Anwenden einer Korrelation unter Verwendung von Paaren von verschiedenen der extrahierten Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) des Abschnitts in diesem Bereich umfasst.

2. Verfahren zur Ableitung einer Reihe von Eigenschaften (S) eines Audioeingangssignals (M), wobei das Verfahren Folgendes umfasst:

- Unterteilen des Audiosignals in eine Vielzahl von Frequenzteilbändern;
- Identifizieren einer Anzahl von Eigenschaften erster Ordnung (f_1, f_2, ..., f_f) von wenigstens einem der Frequenzteilbänder des Audioeingangssignals (M);
- Erzeugen einer Anzahl von Korrelationswerten (ρ_1, ρ_2, ..., ρ_l) von wenigstens einem Teil der Eigenschaften erster Ordnung (f_1, f_2, ..., f_f);
- und Erstellen der Reihe von Eigenschaften (S) für das Audioeingangssignal (M) unter Verwendung der Korrelationswerte (ρ_1, ρ_2, ..., ρ_l), wobei die Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) aus unterschiedlichen Zeitrahmen (t_1, t_2, ..., t_l) des Audioeingangssignals (M) extrahiert werden, und die Erzeugung eines Korrelationswertes (ρ_1, ρ_2, ..., ρ_l, p) das Anwenden einer Korrelation unter Verwendung von Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) des gleichen Frequenzteilbandes von unterschiedlichen Zeitrahmen (t_1, t_2, ..., t_l) umfasst.

3. Verfahren nach Anspruch 2, wobei für jeden Zeitrahmen (t_1, t_2, ..., t_l) von einer Vielzahl von Zeitrahmen ein Eigenschaftsvektor erster Ordnung (fv_1, fv_2, ..., fv_l) in Abhängigkeit von der Zeit extrahiert wird, und die Erzeugung eines Korrelationswertes (ρ_1, ρ_2, ..., ρ_l) das Anwenden einer Kreuzkorrelation zwischen bestimmten Elementen der Eigenschaftsvektoren (fv_1, fv_2, ..., fv_l) über eine Anzahl von Eigenschaftsvektoren (fv_1, fv_2, ..., fv_l) umfasst.

4. Verfahren nach Anspruch 2, wobei für jeden Zeitrahmen (t_1, t_2, ..., t_l) von einer Vielzahl von Zeitrahmen ein Eigenschaftsvektor erster Ordnung (fv_1, fv_2, ..., fv_l) in Abhängigkeit von der Frequenz extrahiert wird, und die Erzeugung eines Korrelationswertes (ρ_1, ρ_2, ..., ρ_l) das Anwenden einer Kreuzkorrelation zwischen bestimmten Elementen der Eigenschaftsvektoren (fv_1, fv_2, ..., fv_l) von zwei Zeitrahmen (t_i, t_{i+1}) über die Frequenz umfasst.

5. Verfahren nach einem der vorhergehenden Ansprüche, wobei eine Eigenschaft erster Ordnung (f_1, f_2, ..., f_f), die beim Erzeugen eines Korrelationswertes (ρ_1, ρ_2, ..., ρ_l) verwendet wird, durch ein Mittel von entsprechenden Eigenschaften erster Ordnung (f_1, f_2, ..., f_f) vor der Erzeugung des Korrelationswertes (ρ_1, ρ_2, ..., ρ_l) angepasst wird.

6. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Reihe von Eigenschaften (S) eine Anzahl von Korrelationswerten (ρ_1, ρ_2, ..., ρ_r) und eine Ableitung von wenigstens einer Anzahl der Eigenschaften erster Ordnung (f_1, f_2, ..., f_f) umfasst.

7. Verfahren zum Einteilen eines Audioeingangssignals (M) in eine Gruppe und Bestimmen, basierend auf der Reihe von Eigenschaften (S) des Audioeingangssignals (M), der Wahrscheinlichkeit, dass das Audioeingangssignal (M) in eine beliebige einer Anzahl von Gruppen hineinfällt, wobei jede Gruppe eine bestimmte Audioklasse darstellt, wobei die Reihe von Eigenschaften (S) unter Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 6 abgeleitet wurde.

8. Verfahren zum Vergleichen von Audioeingangssignalen (M, M'), um einen Ähnlichkeitsgrad zwischen den Audio-

eingangssignalen (M, M') zu bestimmen, wobei das Verfahren Folgendes umfasst:

- Ableiten einer ersten Reihe von Eigenschaften (S) für ein erstes Audioeingangssignal (M);
- Ableiten einer zweiten Reihe von Eigenschaften (S') für ein zweites Audioeingangssignal (M');
- Berechnen einer Distanz zwischen der ersten und der zweiten Reihe von Eigenschaften (S, S') in einem Eigenschaftsraum gemäß einer definierten Distanzmessung;
- Bestimmen des Ähnlichkeitsgrades zwischen dem ersten und dem zweiten Audiosignal (M, M') basierend auf der berechneten Distanz,

wobei die erste und die zweite Reihe von Eigenschaften (S) unter Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 6 abgeleitet wurden.

9. System (1) zur Ableitung einer Reihe von Eigenschaften (S) eines Audioeingangssignals (M), umfassend:

- eine Eigenschaftsidentifikationseinheit (12, 12') zum Identifizieren einer Anzahl von Eigenschaften erster Ordnung (f_1, f_2, ..., f_f) des Audioeingangssignals (M);
- eine Korrelationswert-Erzeugungseinheit (13, 13') zum Erzeugen einer Anzahl von Korrelationswerten (ρ_1, ρ_2, ..., ρ_l) von wenigstens einem Teil der Eigenschaften erster Ordnung (f_1, f_2, ..., f_f);
- und eine Eigenschaftsreihe-Erstellungseinheit (14, 14') zum Erstellen der Reihe von Eigenschaften (S) für das Audioeingangssignal (M) unter Verwendung der Korrelationswerte (ρ_1, ρ_2, ..., ρ_l), wobei unterschiedliche Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) aus einem Abschnitt (t_1, t_2, ..., t_l) in einem gegebenen Bereich des Audioeingangssignals (M) extrahiert werden, und die Erzeugung eines Korrelationswertes (ρ_1, ρ_2, ..., ρ_l, p) das Anwenden einer Korrelation unter Verwendung von Paaren der Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) des Abschnitts in diesem Bereich umfasst.

10. System (1) zur Ableitung einer Reihe von Eigenschaften (S) eines Audioeingangssignals (M), umfassend:

- eine Eigenschaftsidentifikationseinheit (12, 12') zum Identifizieren einer Anzahl von Eigenschaften erster Ordnung (f_1, f_2, ..., f_f) des Audioeingangssignals (M);
- eine Korrelationswert-Erzeugungseinheit (13, 13') zum Erzeugen einer Anzahl von Korrelationswerten (ρ_1, ρ_2, ..., ρ_l) von wenigstens einem Teil der Eigenschaften erster Ordnung (f_1, f_2, ..., f_f);
- und eine Eigenschaftsreihe-Erstellungseinheit (14, 14') zum Erstellen der Reihe von Eigenschaften (S) für das Audioeingangssignal (M) unter Verwendung der Korrelationswerte (ρ_1, ρ_2, ..., ρ_l), wobei die Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) aus unterschiedlichen Zeitrahmen (t_1, t_2, ..., t_l) des Audioeingangssignals (M) extrahiert werden, und die Erzeugung eines Korrelationswertes (ρ_1, ρ_2, ..., ρ_l, ρ) das Anwenden einer Korrelation unter Verwendung von Eigenschaften erster Ordnung (f_1, f_2, ..., f_f, f_a, f_b) des gleichen Frequenzteilbandes von unterschiedlichen Zeitrahmen (t_1, t_2, ..., t_l) umfasst.

11. Einteilungssystem (4) zum Einteilen eines Audioeingangssignals (M) in eine Gruppe, umfassend eine Wahrscheinlichkeitsbestimmungseinheit (43) zum Bestimmen, basierend auf der Reihe von Eigenschaften (S) des Audioeingangssignals (M), der Wahrscheinlichkeit, dass das Audioeingangssignal (M) in eine beliebige einer Anzahl von Gruppen hineinfällt, wobei jede Gruppe eine bestimmte Audioklasse darstellt, wobei die Reihe von Eigenschaften (S) unter Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 6 abgeleitet wurde.

12. Vergleichssystem (5) zum Vergleichen von Audioeingangssignalen (M, M'), um einen Ähnlichkeitsgrad zwischen den Audioeingangssignalen (M, M') zu bestimmen, umfassend:

- eine Vergleichseinheit (52) zum Berechnen einer Distanz zwischen einer ersten und zweiten Reihe von Eigenschaften (S, S') in einem Eigenschaftsraum gemäß einer definierten Distanzmessung, und zum Bestimmen des Ähnlichkeitsgrades zwischen den Audioeingangssignalen (M, M') basierend auf der berechneten Distanz, wobei die erste und die zweite Reihe von Eigenschaften (S) unter Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 6 abgeleitet wurden.

13. Audioverarbeitungsvorrichtung, umfassend ein Einteilungssystem (4) nach Anspruch 11 und/oder ein Vergleichssystem (5) nach Anspruch 12.

14. Computerprogrammprodukt, das direkt in den Speicher einer programmierbaren Audioverarbeitungsvorrichtung geladen werden kann, umfassend Softwarecodeteile zum Ausführen der Schritte eines Verfahrens zur Ableitung

einer Reihe von Eigenschaften (S) nach den Ansprüchen 1 bis 6 oder zum Ausführen der Schritte eines Verfahrens zum Einteilen eines Audioeingangssignals (M) nach Anspruch 7 oder zum Ausführen der Schritte eines Verfahrens zum Vergleichen von Audioeingangssignalen (M, M') nach Anspruch 8, wenn das Programm auf der Audioverarbeitungsvorrichtung läuft.

15. Datenbank, umfassend eine Reihe von Eigenschaften (S), die von einem Audioeingangssignal (M) abgeleitet werden, wobei die Reihe von Eigenschaften (S) unter Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 6 abgeleitet wurde.

Revendications

1. Procédé d'obtention d'un jeu de caractéristiques (S) d'un signal d'entrée audio (M), lequel procédé comprend :

- l'identification d'un nombre de caractéristiques de premier ordre (f_1, f_2, ..., f_f) du signal d'entrée audio (M) ;
- la génération d'un nombre de valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l) à partir d'au moins une partie des caractéristiques de premier ordre (f_1, f_2, ..., f_f) ; et
- la compilation du jeu de caractéristiques (S) pour le signal d'entrée audio (M) en utilisant les valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l), dans lequel différentes caractéristiques de premier ordre (f_1, f_2, ..., f_f, f_a, f_b) sont extraites d'une section (t_1, t_2, ..., t_l) dans un domaine donné du signal d'entrée audio (M), et la génération d'une valeur de corrélation (ρ_1, ρ_2, ..., ρ_l, ρ) comprend la réalisation d'une corrélation en utilisant des paires de différentes caractéristiques parmi les caractéristiques de premier ordre extraites (f_1, f_2, ..., f_f, f_a, f_b) de la section dans ce domaine.

2. Procédé d'obtention d'un jeu de caractéristiques (S) d'un signal d'entrée audio (M), lequel procédé comprend :

- la division du signal audio en une pluralité de sous-bandes de fréquence ;
- l'identification d'un nombre de caractéristiques de premier ordre (f_1, f_2, ..., f_f) d'au moins une desdites sous-bandes de fréquence du signal d'entrée audio (M) ;
- la génération d'un nombre de valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l) à partir d'au moins une partie des caractéristiques de premier ordre (f_1, f_2, ..., f_f) ; et
- la compilation du jeu de caractéristiques (S) pour le signal d'entrée audio (M) en utilisant les valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l), dans lequel les caractéristiques de premier ordre (f_1, f_2, ..., f_f, f_a, f_b) sont extraites à partir de différentes périodes (t_1, t_2, ..., t_l) du signal d'entrée audio (M), et la génération d'une valeur de corrélation (ρ_1, ρ_2, ..., ρ_l, ρ) comprend la réalisation d'une corrélation en utilisant des caractéristiques de premier ordre (f_1, f_2, ..., f_f, f_a, f_b) de la même sous-bande de fréquence de différentes périodes (t_1, t_2, ..., t_l).

3. Procédé selon la revendication 2, dans lequel, pour chaque période (t_1, t_2, ..., t_l) parmi une pluralité de périodes, un vecteur de caractéristique de premier ordre (fv_1, fv_2, ..., fv_l) est extrait en fonction du temps, et la génération d'une valeur de corrélation (ρ_1, ρ_2, ..., ρ_l) comprend la réalisation d'une corrélation croisée entre certains éléments des vecteurs de caractéristique (fv_1, fv_2, ..., fv_l) sur un nombre des vecteurs de caractéristique (fv_1, fv_2, ..., fv_l).

4. Procédé selon la revendication 2, dans lequel, pour chaque période (t_1, t_2, ..., t_l) parmi une pluralité de périodes, un vecteur de caractéristique de premier ordre (fv_1, fv_2, ..., fv_l) est extrait en fonction de la fréquence, et la génération d'une valeur de corrélation (ρ_1, ρ_2, ..., ρ_l) comprend la réalisation d'une corrélation croisée entre certains éléments des vecteurs de caractéristique (fv_1, fv_2, ..., fv_l) de deux périodes (t_i, t_{i+1}) sur la fréquence.

5. Procédé selon l'une quelconque des revendications précédentes, dans lequel une caractéristique de premier ordre (f_1, f_2, ..., f_f) utilisée dans la génération d'une valeur de corrélation (ρ_1, ρ_2, ..., ρ_l) est ajustée par une moyenne de caractéristiques de premier ordre correspondantes (f_1, f_2, ..., f_f) avant la génération de la valeur de corrélation (ρ_1, ρ_2, ..., ρ_l).

6. Procédé selon l'une quelconque des revendications précédentes, dans lequel le jeu de caractéristiques (S) comprend un nombre de valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l) et une dérivée d'au moins un nombre des caractéristiques de premier ordre (f_1, f_2, ..., f_f).

7. Procédé de classification d'un signal d'entrée audio (M) en un groupe et de détermination, en fonction du jeu de caractéristiques (S) du signal d'entrée audio (M), de la probabilité que le signal d'entrée audio (M) est à l'intérieur

d'un quelconque parmi un nombre de groupes, où chaque groupe représente une classe audio particulière, dans lequel le jeu de caractéristiques (S) a été dérivé en utilisant un procédé selon l'une quelconque des revendications 1 à 6.

8. Procédé de comparaison de signaux d'entrée audio (M, M') pour déterminer un degré de similarité entre les signaux d'entrée audio (M, M'), lequel procédé comprend :

- l'obtention d'un premier jeu de caractéristiques (S) pour un premier signal d'entrée audio (M) ;
- l'obtention d'un second jeu de caractéristiques (S') pour un second signal d'entrée audio (M') ;
- le calcul d'une distance entre les premier et second jeux de caractéristiques (S, S') dans un espace de caractéristique selon une mesure de distance définie ;
- la détermination du degré de similarité entre les premier et second signaux audio (M, M') en fonction de la distance calculée,

dans lequel les premier et second jeux de caractéristiques (S) ont été obtenus en utilisant un procédé selon l'une quelconque des revendications 1 à 6.

9. Système (1) pour obtenir un jeu de caractéristiques (S) d'un signal d'entrée audio (M), comprenant :

- une unité d'identification de caractéristiques (12, 12') pour identifier un nombre de caractéristiques de premier ordre (f_1, f_2, ..., f_f) du signal d'entrée audio (M) ;
- une unité de génération de valeurs de corrélation (13, 13') pour générer un nombre de valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l) à partir d'au moins une partie des caractéristiques de premier ordre (f_1, f_2, ..., f_f) ; et
- une unité de compilation de jeu de caractéristiques (14,14') pour compiler le jeu de caractéristiques (S) pour le signal d'entrée audio (M) en utilisant les valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l), dans lequel différentes caractéristiques de premier ordre (f_1, f_2, ..., f_f, f_a, f_b) sont extraites d'une section (t_1, t_2, ..., t_l) dans un domaine donné du signal d'entrée audio (M), et la génération d'une valeur de corrélation (ρ_1, ρ_2, ..., ρ_l, ρ) comprend la réalisation d'une corrélation en utilisant des paires des caractéristiques de premier ordre (f_1, f_2, ..., f_f, f_a, f_b) de la section dans ce domaine.

10. Système (1) pour obtenir un jeu de caractéristiques (S) d'un signal d'entrée audio (M), comprenant :

- une unité d'identification de caractéristiques (12, 12') pour identifier un nombre de caractéristiques de premier ordre (f_1, f_2, ..., f_f) du signal d'entrée audio (M) ;
- une unité de génération de valeurs de corrélation (13, 13') pour générer un nombre de valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l) à partir d'au moins une partie des caractéristiques de premier ordre f_1, f_2, ..., f_f) ; et
- une unité de compilation de jeu de caractéristiques (14, 14') pour compiler le jeu de caractéristiques (S) pour le signal d'entrée audio (M) en utilisant les valeurs de corrélation (ρ_1, ρ_2, ..., ρ_l), dans lequel les caractéristiques de premier ordre (f_1, f_2, ..., f_f, f_a, f_b) sont extraites à partir de différentes périodes (t_1, t_2, ..., t_l) du signal d'entrée audio (M), et la génération d'une valeur de corrélation (ρ_1, ρ_2, ..., ρ_l, ρ) comprend la réalisation d'une corrélation en utilisant des caractéristiques de premier ordre (f_1, f_2, ..., f_f, f_a, f_b) de la même sous-bande de fréquence de différentes périodes (t_1, t_2, ..., t_l).

11. Système de classification (4) pour classifier un signal d'entrée audio (M) en un groupe, comprenant une unité de détermination de probabilité (43) pour déterminer, en fonction du jeu de caractéristiques (S) du signal d'entrée audio (M), la probabilité que le signal d'entrée audio (M) est à l'intérieur d'un quelconque parmi un nombre de groupes, où chaque groupe représente une classe audio particulière, dans lequel le jeu de caractéristiques (S) a été obtenu en utilisant un procédé de l'une quelconque des revendications 1 à 6.

12. Système de comparaison (5) pour comparer des signaux d'entrée audio (M, M') pour déterminer un degré de similarité entre les signaux d'entrée audio (M, M'), comprenant :

- une unité comparatrice (52) pour calculer une distance entre des premier et second jeux de caractéristiques (S, S') dans un espace de caractéristique selon une mesure de distance définie, et pour déterminer le degré de similarité entre les signaux d'entrée audio (M, M') en fonction de la distance calculée, dans lequel les premier et second jeux de caractéristiques (S) ont été obtenus en utilisant un procédé selon l'une quelconque des revendications 1 à 6.

13. Dispositif de traitement audio comprenant un système de classification (4) de la revendication 11 et/ou un système de comparaison (5) de la revendication 12.

14. Produit programme d'ordinateur directement chargeable dans la mémoire d'un dispositif de traitement audio programmable comprenant des portions de code logiciel pour réaliser les étapes d'un procédé d'obtention d'un jeu de caractéristiques (S) des revendications 1 à 6 ou pour réaliser les étapes d'un procédé de classification d'un signal d'entrée audio (M) des revendications 7 ou pour réaliser les étapes d'un procédé de comparaison de signaux d'entrée audio (M, M') de la revendication 8, lorsque ledit programme est exécuté sur le dispositif de traitement audio.

15. Base de données comprenant un jeu de caractéristiques (S) obtenu d'un signal d'entrée audio (M), dans laquelle le jeu de caractéristiques (S) a été obtenu en utilisant un procédé de l'une quelconque des revendications 1 à 6.

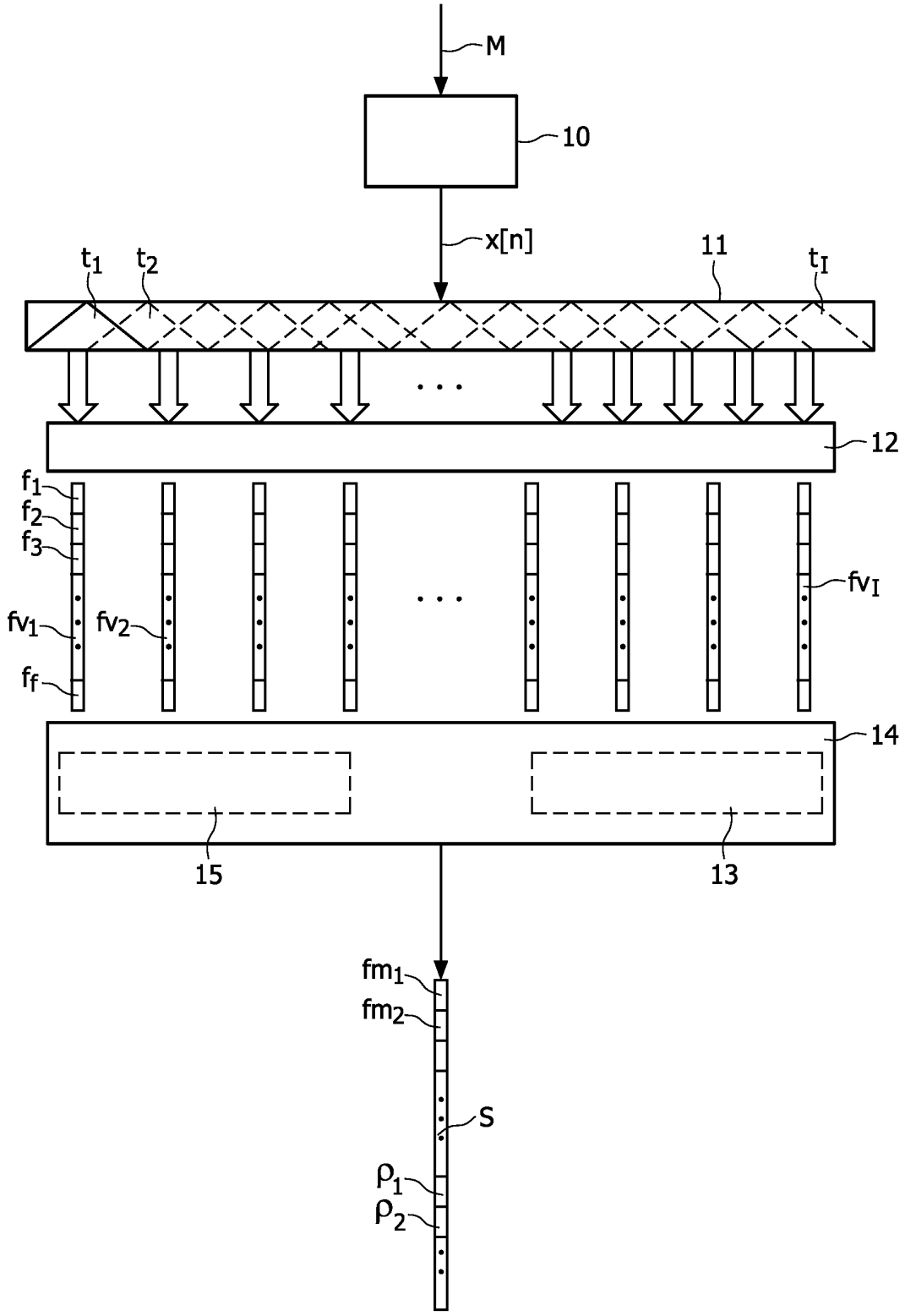

FIG. 1

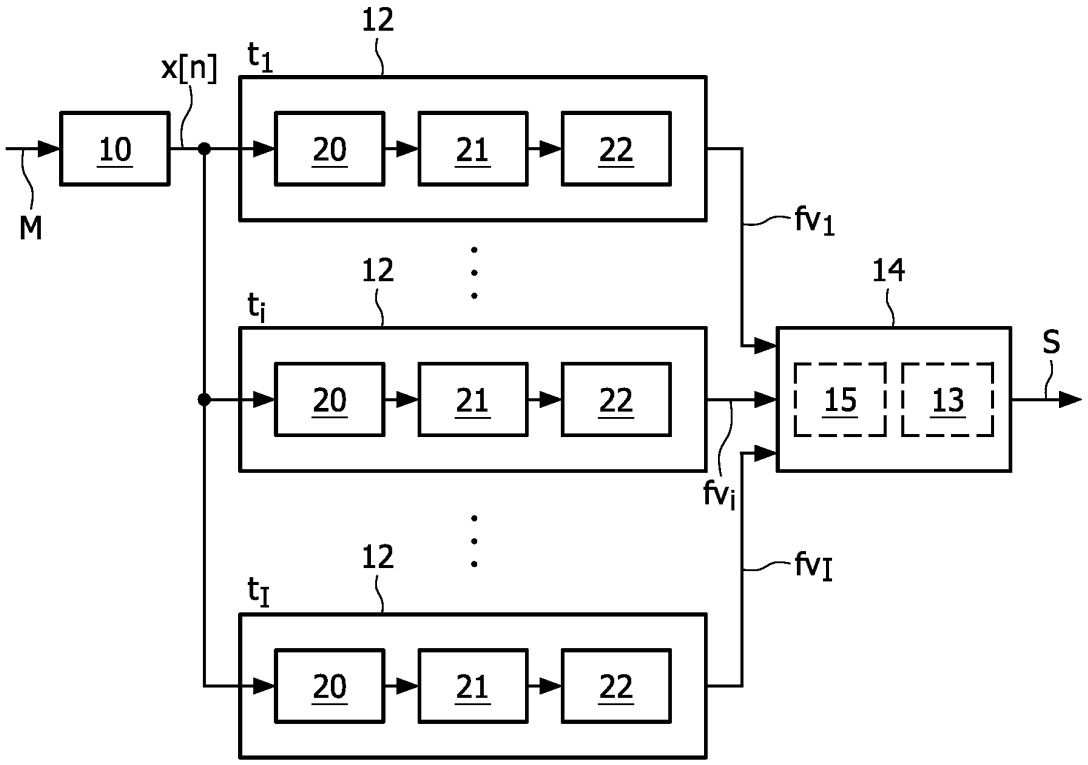
12
x[n]
t1
10
20
21
22
M
fv1
12
14
ti
20
21
22
15
13
S
fvi
fvI
12
tI
20
21
22

FIG. 2a

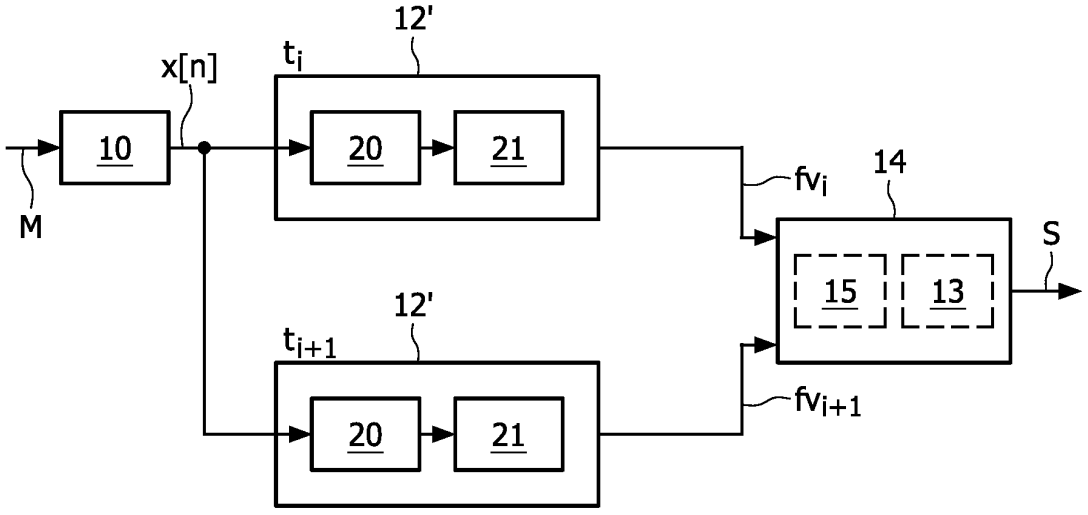
12'
x[n]
ti
10
20
21
M
fvi
14
15
13
S
12'
ti+1
20
21
fvi+1

FIG. 2b

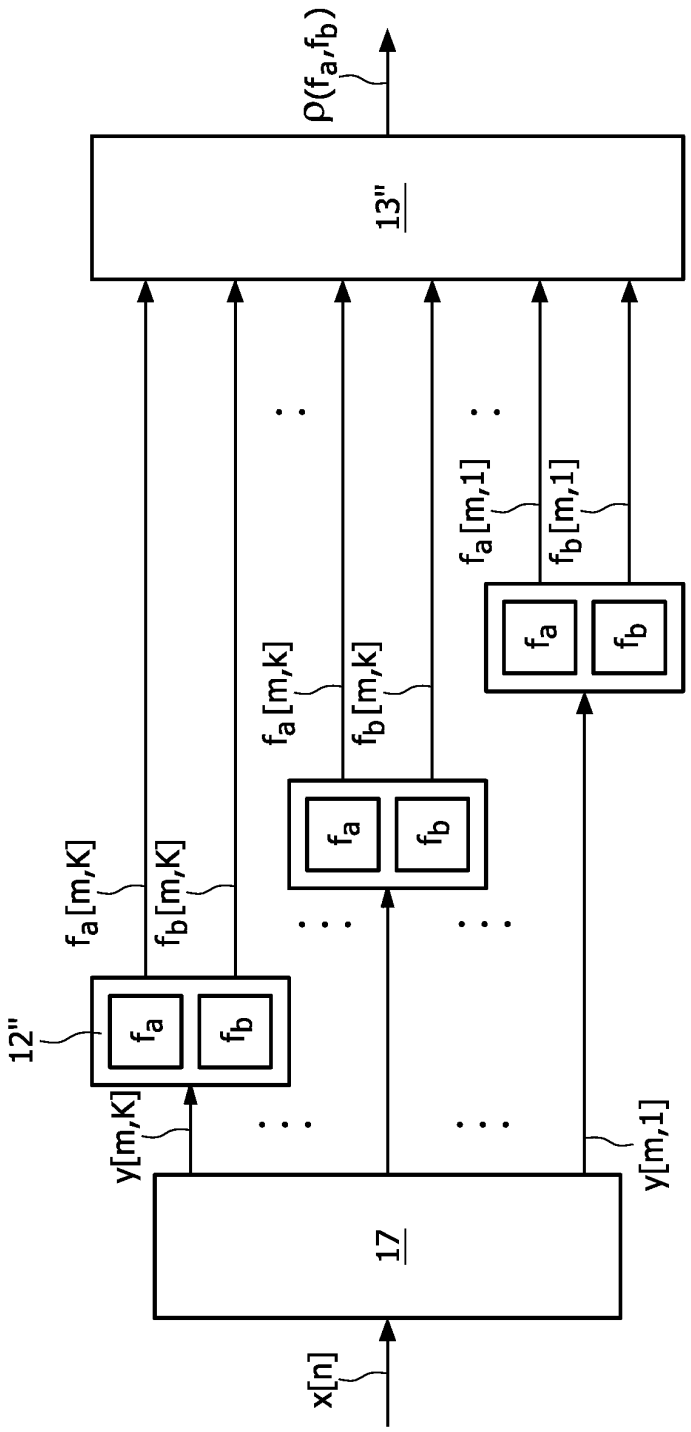

FIG. 3

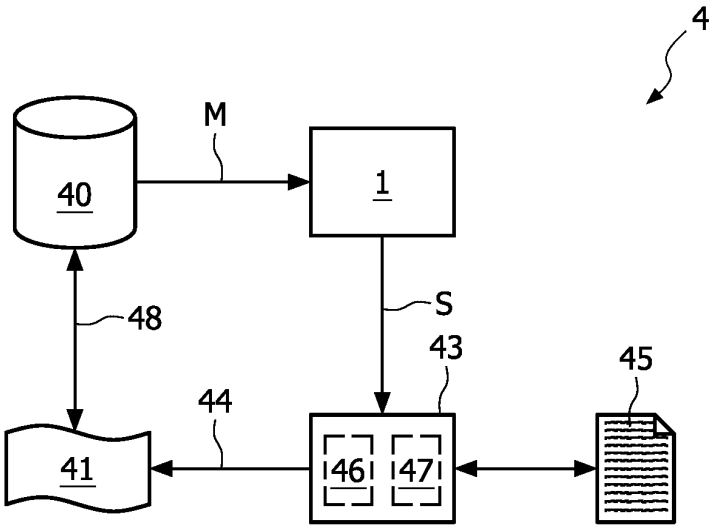

FIG. 4

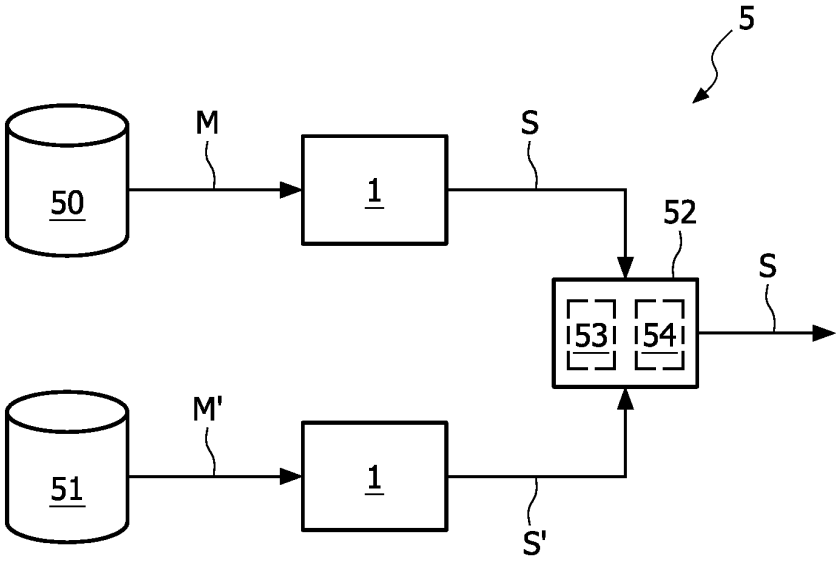

FIG. 5

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- WO 0120609 A2 **[0004]**
- US 5918223 A **[0005]**

Non-patent literature cited in the description

- **AHRENDT et al.** Decision time horizon for music genre classification using short time features. *Proceedings of EUPSICO,* 2004 **[0005]**