(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 April 2007 (26.04.2007)

PCT

(10) International Publication Number
WO 2007/046048 A1

(51) International Patent Classification:
G10H 1/00 (2006.01)

(21) International Application Number:
PCT/IB2006/053787

(22) International Filing Date: 16 October 2006 (16.10.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
05109648.5 17 October 2005 (17.10.2005) EP

(71) Applicant *(for all designated States except US)*: **KONINKLIJKE PHILIPS ELECTRONICS N.V.** [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BREEBAART, Dirk, J.** [NL/NL]; C/o Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). **MCKINNEY, Martin, F.** [US/NL]; C/o Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

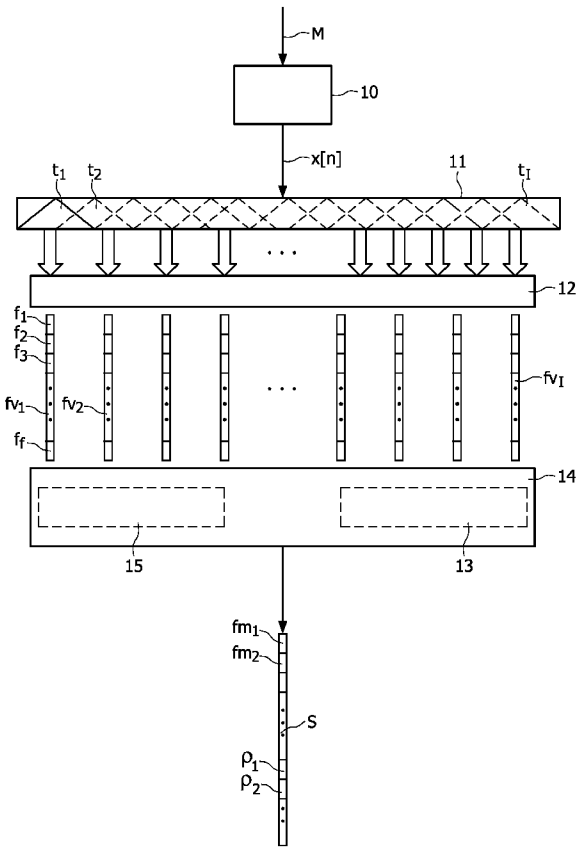
(74) Agents: **GROENENDAAL, Antonius, W., M.** et al.; Philips Intellectual Property & Standards, P.o. Box 220, NL-5600 AE Eindhoven (NL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

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



(57) Abstract: The invention describes 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_I) 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_I). The invention further describes a method of classifying an audio input signal (M) into a group, and a method of comparing audio input signals (M, M') to determine a degree of similarity between the audio input signals (M, M'). The invention also describes a system (1) for deriving a set of features (S) of an audio input signal (M), a classifying system (4) for classifying an audio input signal (M) into a group, and a comparison system (5) for comparing audio input signals (M, M') to determine a degree of similarity between the audio input signals (M, M').

WO 2007/046048 A1

Declaration under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Method of deriving a set of features for an audio input signal

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.

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.

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.

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.

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

To this end, the present invention 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.

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.

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 co-variance 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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

In the diagrams, like numbers refer to like objects throughout.

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

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 $t_1, t_2, \ldots, t_I$, so that each time frame $t_1, t_2, \ldots, t_I$ 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.

For each of these time-frames $t_1, t_2, \ldots, t_I$, a number of first-order features $f_1, f_2, \ldots, f_f$ is extracted in a feature extraction unit 12. These first-order features $f_1, f_2, \ldots, f_f$ 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 $f_1, f_2, \ldots, f_f$ for a time/frequency tile or time-frame is referred to as a first-order feature vector, so that feature vectors $fv_1, fv_2, \ldots, fv_I$ are extracted for the tiles $t_1, t_2, \ldots, t_I$.

In a correlation value generation unit 13, correlation values are generated for certain pairs of first-order features $f_1, f_2, \ldots, f_f$. The pairs of features may be taken from single feature vectors $fv_1, fv_2, \ldots, fv_I$ or from across different feature vectors $fv_1, fv_2, \ldots, fv_I$. 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.

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

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.

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:

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

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]:

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

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):

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

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:

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

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:

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

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.

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):

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

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 $MFCC_i[z]$.

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.

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.

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.

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_i, 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:

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

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.

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 $\rho(f_a,f_b)$:

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

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.

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.

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.

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'.

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'.

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.

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_I) 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_I).

2. A method according to claim 1, wherein the first-order features (f_1, f_2, …, f_f, f_a, f_b) are extracted from one or more sections (t_1, t_2, …, t_I) in a given domain of the audio input signal (M), and the generation of a correlation value (ρ_1, ρ_2, …, ρ_I, ρ) comprises performing a correlation using pairs of the first-order features (f_1, f_2, …, f_f, f_a, f_b) of corresponding sections in this domain.

3. A method according to claim 2, 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_I) of the audio input signal (M), and the generation of a correlation value (ρ_1, ρ_2, …, ρ_I, ρ) comprises performing a correlation using first-order features (f_1, f_2, …, f_f, f_a, f_b) of different time-frames (t_1, t_2, …, t_I).

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

5. A method according to claim 3, wherein, for each time-frame (t_1, t_2, …, t_I) of a plurality of time-frames, a first-order feature vector (fv_1, fv_2, …, fv_I) is extracted as a function of frequency, and generation of a correlation value (ρ_1, ρ_2, …, ρ_I) comprises

performing a cross-correlation between certain elements of the feature vectors (fv_1, fv_2, …, fv_I) of two time-frames (t_i, t_{i+1}) over frequency.

6. 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_I) 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_I).

7. A method according to any of the preceding claims, wherein the set of features (S) comprises a number of correlation values (ρ_1, ρ_2, …, ρ_I) and a derivative of at least a number of the first-order features (f_1, f_2, …, f_f).

8. 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 7.

9. 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 7.

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_1, ρ_2, ..., ρ_I) 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_I).

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 7.

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 7.

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 7 or for performing the steps of a method of classifying an audio input signal (M) according to claims 8 or for performing the steps of a method of comparing audio input signals (M, M') according to claim 9, 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 7.

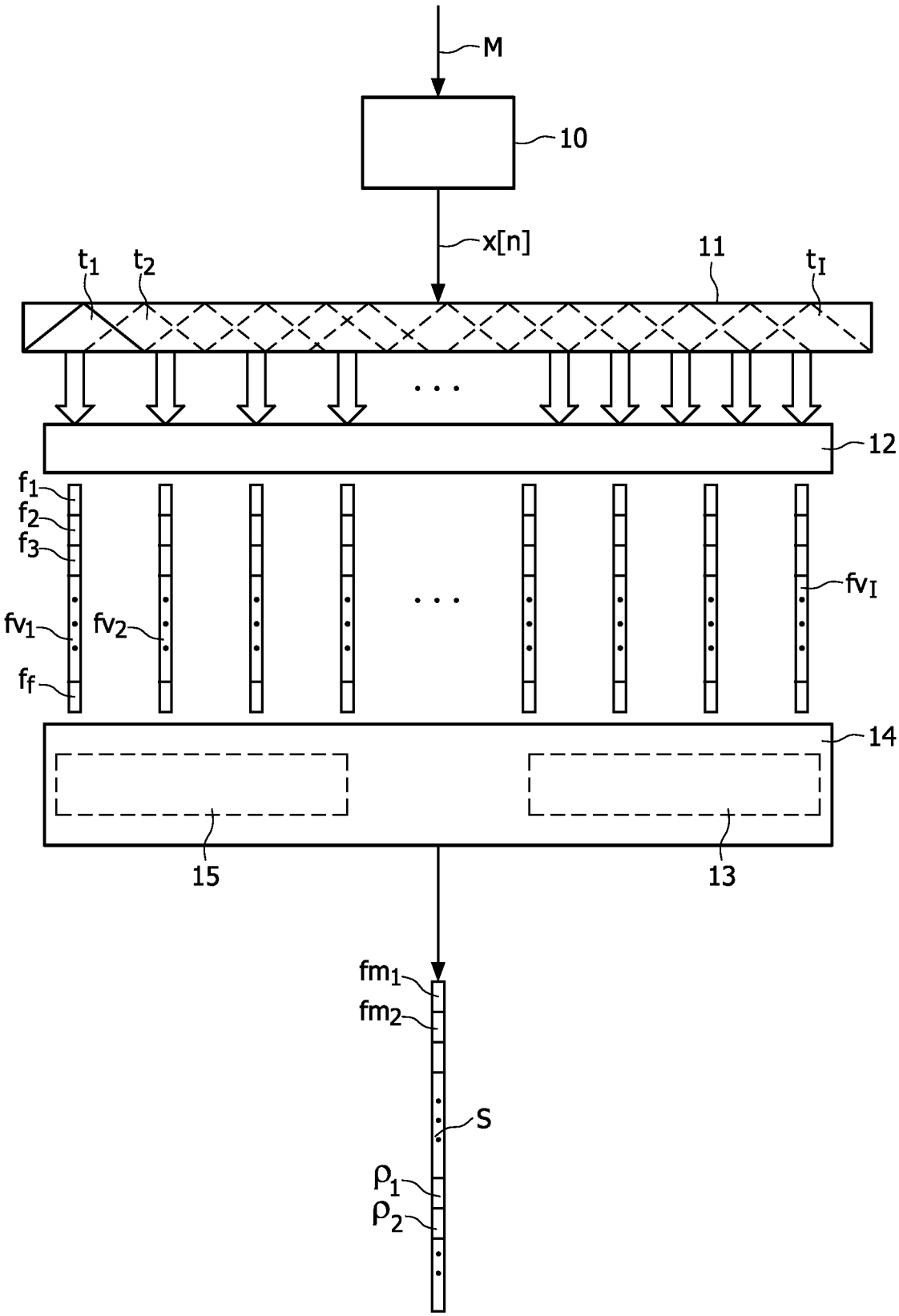

M
10
x[n]
11
t_1
t_2
t_I
12
f_1
f_2
f_3
fv_1
fv_2
f_f
fv_I
14
15
13
fm_1
fm_2
S
ρ_1
ρ_2

FIG. 1

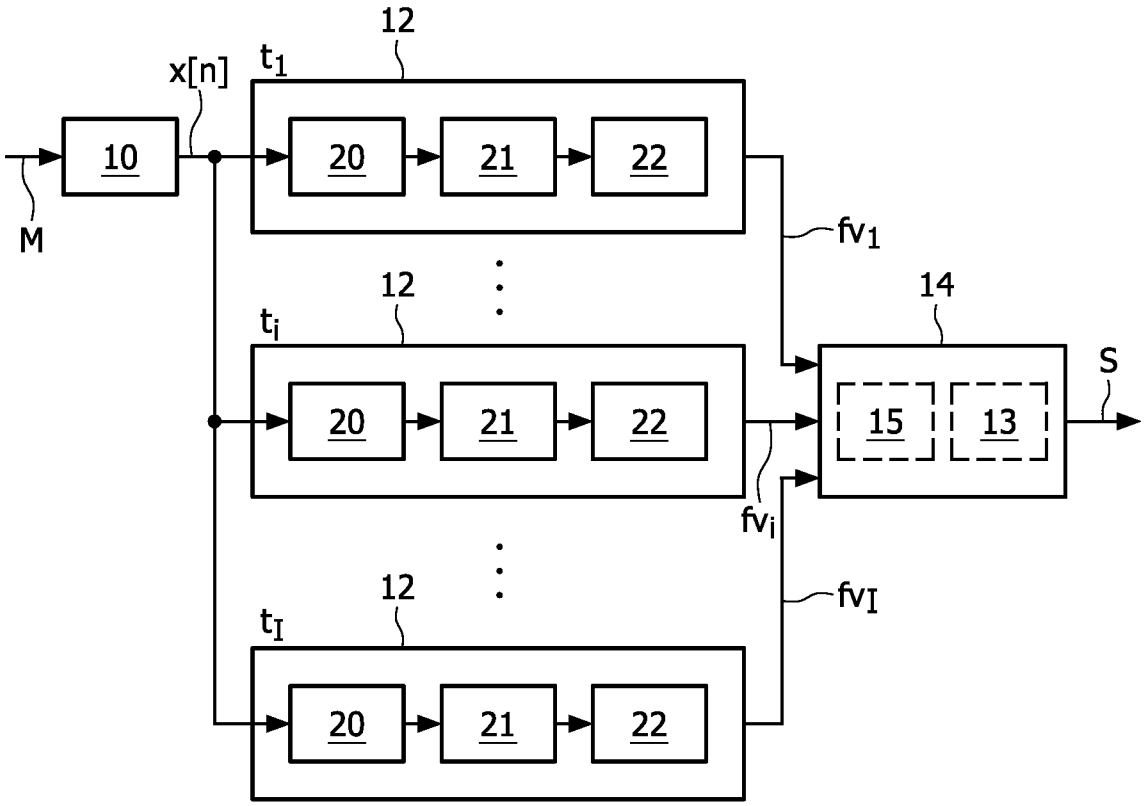

12
x[n]
t1
10
20
21
22
M
fv1
12
14
ti
20
21
22
15
13
S
fvi
fvI
12
tI
20
21
22

FIG. 2a

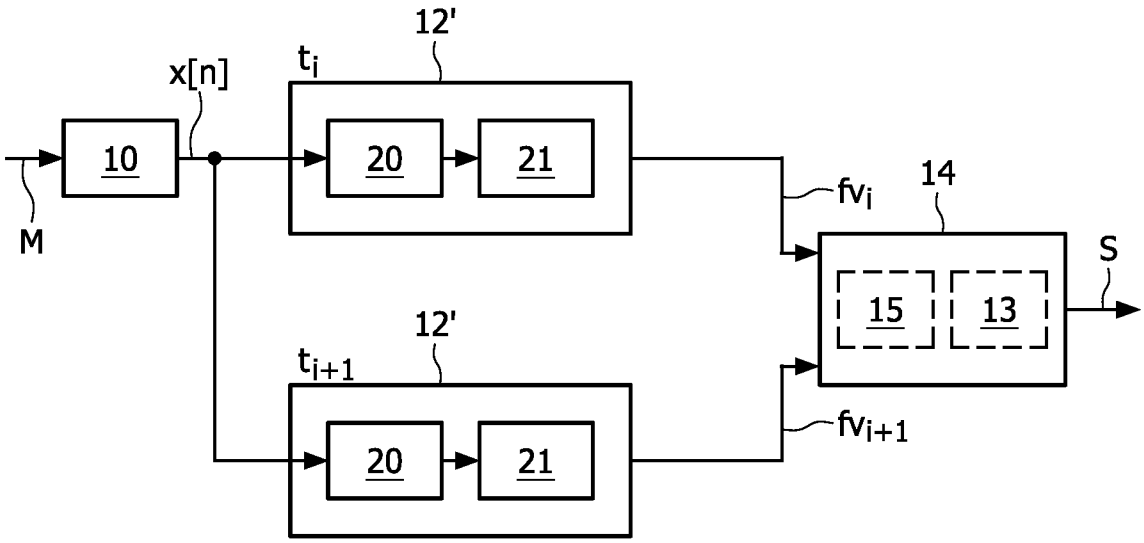

12'
x[n]
ti
10
20
21
M
fvi
14
15
13
S
12'
ti+1
20
21
fvi+1

FIG. 2b

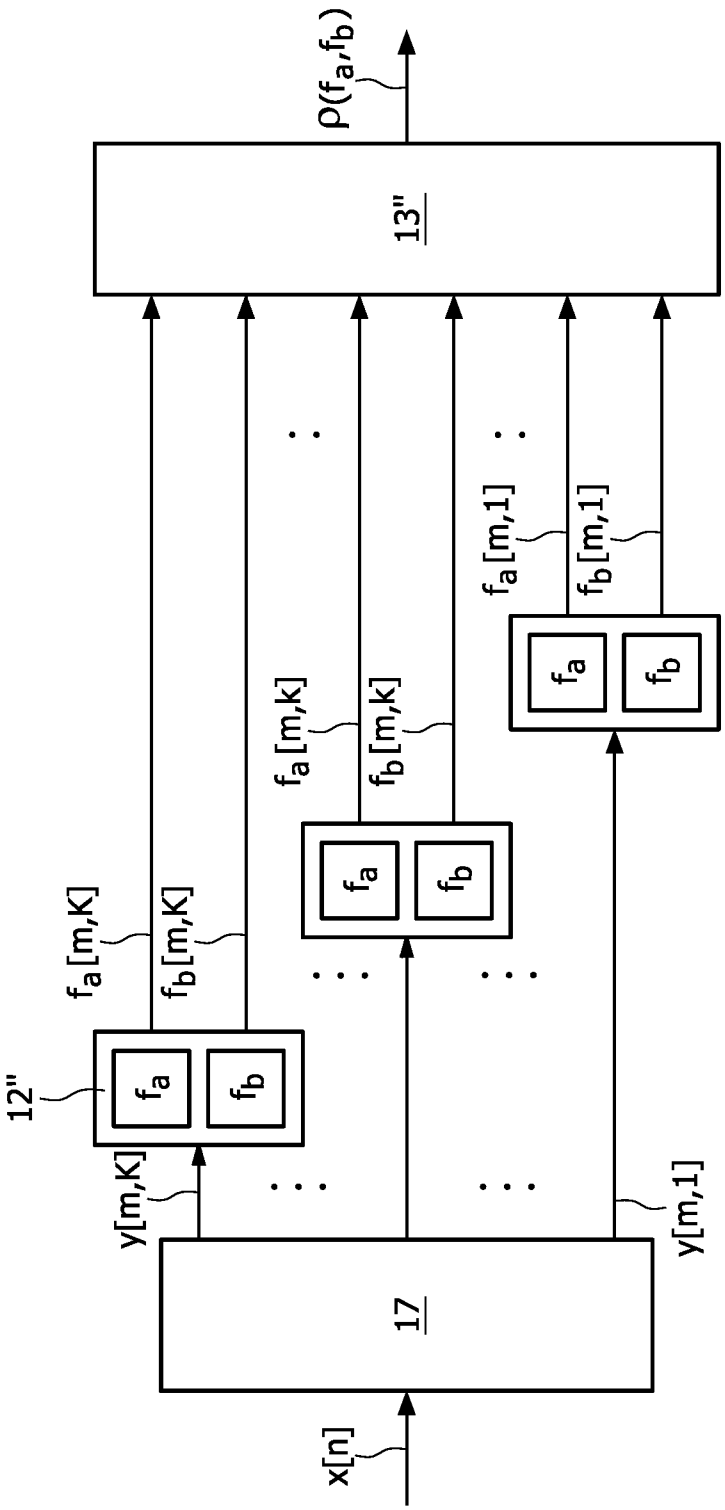


FIG. 3

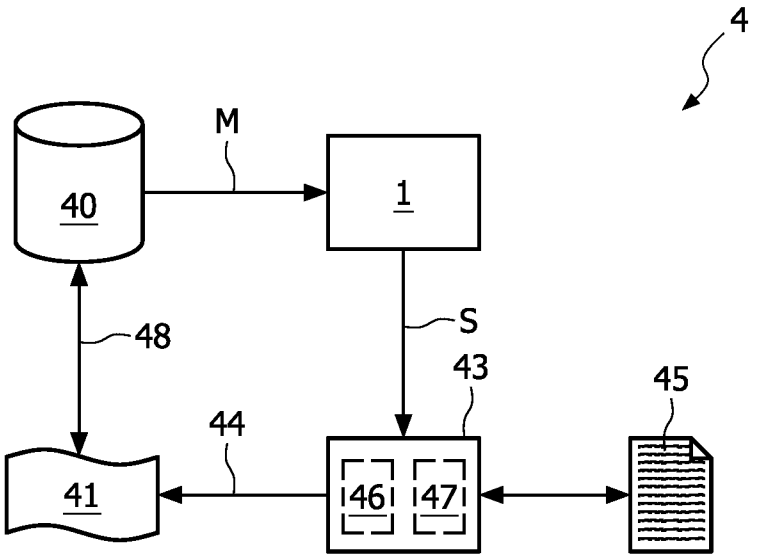


FIG. 4

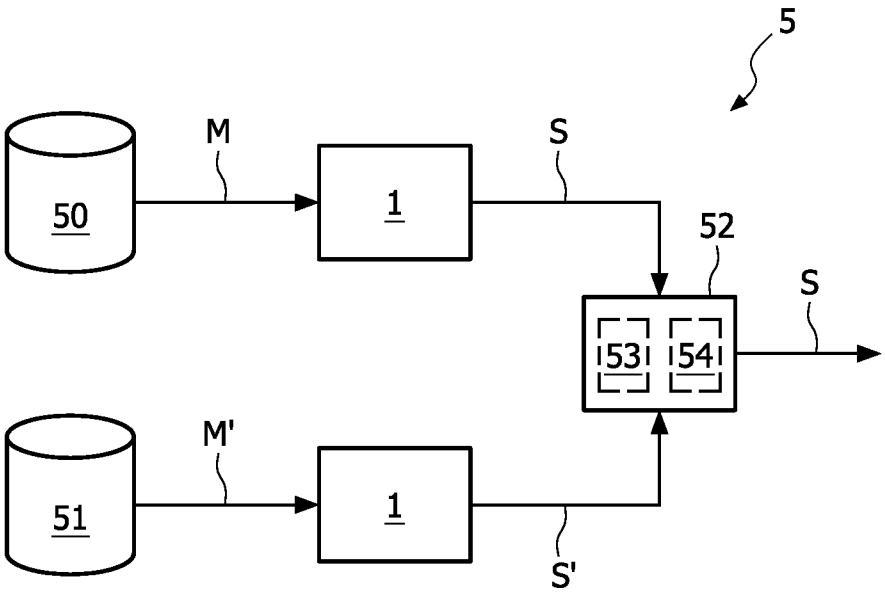


FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2006/053787

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10H1/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10H

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 918 223 A (BLUM THOMAS L [US] ET AL) 29 June 1999 (1999-06-29) abstract; claim 1; figures 1,3,6,12,17,18 column 6, lines 13-53 columns 17,22,23	1-3,5-15
X	WO 88/10540 A (MCS PARTNERS [US]) 29 December 1988 (1988-12-29) abstract; claim 1; figures 1,2,9,14 page 8, line 25 - page 11, line 14 page 17, line 25 - page 32, line 20	1-15
X	WO 98/27543 A2 (INTERVAL RESEARCH CORP [US]) 25 June 1998 (1998-06-25) page 12, line 14 - page 13, line 30; figures 1,13,16	1-6, 8-11, 13-15

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
1 March 2007	23/03/2007

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Feron, Marc

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/IB2006/053787

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5918223 A	29-06-1999	NONE	
WO 8810540 A	29-12-1988	AR 246827 A1	30-09-1994
		AU 2075288 A	19-01-1989
		BR 8807105 A	17-10-1989
		CA 1290063 C	01-10-1991
		EP 0319567 A1	14-06-1989
		ES 2012834 A6	16-04-1990
		HK 1000982 A1	15-05-1998
		IL 86701 A	25-05-1992
		JP 2500875 T	22-03-1990
		JP 3001896 B2	24-01-2000
		MX 166116 B	21-12-1992
		US 4843562 A	27-06-1989
WO 9827543 A2	25-06-1998	AU 5589398 A	15-07-1998
		US 6570991 B1	27-05-2003