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

(11) EP 1 542 206 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
15.06.2005 Bulletin 2005/24

(51) Int Cl.[7]: **G10L 11/00**

(21) Application number: **03028573.8**

(22) Date of filing: **11.12.2003**

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

(71) Applicant: **Sony International (Europe) GmbH**
10785 Berlin (DE)

(72) Inventors:
- **Lam, Yin Hay Sony Int. (Europe) GmbH**
70327 Stuttgart (DE)
- **Marasek, Krzysztof Sony Int. (Europe) GmbH**
70327 Stuttgart (DE)
- **Schaaf, Thomas Sony Int. (Europe) GmbH**
70327 Stuttgart (DE)
- **Schimanowski, Jürgen Sony Int. (Europe) GmbH**
70327 Stuttgart (DE)
- **Kemp, Thomas Sony Int. (Europe) GmbH**
70327 Stuttgart (DE)

(74) Representative: **Körber, Martin, Dipl.-Phys. et al**
Mitscherlich & Partner,
Patent- und Rechtsanwälte,
Sonnenstrasse 33
80331 München (DE)

(54) Apparatus and method for automatic classification of audio signals

(57) The present invention relates to an apparatus and a method for automatic classification of audio signals.

Such an apparatus comprises:

- signal input means (3) for supplying audio signals;
- audio signal fragmenting means (4) for partitioning audio signals supplied by the signal input means (3) into audio fragments of a predetermined length;
- feature extracting means (5) for analysing acoustic characteristics of the audio signals comprised in the audio fragments; and
- classifying means (6) for discriminating the audio fragments provided by the audio signal fragmenting means (4) into a predetermined audio class based on predetermined audio class classsifying models (71,72,73) by using acoustic characteristics of the audio signals comprised in the audio fragments, wherein a predetermined audio class classifying model (71,72,73) is provided for each audio class and each audio class represents a respective kind of audio signals comprised in the corresponding audio fragment.

It is a disadvantage that singing voice included in the audio signal frequently is misclassified as speech, particularly when the singing voice is the dominant signal component. The reason is that singing voice is more similar to speech than to music.

To solve this problem, according to the present invention an individual predetermined audio class classifying model (71,72,73) is provided for at least each audio class "speech", "music" and "singing voice".

Furthermore, the above disadvantage is overcome by the inventive method and the inventive software product.

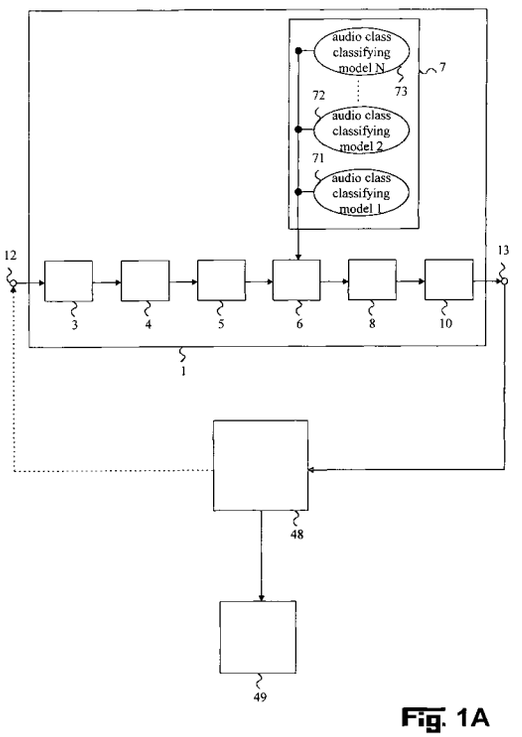


Fig. 1A

EP 1 542 206 A1

Description

[0001] The present invention relates to an apparatus and a method for automatic classification of audio signals comprising the combination of features of independent claims 1 and 12, respectively.

[0002] There is a growing amount of both audio and video data / signals available on the Internet and in a variety of storage media e.g. CDs or digital video discs. Furthermore, said audio and video data is provided by a huge number of telestations as an analogue or digital audio or video signal.

[0003] Currently, there is a desire for the possibility to search for segments of interest / important events (e.g. certain topics, persons, events or plots etc.) in said audio and/or video signal.

[0004] In this regard, only self-contained activities (events) having a certain minimum importance (important events) are accounted for.

[0005] Said self-contained activities / important events might be the different notices mentioned in a newsmagazine or different pieces of music reproduced in a radio show, for example. If the programme is a certain football match, for example, said self-contained activities / important events might be kick-off, penalty kick, throw-in etc..

[0006] In the following, said self-contained activities (events) that are included in a certain programme and meet a minimum importance are called "important events" or "contents".

[0007] The traditional audio / video tape recorder sample playback mode for browsing and skimming an analogue audio / video signal is cumbersome and inflexible. The reason for this problem is that the signal is treated as a linear block of samples. No searching functionality (except fast forward and fast reverse) is provided.

[0008] To address this problem some modem audio / video tape recorder offer the possibility to set indexes either manually or automatically each time a recording operation is started to allow automatic recognition of certain sequences of video signals. It is a disadvantage with said indexes that the indexes are not adapted to individually identify a certain sequence of audio / video signals.

[0009] On the other hand, digital audio / video discs contain digital data (digitised audio / video signals), wherein tracks or chapters are added to the digital data during the production of the digital disc. Said tracks / chapters normally allow identification of separate portions of data / the story line, only. Especially, said chapters do not allow identification of certain important events / contents (self-contained activities / events having a certain minimum importance) contained in the data. Furthermore, said tracks / chapters are not neutral since they are provided by the manufacturer of the digital disc.

[0010] An obvious solution for the problem of handling large amounts of audio / video signals would be to manually segment the signals of each programme into segments according to its important events and to provide detailed information with respect to the signal included in said segments.

[0011] Due to the immense amount of e.g. sequences comprised in the available audio / video signals, manual segmentation is extremely time-consuming and thus expensive. Therefore, this approach is not practicable to process a huge amount of audio / video signals.

[0012] To solve the above problem approaches for automatic segmentation of audio / video signals with respect to important events / contents comprised in the signals have been recently proposed.

[0013] Possible application areas for such an automatic segmentation of audio / video signals are digital libraries or the Internet, for example.

[0014] The known approaches for the segmentation process comprise fragmenting, automatic classification and automatic segmentation of the raw signals.

[0015] "Fragmenting" is performed to partition the raw signals into fragments of a suitable length for further processing. The fragments comprise a suitable amount of signals, each. Thus, the accuracy of the following classification and segmentation process is depending on the length of said fragments.

[0016] "Classification" stands for a raw discrimination of the signals comprised in the fragments with respect to the origin of the signals (e.g. speech, music, noise, silence and gender of speaker). Classification usually is performed by signal analysis techniques based on audio class classifying models. Thus, classification results in a sequence of fragments, which are discriminated with respect to the origin of the signals comprised in the fragments.

[0017] "Segmentation" stands for segmenting the raw signal into individual sequences of cohesive fragments wherein each sequence contains a content (self-contained activity of a minimum importance) included in the signals of said sequence. Segmentation can be performed based on content classifying rules.

[0018] Each content comprises all the fragments, which belong to the respective self-contained activity comprised in the raw signal (e.g. a goal, a penalty kick of a football match or different news during a news magazine or different pieces of music of a music sampler).

[0019] A segmentation apparatus 40 for automatic segmentation of audio signals according to the prior art is shown in Fig. 5.

[0020] The effect of said segmentation apparatus 40 on an audio signal 50 is shown in Fig. 6.

[0021] The segmentation apparatus 40 comprises audio signal input means 42 for supplying a raw audio signal 50 via an audio signal entry port 41.

[0022] In the present example, said raw audio signal 50 is part of a video signal stored in a suitable video format in a hard disc 48.

[0023] Alternatively, said raw audio signal might be a

real time signal (e.g. an audio signal of a conventional television channel), for example.

[0024] The audio signals 50 supplied by the audio signal input means 42 are transmitted to audio signal fragmenting means 43. The audio signal fragmenting means 43 partitions the audio signals 50 (and the respective video signals) into audio fragments 51 (and corresponding video fragments) of a predetermined length.

[0025] The audio fragments 51 generated by the audio signal fragmenting means 43 are further transmitted to classifying means 44.

[0026] The classifying means 44 discriminates the audio clips 51 into predetermined audio classes 52 based on predetermined audio class classifying models by analysing acoustic characteristics of the audio signal 50 comprised in the audio fragments 51, whereby each audio class identifies a kind of audio signals included in the respective audio fragment.

[0027] Each of the audio class classifying models allocates a combination of certain acoustic characteristics of an audio signal to a certain kind of audio signal.

[0028] Here, the acoustic characteristics for the audio class classifying model identifying the kind of audio signals "silence" are "low energy level" and "low zero cross rate" of the audio signal comprised in the respective audio clip, for example.

[0029] In the present example an audio class and a corresponding audio class classifying model for each "silence" (class 1), "speech" (class 2), "cheering/clapping" (class 3) and "music" (class 4) are provided.

[0030] Said audio class classifying models are stored in the classifying means 44.

[0031] The audio clips 52 discriminated into audio classes by the classifying means 44 are supplied to segmenting means 45.

[0032] A plurality of predetermined content classifying rules is stored in the segmenting means 45. Each content classifying rule allocates a certain sequence of audio classes of consecutive audio clips to a certain content / important event.

[0033] In the present example a content classifying rule for each a "free kick" (content 1), a "goal" (content 2), a "foul" (content 3) and "end of game" (content 4) are provided.

[0034] It is evident that the contents comprised in the audio signals are composed of a sequence of consecutive audio fragments, each. This is visualised by the segmented signal 53 of Fig. 6.

[0035] Since each audio fragment can be discriminated into an audio class each content / important event comprised in the audio signals is composed of a sequence of corresponding audio classes of consecutive audio fragments, too.

[0036] Therefore, by comparing a certain sequence of audio classes of consecutive audio fragments that belongs to the audio signals with the sequences of audio classes of consecutive audio fragments that belong to the content classifying rules the segmenting means 45 detects a rule that meets the respective sequence of audio classes.

[0037] In consequence, the content allocated to said rule is allocated to the respective sequence of consecutive audio fragments that belongs to the audio signals.

[0038] Thus, based on said content classifying rules the segmenting means 45 segments the classified audio signals provided by the discrimination means 44 into a sequence of contents 53 (self-contained activities).

[0039] In the present example, an output file generation means 46 is used to generate a video output file containing the audio signals 50, the corresponding video signals and information regarding the corresponding sequence of contents 53.

[0040] Said output file is stored via a signal output port 47 into a hard disc 48.

[0041] By using a video playback apparatus 49 the video output files stored in the hard disc 48 can be played back.

[0042] In the present example, the video playback apparatus 49 is a digital video recorder which is further capable to extract or select individual contents comprised in the video output file based on the information regarding the sequence of contents 53 comprised in the video output file.

[0043] Thus, segmentation of audio signals with respect to its contents / important events is performed by the segmentation apparatus 40 shown in Fig. 5.

[0044] A stochastic signal model frequently used with classification of audio signals / data is the HIDDEN MARKOV MODEL, which is explained in detail in the essay "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" by Lawrence R. RABINER published in the Proceedings of the IEEE, Vol. 77, No.2, February 1989.

[0045] Different approaches for audio-classification-segmentation with respect to speech, music, silence and gender are disclosed in the paper "Speech/Music/Silence and Gender Detection Algorithm" of Hadi HARB, Liming CHEN and Jean-Yves AULOGE published by the Lab. ICTT Dept. Mathematiques - Informatiques, ECOLE CENTRALE DE LYON. 36, avenue Guy de Collongue B.P. 163, 69131 ECULLY Cedex, France.

[0046] In general, the above paper is directed to discrimination of an audio channel into speech/music/silence/and noise that helps improving scene segmentation. Four approaches for audio class discrimination are proposed: A "model-based approach" where models for each audio class are created, the models being based on low level features of the audio data such as cepstrum and MFCC. A "metric-based segmentation approach" uses distances between neighbouring windows for segmentation. A "rule-based approach" comprises creation of individual rules for each class wherein the rules are based on high and low level features. Finally, a "decoder-based approach" uses the hidden Makrov model of a speech recognition system wherein the hidden

Makrov model is trained to give the class of an audio signal.

[0047] Furthermore, this paper describes in detail speech, music and silence properties to allow generation of rules describing each class according to the "rule based approach" as well as gender detection to detect the gender of a speech signal.

[0048] Zhu LIU and Yao WANG of the Polytechnic University Brooklyn, USA together with Tsuhan CHEN of the Carnegie Mellon University, Pittsburg, USA disclose "Audio Feature Extraction and Analysis for Scene Segmentation and Classification". This paper describes the use of associated audio information for video scene analysis of video data to discriminate five types of TV programs, namely commercials, basketball games, football games, news report and weather forecast.

[0049] According to this paper the audio data is divided into a plurality of clips, each clip comprising a plurality of frames.

[0050] A set of low level audio features comprising analysis of volume contour, pitch contour and frequency domain features as bandwidth are proposed for classification of the audio data contained in each clip.

[0051] Using a clustering analysis, the linear separability of different classes is examined to separate the video sequence into the above five types of TV programs.

[0052] Three layers of audio understanding are discriminated in this paper: In a "low-level acoustic characteristics layer" low level generic features such as loudness, pitch period and bandwidth of an audio signal are analysed. In an "intermediate-level acoustic signature layer" the object that produces a particular sound is determined by comparing the respective acoustic signal with signatures stored in a database. In a "high level semantic-model" some α *priori* known semantic rules about the structure of audio in different scene types (e. g. only speech in news reports and weather forecasts; speech together with noisy background in commercials) are used.

[0053] To segment the audio data, sequences of audio classes of consecutive audio clips are used. Thus, depending on the sequence of audio classes of consecutive audio clips (e.g. speech-silence-cheering/clapping-music) a suitable number of consecutive audio clips (e.g. 4) is allocated to a segment comprising one important event (e.g. "goal").

[0054] To further enhance accuracy of the above-described method, it is proposed to combine the analysis of the audio data of video data with an analysis of the visual information comprised in the video data (e.g. the respective colour patterns and shape of imaged objects).

[0055] The patent US 6,185,527 discloses a system and method for indexing an audio stream for subsequent information retrieval and for skimming, gisting and summarising the audio stream. The system and method includes use of special audio prefiltering such that only relevant speech segments that are generated by a speech recognition engine are indexed. Specific indexing features are disclosed that improve the precision and recall of an information retrieval system used after indexing for word spotting. The described method includes rendering the audio stream into intervals, with each interval including one or more segments. For each segment of an interval it is determined whether the segment exhibits one or more predetermined audio features such as a particular range of zero crossing rates, a particular range of energy, and a particular range of spectral energy concentration. The audio features are heuristically determined to represent respective audio events, including silence, music, speech, and speech on music. Also, it is determined whether a group of intervals matches a heuristically predefined meta pattern such as continuous uninterrupted speech, concluding ideas, hesitations and emphasis in speech, and so on, and the audio stream is then indexed based on the interval classification and meta pattern matching, with only relevant features being indexed to improve subsequent precision of information retrieval. Also, alternatives for longer terms generated by the speech recognition engine are indexed along with respective weights, to improve subsequent recall.

[0056] Thus, it is inter alia proposed to automatically provide a summary of an audio stream or to gain an understanding of the gist of an audio stream.

[0057] Don KIMBER and Lynn WILCOX describe algorithms, which generate indices from automatic acoustic segmentation, in the essay "Acoustic Segmentation for Audio Browsers". These algorithms use hidden Markov models to segment audio into segments corresponding to different speakers or acoustic classes. Types of proposed acoustic classes include speech, silence, laughter, non-speech sounds and garbage, wherein garbage is defined as non-speech sound not explicitly modelled by the other class models.

[0058] An implementation of the known methods is proposed by George TZANETAKIS and Perry COOK in the essay "MARSYAS: A framework for audio analysis" wherein a client-server architecture is used.

[0059] A summary and definition of acoustic characteristics frequently used to discriminate audio signals into audio classes is given in the paper "Content Analysis for Audio Classification and Segmentation" of Lie LU and Hong-Jiang ZHANG which was published in the IEEE transactions on speech and audio processing, vol. 10, No. 7 of October 2002. The definitions given in this paper apply to the definitions of the acoustic characteristics used in the present patent application.

[0060] Further approaches regarding classification and segmentation of audio signals are described in the essays "ROBUST HMM-BASED SPEECH/MUSIC SEGMENATIONG" of Iitendra AJMERA, Iain A. McCOWAN and Hervé BOURLARD, Dalle Molle Institute for Perceptual Artificial Intelligence, P.O. Box 592, CH-1920 martigny, Switzerland and "A Robust Audio Classification and Segmentation Method" of Lie LU, Hao

JIANG and HongJiang ZHANG, Microsoft research, China.

[0061] It is a disadvantage with the above described classification apparatus and methods that music containing singing voice frequently is misclassified as speech, particularly when the singing voice is the dominant signal component.

[0062] Music is a very general term that covers a huge variety of audio signals such as different instrumental sounds, singing voice with instrumental sound and also pure singing voice although in real life application, pure singing voice is not common. A robust speech/music classification should be able to distinguish speech from music regardless of the type of music, i.e. pure instrumental sound, singing voice etc..

[0063] However, as singing voice is more similar to speech rather than to music in general, state-of-the-art speech/music classification system usually fail to classify an audio signal containing music correctly when there is only a singing voice, or a dominant singing voice in the signal.

[0064] Although it is possible to add singing voice in the training material to train the audio class classifying model for the audio class "music", due to the presence of many other music signals, such as instrumental music, orchestra, pop etc. such an approach usually does not improve the classification performance. This is a crucial drawback of the prior art since music comprising singing voice is among the most common music available in real-live applications.

[0065] It is the object of the present invention to overcome the above-cited disadvantage and to provide an apparatus and a method for automatic classification of audio signals that provides an enhanced accuracy when classifying an audio signal comprising a singing voice.

[0066] The above object is solved by an apparatus for automatic classification of audio signals comprising the combination of features of independent claim 1.

[0067] Furthermore, the above object is solved by a method for automatic classification of audio signals comprising the combination of features of independent claim 12.

[0068] Further developments are set forth in the respective dependent claims.

[0069] According to a preferred embodiment of the present invention an apparatus for automatic classification of audio signals comprises signal input means for supplying audio signals, audio signal fragmenting means for partitioning audio signals supplied by the signal input means into audio fragments of a predetermined length, feature extracting means for analysing acoustic characteristics of the audio signals comprised in the audio fragments and classifying means for discriminating the audio fragments provided by the audio signal fragmenting means into a predetermined audio class based on predetermined audio class classifying models by using acoustic characteristics of the audio signals comprised in the audio fragments, wherein a predetermined audio class classifying model is provided for each audio class and each audio class represents a respective kind of audio signals comprised in the corresponding audio fragment, wherein an individual predetermined audio class classifying model is provided for at least each audio class "speech", "music" and "singing voice".

[0070] Since an individual predetermined audio class classifying model is provided for at least each audio class "speech", "music" and "singing voice", an audio class classifying model specialised in singing voice included in the raw audio signal is provided. Thus, a singing voice can be identified in a raw audio signal with high accuracy.

[0071] Advantageously, the inventive apparatus for automatic classification of audio signals further comprises a classifier database comprising the predetermined audio class classifying models, wherein the classifying means discriminates the audio fragments provided by the audio signal fragmenting means into predetermined audio classes based on the audio class classifying models stored in the classifier database.

[0072] By the provision of a classifier database comprising audio class classifying models, audio class classifying models that are specialised (trained) for a certain kind of audio signal might be used. The usage of specialised audio class classifying models significantly enhances accuracy of the classification of the audio signals.

[0073] Favourably, the classifying means further allocates audio fragments discriminated into the audio class "singing voice" to the audio class "music".

[0074] Thus, once an audio fragment has been discriminated into the audio class "singing voice" said fragment additionally or finally is allocated into the audio class "music".

[0075] Alternatively, as a "singing voice" comprised in an audio signal is very similar to the audio signal "speech", the accuracy of the inventive apparatus is significantly enhanced by further discriminating audio fragments allocated to the audio class "speech" into the audio classes "speech" and "singing voice".

[0076] Furthermore, it is profitable that the acoustic characteristics analysed in the audio signals comprised in the audio fragments by the feature extracting means include volume standard deviation and/or volume dynamic range and/or high zero crossing rate ratio and/or low short-term energy ratio and/or spectral flux and/or zero crossing rate and/or energy/loudness and/or sub-band energy rate and/or mel-cepstral frequency components and/or frequency centroid and/or bandwidth and/or line spectrum frequencies and/or roll-off.

[0077] It is preferred that the audio class classifying models are provided as hidden Markov models and/or Neuronal Networks and/or Gaussian Mixture Models and/or decision trees.

[0078] Advantageously, the audio class model for the audio class "singing voice" is trained by a training audio

signal comprising pure singing voice, only.

[0079] Thus, a suitable audio class model for the audio class "singing voice" can be achieved in a very easy and reliable way.

[0080] According to a preferred embodiment, the apparatus for automatic classification of audio signals further comprises segmentation means for segmenting classified audio signals into individual audio windows consisting of sequences of cohesive audio fragments based on predetermined content classifying rules by analysing a sequence of audio classes of cohesive audio fragments provided by the classifying means, wherein each sequence of cohesive audio fragments segmented by the segmentation means corresponds to an individual content included in the audio signal.

[0081] According to this embodiment, it is further profitable if the segmentation means allocates a predefined number of audio fragments to an audio window, determines the number of audio fragments of each audio class comprised in the audio window and allocates the majority audio class to the respective audio window.

[0082] Thus, the allocation of audio classes in the audio window is used to segment the audio signal. Complicated content classifying rules can be avoided.

[0083] It is beneficial if each audio fragment generated by the audio signal fragmenting means corresponds to a frame consisting of a predefined number N of signal samples.

[0084] Furthermore, it is preferred that the inventive apparatus for automatic classification of audio signals further comprises signal output means for generating an output file, wherein the output file contains the raw audio signal supplied to the signal input means and an information signal comprising information regarding to the audio classes and / or the audio windows and / or contents included in the raw signal.

[0085] Provision of such an information signal allows a distinct identification of the audio classes and audio windows extracted from the raw audio signals. Search engines and signal playback means can handle such an output file with ease. Therefore, a research for an audio window of a certain content comprised in the output file can be performed with ease.

[0086] Furthermore, the above object is solved by a method for automatic classification of audio signals comprising the following steps:

- partitioning audio signals into audio fragments of a predetermined length;
- analysing acoustic characteristics of the audio signals comprised in the audio fragments; and
- discriminating the audio fragments into a predetermined audio class based on predetermined audio class classifying models by using acoustic characteristics of the audio signals comprised in the audio fragments, wherein a predetermined audio class classifying model is provided for each audio class and each audio class represents a respective kind of audio signals comprised in the corresponding audio fragment;

wherein the step of discriminating the audio fragments into a predetermined audio class is performed by using an individual predetermined audio class classifying model for at least each audio class "speech", "music" and "singing voice".

[0087] Preferably, the method further comprises the step of providing a classifier database comprising the predetermined audio class classifying models, wherein the step of discriminating the audio fragments into a predetermined audio class is performed by using the audio class classifying models stored in the classifier database.

[0088] Favourably, the method further comprises the step of allocating the audio fragments discriminated into the audio class "singing voice" to the audio class "music".

[0089] Alternatively, it is beneficial if the method further comprises the step of discriminating the audio fragments allocated to the audio class "speech" into the audio classes "speech" and "singing voice".

[0090] Moreover, it is preferred that the step of analysing acoustic characteristics in the audio signals comprised in the audio fragments includes analysis of volume standard deviation and/or volume dynamic range and/or high zero crossing rate ratio and/or low short-term energy ratio and/or spectral flux and/or zero crossing rate and/or energy/loudness and/or sub-band energy rate and/or mel-cepstral frequency components and/or frequency centroid and/or bandwidth and/or line spectrum frequencies and/or roll-off.

[0091] Favourably, the audio class classifying models are provided as hidden Markov models and/or Neuronal Networks and/or Gaussian Mixture Models and/or decision trees.

[0092] Moreover, it is beneficial if the method further comprises the step of training the audio class model for the audio class "singing voice" by a training audio signal comprising pure singing voice, only.

[0093] According to a preferred embodiment of the present invention the method further comprises the steps of analysing a sequence of audio classes of cohesive audio fragments and segmenting classified audio signals into individual audio windows consisting of sequences of cohesive audio fragments based on predetermined content classifying rules by using the analyses of said sequence of audio classes of cohesive audio fragments, wherein each sequence of cohesive audio fragments corresponds to an individual content included in the audio signal.

[0094] It is further preferred that the method further comprises the steps of allocating a predefined number of audio fragments to an audio window, determining the number of audio fragments of each audio class comprised in the audio window and allocating the majority audio class to the respective audio window.

[0095] Furthermore, the above object is solved by a software product comprising a series of state elements that are adapted to be processed by a data processing means of a terminal such, that a method according to one of the claims 12 to 20 may be executed thereon.

[0096] In the following detailed description, the present invention is explained by reference to the accompanying drawings, in which like reference characters refer to like parts throughout the views, wherein:

Fig. 1A shows a block diagram of an apparatus for automatic classification of audio signals according to a preferred embodiment of the present invention;

Fig. 1B schematically shows the effect the inventive apparatus for automatic classification of audio signals has on audio signals;

Fig. 2 shows a flow diagram of a preferred embodiment of the inventive method for automatic classification of audio signals;

Fig. 3 shows a flow diagram of an alternative embodiment of the inventive method for automatic classification of audio signals;

Fig. 4 shows a flow diagram of a further embodiment of the inventive method for automatic classification of audio signals;

Fig. 5 shows a block diagram of a segmentation apparatus according to the prior art; and

Fig. 6 schematically shows the effect the segmentation apparatus according to the prior art has on audio signals.

[0097] Fig. 1A shows a block diagram of an apparatus for automatic classification of audio signals according to the one preferred embodiment of the present invention.

[0098] Fig. 1B schematically shows the effect the inventive apparatus for automatic classification of audio signals has on audio signals.

[0099] In the present embodiment, a raw audio signal 2 is supplied via an input port 12 to signal input means 3 of the inventive apparatus 1 for automatic classification of audio signals.

[0100] In the present example, the raw audio signal 2 provided to the signal input means 3 is a digital video data file which is stored on a suitable recording medium 48 (e.g. a hard disc or a digital video disc).

[0101] The digital video data file is composed of at least an audio signal and a picture signal and an information signal.

[0102] Alternatively, the raw signals 2 provided to the signal input means 3 might be real time video signals of a conventional television channel or audio signals of a radio broadcasting station.

[0103] According to this preferred embodiment, the inventive apparatus 1 for automatic classification of audio signals is included into a digital video recorder, which is not shown in the figures.

[0104] Alternatively, the apparatus for automatic classification of audio signals might be included in a different digital audio / video apparatus, such as a personal computer or workstation or might even be provided as a separate equipment (e.g. a top set box).

[0105] The signal input means 3 converts the raw signals 2 into a suitable format.

[0106] Audio signals comprised in the raw signal 2 provided to signal input means 3 via the input port 12 are read out by the signal input means 3 and transmitted to audio signal fragmenting means 4.

[0107] The audio signal fragmenting means 4 partitions said audio signals 2 into audio fragments 41, 42, 43, ..., 4N of a predetermined length.

[0108] Said audio fragments 41, 42, 43, ..., 4N preferably are the smallest unit of audio signal analysis, respectively.

[0109] In the present embodiment, one audio fragment comprises one frame of audio (video) signals and is about 10 milliseconds in length.

[0110] It is obvious for a skilled person that the audio fragments alternatively might comprise more than one frame of audio (video) signals.

[0111] Alternatively, one frame might comprise more or less than 10 milliseconds of audio signals (preferably between 4 and 20 milliseconds of audio signals, e.g. 6, 8, 12 or 14 milliseconds of audio signals).

[0112] According to an alternative embodiment more than one frame is comprised in an audio fragment. In this case it is evident for a man skilled in the art that the audio signals comprised in each audio fragment might be further divided into a plurality of frames of e.g. 512 samples. In this case it is profitable if consecutive frames are shifted by 180 samples with respect to the respective antecedent frame. This subdivision allows a precise and easy processing of the audio signals comprised in each audio fragment.

[0113] It is important to emphasise that the audio signal fragmenting means 4 do not necessarily subdivide the audio signals 2 into audio fragments 41, 42, 43, ..., 4N in a literal sense. In the present embodiment, the audio signal fragmenting means 4 defines fragments of audio signals comprising a suitable amount of audio signals within the audio signals, only.

[0114] In the present example, the audio signal fragmenting means 4 generates a meta data file defining audio fragments 41, 42, 43, ..., 4N in the audio signal 2 while the audio signal itself remains unamended.

[0115] The audio fragments defined by the audio signal fragmenting means 4 are transmitted to feature extracting means 5.

[0116] The feature extracting means 5 analyses acoustic characteristics of audio signals comprised in

the audio fragments 41, 42, 43,..., 4N.

[0117] In the present embodiment, volume standard deviation and volume dynamic range and high zero crossing rate ratio and low short-term energy ratio and spectral flux and zero crossing rate and energy/loudness and sub-band energy rate and mel-cepstral frequency components and frequency centroid and bandwidth and line spectrum frequencies and roll-off of the signals comprised in the audio fragments 41, 42, 43, ..., 4N are analysed by the feature extracting means 5.

[0118] It is obvious for a skilled person that it might be sufficient to analyse a subset of the above acoustic characteristics.

[0119] The acoustic characteristics of audio signals comprised in the audio fragments 41, 42, 43, ..., 4N are output to classifying means 6 by the feature extracting means 5.

[0120] The classifying means 6 automatically discriminates the audio fragments 41, 42, 43 provided by the audio signal fragmenting means 4 into a predetermined audio class 61, 62, 63 by using the acoustic characteristics of the audio signals comprised in the audio fragments 41, 42, 43 analysed by the feature extracting means 5. Each audio class 61, 62, 63 represents a respective kind of audio signals comprised in the corresponding audio fragment 41, 42, 43.

[0121] Discrimination is performed by the classifying means 6 based on predetermined audio class classifying models 71, 72, 73 which are stored in a classifier database 7.

[0122] A predetermined audio class classifying model 71, 72, 73 is provided in the classifier database 7 for each audio class 61, 62, 63. According to the present invention, an individual predetermined audio class classifying model 71, 72, 73 is provided for at least each audio class 61, 62, 63 "speech", "music" and "singing voice". The audio class 63 "singing voice" alternatively might be referred to as "a capella music".

[0123] Since an individual predetermined audio class classifying model 71, 72, 73 is provided for at least each audio class 61, 62, 63 "speech", "music" and "singing voice", an audio class classifying model 73 specialised in singing voice included in the raw audio signal is provided. Thus, a singing voice can be identified in a raw audio signal 2 with high accuracy.

[0124] Furthermore, by the provision of a classifier database 7 comprising audio class classifying models 71, 72, 73, audio class classifying models 71, 72, 73 that are specialised (trained) for a certain kind of audio signal 2 might be used. The usage of specialised audio class classifying models 71, 72, 73 significantly enhances accuracy of the classification of the audio signals 2.

[0125] In the present example, the predetermined audio class classifying models 71, 72, 73 are stored in the classifier database 7 as Gaussian Mixture Models (GMM).

[0126] Alternatively, the audio class classifying models might even be provided e.g. as Neuronal Networks and/or hidden Markov models and/or decision trees.

[0127] To achieve a suitable audio class model for the audio class "singing voice" in a very easy and reliable way, the audio class model 73 for the audio class 63 "singing voice" is trained by a training audio signal comprising pure singing voice, only.

[0128] In this respect, training is performed by analysing a plurality of raw signals consisting of "singing voice", only, and varying parameters of the audio class model 73 for the audio class "singing voice" till a satisfying accuracy for a correct identification of "singing voice" in the raw signal is achieved by the audio class model 73.

[0129] By using self-learning models the variation of the parameters can be automated.

[0130] The training signal might be provided by a large database (not shown in the Figures).

[0131] In case the audio class models are Gaussian Mixture Models (GMM) a linear combination of the Gaussian probability density functions (pdf) is used to model the pdf's of the signal belonging to a given audio class. Component Gaussians can have full or diagonal covariance matrices. GMM parameters such as individual Gaussians and their weight factors are tuned to suit the training signals. A GMM can approximate well any continuous pdfs. The dimensions of the component Gaussians depend on the parametrization of the underlying acoustic signal. It can be, for example, set of Mel Frequency Cepstral Coefficients (MFCC) and their derivatives computed over signal frames and windows which are sequences of frames, as well as other spectral and time characteristics such as spectral centroid, spectral flux, zero crossing rate etc..

[0132] In the present embodiment, said classifier database 7 is a convention hard disc. Alternatively, e.g. an EEPROM or a FLASH-memory might be used.

[0133] It is obvious for a skilled person that the discrimination of the audio fragments 41, 42, 43 is not necessarily performed in a literal sense, but might be performed e.g. by automatically generating a meta file (information signal) dedicated to the (raw) audio signal 2, the meta file comprising e.g. pointers to identify the audio fragments 41, 42, 43 and the corresponding audio classes 61, 62, 63 in the audio signal 2.

[0134] In the present embodiment, said pointers contained in the meta file identify both the location and the audio class 61, 62, 63 of the fragments 41, 42, 43 comprised in the audio signals 2.

[0135] Favourably, the classifying means 6 allocates audio fragments 41, 42, 43 discriminated into the audio class 63 "singing voice" to the audio class 62 "music".

[0136] Thus, once an audio fragment 41, 42, 43 has been discriminated into the audio class 63 "singing voice" it is additionally allocated / re-categorised into the audio class 62 "music" as it is shown in Fig.2.

[0137] Alternatively, as it is shown in Fig. 3, the audio fragments allocated to the audio class 61 "speech" are further discriminated into the audio classes 61 "speech"

and 63 "singing voice" to increase the accuracy of the inventive apparatus 1. The reasons is that a "singing voice" comprised in an audio signal 2 is very similar to the audio signal "speech". Thus, a signal that seems to contain "speech" has to be further examined to detect whether the pretended "speech" is real "speech" or indeed a "singing voice".

[0138] After discrimination into audio classes 61, 62, 63 by the classifying means 6, the classified audio fragments 9 are transmitted to a segmentation means 8.

[0139] Said segmentation means 8 segments the classified audio signals 9 into individual audio windows 81, 82, 83 consisting of sequences of cohesive audio fragments 41, 42, 43 based on predetermined content classifying rules by analysing a sequence of audio classes 61, 62, 63 of cohesive audio fragments 41, 42, 43 provided by the classifying means 6. Each sequence of cohesive audio fragments 41, 42, 43 segmented by the segmentation means 8 corresponds to an individual content included in the audio signal 2.

[0140] Contents are self-contained activities comprised in the audio signals of a certain programme that meet a certain minimum importance.

[0141] The length of time of the contents comprised in the audio signals of a programme usually differs. Thus, each content comprises a certain individual number of cohesive audio fragments 41, 42, 43.

[0142] If the programme is news, for example, the contents are the different notices mentioned in the news. If the programme is football, for example, said contents are kick-off, penalty kick, throw-in, goal, etc. If the programme is a music sampler, said contents are the individual pieces of music, for example.

[0143] As said before, the contents comprised in the audio signal are composed of a sequence of consecutive audio fragments 41, 42, 43, each. Since each audio fragment 41, 42, 43 is discriminated into an audio class 61, 62, 63 each content is composed of a sequence of corresponding audio classes 61, 62, 63 of consecutive the audio fragments 41, 42, 43, too.

[0144] Therefore, by comparing the sequences of audio classes 61, 62, 63 of consecutive audio fragments 41, 42, 43 which belong to the contents of the respective audio signal with the sequences of audio classes 61, 62, 63 of consecutive audio fragments 41, 42, 43 which belong to the content classifying rules it is possible to find content classifying rules which are adapted to identify the respective content.

[0145] The function of the content classifying rules will become more apparent by the following example:

[0146] The sequence of audio classes of cohesive audio fragments 41, 42, 43 for the content classifying rule identifying the content "goal" might be "speech" 61, "singing voice" 63 and "music" 62.

[0147] Thus, in case the sequence of audio classes of cohesive audio fragments 41, 42, 43 "speech" 61, "singing voice" 63 and "music" 62 is to be segmented by the segmentation means 8, the content 1 "goal" will be allocated to said sequence of audio fragments 41, 42, 43 (window 81 of Fig. 1B).

[0148] According to an alternative embodiment, the segmentation means 8 allocates a predefined number of audio fragments 41, 42, 43, ..., 4N to an audio window 81, 82, 83, ..., 8N determines the number of audio fragments 41, 42, 43, ..., 4N of each audio class 61, 62, 63 comprised in the audio window and allocates the majority audio class to the respective audio window.

[0149] Thus, the allocation of audio classes 61, 62, 63 in the audio window 81, 82, 83 is used to segment the audio signal. In consequence, complicated content classifying rule can be avoided.

[0150] This will become more apparent by the following example:

[0151] In case an audio window 8N comprises a sequence of audio fragments 4N-3, 4N-2, 4N-1, 4N of the audio classes 62, 63, 61, 61 the audio class 61 automatically will be determined as being the majority audio class by the segmentation means 8. Thus, the audio class 61 automatically will be allocated by the segmentation means 8 to the respective audio window 8N as content k (see Fig. 1B).

[0152] In case no majority audio class can be identified due to a tie-situation, additional rules may be provided to rank preferred audio classes.

[0153] Furthermore, the inventive apparatus 1 for automatic classification of audio signals 2 comprises signal output means 10.

[0154] Said signal output means 10 automatically generates an output file 11 containing the raw signal 2 supplied by the signal input means 3 and an information signal (meta file) comprising information regarding to the fragments, audio classes, windows and contents included in the raw signal 2.

[0155] Search engines and signal playback means can handle a correspondingly processed signal 11 with ease. Therefore, a research for the audio classes and contents comprised in the output file 11 is facilitated.

[0156] The output file 11 is output by the signal output means 10 via an output port 13.

[0157] The signal output via said output port 13 might be stored into a suitable recording medium 48 which might be a conventional hard disc or optical disc, for example.

[0158] In the following, preferred methods for automatic classification of audio signals are explained in detail by reference to Figs. 2, 3 and 4:

[0159] In a first step S1 (not shown in the Figures) audio signals 2 automatically are partitioned into audio fragments 41, 42, 43 of a predetermined length. In the present preferred embodiment, the length of the fragments 41, 42, 43 is one frame, each.

[0160] In the following step S2 acoustic characteristics of the audio signals comprised in the audio fragments 41, 42, 43 are analysed.

[0161] Said acoustic characteristics include volume standard deviation and/or volume dynamic range and/

or high zero crossing rate ratio and/or low short-term energy ratio and/or spectral flux and/or zero crossing rate and/or energy/loudness and/or sub-band energy rate and/or mel-cepstral frequency components and/or frequency centroid and/or bandwidth and/or line spectrum frequencies and/or roll-off.

[0162] A classifier database 7 containing predetermined audio class classifying models 71, 72, 73 is provided in the following method step S4. A predetermined audio class classifying model 71, 72, 73 is provided for each audio class 61, 62, 63 and each audio class 61, 62, 63 represents a respective kind of audio signals comprised in the corresponding audio fragment 41, 42, 43 of the raw audio signal 2.

[0163] According to the present invention, an individual predetermined audio class classifying model 71, 72, 73 for at least each audio class 61, 62, 63 "speech", "music" and "singing voice" is provided.

[0164] In the present embodiment, the audio class classifying models 71, 72, 73 are provided as hidden Markov models. Alternatively Neuronal Networks or Gaussian Mixture Models or decision trees might be used.

[0165] To achieve a suitable audio class model 73 for the audio class 63 "singing voice", in a preparatory step S7 the audio class model 73 for the audio class 63 "singing voice" has been trained by a training audio signal consisting of pure singing voice, only.

[0166] By using acoustic characteristics of the audio signals comprised in the audio fragments 41, 42, 43 analysed in the preceding method step S2 the audio fragments 41, 42, 43 are discriminated into predetermined audio classes 61, 62, 63 based on predetermined audio class classifying models 71, 72, 73 stored in the classifier database 7 that has been provided in method step S4.

[0167] In case an audio fragment 41, 42, 43 is discriminated in step S3 into the audio class 63 "singing voice", said audio fragment 41, 42, 43 is further allocated to the audio class 62 "music" in the following step S5 as it is shown in Fig. 2.

[0168] Alternatively, in case an audio fragment 41, 42, 43 is discriminated in step S3 into the audio class 61 "speech", said audio fragment 41, 42, 43 is further discriminated into the audio classes 61 "speech" and 63 "singing voice" in the following step S6 as it is shown in Fig. 3.

[0169] A sequence of audio classes 61, 62, 63 of cohesive audio fragments 41, 42, 43 is analysed in the following method step S8.

[0170] In the following step S9, classified audio signals 9 provided by method steps S3, S5 and S6 are segmented into individual audio windows 81, 82, 83 consisting of sequences of cohesive audio fragments 41, 42, 43 based on predetermined content classifying rules. Said segmentation is performed by using the sequence of audio classes 61, 62, 63 of cohesive audio fragments 41, 42, 43 analysed in method step S8. Each sequence of cohesive audio fragments 41, 42, 43 corresponds to an individual content included in the audio signal.

[0171] According to a preferred embodiment said segmentation might be performed as follows:

[0172] In a first sub-step S10a predefined number of audio fragments 41, 42, 43 is allocated to an audio window 81, 82, 83.

[0173] In a second sub-step S11 the number of audio fragments 41, 42, 43 of each audio class 61, 62, 63 comprised in the audio window 81, 82, 83 is determined.

[0174] In a third sub-step S12 the majority audio class 61, 62, 63 is allocated to the respective audio window 81, 82, 83.

[0175] Thus, a segmented audio signal is provided (see Fig. 4).

[0176] According to a further alternative embodiment of the inventive method (not shown in the figures), a Viterbi algorithm is used for step 9 of segmenting classified audio signals.

[0177] In this case, the steps S3, S5 and S6 of discriminating audio fragments and the steps S8 and S9 of analysing a sequence of audio classes and segmenting classified audio signals are combined into a joint optimisation of the best state sequence of audio fragments that explains the observation with the highest possible likelihood, given the model.

[0178] To enhance clarity of Fig. 1, supplementary means as power supply, buffer memories etc. are not shown.

[0179] In the one embodiment of the present invention shown in Fig. 1, separate microprocessors are used for the signal input means 3, the audio signal fragmenting means 4, the feature extracting means 5, the classifying means 6, the segmentation means 8 and the signal output means 10.

[0180] Alternatively, one single microcomputer might be used to incorporate the signal input means, the audio signal fragmenting means, the feature extracting means, the classifying means, the segmentation means and the signal output means.

[0181] Further alternatively, the signal input means and the signal output means might be incorporated in one common microcomputer and the audio signal fragmenting means, the feature extracting means, the classifying means and the segmentation means might be incorporated in another common microcomputer.

[0182] Preferably, the inventive apparatus for automatic classification of audio signals might be integrated into a digital audio / video recorder or top set box or realised by use of a conventional personal computer or workstation.

[0183] According to a further embodiment of the present invention (which is not shown in the figures), the above object is solved by a software product comprising a series of state elements that are adapted to be processed by a data processing means of a terminal such, that a method according to one of the claims 12 to 20

may be executed thereon.

[0184] Said terminal might be a personal computer or video recording/reproducing apparatus, for example.

[0185] In summary, the inventive apparatus and method for automatic classification of audio signals uses an easy and reliable way for classification of audio signals comprising a singing voice.

[0186] Since an individual predetermined audio class classifying model is provided for at least each audio class "speech", "music" and "singing voice", an audio class classifying model specialised in singing voice included in the raw audio signal is provided. Thus, singing voice can be identified in a raw audio signal with high accuracy.

Claims

1. Apparatus (1) for automatic classification of audio signals comprising:

   - signal input means (3) for supplying audio signals (2);
   - audio signal fragmenting means (4) for partitioning audio signals (2) supplied by the signal input means (3) into audio fragments (41, 42, 43) of a predetermined length;
   - feature extracting means (5) for analysing acoustic characteristics of the audio signals comprised in the audio fragments (41, 42, 43); and
   - classifying means (6) for discriminating the audio fragments (41, 42, 43) provided by the audio signal fragmenting means (4) into a predetermined audio class (61, 62, 63) based on predetermined audio class classifying models (71, 72, 73) by using acoustic characteristics of the audio signals comprised in the audio fragments (41, 42, 43), wherein a predetermined audio class classifying model (71, 72, 73) is provided for each audio class (61, 62, 63) and each audio class (61, 62, 63) represents a respective kind of audio signals comprised in the corresponding audio fragment (41, 42, 43);

   characterised in that
   an individual predetermined audio class classifying model (71, 72, 73) is provided for at least each audio class (61, 62, 63) "speech", "music" and "singing voice".

2. Apparatus for automatic classification of audio signals according to claim 1,
   characterised in that
   the apparatus (1) for automatic classification of audio signals further comprises:

   - a classifier database (7) comprising the predetermined audio class classifying models (71, 72, 73);

   wherein the classifying means (6) discriminates the audio fragments (41, 42, 43) provided by the audio signal fragmenting means (4) into predetermined audio classes (61, 62, 63) based on the audio class classifying models (71, 72, 73) stored in the classifier database (7).

3. Apparatus for automatic classification of audio signals according to claim 1 or 2,
   characterised in that
   the classifying means (6) further allocates audio fragments (41, 42, 43) discriminated into the audio class (63) "singing voice" to the audio class (62) "music".

4. Apparatus for automatic classification of audio signals according to claim 1 or 2,
   characterised in that
   the classifying means (6) further discriminates audio fragments (41, 42, 43) allocated to the audio class (61) "speech" into the audio classes (61, 63) "speech" and "singing voice".

5. Apparatus for automatic classification of audio signals according to one of the preceding claims,
   characterised in that
   the acoustic characteristics analysed in the audio signals comprised in the audio fragments (41, 42, 43) by the feature extracting means (5) include volume standard deviation and/or volume dynamic range and/or high zero crossing rate ratio and/or low short-term energy ratio and/or spectral flux and/or zero crossing rate and/or energy/loudness and/or sub-band energy rate and/or mel-cepstral frequency components and/or frequency centroid and/or bandwidth and/or line spectrum frequencies and/or roll-off.

6. Apparatus for automatic classification of audio signals according to one of the preceding claims,
   characterised in that
   the audio class classifying models (71, 72, 73) are provided as hidden Markov models and/or Neuronal Networks and/or Gaussian Mixture Models and/or decision trees.

7. Apparatus for automatic classification of audio signals according to one of the preceding claims,
   characterised in that
   the audio class model (73) for the audio class (63) "singing voice" is trained by a training audio signal comprising pure singing voice, only.

8. Apparatus for automatic classification of audio signals according to one of the preceding claims,

characterised in that the apparatus (1) for automatic classification of audio signals further comprises:

- segmentation means (8) for segmenting classified audio signals (9) into individual audio windows (81, 82, 83) consisting of sequences of cohesive audio fragments (41, 42, 43) based on predetermined content classifying rules by analysing a sequence of audio classes (61, 62, 63) of cohesive audio fragments (41, 42, 43) provided by the classifying means (6), wherein each sequence of cohesive audio fragments (41, 42, 43) segmented by the segmentation means (8) corresponds to an individual content included in the audio signal.

9. Apparatus for automatic classification of audio signals according to claim 8,
characterised in that the segmentation means (8)

- allocates a predefined number of audio fragments (41, 42, 43) to an audio window (81, 82, 83),
- determines the number of audio fragments (41, 42, 43) of each audio class (61, 62, 63) comprised in the audio window (81, 82, 83) and
- allocates the majority audio class (61, 62, 63) to the respective audio window (81, 82, 83).

10. Apparatus for automatic classification of audio signals according to one of the preceding claims,
characterised in that
each audio fragment (41, 42, 43) generated by the audio signal fragmenting means (4) corresponds to a frame consisting of a predefined number N of signal samples.

11. Apparatus for automatic classification of audio signals according to one of the preceding claims,
characterised in that the apparatus (1) for automatic classification of audio signals further comprises:

- signal output means (10) for generating an output file (11);

wherein the output file contains the raw audio signal (2) supplied to the signal input means (3) and an information signal comprising information regarding to the audio classes (61, 62, 63) and / or the audio windows (81, 82, 83) and / or contents included in the raw signal (2).

12. Method for automatic classification of audio signals comprising the following steps:

- (S1) partitioning audio signals (2) into audio fragments (41, 42, 43) of a predetermined length;
- (S2) analysing acoustic characteristics of the audio signals comprised in the audio fragments (41, 42, 43); and
- (S3) discriminating the audio fragments (41, 42, 43) into a predetermined audio class (61, 62, 63) based on predetermined audio class classifying models (71, 72, 73) by using acoustic characteristics of the audio signals comprised in the audio fragments (41, 42, 43), wherein a predetermined audio class classifying model (71, 72, 73) is provided for each audio class (61, 62, 63) and each audio class (61, 62, 63) represents a respective kind of audio signals comprised in the corresponding audio fragment (41, 42, 43);

characterised in that
the step (S3) of discriminating the audio fragments (41, 42, 43) into a predetermined audio class (61, 62, 63) is performed by using an individual predetermined audio class classifying model (71, 72, 73) for at least each audio class (61, 62, 63) "speech", "music" and "singing voice".

13. Method for automatic classification of audio signals according to claim 12,
characterised in that the method further comprises the step of

- (S4) providing a classifier database (7) comprising the predetermined audio class classifying models (71, 72, 73);

wherein the step (S3) of discriminating the audio fragments (71, 72, 73) into a predetermined audio class (61, 62, 63) is performed by using the audio class classifying models (71, 72, 73) stored in the classifier database (7).

14. Method for automatic classification of audio signals according to claim 12 or 13,
characterised in that the method further comprises the step of

- (S5) allocating the audio fragments (41, 42, 43) discriminated into the audio class (63) "singing voice" to the audio class (62) "music".

15. Method for automatic classification of audio signals according to one of the claims 12 to 13,
characterised in that the method further comprises the step of

- (S6) discriminating the audio fragments (41, 42, 43) allocated to the audio class (61) "speech" into the audio classes (61, 63) "speech" and

"singing voice".

16. Method for automatic classification of audio signals according to one of the claims 12 to 15, **characterised in that** the step (S2) of analysing acoustic characteristics in the audio signals comprised in the audio fragments (41, 42, 43) includes analysis of volume standard deviation and/or volume dynamic range and/or high zero crossing rate ratio and/or low short-term energy ratio and/or spectral flux and/or zero crossing rate and/or energy/loudness and/or sub-band energy rate and/or mel-cepstral frequency components and/or frequency centroid and/or bandwidth and/or line spectrum frequencies and/or roll-off.

17. Method for automatic classification of audio signals according to one of the claims 12 to 16, **characterised in that** the audio class classifying models (71, 72, 73) are provided as hidden Markov models and/or Neuronal Networks and/or Gaussian Mixture Models and/or decision trees.

18. Method for automatic classification of audio signals according to one of the claims 12 to 17, **characterised in that** the method further comprises the step of

- (S7) training the audio class model (73) for the audio class (63) "singing voice" by a training audio signal comprising pure singing voice, only.

19. Method for automatic classification of audio signals according to one of the claims 12 to 18, **characterised in that** the method further comprises the steps of

- (S8) analysing a sequence of audio classes (61, 62, 63) of cohesive audio fragments (41, 42, 43); and
- (S9) segmenting classified audio signals (9) into individual audio windows (81, 82, 83) consisting of sequences of cohesive audio fragments (41, 42, 43) based on predetermined content classifying rules by using the analyses of said sequence of audio classes (61, 62, 63) of cohesive audio fragments (41, 42, 43), wherein each sequence of cohesive audio fragments (41, 42, 43) corresponds to an individual content included in the audio signal.

20. Method for automatic classification of audio signals according to claim 19, **characterised in that** the method further comprises the steps of

- (S10) allocating a predefined number of audio fragments (41, 42, 43) to an audio window (81, 82, 83);
- (S11) determining the number of audio fragments (41, 42, 43) of each audio class (61, 62, 63) comprised in the audio window (81, 82, 83); and
- (S12) allocating the majority audio class (61, 62, 63) to the respective audio window (81, 82, 83).

21. Software product comprising a series of state elements that are adapted to be processed by a data processing means of a terminal such, that a method according to one of the claims 12 to 20 may be executed thereon.

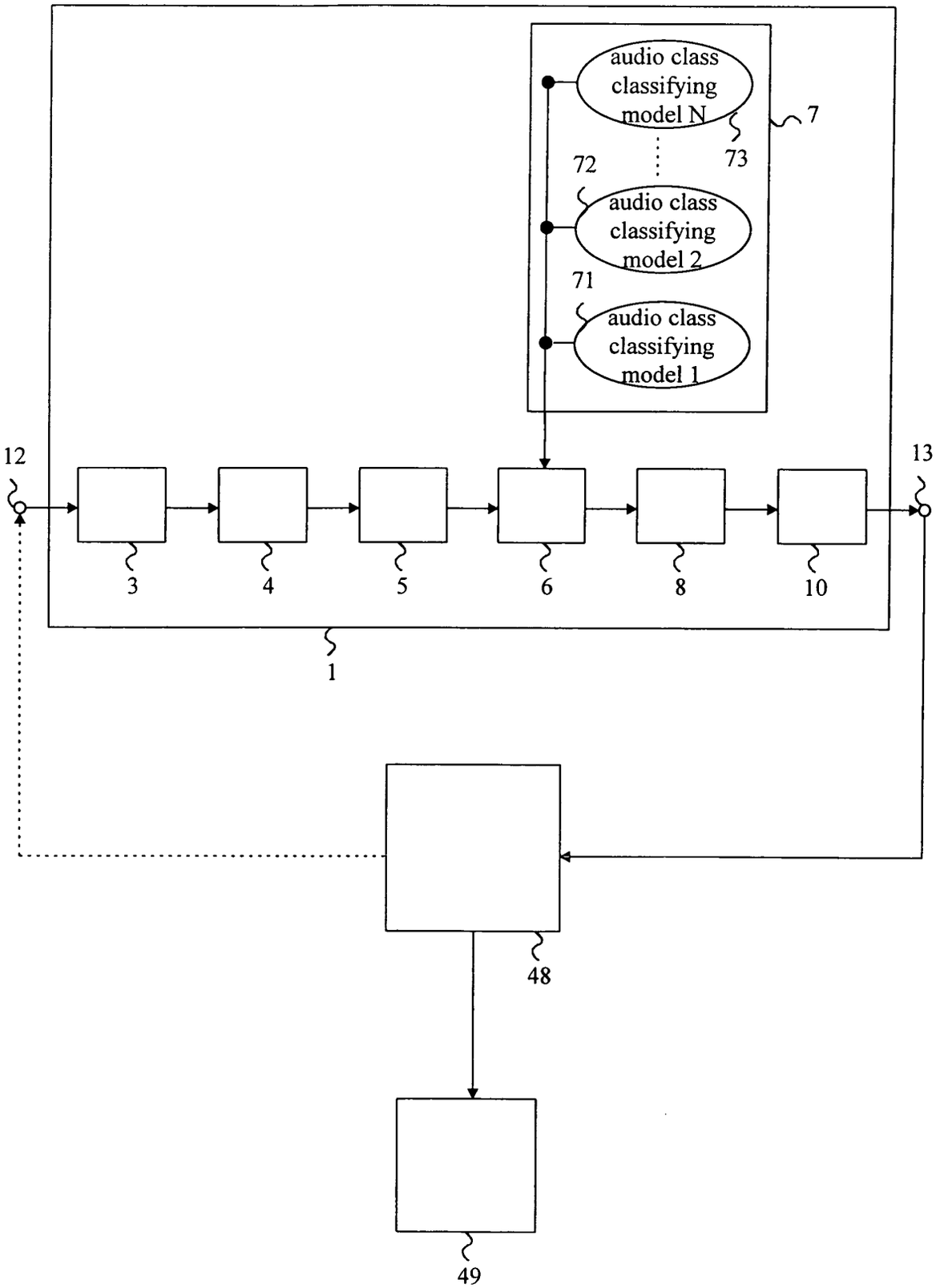

audio class classifying model N
7
73
72
audio class classifying model 2
71
audio class classifying model 1
12
13
3
4
5
6
8
10
1
48
49

Fig. 1A

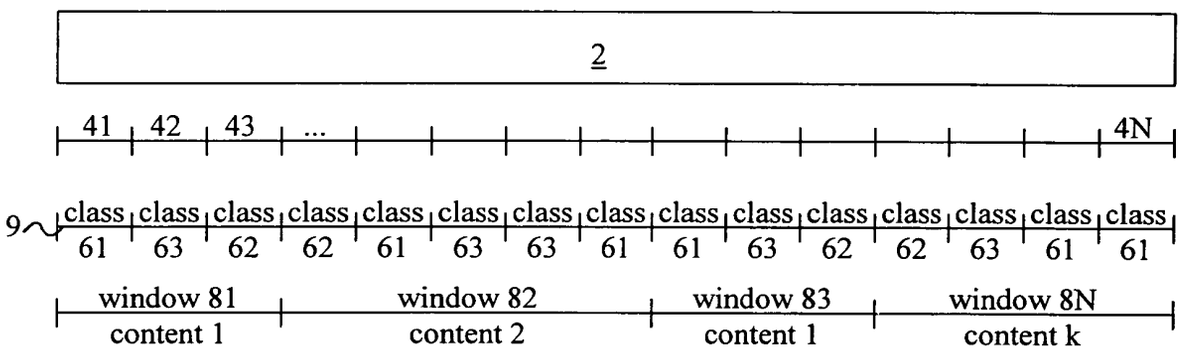

2
41
42
43
...
4N
9
class 61
class 63
class 62
class 62
class 61
class 63
class 63
class 61
class 61
class 63
class 62
class 62
class 63
class 61
class 61
window 81
content 1
window 82
content 2
window 83
content 1
window 8N
content k

Fig. 1B

Fig. 2

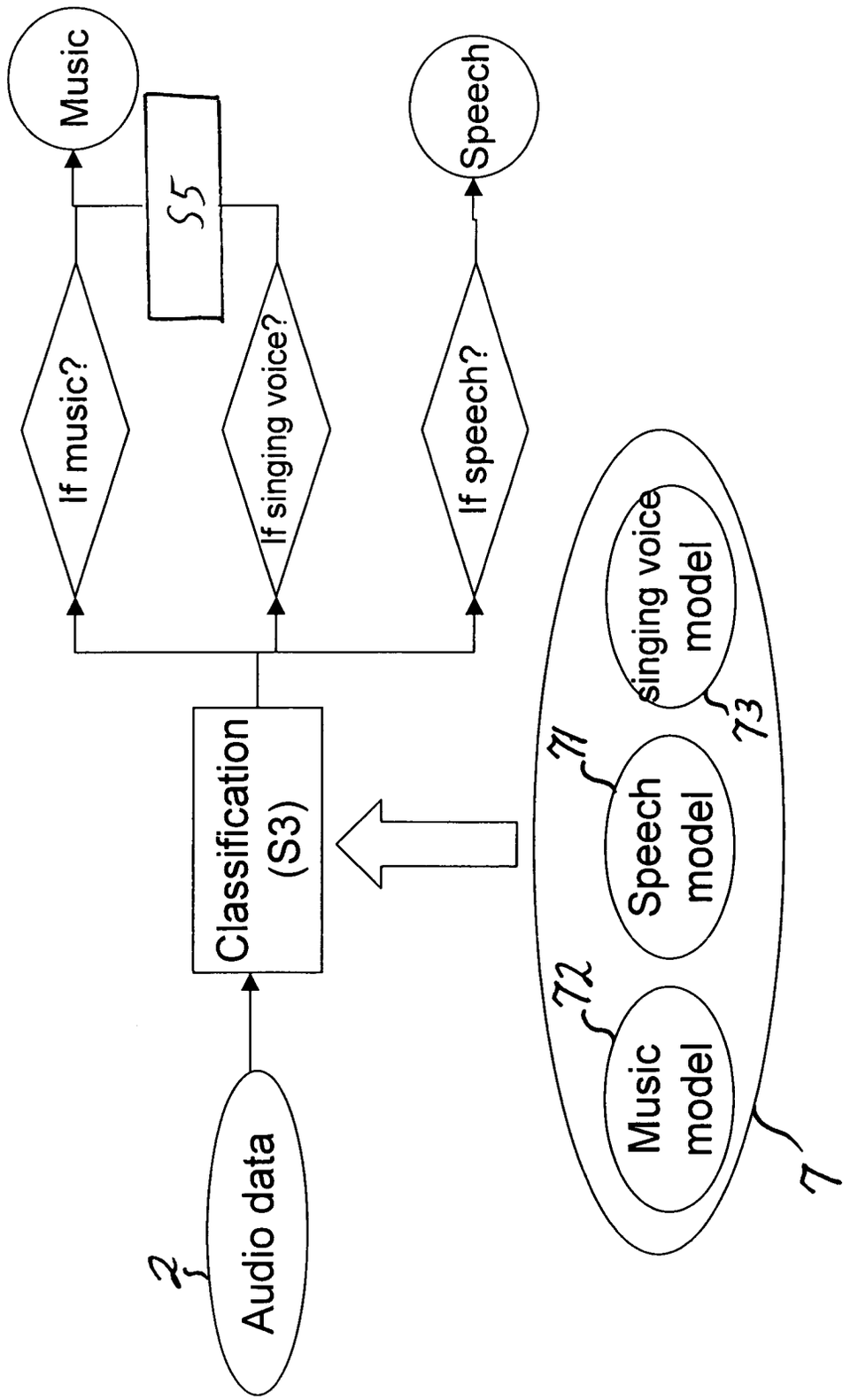

Music
55
Speech
If music?
If singing voice?
If speech?
Classification (S3)
singing voice model
Speech model
Music model
71
72
73
7
2
Audio data

Fig. 3

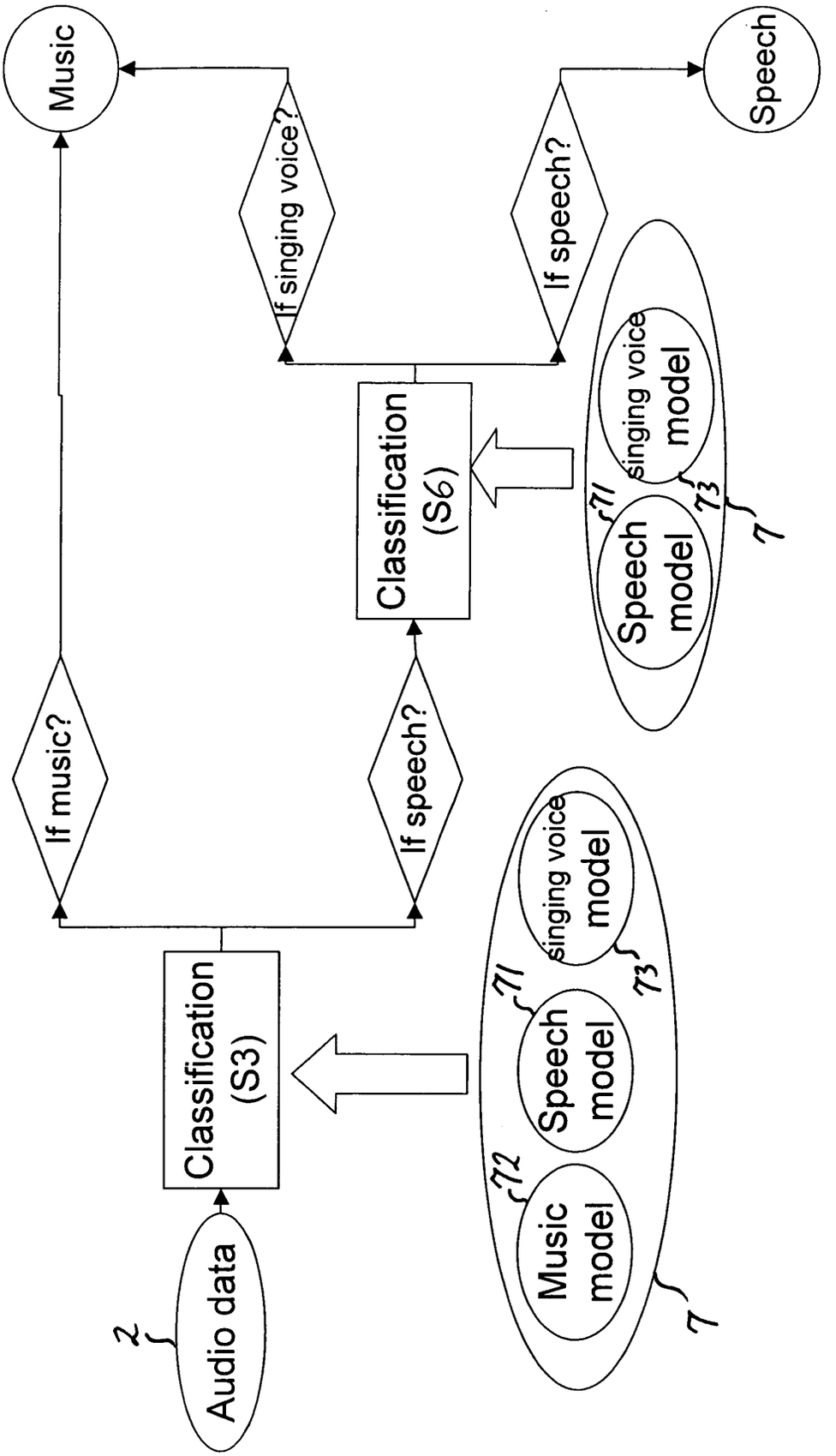
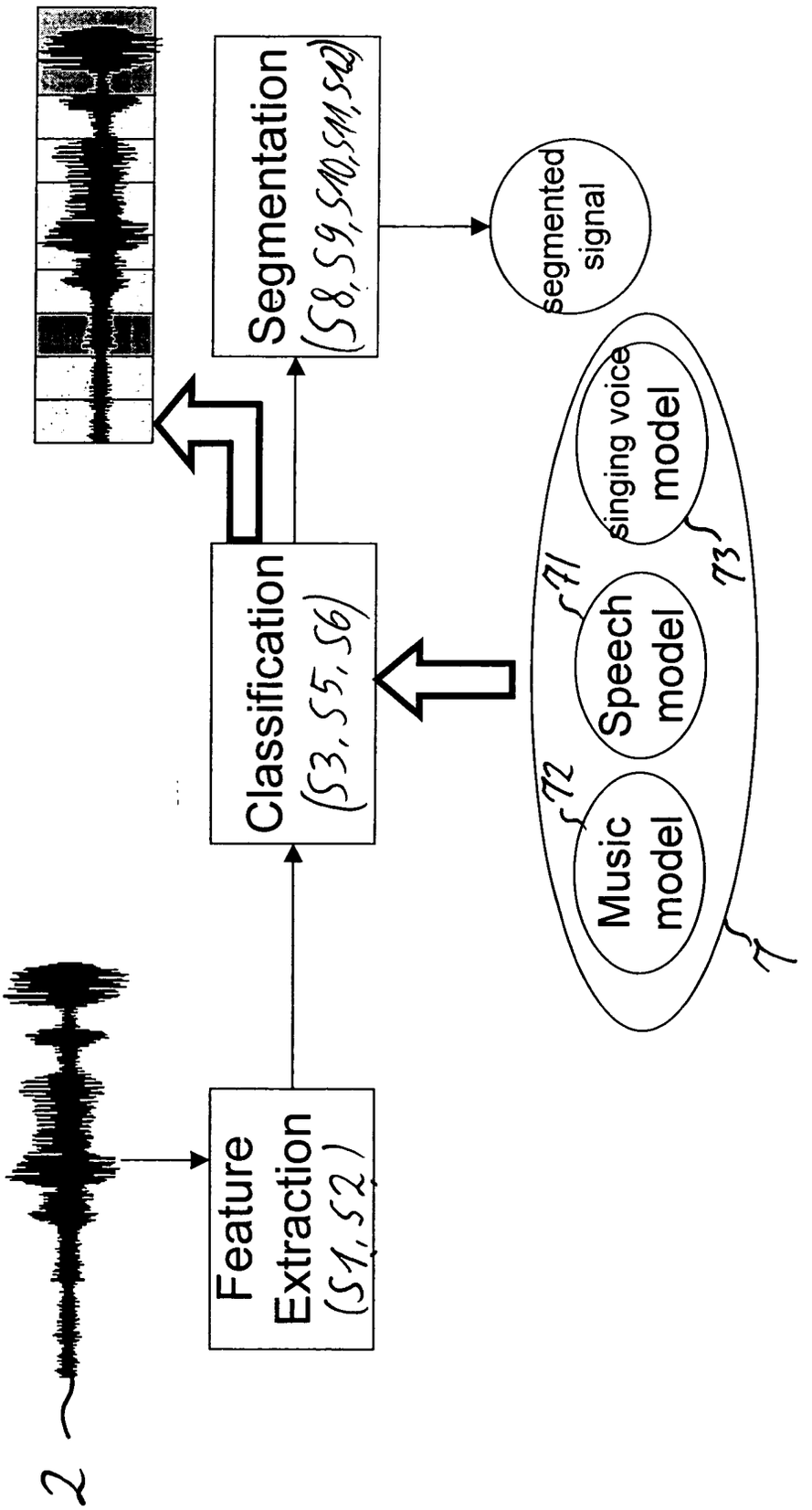

Music
If singing voice?
If speech?
Speech
Classification (S6)
singing voice model
Speech model
71
73
7
If music?
If speech?
Classification (S3)
singing voice model
Speech model
Music model
71
72
73
7
Audio data
2

Fig. 4


Segmentation
(S8, S9, S10, S11, S12)
segmented signal
Classification
(S3, S5, S6)
singing voice model
Speech model
Music model
71
72
73
7
Feature Extraction
(S1, S2)
2

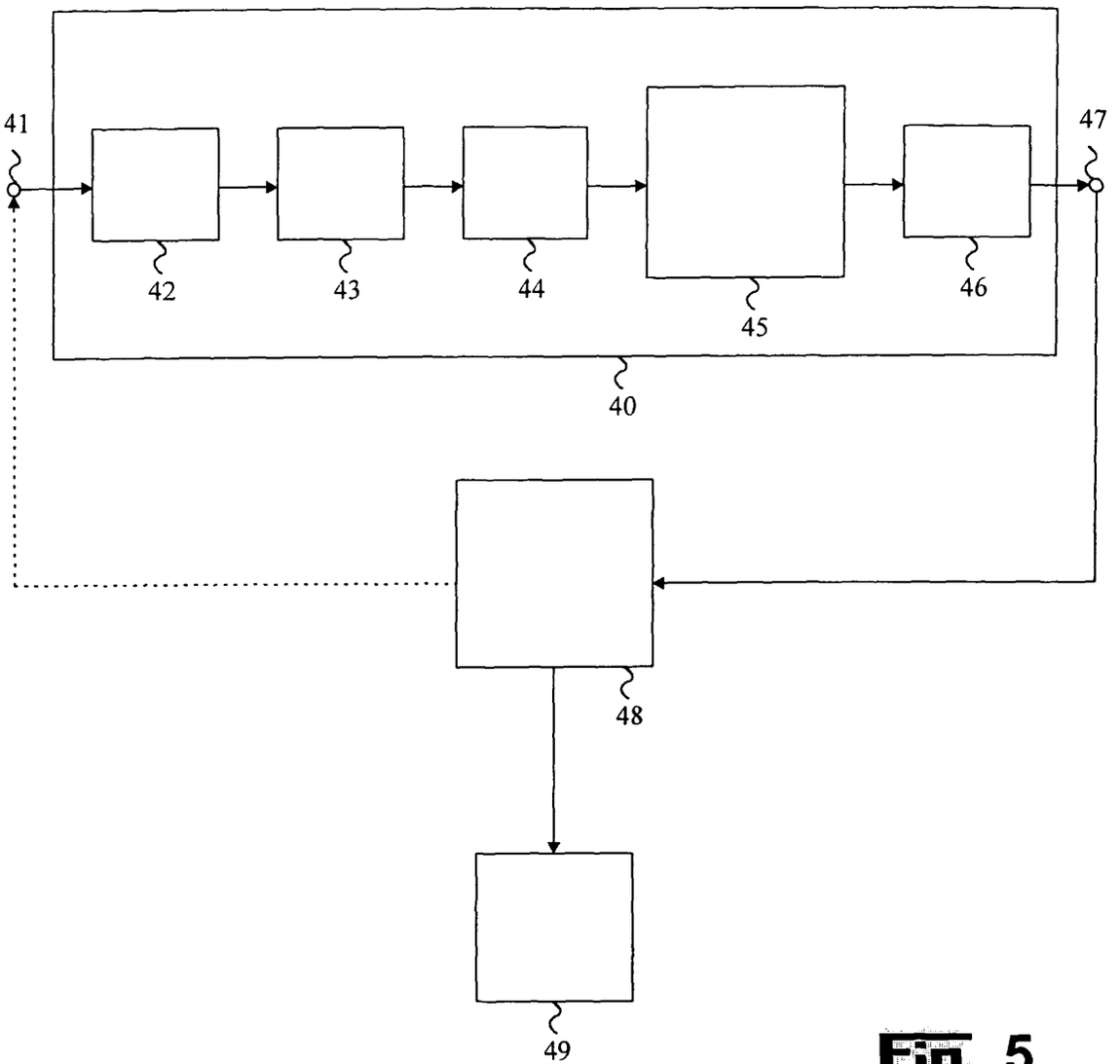


Fig. 5
PRIOR ART

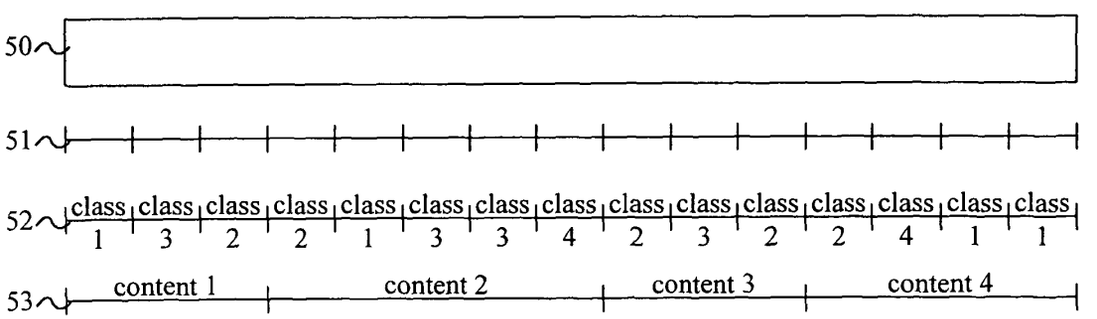


Fig. 6
PRIOR ART

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 03 02 8573

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	ZHANG T ET AL: "Audio content analysis for online audiovisual data segmentation and classification" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, IEEE INC. NEW YORK, US, vol. 9, no. 4, May 2001 (2001-05), pages 441-457, XP001164214 ISSN: 1063-6676 * abstract; figure 1 * * page 443, left-hand column, line 20 - right-hand column, line 11 * ---	1-21	G10L11/00
X	WU CHOU ET AL: "Robust singing detection in speech/music discriminator design" 2001 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. PROCEEDINGS (CAT. NO.01CH37221), 2001 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. PROCEEDINGS, SALT LAKE CITY, UT, USA, 7-11 MAY 2001, pages 865-868 vol.2, XP002278343 2001, Piscataway, NJ, USA, IEEE, USA ISBN: 0-7803-7041-4 * abstract; figure 3 * ---	1,12,21	TECHNICAL FIELDS SEARCHED (Int.Cl.7) G10L
X	US 2002/163533 A1 (LI DONGGE ET AL) 7 November 2002 (2002-11-07) * abstract; figures 1,3 * * paragraph [0043] * ---	1,12,21	
X	WO 01/16937 A (WAVEMAKERS RES INC ;ZAKARAUSKAS PIERRE (US)) 8 March 2001 (2001-03-08) * abstract * * page 2, line 18 - page 3, line 29 * -----	1,12,21	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
MUNICH	29 April 2004	Zimmermann, E

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
.....
& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 03 02 8573

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

29-04-2004

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2002163533 A1	07-11-2002	CN 1463419 T	24-12-2003
		EP 1374219 A2	02-01-2004
		WO 02077966 A2	03-10-2002
		TW 550539 B	01-09-2003
		WO 02078331 A1	03-10-2002
		US 2002138851 A1	26-09-2002
WO 0116937 A	08-03-2001	AU 7471600 A	26-03-2001
		CA 2382122 A1	08-03-2001
		EP 1210711 A1	05-06-2002
		JP 2003508804 T	04-03-2003
		WO 0116937 A1	08-03-2001

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82