(12) UK Patent Application (19) GB (11) 2533654 (13) A

(43) Date of A Publication 29.06.2016

(21) Application No: **1503467.1**

(22) Date of Filing: **02.03.2015**

(30) Priority Data:
(31) **FI2014051036** (32) **22.12.2014** (33) **WO**

(71) Applicant(s):
Nokia Technologies Oy
Karaportti 3, 02610 Espoo, Finland

(72) Inventor(s):
Antti Johannes Eronen
Arto Juhani Lehtiniemi
Jussi Artturi Leppänen
Pasi Saari

(74) Agent and/or Address for Service:
Nokia Technologies Oy
IPR Department, Karakaari 7, 02610 Espoo, Finland

(51) INT CL:
G06F 17/30 (2006.01)

(56) Documents Cited:
WO 2006/129274 A1 **JP 2008170991 A**
US 20070174274 A1 **US 20020002899 A1**

(58) Field of Search:
INT CL **G06F, G10H**
Other: **EPODOC, WPI**

(54) Title of the Invention: **Analysing audio data**
Abstract Title: **Analysing audio data to determine dominance of an audible characteristic**

(57) An apparatus is configured to determine one or more acoustic features of an audio track, determine dominance of an audible characteristic in the audio track based at least partly on said one or more acoustic features and store metadata for the audio track indicating said dominance of the audible characteristic. The apparatus may select one or more audio tracks from a catalogue, or store, having a dominance of the audible characteristic within a range of values based on the dominance of the audible characteristic of the audio track and, optionally, user preferences. Information may be output identifying the one or more selected tracks. The audible characteristic may be a contributing musical instrument or a genre. The determined dominance may include one or more of overall dominance for the entire audio track, varying dominance information and information regarding the dominance of an instrument relative to other instruments in the audio track.

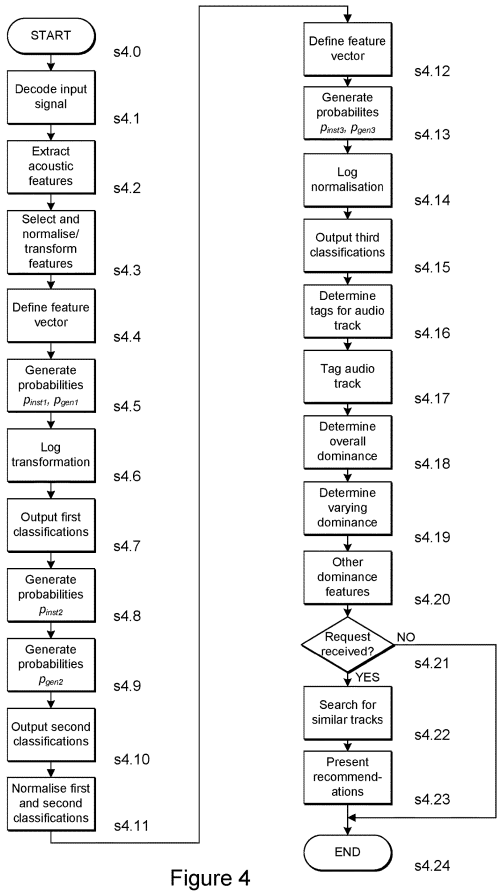

Figure 4

GB 2533654 A

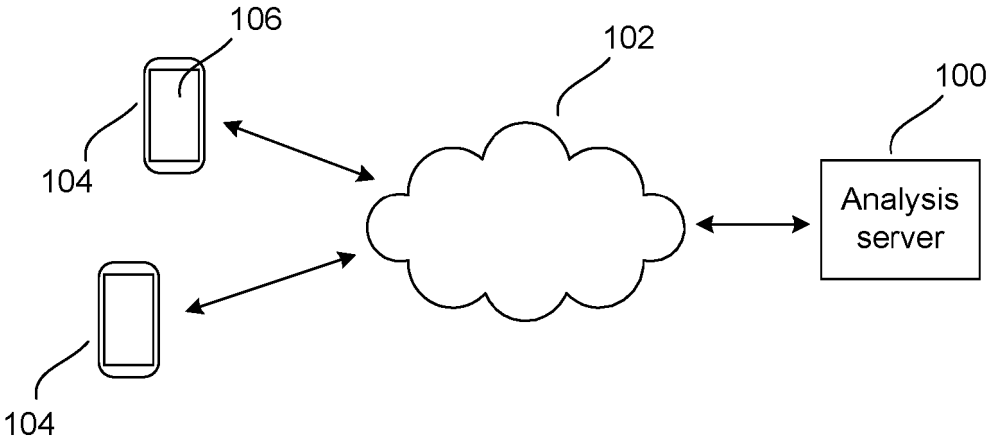

Figure 1

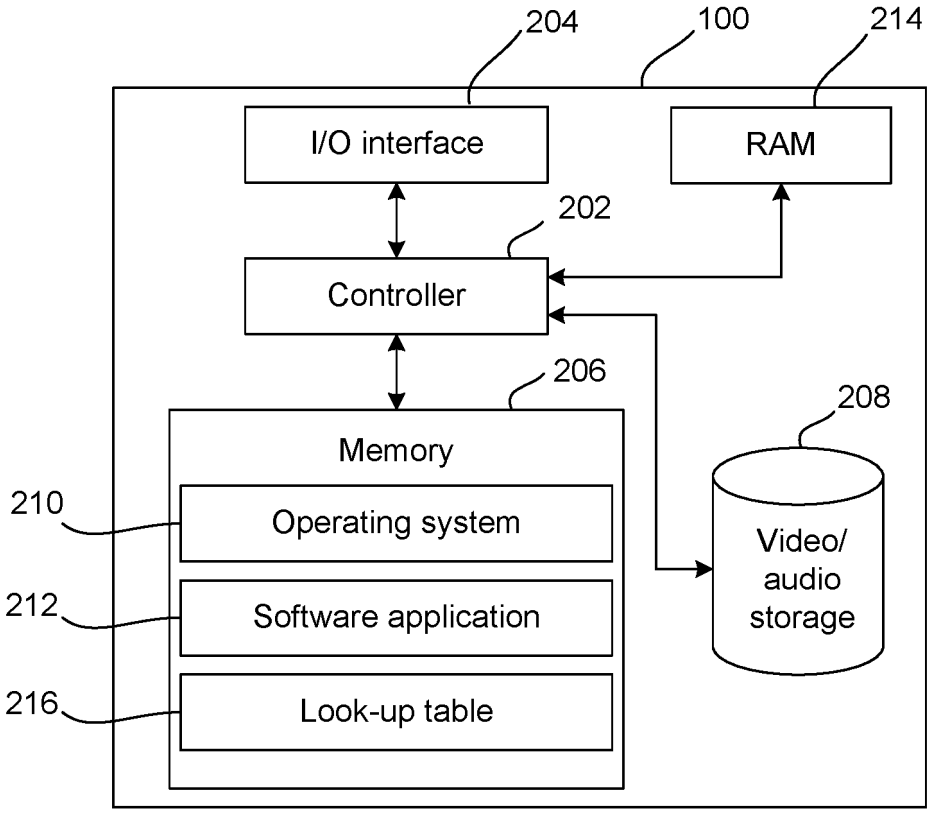

Figure 2

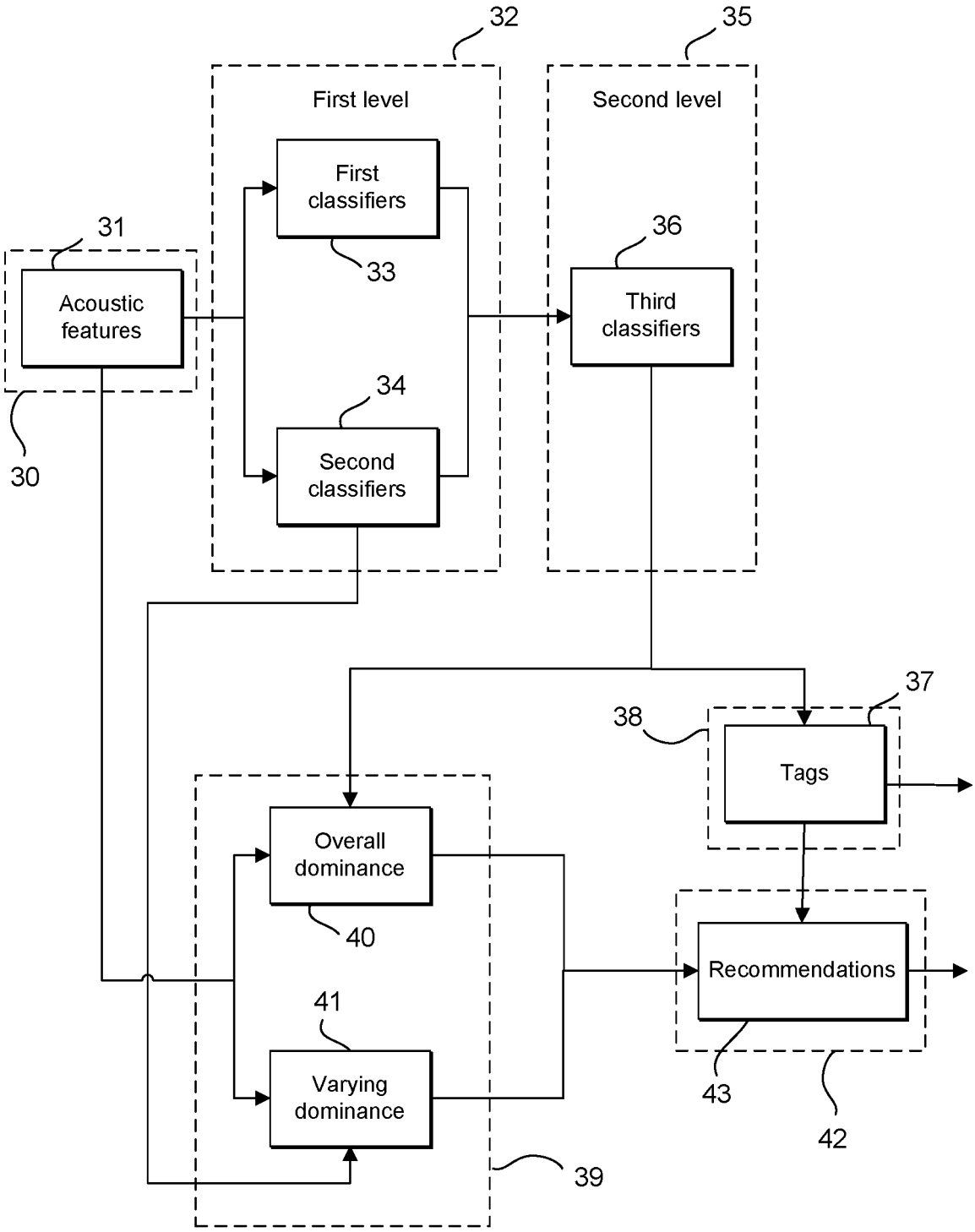
32
35
First level
Second level
First classifiers
33
31
36
Acoustic features
Third classifiers
34
Second classifiers
30
37
38
Tags
Overall dominance
40
Recommendations
41
Varying dominance
43
42
39

Figure 3

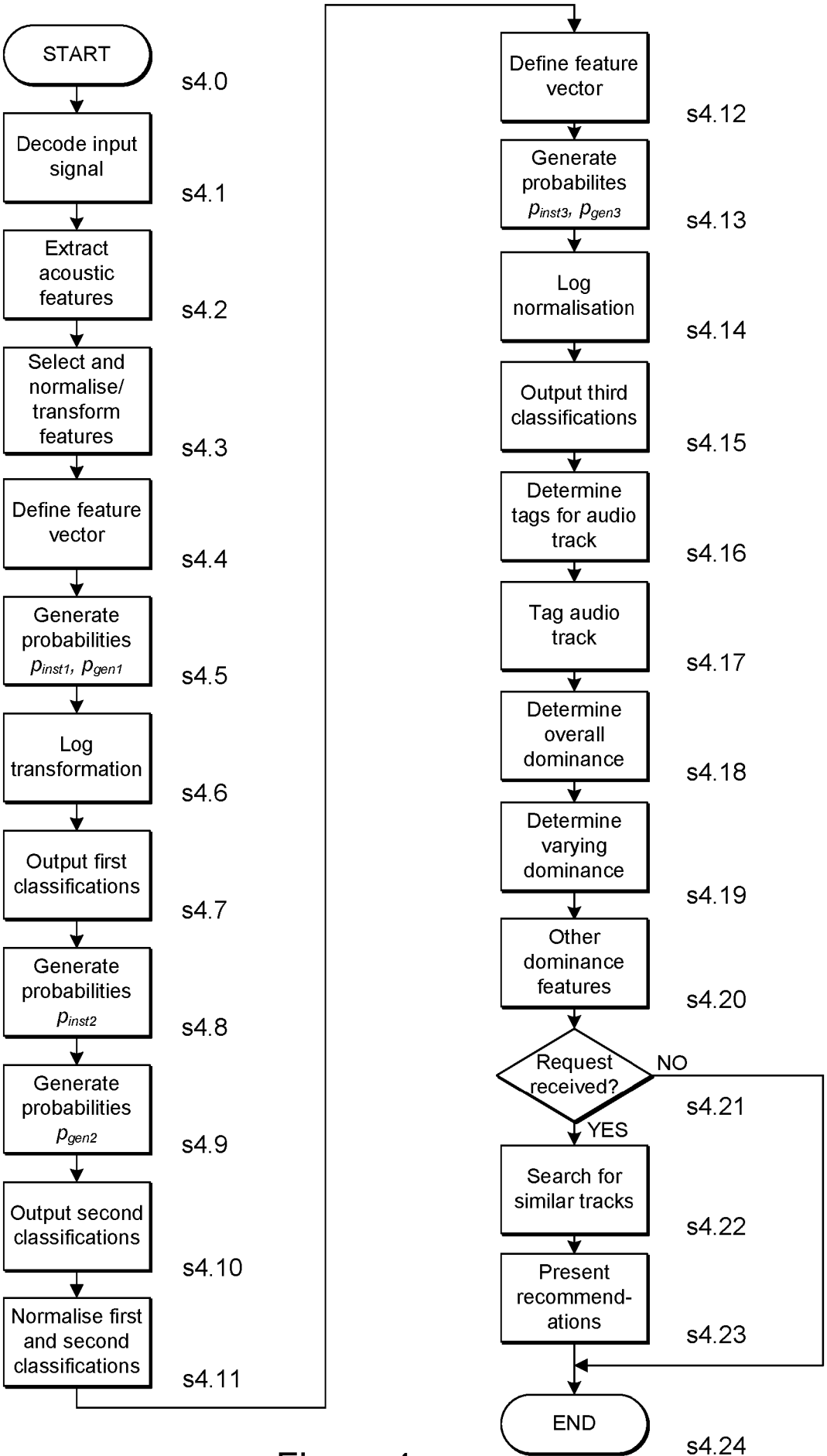
START
s4.0
Decode input signal
s4.1
Extract acoustic features
s4.2
Select and normalise/ transform features
s4.3
Define feature vector
s4.4
Generate probabilities p_{inst1}, p_{gen1}
s4.5
Log transformation
s4.6
Output first classifications
s4.7
Generate probabilities p_{inst2}
s4.8
Generate probabilities p_{gen2}
s4.9
Output second classifications
s4.10
Normalise first and second classifications
s4.11
Define feature vector
s4.12
Generate probabilites p_{inst3}, p_{gen3}
s4.13
Log normalisation
s4.14
Output third classifications
s4.15
Determine tags for audio track
s4.16
Tag audio track
s4.17
Determine overall dominance
s4.18
Determine varying dominance
s4.19
Other dominance features
s4.20
Request received?
NO
s4.21
YES
Search for similar tracks
s4.22
Present recommend-ations
s4.23
END
s4.24

Figure 4

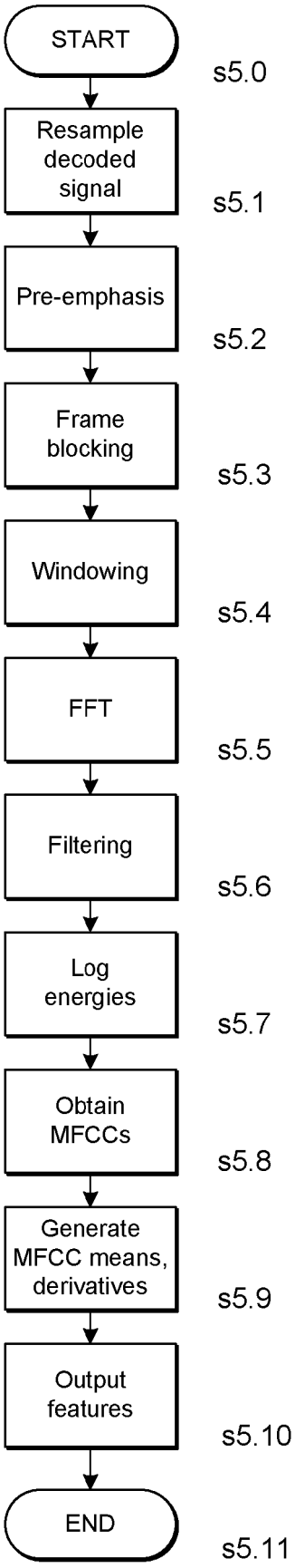
START
s5.0
Resample decoded signal
s5.1
Pre-emphasis
s5.2
Frame blocking
s5.3
Windowing
s5.4
FFT
s5.5
Filtering
s5.6
Log energies
s5.7
Obtain MFCCs
s5.8
Generate MFCC means, derivatives
s5.9
Output features
s5.10
END
s5.11

Figure 5

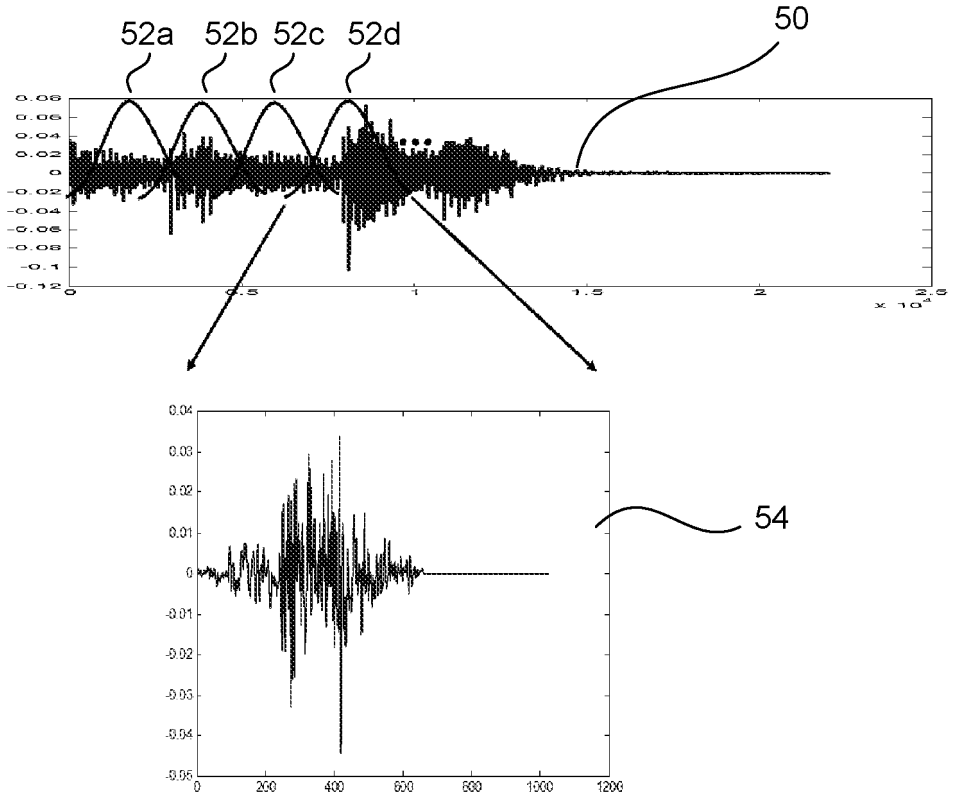

Figure 6

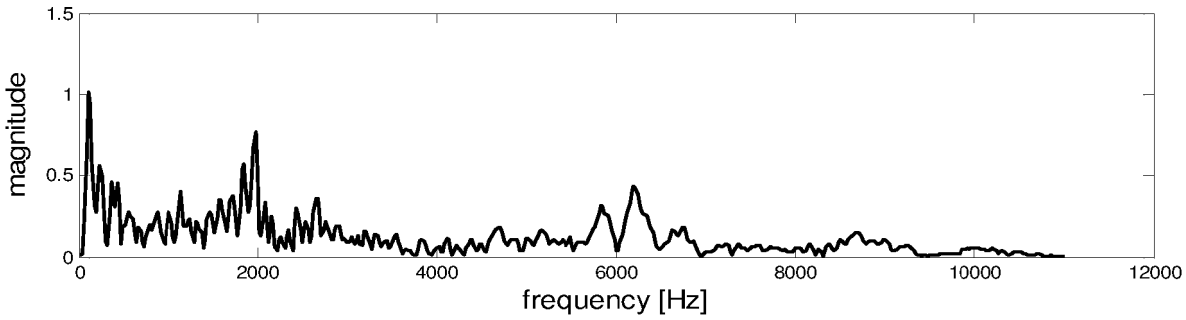

Figure 7

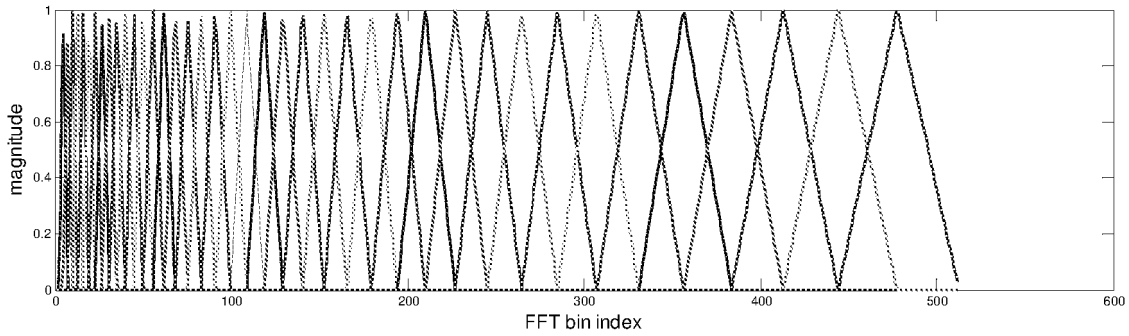

Figure 8

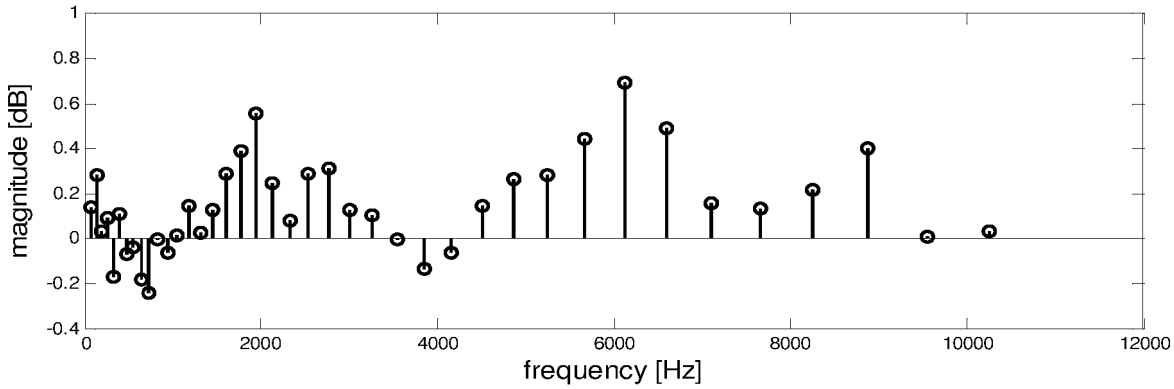

Figure 9

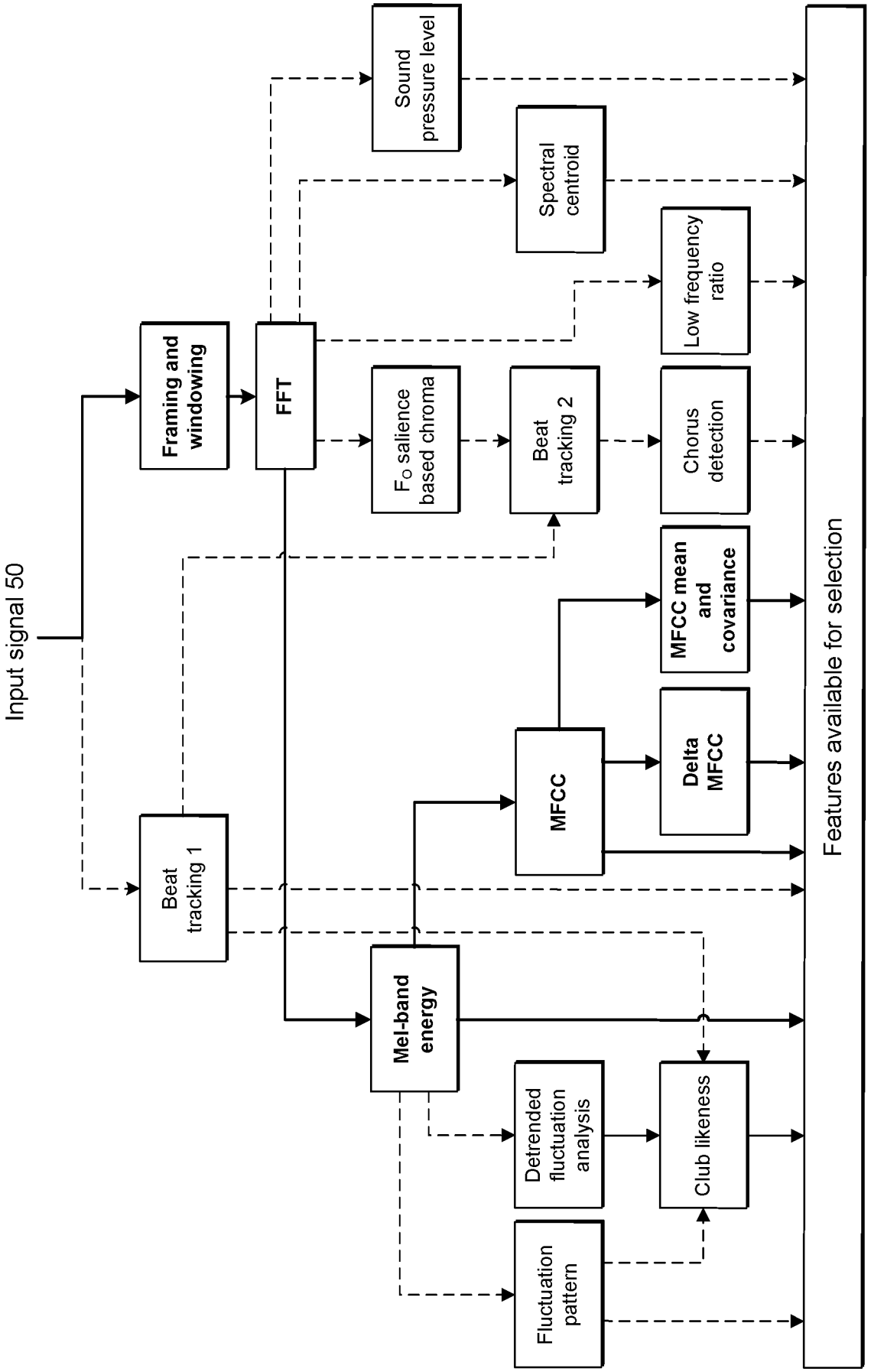
Input signal 50
Framing and windowing
FFT
Beat tracking 1
Mel-band energy
Fluctuation pattern
Detrended fluctuation analysis
Club likeness
MFCC
Delta MFCC
MFCC mean and covariance
F_O salience based chroma
Beat tracking 2
Chorus detection
Low frequency ratio
Spectral centroid
Sound pressure level
Features available for selection

Figure 10

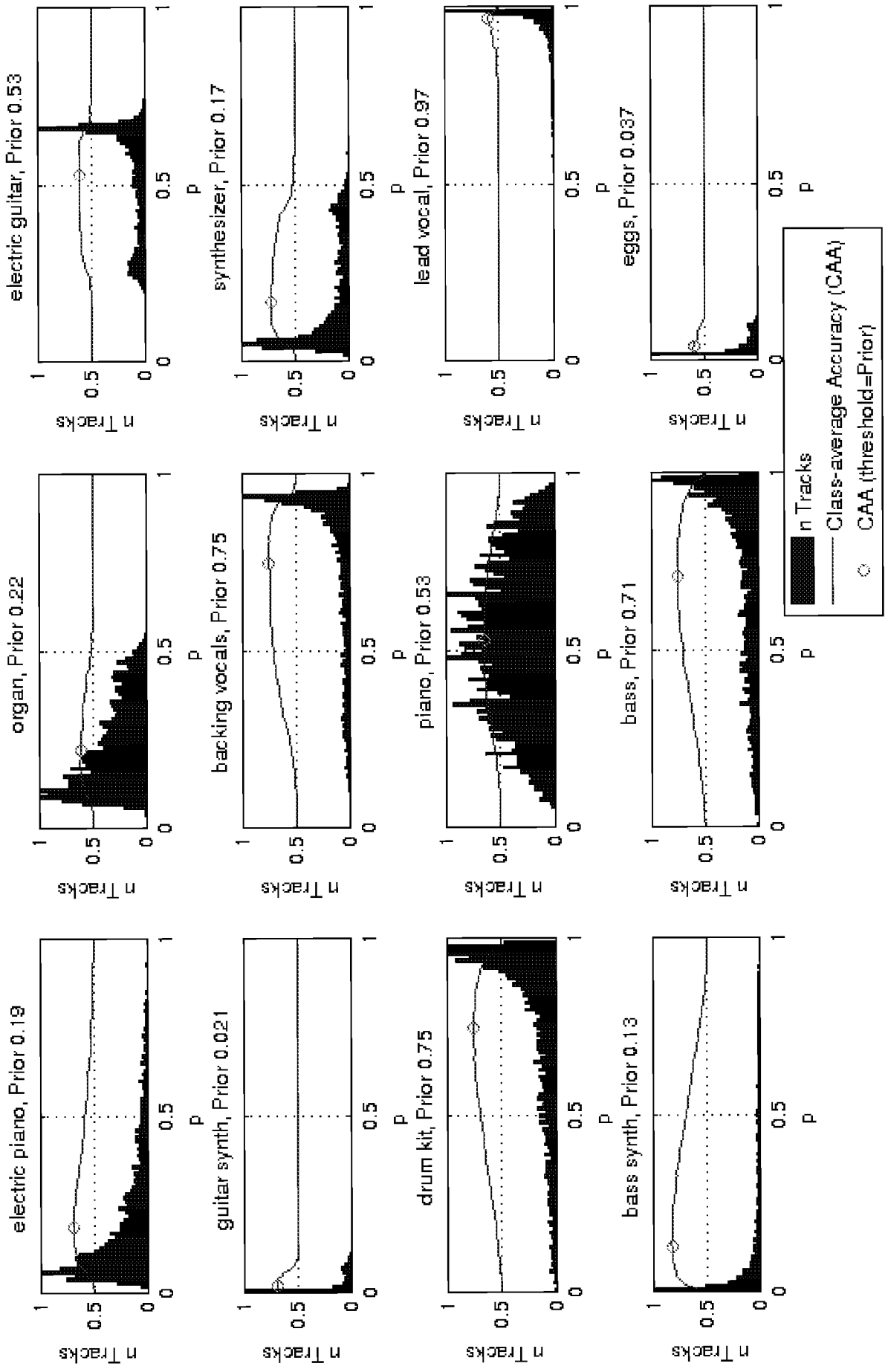
electric guitar, Prior 0.53
synthesizer, Prior 0.17
lead vocal, Prior 0.97
eggs, Prior 0.037
organ, Prior 0.22
backing vocals, Prior 0.75
piano, Prior 0.53
bass, Prior 0.71
electric piano, Prior 0.19
guitar synth, Prior 0.021
drum kit, Prior 0.75
bass synth, Prior 0.13
n Tracks
p
n Tracks
Class-average Accuracy (CAA)
CAA (threshold=Prior)

Figure 11

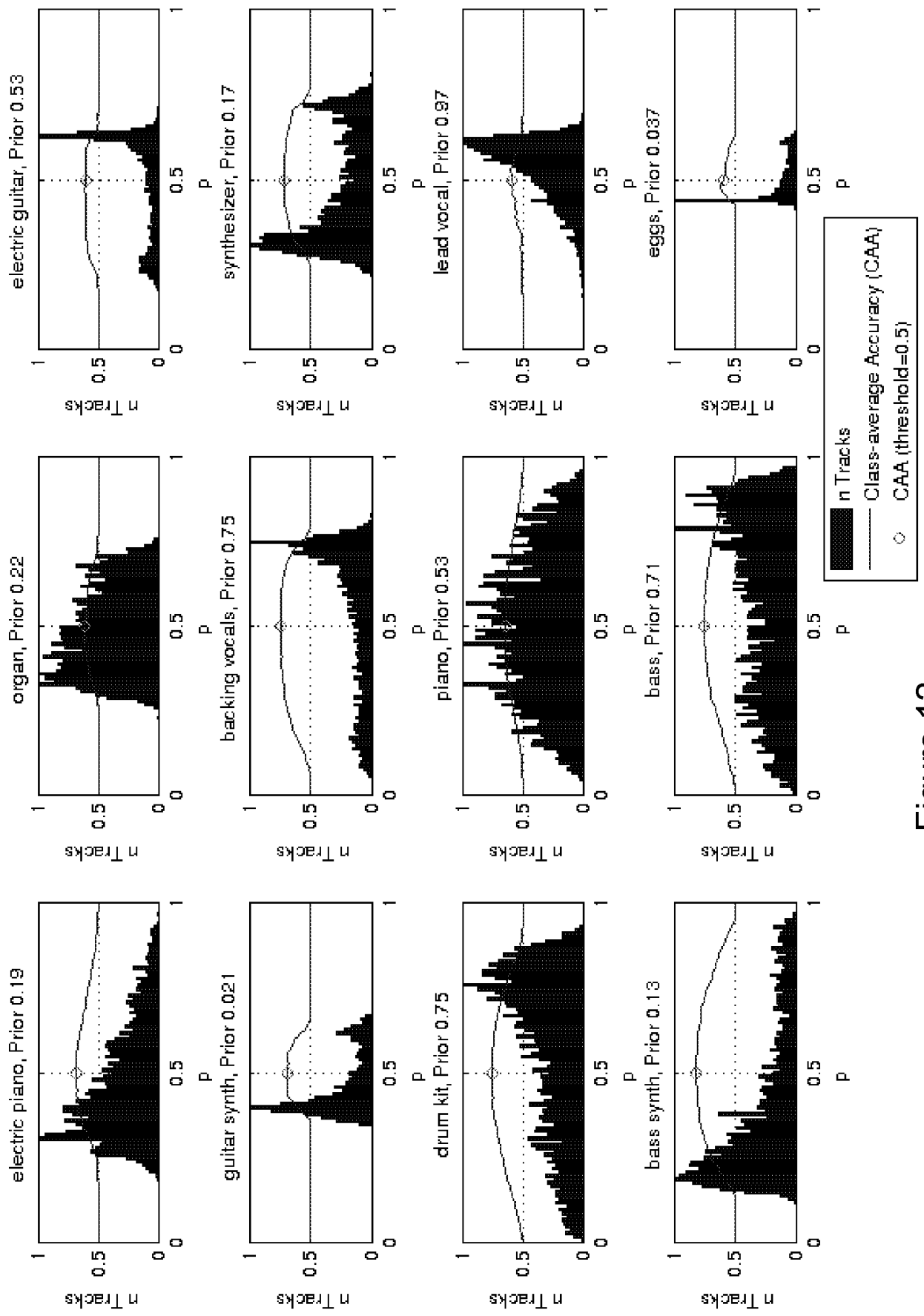
electric guitar, Prior 0.53
synthesizer, Prior 0.17
lead vocal, Prior 0.97
eggs, Prior 0.037
organ, Prior 0.22
backing vocals, Prior 0.75
piano, Prior 0.53
bass, Prior 0.71
electric piano, Prior 0.19
guitar synth, Prior 0.021
drum kit, Prior 0.75
bass synth, Prior 0.13
n Tracks
p
0
0.5
1
n Tracks
Class-average Accuracy (CAA)
CAA (threshold=0.5)

Figure 12

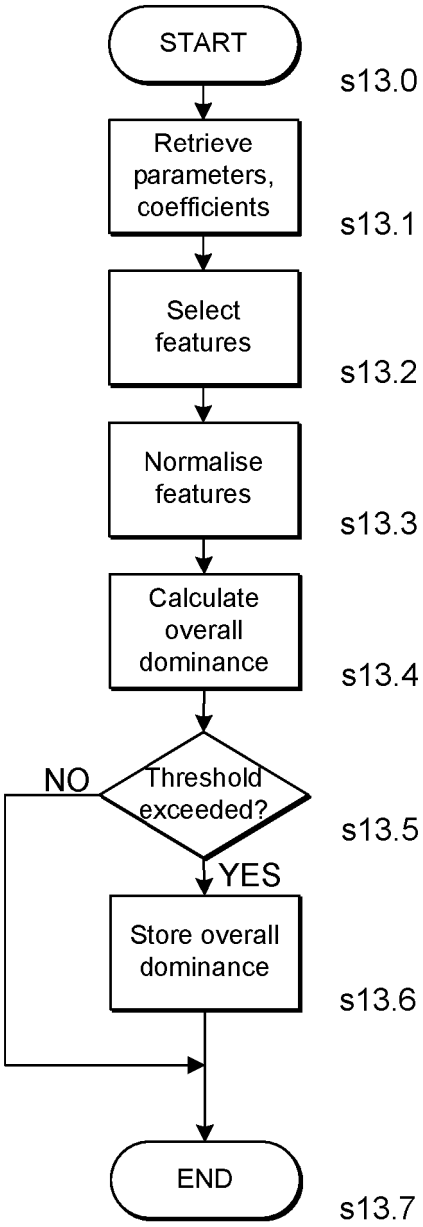

Figure 13

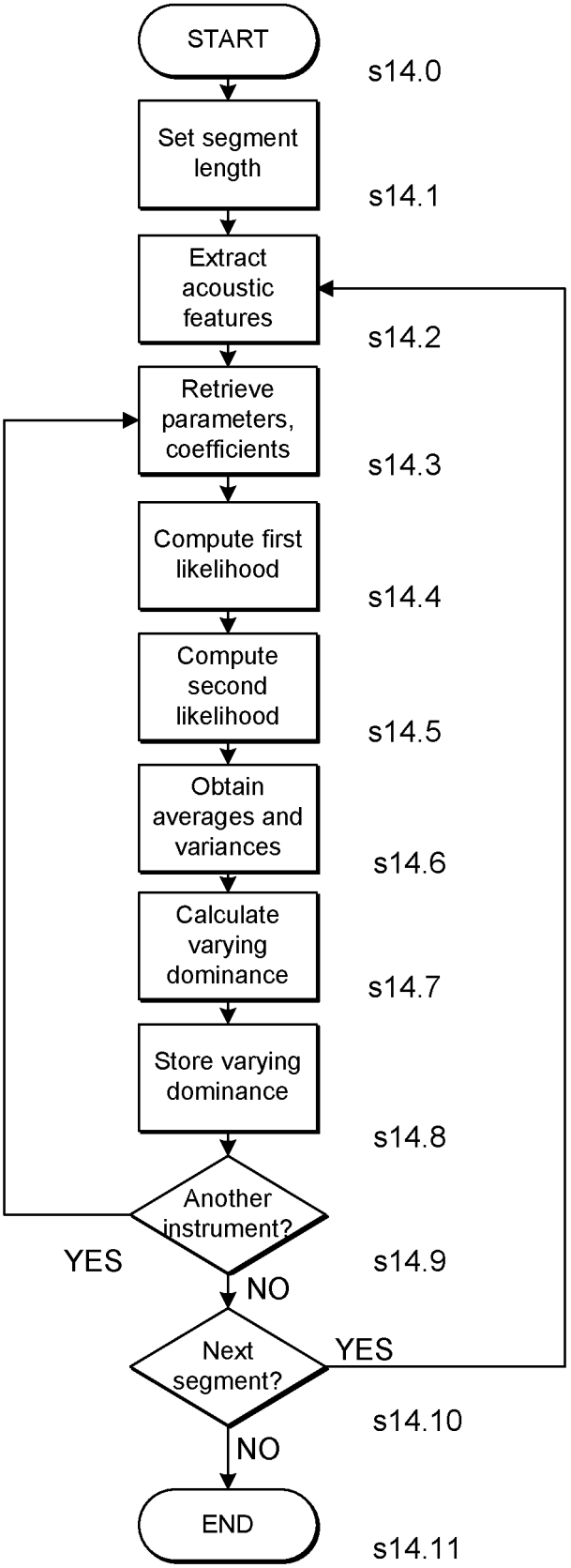

Figure 14

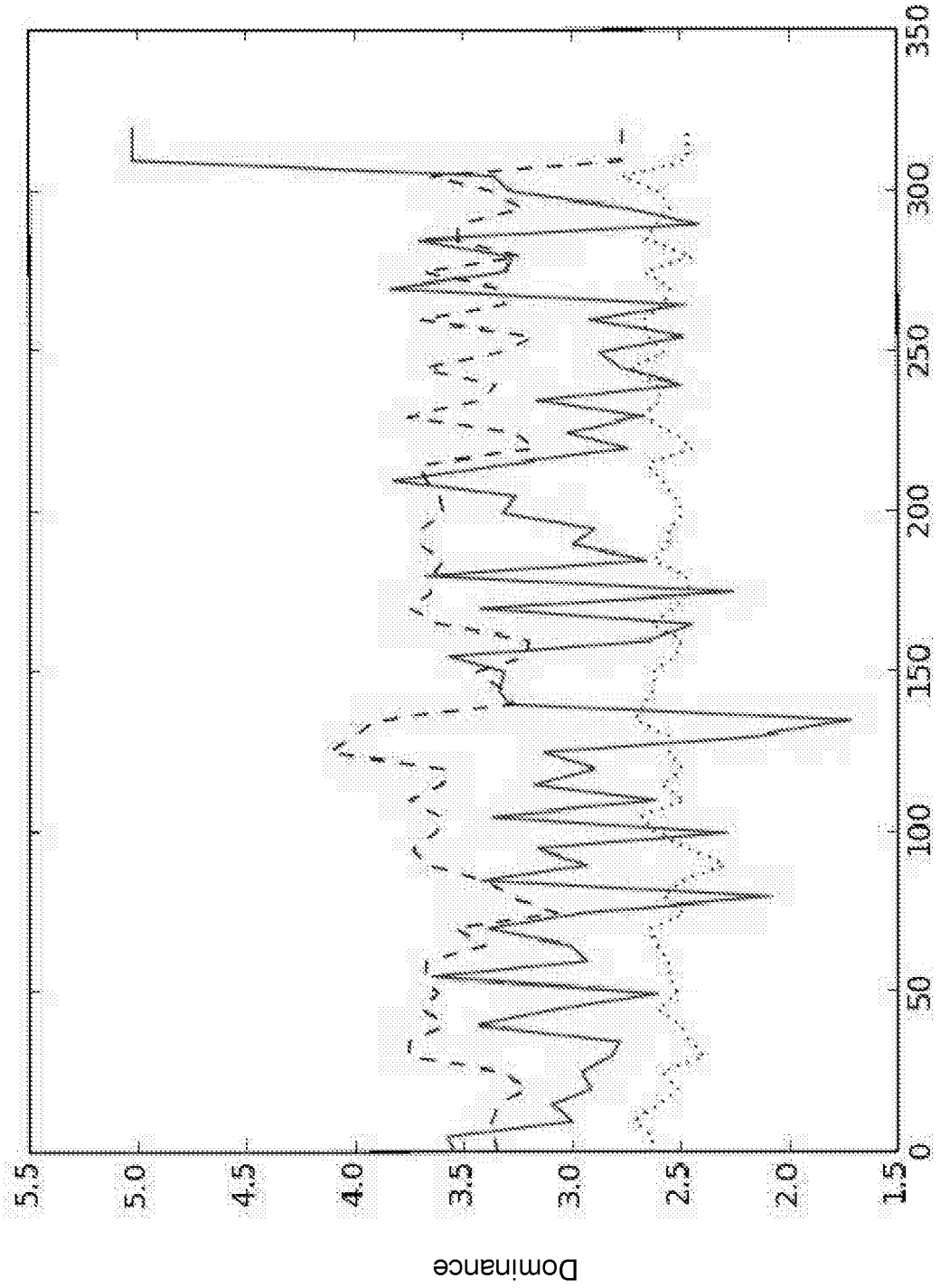
Dominance
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
50
100
150
200
250
300
350
Elapsed time (seconds)

Figure 15

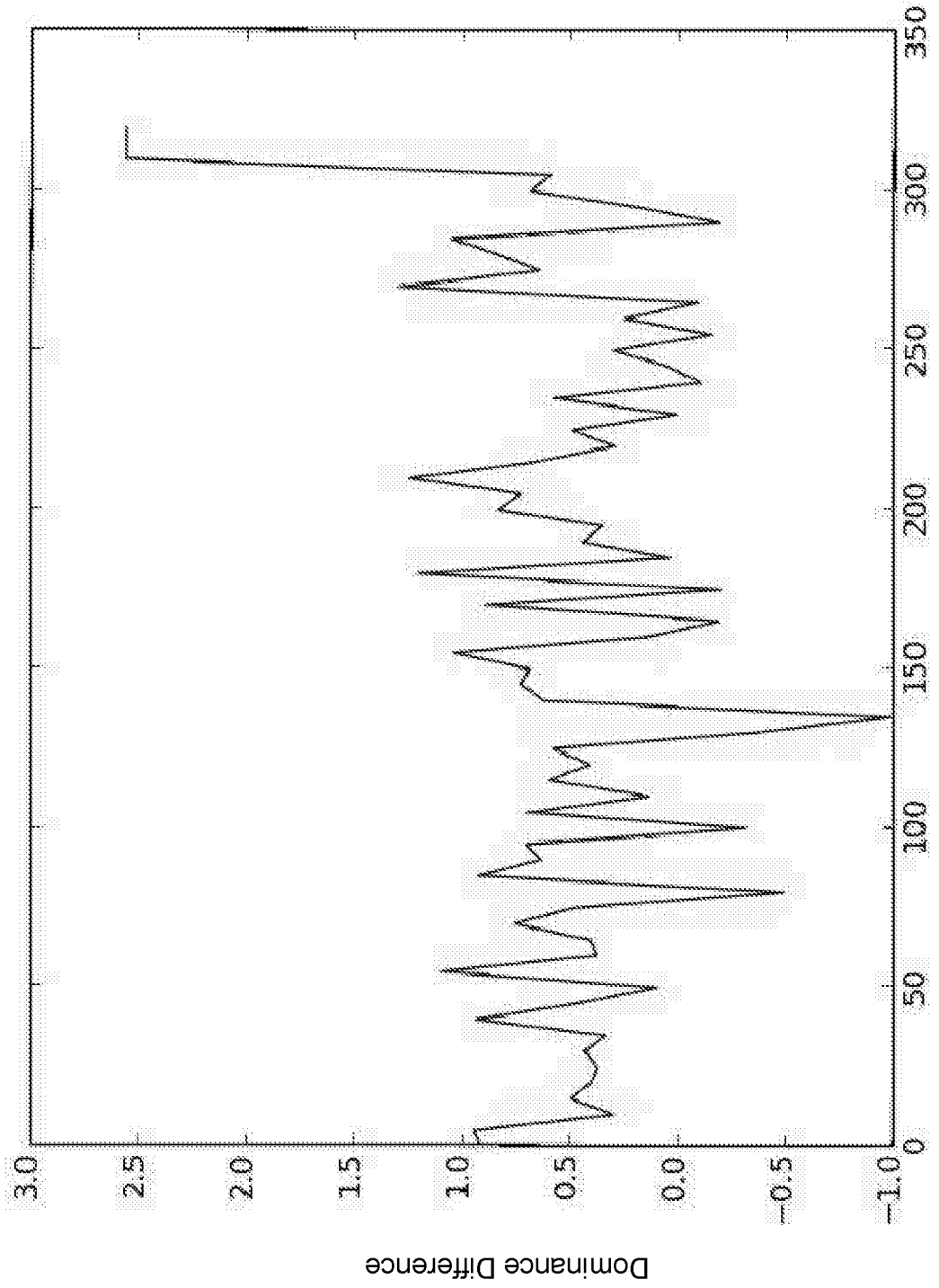
Dominance Difference
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
0
50
100
150
200
250
300
350
Elapsed time (seconds)

Figure 16

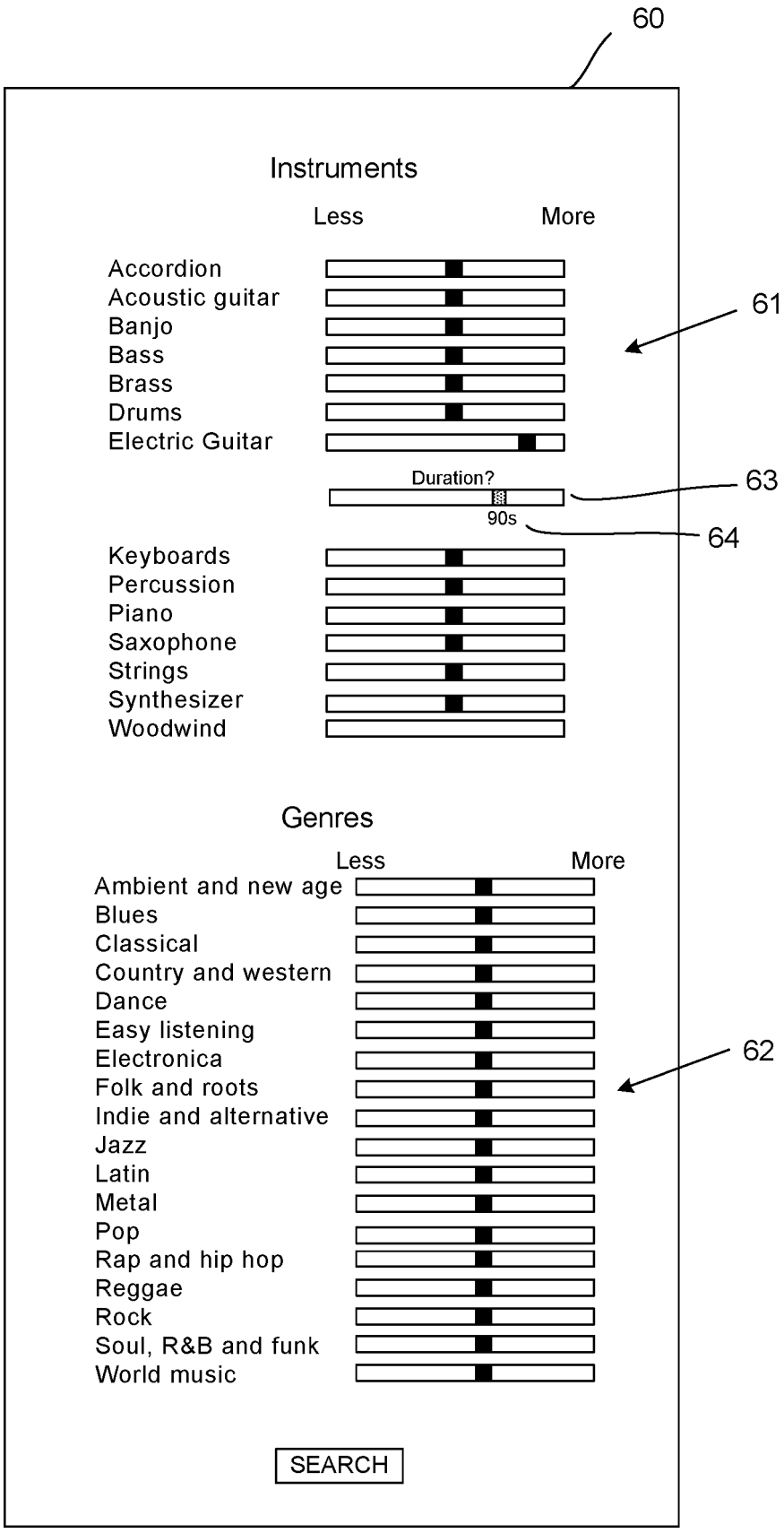
60
Instruments
Less
More
Accordion
Acoustic guitar
Banjo
Bass
Brass
Drums
Electric Guitar
61
Duration?
63
90s
64
Keyboards
Percussion
Piano
Saxophone
Strings
Synthesizer
Woodwind
Genres
Less
More
Ambient and new age
Blues
Classical
Country and western
Dance
Easy listening
Electronica
Folk and roots
Indie and alternative
Jazz
Latin
Metal
Pop
Rap and hip hop
Reggae
Rock
Soul, R&B and funk
World music
62
SEARCH

Figure 17

Analysing audio data

Field

This disclosure relates to analysing audio data. In particular, this disclosure relates to determining dominance of characteristics of audio tracks, such as classification information and/or tags for a piece of music.

Background

Audio content databases, streaming services, online stores and media player software applications often include genre classifications, to allow a user to search for tracks to play stream and/or download.

Some databases, services, stores and applications also include a facility for recommending music tracks to a user based on a history of music that they have accessed in conjunction with other data, such as rankings of tracks or artists from the user, history data from other users who have accessed the same or similar tracks in the user's history or otherwise have similar user profiles, metadata assigned to the tracks by experts and/or users, and so on.

Summary

According to an aspect, an apparatus includes a controller and a memory in which is stored computer-readable instructions which, when executed by the controller, cause the controller to determine one or more acoustic features of an audio track, determine dominance of an audible characteristic in the audio track based at least in part on said one or more acoustic features, and store metadata for the audio track indicating said dominance of the audible characteristic.

Optionally, the computer-readable instructions, when executed by the controller, may further cause the controller to select one or more tracks from a catalogue having a dominance of the audible characteristic within a range of dominance values defined at least in part based on the dominance of the audible characteristic of the audio track and output information identifying said one or more selected tracks.

The apparatus may further include a user interface to receive input indicating a preferred dominance of the audible characteristic, wherein said range of dominance values is further based on the received input.

The dominance may include an overall dominance indicating a level of audible distinguishment of a musical instrument in the audio track. Alternatively, or additionally, the dominance may include an overall dominance indicates a degree of conformity to a musical genre of the audio track and/or a varying dominance indicating a level of audible distinguishability of a musical instrument in one or more temporal segments of the audio track.

Where the dominance includes a varying dominance, the computer-readable instructions may, when executed by the controller, further cause the controller to determine at least one of a difference in dominance of the musical instrument and an average of other musical instruments in the audio track, a frequency of changes in dominance for the musical instrument and a duration of at least one section of the audio track for which the musical instrument is dominant.

According to another aspect, an apparatus includes a controller and a memory in which is stored computer-readable instructions which, when executed by the controller, cause the controller to select one or more tracks from a catalogue having a dominance of an audible characteristic within a range of dominance values and output information identifying said one or more selected tracks.

The range of dominance values may be based on a dominance for the audible characteristic in a first audio track. For example, the apparatus may be configured to determine a dominance of the first audio track and then set the range of dominance values based, at least in part, on the determined dominance.

The apparatus may further include a user interface to receive input indicating a preferred dominance of the audible characteristic, wherein said range of dominance values is further based on the received input.

The dominance may include an overall dominance indicating a level of audible distinguishability of a musical instrument in the audio track. Alternatively, or additionally, the dominance may include an overall dominance indicates a degree of conformity to a musical genre of the audio track and/or a varying dominance indicating a level of audible distinguishability of a musical instrument in one or more temporal segments of the audio track.

Where the dominance includes a varying dominance, the computer-readable instructions may, when executed by the controller, further cause the controller to determine at least one of a difference in dominance of the musical instrument and an average of other musical instruments in the audio track, a frequency of changes in dominance for the musical instrument and a duration of at least one section of the audio track for which the musical instrument is dominant.

According to yet another aspect, a method includes determining one or more acoustic features of an audio track, determining dominance of an audible characteristic in the audio track based at least in part on said one or more acoustic features, and storing metadata for the audio track indicating said dominance of the audible characteristic.

The method may further include selecting one or more tracks from a catalogue having a dominance of the audible characteristic within a range of dominance values defined at least in part based on the dominance of the audible characteristic of the audio track and outputting information identifying said one or more selected tracks.

Alternatively, or additionally, the method may further include receiving input indicating a preferred dominance of the audible characteristic, wherein said range of dominance values is further based on the received input.

The dominance may include at least one of an overall dominance indicating a level of audible distinguishability of a musical instrument in the audio track, an overall dominance indicating a degree of conformity to a musical genre of the audio track and a varying dominance indicating a level of audible distinguishability of a musical instrument in one or more temporal segments of the audio track.

Where the dominance includes a varying dominance, the method may include determining at least one of a difference in dominance of the musical instrument and an average of other musical instruments in the audio track, a frequency of changes in dominance for the musical instrument and a duration of at least one section of the audio track for which the musical instrument is dominant.

According to a further aspect, a method includes selecting one or more tracks from a catalogue having a dominance of an audible characteristic within a range of

dominance values and outputting information identifying said one or more selected tracks.

Alternatively, or additionally, the method may further include receiving input indicating a preferred dominance of the audible characteristic, wherein said range of dominance values is further based on the received input.

The dominance may include at least one of an overall dominance indicating a level of audible distinguishability of a musical instrument in the audio track, an overall dominance indicating a degree of conformity to a musical genre of the audio track and a varying dominance indicating a level of audible distinguishability of a musical instrument in one or more temporal segments of the audio track.

Where the dominance includes a varying dominance, the method may include determining at least one of a difference in dominance of the musical instrument and an average of other musical instruments in the audio track, a frequency of changes in dominance for the musical instrument and a duration of at least one section of the audio track for which the musical instrument is dominant.

According to another further aspect, an apparatus includes a feature extractor to determine one or more acoustic features of an audio track, a dominance determination module to determine dominance of an audible characteristic in the audio track based on said one or more acoustic features, and a memory to store metadata for the audio track indicating said dominance of the audible characteristic. Optionally, the apparatus may further include a recommendations module to select one or more tracks from a catalogue having a dominance of the audible characteristic within a range of dominance values defined at least in part based on the dominance of the audible characteristic of the audio track and output information identifying said one or more selected tracks.

According to yet another further aspect, an apparatus includes a recommendations module to select one or more tracks from a catalogue having a dominance of an audible characteristic within a range of dominance values and output information identifying said one or more selected tracks.

According to an additional aspect, an apparatus includes means for determining one or more acoustic features of an audio track, means for determining dominance

of an audible characteristic in the audio track based at least in part on said one or more acoustic features, and means for storing metadata for the audio track indicating said dominance of the audible characteristic.

According to another additional aspect, an apparatus includes means for selecting one or more tracks from a catalogue having a dominance of an audible characteristic within a range of dominance values and means for outputting information identifying said one or more selected tracks.

Brief description of the drawings

Embodiments will now be described by way of non-limiting examples with reference to the accompanying drawings, of which:

Figure 1 is a schematic diagram of a system in which an embodiment may be included;

Figure 2 is a schematic diagram of components of an analysis server according to an embodiment, in the system of Figure 1;

Figure 3 is an overview of a method of determining dominance information for an audio track and recommending further audio tracks, according to an embodiment;

Figure 4 is a flowchart of a method according to Figure 3, which may be performed by the analysis server of Figure 2;

Figure 5 is a flowchart of a method of calculating mel-frequency cepstral coefficients in part of the method of Figure 4;

Figure 6 depicts an example of frame blocking and windowing in the method of Figure 5;

Figure 7 is an example of a spectrum generated by transforming a portion of a frame in the method of Figure 5;

Figure 8 depicts a bank of weighted mel-frequency filters used in the method of Figure 5;

Figure 9 depicts a spectrum of log mel-band energies in the method of Figure 5;

Figure 10 is an overview of a process for obtaining multiple types of acoustic features in the method of Figure 4;

Figure 11 shows example probability distributions for a number of first classifications;

Figure 12 shows the example probability distributions of Figure 11 after logarithmic transformation;

Figure 13 is a flowchart of an example method of determining overall dominance in the method of Figure 4;

Figure 14 is a flowchart of an example method of determining varying dominance in the method of Figure 4;

Figure 15 is a graph of showing varying dominance values for various musical instruments in an example audio track;

Figure 16 is a graph showing varying dominance values for a selected musical instrument relative to other musical instruments in the example audio track; and

Figure 17 depicts an example of a user interface that may be used in the method of Figure 4.

Detailed description

Embodiments described herein concern determining dominance of characteristics indicated by classification information, such as tags, for audio data and/or selecting audio data based on such dominance. Embodiments of the present invention are described in the context of music tracks.

Referring to Figure 1, an analysis server 100 is shown connected to a network 102, which can be any data network such as a Local Area Network (LAN), Wide Area Network (WAN) or the Internet. The analysis server 100 is configured to receive and process requests for audio content from one or more terminals 104 via the network 102.

In the present example, three terminals 104 are shown, each incorporating media playback hardware and software, such as a speaker (not shown) and/or audio output jack (not shown) and a processor (not shown) that executes a media player software application to stream and/or download audio content over the network 102 and to play audio content through the speaker. As well as audio content, the terminals 104 may be capable of streaming or downloading video content over the network 102 and presenting the video content using the speaker and a display 106. Suitable terminals 104 will be familiar to persons skilled in the art. For instance a smart phone could serve as a terminal 104 in the context of this application although a laptop, tablet or desktop computer may be used instead. Such devices include music and video playback and data storage functionality and can be connected to the music analysis sever 100 via a cellular network, Wi-fi, Bluetooth® or any other suitable connection such as a cable or wire.

As shown in Figure 2, the analysis server 100 includes a controller 202, an input and output interface 204 configured to transmit and receive data via the network 102, a memory 206 and a mass storage device 208 for storing video and audio data.

The controller 202 is connected to each of the other components in order to control operation thereof. The controller 202 may take any suitable form. For instance, it may be a processing arrangement that includes a microcontroller, plural microcontrollers, a processor, or plural processors.

The memory 206 and mass storage device 208 may be in the form of a non-volatile memory such as read only memory (ROM) a hard disk drive (HDD) or a solid state drive (SSD). The memory 206 stores, amongst other things, an operating system 210 and at least one software application 212 to be executed by the controller 202. Random Access Memory (RAM) 214 is used by the controller 202 for the temporary storage of data.

The operating system 210 may contain code which, when executed by the controller 202 in conjunction with the RAM 214, controls operation of analysis server 100 and provides an environment in which the or each software application 212 can run.

Software application 212 is configured to control and perform audio and video information processing by the controller 202 of the analysis server 100. The operation of this software application 212 according to a first embodiment will now be described in detail, with reference to Figures 3 and 4. In the following, the accessed audio track is referred to as the input signal.

Figure 3 is an overview of a determination of tag and dominance information for the audio track by the controller 202 of the analysis server 100, in which the controller 202 acts as a feature extractor 30, first level classifiers 32, second level classifiers 33, a tagging module 38 and a dominance determination module 39. Acoustic features 31 of the audio are extracted and input to first level classifiers 32 to generate first level classifications for the audio track. In this example, first classifiers 33 and second classifiers 34 are used to generate first and second classifications respectively. In the embodiments to be described below, the first classifiers 33 are non-probabilistic classifiers, while the second classifiers 34 are probabilistic classifiers.

The first and second classifications generated by the first level classifiers 32 are provided as inputs to a second level classifier 35. One or more second level classifications are generated by the second level classifier 35, based at least in part on the first and second classifications. In the embodiments to be described below, the second level classifiers 35 include a third classifier 36, which outputs a third classification.

One or more tags 37 are generated, based on the second level classifications. Such tags 37 may be stored by the tagging module 38 to characterise the audio track in a database, organise or search a database of audio tracks and/or determine a similarity between the audio track and other audio tracks, for example, to select other audio tracks for playback or purchase by a user.

The dominance determination module 39 is configured to calculate dominances 40, 41 of one or more of the characteristics indicated by the tags 37 for the audio track. For a tag 37 based on the inclusion of a musical instrument, its overall dominance indicates how audibly distinguishable the particular instrument is when compared with the other instruments in the mix of the audio track. The dominance may reflect the significance of the role played by the instrument in a musical composition. For example, a leading instrument, such as lead vocal, would be expected to be more audibly distinguishable and, therefore, more dominant than an accompanying instrument, while a solo instrument would be expected to display even greater dominance. For a tag 37 based on a particular musical genre, its dominance relates to the strength or salience of the tag 37 for the audio track to indicate a degree of conformity, that is how closely the audio track conforms, to that particular genre.

The dominance of a tag 37 may be stable over the duration of an audio track or may vary. Hence, the dominances 40, 41 calculated by the determination module 39 include an overall dominance 40, which may be a single value associated with the audio track, and a varying dominance 41, which provides information showing how the dominance of the tag 37 changes over the duration of the audio track. The varying dominance 41 may be used, for example, to identify sections of the audio track dominated by a particular musical instrument, such as a guitar solo in a rock song.

The controller may further act as a recommendation module 42, configured to select further audio tracks from a catalogue for presentation as recommendations 43 for a user, based at least in part on results output by the dominance determination module 39.

The method will now be described in more detail, with reference to Figures 4 to 14. Parts of such a method, relating to extraction of acoustic features and determinations of probabilities, classifications and tags, were discussed in the applicant's co-pending patent application PCT/FI2014/051036, filed on 22 December 2014, the disclosure of which is incorporated herein by reference.

Beginning at s4.0 of Figure 4, if the received input signal is in a compressed format, such as MPEG-1 Audio Layer 3 (MP3), Advanced Audio Coding (AAC) and so on, the input signal is decoded into pulse code modulation (PCM) data (step s4.1). In this particular example, the samples for decoding are taken at a rate of 44.1 kHz and have a resolution of 16 bits.

Next, the software application 212 causes the controller 202 to extract acoustic features 31 or descriptors which indicate characteristics of the audio track (s4.2). In this particular embodiment, the features 31 are based on mel-frequency cepstral coefficients (MFCCs). In other embodiments, other features such as fluctuation pattern and danceability features, beats per minute (BPM) and related features, chorus features and other features may be used instead of, or as well as MFCCs.

An example method for extracting acoustic features 31 from the input signal at s4.2 will now be described, with reference to Figure 5.

Starting at s5.0, the controller 202 may, optionally, resample the decoded input signal at a lower rate, such as 22050 kHz (s5.1).

An optional "pre-emphasis" process is shown as s5.2. Since audio signals conveying music tend to have a large proportion of their energy at low frequencies, the pre-emphasis process filters the decoded input signal to flatten the spectrum of the decoded input signal. The relatively low sensitivity of the human ear to low frequency sounds may be modelled by such flattening. One example of a suitable filter for this purpose is a first-order Finite Impulse Response (FIR) filter with a transfer function of $1-0.98z^{-1}$.

At s5.3, the controller 202 blocks the input signal into frames. The frames may include, for example, 1024 or 2048 samples of the input signal, and the subsequent frames may be overlapping or they may be adjacent to each other according to a hop-size of, for example, 50% and 0%, respectively. In other examples, the frames may be non-adjacent so that only part of the input signal is formed into frames.

Figure 6 depicts an example in which an input signal 50 is divided into blocks to produce adjacent frames of about 30 ms in length which overlap one another by 25%. However, frames of other lengths and/or overlaps may be used.

A Hamming window, such as windows 52a to 52d, is applied to the frames at s5.4, to reduce windowing artifacts. An enlarged portion in Figure 6 depicts a frame 54 following the application of the window 52d to the input signal 50.

At s5.5, a Fast Fourier Transform (FFT) is applied to the windowed signal to produce a magnitude spectrum of the input signal. An example FFT spectrum is shown in Figure 7. Optionally, the FFT magnitudes may be squared to obtain a power spectrum of the signal for use in place of the magnitude spectrum in the following.

The spectrum produced by the FFT at s5.5 may have a greater frequency resolution at high frequencies than is necessary, since the human auditory system is capable of better frequency resolution at lower frequencies but is capable of lower frequency resolution at higher frequencies. So, at s5.6, the spectrum is filtered to simulate non-linear frequency resolution of the human ear.

In this example, the filtering at s5.6 is performed using a filter bank having channels of equal bandwidths on the mel-frequency scale. The mel-frequency scaling may be achieved by setting the channel centre frequencies equidistantly on a mel-frequency scale, given by the Equation (1),

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (1)$$

where f is the frequency in Hertz.

The output of each filtered channel is a sum of the FFT frequency bins belonging to that channel, weighted by a mel-scale frequency response. The weights for filters in an example filter bank are shown in Figure 8. In Figure 8, 40 triangular-shaped bandpass filters are depicted whose center frequencies are evenly spaced on a perceptually motivated mel-frequency scale. The filters may span frequencies from 30 hz to 11025 Hz, in the case of the input signal having a sampling rate of 22050 Hz. For the sake of example, the filter heights in Figure 8 have been scaled to unity.

Variations may be made in the filter bank in other embodiments, such as spanning the band centre frequencies linearly below 1000 Hz, scaling the filters such that they have unit area instead of unity height, varying the number of frequency bands, or changing the range of frequencies spanned by the filters.

The weighted sum of the magnitudes from each of the filter bank channels may be referred to as mel-band energies $\tilde{m}_j$,where $j=1...N$, N being the number of filters.

In s5.7, a logarithm, such as a logarithm of base 10, may be taken from the mel-band energies $\tilde{m}_j$, producing log mel-band energies m_j. An example of a log mel-band energy spectrum is shown in Figure 9.

Next, at s5.8, a Discrete Cosine Transform is applied to a vector of the log mel-band energies m_j to obtain the MFCCs according to Equation (2),

$$c_{mel}(i) = \sum_{j=1}^{N} m_j \cos\left(\frac{\pi \cdot i}{N}\left(j - \frac{1}{2}\right)\right) \qquad (2)$$

where N is the number of filters, $i=0,...,I$ and I is the number of MFCCs. In an exemplary embodiment, $I=20$.

At s5.9, further mathematical operations may be performed on the MFCCs produced at s5.8, such as calculating a mean of the MFCCs and/or time derivatives of the MFCCs to produce the required audio features 31 on which the calculation of the first and second classifications by the first and second classifiers 33, 34 will be based.

In this particular embodiment, the audio features 31 produced at s5.9 include one or more of:

- a MFCC matrix for the audio track;
- first and, optionally, second time derivatives of the MFCCs, also referred to as "delta MFCCs";
- a mean of the MFCCs of the audio track;
- a covariance matrix of the MFCCs of the audio track;
- an average of mel-band energies over the audio track, based on output from the channels of the filter bank obtained in s5.6;
- a standard deviation of the mel-band energies over the audio track;
- an average logarithmic energy over the frames of the audio track, obtained as an average of $c_{mel}(0)$ over a period of time obtained, for example, using Equation (2) at s4.8; and
- a standard deviation of the logarithmic energy.

The extracted features 31 are then output (s5.10) and the feature extraction method ends (s5.11).

As noted above, the features 31 extracted at s4.2 may also include a fluctuation pattern and danceability features for the track, such as:

- a median fluctuation pattern over the song;
- a fluctuation pattern bass feature;
- a fluctuation pattern gravity feature;
- a fluctuation pattern focus feature;
- a fluctuation pattern maximum feature;
- a fluctuation pattern sum feature;
- a fluctuation pattern aggressiveness feature;
- a fluctuation pattern low-frequency domination feature;
- a danceability feature (detrended fluctuation analysis exponent for at least one predetermined time scale);and
- a club-likeness value.

The mel-band energies calculated in s5.8 may be used to calculate one or more of the fluctuation pattern features listed above. In an example method of fluctuation pattern analysis, a sequence of logarithmic domain mel-band magnitude frames are arranged into segments of a desired temporal duration and the number of frequency bands is reduced. A FFT is applied over coefficients of each of the frequency bands across the frames of a segment to compute amplitude modulation frequencies of loudness in a described range, for example, in a range of 1 to 10 Hz. The amplitude modulation frequencies may be weighted and smoothing filters applied. The results of the fluctuation pattern analysis for each segment may take the form of a matrix with rows corresponding to modulation frequencies and columns corresponding to the reduced frequency bands and/or a vector based on those parameters for the segment. The vectors for multiple segments may be averaged to generate a fluctuation pattern vector to describe the audio track.

Danceability features and club-likeness values are related to beat strength, which may be loosely defined as a rhythmic characteristic that allows discrimination between pieces of music, or segments thereof, having the same tempo. Briefly, a piece of music characterised by a higher beat strength would be assumed to exhibit perceptually stronger and more pronounced beats than another piece of music having a lower beat strength. As noted above, a danceability feature may be obtained by detrended fluctuation analysis, which indicates correlations across different time scales, based on the mel-band energies obtained at s5.8.

Examples of techniques of club-likeness analysis, fluctuation pattern analysis and detrended fluctuation analysis are disclosed in British patent application no. 1401626.5, as well as example methods for extracting MFCCs. The disclosure of GB 1401626.5 is incorporated herein by reference in its entirety.

The features 31 extracted at s4.2 may include features relating to tempo in beats per minute (BPM), such as:

- an average of an accent signal in a low, or lowest, frequency band;
- a standard deviation of said accent signal;
- a maximum value of a median or mean of periodicity vectors;
- a sum of values of the median or mean of the periodicity vectors;

- tempo indicator for indicating whether a tempo identified for the input signal is considered constant, or essentially constant, or is considered non-constant, or ambiguous;
- a first BPM estimate and its confidence;
- a second BPM estimate and its confidence;
- a tracked BPM estimate over the audio track and its variation;
- a BPM estimate from a lightweight tempo estimator.

Example techniques for beat tracking, using accent information, are disclosed in US published patent application no. 2007/240558 A1, US patent application no. 14/302,057, International (PCT) published patent application nos. WO2013/164661 A1 and WO2014/001849 A1, the disclosures of which are hereby incorporated by reference in their entireties.

In one example beat tracking method, described in GB 1401626.5, one or more accent signals are derived from the input signal 50, for detection of events and/or changes in the audio track. A first one of the accent signals may be a chroma accent signal based on fundamental frequency F_0 salience estimation, while a second one of the accent signals may be based on a multi-rate filter bank decomposition of the input signal 50.

A BPM estimate may be obtained based on a periodicity analysis for extraction of a sequence of periodicity vectors on the basis of the accent signals, where each periodicity vector includes a plurality of periodicity values, each periodicity value describing the strength of periodicity for a respective period length, or "lag". A point-wise mean or median of the periodicity vectors over time may be used to indicate a single representative periodicity vector over a time period of the audio track. For example, the time period may be over the whole duration of the audio track. Then, an analysis can be performed on the periodicity vector to determine a most likely tempo for the audio track. One example approach comprises performing k-nearest neighbours regression to determine the tempo. In this case, the system is trained with representative music tracks with known tempo. The k-nearest neighours regression is then used to predict the tempo value of the audio track based on the tempi of k-nearest representative tracks. More details of such an approach have been described in Eronen, Klapuri, "Music Tempo Estimation With k -NN Regression", IEEE Transactions on Audio, Speech, and Language

Processing, Vol. 18, Issue 1, pages 50-57, the disclosure of which is incorporated herein by reference.

Chorus related features that may be extracted at s3.2 include:

- a chorus start time; and
- a chorus end time.

Example systems and methods that can be used to detect chorus related features are disclosed in US 2008/236371 A1, the disclosure of which is hereby incorporated by reference in its entirety.

Other features that may be used as additional input include:

- a duration of the audio track in seconds,
- an A-weighted sound pressure level (SPL);
- a standard deviation of the SPL;
- an average brightness, or spectral centroid (SC), of the audio track, calculated as a spectral balancing point of a windowed FFT signal magnitude in frames of, for example, 40 ms in length;
- a standard deviation of the brightness;
- an average low frequency ratio (LFR), calculated as a ratio of energy of the input signal below 100Hz to total energy of the input signal, using a windowed FFT signal magnitude in 40 ms frames; and
- a standard deviation of the low frequency ratio.

Figure 10 is an overview of a process of extracting multiple acoustic features 31, some or all of which may be obtained in s4.2. Figure 10 shows how some input features are derived, at least in part, from computations of other input features. The features 31 shown in Figure 10 include the MFCCs, delta MFCCs and mel-band energies discussed above in relation to Figure 5, indicated using bold text and solid lines.

The dashed lines and standard text in Figure 10 indicate other features that may be extracted and made available alongside, or instead of, the MFCCs, delta MFCCs and mel-band energies, for use in calculating the first level classifications. For example, as discussed above, the mel-band energies may be used to calculate

fluctuation pattern features and/or danceability features through detrended fluctuation analysis. Results of fluctuation pattern analysis and detrended fluctuation analysis may then be used to obtain a club-likeness value. Also as noted above, beat tracking features, labeled as "beat tracking 2" in Figure 10, may be calculated based, in part, on a chroma accent signal from a F_0 salience estimation.

Returning to Figure 4, in s4.3 to s4.10, the software application 212 causes the controller 202 to produce the first level classifications, that is the first classifications and the second classifications, based on the features 31 extracted in s4.2. Although Figure 4 shows s4.3 to s4.10 being performed sequentially, in other embodiments, s4.3 to s4.7 may be performed after, or in parallel with, s4.8 to s4.10.

The first and second classifications are generated using the first classifiers 33 and the second classifiers 34 respectively, where the first and second classifiers 33, 34 are different from one another. For instance, the first classifiers 33 may be non-probabilistic and the second classifiers 34 may be probabilistic classifiers, or vice versa. In this particular embodiment, the first classifiers 33 are support vector machine (SVM) classifiers, which are non-probabilistic. Meanwhile, the second classifiers 34 are based on one or more Gaussian Mixture Models (GMMs).

In s4.3, one, some or all of the features 31 or descriptors extracted in s4.2, to be used to produce the first classifications 33, are selected and, optionally, normalised. For example, a look up table 216 or database may be stored in the memory 206 of the for each of the first classifications to be produced by the analysis server 100, that provides a list of features to be used to generate each first classifier and statistics, such as mean and variance of the selected features, that can be used in normalisation of the extracted features 31. In such an example, the controller 202 retrieves the list of features from the look up table 216, and accordingly selects and normalises the listed features for each of the first classifications to be generated. The normalisation statistics for each first classification in the database may be determined during training of the first classifiers 33.

As noted above, in this example, the first classifiers 33 are SVM classifiers. The SVM classifiers 33 are trained using a database of audio tracks for which information regarding musical instruments and genre is already available. The

database may include tens of thousands of tracks for each particular musical instrument that might be tagged.

Examples of musical instruments for which information may be provided in the database include:

- Accordion;
- Acoustic guitar;
- Backing vocals;
- Banjo;
- Bass synthesizer;
- Brass instruments;
- Glockenspiel;
- Drums;
- Eggs;
- Electric guitar;
- Electric piano;
- Guitar synthesizer;
- Keyboards;
- Lead vocals;
- Organ;
- Percussion;
- Piano;
- Saxophone;
- Stringed instruments;
- Synthesizer; and
- Woodwind instruments.

The training database includes indications of genres that the audio tracks belong to, as well as indications of genres that the audio tracks do not belong to. Examples of musical genres that may be indicated in the database include:

- Ambient and new age;
- Blues;
- Classical;
- Country and western;
- Dance;
- Easy listening;
- Electronica;
- Folk and roots;
- Indie and alternative;
- Jazz;
- Latin;
- Metal;
- Pop;
- Rap and hip hop;
- Reggae;
- Rock;
- Soul, R&B and funk; and
- World music.

By analysing features 31 extracted from the audio tracks in the training database, for which instruments and/or genre are known, a SVM classifier 33 can be trained to determine whether or not an audio track includes a particular instrument, for example, an electric guitar. Similarly, another SVM classifier 33 can be trained to determine whether or not the audio track belongs to a particular genre, such as Metal.

In this embodiment, the training database provides a highly imbalanced selection of audio tracks, in that a set of tracks for training a given SVM classifier 33 includes many more positive examples than negative ones. In other words, for training a SVM classifier 33 to detect the presence of a particular instrument, a set of audio tracks for training in which the number of tracks that include that instrument is significantly greater than the number of tracks that do not include that instrument will be used. Similarly, in an example where a SVM classifier 33 is being trained to

determine whether an audio track belongs to a particular genre, the set of audio tracks for training might be selected so that the number of tracks that belong to that genre is significantly greater than the number of tracks that do not belong to that genre.

An error cost may be assigned to the different first classifications 33 to take account of the imbalances in the training sets. For example, if a minority class of the training set for a particular first classification includes 400 songs and an associated majority class contains 10,000 tracks, an error cost of 1 may be assigned to the minority set and an error cost of 400/10,000 may be assigned to the majority class. This allows all of the training data to be retained, instead of downsampling data of the negative examples.

New SVM classifiers can be added by collecting new training data and training the new classifiers. Since the SVM classifiers 33 are binary, new classifiers can be added alongside existing classifiers.

As mentioned above, the training process can include determining a selection of one or more features 31 to be used as a basis for particular first classifications and statistics for normalising those features 31. The number of features available for selection, M, may be much greater than the number of features selected for determining a particular first classification, N; that is, $M >> N$. The selection of features 31 to be used is determined iteratively, based on a development set of audio tracks for which the relevant instrument or genre information is available, as follows.

Firstly, to reduce redundancy, a check is made as to whether two or more of the features are so highly correlated that the inclusion of more than one of those features would not be beneficial. For example, pairwise correlation coefficients may be calculated for pairs of the available features and, if it is found that two of the features have a correlation coefficient that is larger than 0.9, then only one of that pair of features is considered available for selection.

The feature selection training starts using an initial selection of features, such as the average MFCCs for audio tracks in the development set or a single “best” feature for a given first classification. For instance, a feature that yields the largest

classification accuracy when used individually may be selected as the "best" feature and used as the sole feature in an initial feature selection.

An accuracy of the first classification based on the initial feature selection is determined. Further features are then added to the feature selection to determine whether or not the accuracy of the first classification is improved by their inclusion.

Features to be tested for addition to the selection of features may be chosen using a method that combines forward feature selection and backward feature selection in a sequential floating feature selection. Such feature selection may be performed during the training stage, by evaluating the classification accuracy on a portion of the training set.

In each iteration, each of the features available for selection is added to the existing feature selection in turn, and the accuracy of the first classification with each additional feature is determined. The feature selection is then updated to include the feature that, when added to the feature selection, provided the largest increase in the classification accuracy for the development set.

After a feature is added to the feature selection, the accuracy of the first classification is reassessed, by generating first classifications based on edited features selections in which each of the features in the feature selection is omitted in turn. If it is found that the omission of one or more features provides an improvement in classification accuracy, then the feature that, when omitted, leads to the biggest improvement in classification accuracy is removed from the feature selection. If no improvements are found when any of the existing features are left out, but the classification accuracy does not change when a particular feature is omitted, that feature may also be removed from the feature selection in order to reduce redundancy.

The iterative process of adding and removing features to and from the feature selection continues until the addition of a further feature no longer provides a significant improvement in the accuracy of the first classification. For example, if the improvement in accuracy falls below a given percentage, the iterative process may be considered complete, and the current selection of features is stored in the lookup table 216, for use in selecting features in s4.3.

The selected features 31 may be normalised, for example, by subtracting a mean value for the feature and normalising the standard deviation. However, it is noted that the normalisation of the selected features 31 at s4.3 is optional. Where provided, the normalisation of the selected features 31 in s4.3 may potentially improve the accuracy of the first classifications. Where normalisation is used, the features may be normalised before or after the selection is performed.

In another embodiment, at s4.3, a linear feature transform may be applied to the available features 31 extracted in s4.2, instead of performing the feature selection procedure described above. For example, a Partial Least Squares Discriminant Analysis (PLS-DA) may be used to obtain a linear combination of features for calculating a corresponding first classification. Instead of using the above iterative process to select N features from the set of M features, a linear feature transform is applied to an initial high-dimensional set of features to arrive at a smaller set of features which provides a good discrimination between classes. The initial set of features may include some or all of the available features, such as those shown in Figure 10, from which a reduced set of features can be selected based on the result of the transform.

The PLS-DA transform parameters may be optimized and stored in a training stage. During the training stage, the transform parameters and its dimensionality may be optimized for each tag or output classification, such as an indication of an instrument or a genre. More specifically, the training of the system parameters can be done in a cross-validation manner, for example, as five-fold cross-validation, where all the available data is divided into five non-overlapping sets. At each fold, one of the sets is held out for evaluation and the four remaining sets are used for training. Furthermore, the division of folds may be specific for each tag or classification.

For each fold and each tag or classification, the training set is split into 50%-50% inner training-test folds. Then, the PLS-DA transform may be trained on the inner training-test folds and the SVM classifier 33 may be trained on the obtained dimensions. The accuracy of the SVM classifier 33 using the transformed features transformed may be evaluated on the inner test fold. It is noted that, when a feature vector (track) is tested, it is subjected to the same PLS-DA transform, the parameters of which were obtained during training. This manner, an optimal dimensionality for the PLS-DA transform may be selected. For example, the

dimensionality may be selected such that the area under the receiver operating characteristic (ROC) curve is maximized. In one example embodiment, an optimal dimensionality is selected among candidates between 5 to 40 dimensions. Hence, the PLS-DA transform is trained on the whole of the training set, using the optimal number of dimensions, and then used in training the SVM classifier 33.

In the following, an example is discussed in which the selected features 31 on which the first classifications are based are the mean of the MFCCs of the audio track and the covariance matrix of the MFCCs of the audio track, although in other examples alternative and/or additional features, such as the other features shown in Figure 10, may be used.

At s4.4, the controller 202 defines a single "feature vector" for each set of selected features 31 or selected combination of features 31.

The feature vectors may then be normalized to have a zero mean and a variance of 1, based on statistics determined and stored during the training process.

At s4.5, the controller 202 generates one or more first probabilities that the audio track has a certain characteristic, corresponding to a potential tag 37, based on the normalized transformed feature vector or vectors. A first classifier 33 is used to calculate a respective probability for each feature vector defined in s4.4. In this manner, the number of SVM classifiers 33 corresponds to the number of characteristics or tags 37 to be predicted for the audio track.

In this particular example, a probability is generated for each instrument tag and for each musical genre tag to be predicted for the audio track, based on the mean MFCCs and the MFCC covariance matrix. In other embodiments, the controller may generate only one or some of these probabilities and/or calculate additional probabilities at 4.5. The different classifications may be based on respective selections of features from the available features 31 extracted in s4.2.

The SVM classifiers 33 may use a radial basis function (RBF) kernel K, defined as:

$$K(\vec{u},\vec{v}) = e^{-\gamma\left(\|\vec{u}-\vec{v}\|^2\right)} \qquad (3)$$

where the default γ parameter is the reciprocal of the number of features in the feature vector, $\vec{u}$ is the input feature vector and $\vec{v}$ is a support vector.

The first classifications may be based on an optimal predicted probability threshold that separates a positive prediction from a negative prediction for a particular tag, based on the probabilities output by the SVM classifiers 33. The setting of an optimal predicted probability threshold may be learned in the training procedure to be described later below. Where there is no imbalance in data used to train the first classifiers 33, the optimal predicted probability threshold may be 0.5. However, where there is an imbalance between the number of tracks providing positive examples and the number of tracks provided negative examples in the training sets used to train the first classifiers 33, the threshold p_{thr} may be set to a prior probability of a minority class P_{min} in the first classification, using Equation (4) as follows:

$$p_{thr} = P_{\min} = \frac{n_{\min}}{n_{maj}} \qquad (4)$$

where, in the set of n tracks used to train the SVM classifiers, n_{min} is the number of tracks in the minority class and n_{maj} is the number of tracks in a majority class.

The prior probability P_{min} may be learned as part of the training of the SVM classifier 33.

Probability distributions for examples of possible first classifications, based on an evaluation of a number n of tracks, are shown in Figure 11. The nine examples in Figure 11 suggest a correspondence between a prior probability for a given first classification and its probability distribution based on the n tracks. Such a correspondence is particularly marked where the SVM classifier 33 was trained with an imbalanced training set of tracks. Consequently, the predicted probability threshold for the different examples vary over a considerable range.

Optionally, a logarithmic transformation may be applied to the probabilities output by the SVM classifiers 33 (s4.6), so that the probabilities of all the first classifications are on the same scale and the optimal predicated probability threshold may correspond to a predetermined value, such as 0.5.

Equations (5) to (7) below provide an example normalization which adjusts the optimal predicted probability threshold to 0.5. Where the probability output by a SVM classifier 33 is p and the prior probability P of a particular tag being applicable to a track is greater than 0.5, then the normalized probability p_{norm} is given by:

$$p_{norm} = 1 - (1 - p)^L \tag{5}$$

$$\text{where } L = \frac{\log(0.5)}{\log(1 - P)} \tag{6}$$

Meanwhile, where the prior probability P is less than or equal to 0.5, then the normalised probability p_{norm} is given by:

$$p_{norm} = p^{L'} \tag{7}$$

$$\text{where } L' = \frac{\log(0.5)}{\log(P)} \tag{8}$$

Figure 12 depicts the example probability distributions of Figure 11 after a logarithmic transformation has been applied, on which optimal predicated probability thresholds of 0.5 are marked.

The first classifications are then output (s4.7). The first classifications correspond to the normalised probability p_{norm} that a respective one of the tags 37 to be considered applies to the audio track. The first classifications may include probabilities p_{inst1} that a particular instrument is included in the audio track and probabilities p_{gen1} that the audio track belongs to a particular genre.

Returning to Figure 4, in s4.8 to s4.10, second classifications for the input signal are determined based on the MFCCs and other parameters produced in s4.2, using the second classifiers 34. In this particular example, the features 31 on which the second classifications are based are per-frame MFCC feature vectors for the audio track and their first and second time derivatives.

In s4.8 to s4.10, the probabilities of the audio track including a particular instrument or belonging to a particular genre are assessed using probabilistic models that have been trained to represent the distribution of features extracted from audio signals captured from each instrument or genre. As noted above, in this example the probabilistic models are GMMs. Such models can be trained using an expectation maximisation algorithm that iteratively adjusts the model parameters to maximise the likelihood of the model for a particular instrument or genre generating features matching one or more input features in the captured audio signals for that instrument or genre. The parameters of the trained probabilistic models may be stored in a database, for example, in the database 208 if the analysis server 100, or in remote storage that is accessible to the analysis server 100 via a network, such as the network 102.

For each instrument or genre, at least one likelihood is evaluated that the respective probabilistic model could have generated the selected or transformed features from the input signal. The second classifications correspond to the models which have the largest likelihood of having generated the features of the input signal.

In this example, probabilities are generated for each instrument tag at s4.8 and for each musical genre tag at s4.9. In other embodiments, the controller 202 may generate only one or some of these second classifications and/or calculate additional second classifications at s4.8 and s4.9.

In this embodiment, in s4.8 and s4.9, probabilities p_{inst2} that the instrument tags will apply, or not apply, are produced by the second classifiers 34 using first and second Gaussian Mixture Models (GMMs), based on the MFCCs and their first time derivatives calculated in s4.2. Meanwhile, probabilities p_{gen2} that the audio track belongs to a particular musical genre are produced by the second classifiers 34 using third GMMs. However, the first and second GMMs used to compute the instrument-based probabilities p_{inst2} may be trained and used slightly differently from third GMMs used to compute the genre-based probabilities $_{gen2}$, as will now be explained.

In the following, s4.8 precedes s4.9. However, in other embodiments, s4.9 may be performed before, or in parallel with, s4.8.

In this particular example, first and second GMMs are used to generate the instrument-based probabilities p_{inst2} (s4.8),based on MFCC features 31 obtained in s4.2.

The first and second GMMs used in s4.8 may have been trained with an Expectation-Maximisation (EM) algorithm, using a training set of examples which are known either to include the instrument and examples which are known to not include the instrument. For each track in the training set, MFCC feature vectors and their corresponding first time derivatives are computed. The MFCC feature vectors for the examples in the training set that contain the instrument are used to train a first GMM for that instrument, while the MFCC feature vectors for the examples that do not contain the instrument are used to train a second GMM for that instrument. In this manner, for each instrument to be tagged, two GMMs are produced. The first GMM is for a track that includes the instrument, while the second GMM is for a track that does not include the instrument. In this example, the first and second GMMs each contain 64 component Gaussians.

The first and second GMMs may then be refined by discriminative training using a maximum mutual information (MMI) criterion on a balanced training set where, for each instrument to be tagged, the number of example tracks that contain the instrument is equal to the number of example tracks that do not contain the instrument.

Returning to the determination of the second classifications, two likelihoods are computed based on the first and second GMMs and the MFCCs for the audio track. The first is a likelihood that the corresponding instrument tag applies to the track, referred to as L_{yes}, while the second is a likelihood that the instrument tag does not apply to the track, referred to as L_{no}. The first and second likelihoods may be computed in a log-domain, and then converted to a linear domain.

In this particular embodiment, the first and second likelihoods L_{yes}, L_{no} are assessed for one or more temporal segments, or frames, of the audio track. The duration of a segment may be set at a fixed value, such as 5 seconds. In one example, where a sampling rate of 44100 Hz and an analysis segment length of 1024 samples for the first and second GMMs is used, a 5 second segment would contain 215 likelihood samples over which average likelihoods L_{yes}, L_{no} and, optionally, their standard deviation for that segment can be calculated.

Alternatively, the duration of a segment may be set to correspond to the tempo or bar times of the audio track. For example, the length of a bar may be determined, for example from tempo-related metadata for the audio track, and the segment length set to the duration of one bar. In other examples, the segment length may be set to a duration of multiple bars.

The first and second likelihoods L_{yes}, L_{no} are then mapped to a probability p_{inst2} of the tag applying. An example mapping is as follows:

$$p_{inst2} = \frac{\overline{L}_{yes}}{\left(\overline{L}_{yes} + \overline{L}_{no}\right)} \qquad (9)$$

where $\overline{L}_{yes}$ and $\overline{L}_{no}$ are averages of the first and second likelihoods L_{yes}, L_{no} of the analysed segments of the audio track. In another example, a sum of the first and second likelihoods L_{yes}, L_{no} for the analysed segments of the audio track might be used in Equation (9), instead of the averages $\overline{L}_{yes}$ and $\overline{L}_{no}$.

As noted above, the third GMMs, used for genre-based classification, are trained differently to the first and second GMMs. For each genre to be considered, a third GMM is trained based on MFCCs for a training set of tracks known to belong to that genre. One third GMM is produced for each genre to be considered. In this example, the third GMM includes 64 component Gaussians.

In s4.9, for each of the genres that may be tagged, a likelihood L is computed for the audio track belonging to that genre, based on the likelihood of each of the third GMMs being capable of outputting the MFCC feature vector of the audio track or, alternatively, the MFCC feature vector of a segment of the audio track. For example, to determine which of the eighteen genres in the list hereinabove might apply to the audio track, eighteen likelihoods would be produced.

The genre likelihoods are then mapped to probabilities p_{gen2}, as follows:

$$p_{gen2}(i) = \frac{L(i)}{\sum_{j=1}^{m} L(j)} \qquad (10)$$

where m is the number of genre tags to be considered.

The second classifications, which correspond to the probabilities p_{inst2} and p_{gen2}, are then output (s4.10).

In another embodiment, the first and second GMMs for analysing the instruments included in the audio track may be trained and used in the manner described above for the third GMMs. In yet further embodiments, the GMMs used for analysing genre may be trained and used in the same manner, using either of techniques described in relation to the first, second and third GMMs above.

The first classifications p_{inst1} and p_{gen1} and the second classifications p_{inst2} and p_{gen2} for the audio track are normalized to have a mean of zero and a variance of 1 (s4.11) and collected to form a feature vector for input to one or more second level classifiers 35 (s4.12). In this particular example, the second level classifiers 35 include third classifiers 36. The third classifiers 36 may be non-probabilistic classifiers, such as SVM classifiers.

The third classifiers 36 may be trained in a similar manner to that described above in relation to the first classifiers 33. At the training stage, the first classifiers 33 and the second classifiers 34 may be used to output probabilities for the training sets of example audio tracks from the database. The outputs from the first and second classifiers 33, 34 are then used as input data to train the third classifier 35.

The third classifier 36 generates determine probabilities p_{inst3} for whether the audio track contains a particular instrument and/or probabilities p_{gen3} for whether the audio track belongs to a particular genre (s4.13).

The probabilities p_{inst3} , p_{gen3} are then log normalised (s4.14), as described above in relation to the first classifications, so that a threshold of 0.5 may be applied to generate the third classifications, which are output at s4.15.

The controller 202 then determines whether each instrument tag and each genre tag 37 applies to the audio track based on the third classifications (s4.16).

Where it is determined that an instrument or genre tag 37 applies to the audio track (s4.16), the tag 37 is associated with the track (s4.17), for example, by storing an indication that the tag 37 applies as part of metadata for the audio track. Alternatively, or additionally, the probabilities themselves and/or the features 31 extracted at s4.2 may be output for further analysis and/or storage.

The controller 202 determines and outputs the overall dominance 40 and the varying dominance 41 of one or more of the tags 37 for the audio track (s4.18 to s4.20). It is noted that, while Figure 4 shows s4.18 to s4.20 being performed after the output of the second classifications (s4.10), the determination of the third classifications and tags 37 (s4.11 to s4.16) and the tagging of the audio track (s4.17), the dominances 40, 41 may be determined before, or in parallel, with some or all of s4.10 to s4.17.

Example methods for determining the overall dominance 40 and varying dominance 41 for a tag 37 will now be explained with reference to Figures 13 and 14 respectively. In this particular embodiment, dominance is expressed using numerical values between 0 and 5, where 0 indicates a relatively low dominance and 5 indicates that a characteristic is highly dominant. However, in other embodiments, other scales or values may be used to indicate dominance.

The overall dominance 40 is assessed using an overall dominance model trained to predict an overall dominance value based on acoustic features 31 extracted from an audio track and the probabilities p_{inst3}, p_{gen3} calculated by the third classifiers 36 of Figure 3 extracted from an audio track. The overall dominance model is created and trained using a plurality of T_1 training audio tracks for which dominance for different characteristics, such as instruments and/or genres, are known. For example, the training audio tracks may be music tracks for which one or more listeners have assessed the dominance of particular musical instruments and/or genres and provided annotations indicating the assessed dominances accordingly. The number T_1 of training audio tracks might be of the order of a few thousand. The T_1 training audio tracks may be selected to include a minimum of one hundred tracks, or a few hundred tracks, for each musical instrument or genre corresponding to a tag 37. In general, is the availability of a larger number T_1 of training audio tracks made available allows the model to be trained with greater accuracy.

In the training process, acoustic features are extracted from the training audio tracks in a similar manner to that described with reference to Figure 5 and probabilities p_{inst3}, p_{gen3} for each instrument and genre are generated as described with reference to s4.3 to s4.14 of Figure 4.

For each of the T_1 training audio tracks, selected acoustic features and the relevant probabilities p_{inst3} or p_{gen3} are concatenated to create a feature vector for estimating the dominance of a particular musical instrument or genre. Pairwise correlation coefficients for pairs of the extracted features are calculated. If a correlation coefficient indicates a high level of correlation between two features, for example if the correlation coefficient is greater than 0.9, then only one of the pair of features remains available for selection, in order to avoid redundancy.
The respective feature vector $x_1...x_{T1}$ for each of the T_1 training audio tracks are then created, based on the selected features corresponding to the particular instrument or genre. A T_1 *x* d matrix that includes the feature vectors $x_1...x_{T1}$ for the training audio tracks is compiled, where d is the dimension of the feature vectors. At this stage, the dimension d may be, for example, 250.

The matrix is normalised so that the values in each row have a mean of zero and a variance of unity. The mean and the standard deviation vectors used to normalise the rows of the matrix are stored in the memory 206 for later use when analysing new audio tracks.

Even after the removal of correlated features, the number of features in the feature vectors may be large. To reduce computing requirements, a subset of Q features is selected to form a basis for the model for assessing the overall dominance.

In this particular example, the Q features are selected using univariate linear regression tests, in which the "regressors" are column vectors based on the columns of the T_1 *x* d matrix after normalisation, corresponding to extracted acoustic features and the probabilities p_{inst3} or p_{gen3} corresponding to a particular tag of the T_1 training audio tracks, and the "data" are the dominances provided in the annotations for the training audio tracks. For each of the regressors, the following is performed.

A cross-correlation coefficient for one of the regressors, a "regressor of interest", and the data is computed. The cross-correlation coefficient is then converted to an

F-score, indicating the predictive capability of the cross-correlation, and then to a p-value, indicating its statistical significance.

Q features are then selected, based on the F-scores and p-values for the respective regressors. The value of Q may vary according to the dominance model that is used and a suitable value for Q may be determined as part of the training procedure. For example, regressors may be trained on a subset of the T training audio tracks, their performance assessed using the remaining training audio tracks and the number of features leading to the minimum mean-absolute-error (MAE) selected as Q. Typically, the number Q of features in the subset will be between 1 and 30 for each instrument or genre.

The overall dominance model is then trained using the determined number Q of selected features and the probability p_{inst3} or p_{gen3} corresponding to the relevant instrument or genre. In one particular example, ordinary least squares regression is used to predict dominance, for example using Equation (11) as follows:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_Q x_Q + \mathrm{A} \qquad (11)$$

where $\beta_1 \ldots \beta_Q$ are the regression coefficients and A is an intercept corresponding to a particular instrument or genre.

For each instrument or genre, certain parameters and data regarding the regression are stored in the memory 206, for use in later analysis of audio tracks. In this particular example, where linear regression is used, the stored data may include the indices of the Q selected features, together with the corresponding regression coefficients $\beta_1 \ldots \beta_Q$ and intercept A for the particular instrument or genre.

In other examples, another technique may be used instead of the least squares regression discussed above. Examples of alternatives for least squares regression include epsilon support vector machine (SVM) regression, as discussed in Smola, A.J. and Scholkopf, B., "A tutorial on support vector regression", Statistics and Computing, 2004, vol 14, pages 199–222, 2004, and support vector ordinal regression, described Chu W. and Keerthi, S.S., "New approaches to support vector ordinal regression", in Proceedings of the 22nd International Conference on Machine Learning (ICML-22), 2005, pages 145-152. Where epsilon support vector

machine regression or support vector ordinal regression is used, the dominance may be predicted using Equation (12) in place of Equation (11), as follows:

$$\sum_{i=1}^{s} \alpha_i K(\vec{x}, \vec{x}_i) + b \qquad (12)$$

where K is a kernel function, such as the RBF kernel in Equation (3) above, α_i, i=1,..., s are weights, b is a constant offset, and $\vec{x}_i$ are support vectors.

Moreover, it is not necessary for the the same regression method to be used for training the overall dominance models for different instruments and genres. In other embodiments, the regression method used for a particular instrument or genre can be selected based on the performance of the different regression methods on validations performed on the T_1 training audio tracks. For example, for each value of Q to be evaluated, multiple models may be trained, such as a linear regression model, an epsilon SVM regression model using a radial basis function kernel, and a support vector ordinal regression model with explicit constraints by Chu and Keerthi (2005, cited above) on a subset of the T_1 training audio tracks, and their performance is assessed using the remaining T_1 training audio tracks by evaluating the mean-absolute-error (MAE) between the data and predictions. For each value of Q, the regressor leading to the smallest MAE is selected. Hence, in this example, the dominance of different instruments and/or genres may be determined using different regression methods.

Other examples of regression methods that may be used for determining dominance include random forest regression, neural networks, polynomial regression, general linear models, logistic regression, probit regression, nonlinear regression, principal components analysis, ridge regression, Lasso regression, and so on.

Where such other regression techniques are used, the parameters and data stored in the memory 206 for use in later analysis of audio tracks may differ from those noted above. For example, if an epsilon support vector machine regression or support vector ordinal regression is to be used, their respective parameters, such as support vectors $\vec{x}_i$, a RBF kernel width parameter γ, weights α_i and an offset b may be stored.

Figure 13 depicts an example method of determining the overall dominance 40 at s4.18, using the overall dominance model.

Starting at s13.0, the regression parameters and data stored in the memory 206 for an instrument or genre corresponding to a tag 37 of the audio track are retrieved from the memory 206 (s13.1). In this example, where linear regression is used, the retrieved data includes parameters Q, A, indices of features to be selected and regression coefficients $\beta_1 \ldots \beta_Q$.

The Q features indicated by the retrieved indices are then selected from the audio features 31 extracted from the audio track at s4.2 (s13.2) and normalised (s13.3).

The overall dominance 40 is then calculated, using the retrieved coefficients $\beta_1 \ldots \beta_Q$, the intercept A and the probability p_{inst3} or p_{gen3} corresponding to the instrument or genre being assessed, as calculated by the third classifiers 36 at s4.13 to s4.14 above (s13.4). In this example, where linear regression is used, the dominance is calculated using Equation (11) above. Where epsilon support vector machine regression or support vector ordinal regression is used, the dominance may be calculated using Equation (12) above.

In this particular example, if the overall dominance 40 exceeds a threshold (s13.5), such as 0.5, then it is stored as metadata for the audio track (s13.6). Alternatively, in another embodiment, such a threshold may be omitted and the overall dominance 40 stored at s13.6 regardless of its value.

The procedure for determining the overall dominance 40 is then complete (s13.7).

The varying dominance 41 is assessed using a varying dominance model trained using a plurality of T_2 training audio tracks for which varying dominance values are available. A suitable value for T_2 is at least one hundred, however the model may be trained more accurately if at least a few hundred training audio tracks are provided with varying dominance information for each musical instrument.

The T_2 training audio tracks may be music tracks for which one or more listeners have assessed the dominance of particular musical instruments over one or more time intervals within the music tracks and provided annotations indicating the

assessed dominances for that segment of the music track accordingly. The annotations may indicate one or more first time points or intervals with a relatively low dominance value for a particular musical instrument and one or more other points or time second intervals with a relatively high dominance value for that instrument when compared with the first time points or intervals. While it may be possible to provide annotations for time intervals covering an entire duration of a training audio track, it is not necessary to do so.

Additionally, or alternatively, the T_2 training audio tracks may include music tracks for which annotated dominance information provides only overall dominance values. In some embodiments, the T_2 training audio tracks may be the same as, or may include, the T_1 training audio tracks used to train the overall dominance model.

In the training process, acoustic features are extracted from samples of the training audio tracks and MFCCs are computed in a similar manner to that described with reference to Figure 5. For each musical instrument to be assessed, two likelihoods are computed based on first and second GMMs and the MFCCs for each sample. The first is a likelihood that a particular musical instrument contributes to the sample, referred to as L_{yes}, while the second is a likelihood that the instrument tag does not contribute to the sample, referred to as L_{no}.

The first and second GMMs may be the same as the first and second GMMs trained for use in the second classifiers 34 and the first and second likelihoods L_{yes}, L_{no} may be calculated in the same manner described hereinabove.

Where annotated dominance information has been provided for separate segments of a training audio track, averages of the likelihoods L_{yes}, L_{no} and their standard deviation for each musical instrument in each segment are calculated. If only overall dominance information is available for a training audio track, the averages of the likelihoods L_{yes}, L_{no} and their standard deviation may be calculated over the entire duration of the training audio track.

In this particular example, the varying dominance model is a linear regression model, trained using a least squares criterion. Alternatively, or in addition, to linear regression, the model could use support vector machine regression or one of the other regression techniques mentioned above in relation to the overall

dominance model. The selection of which regression technique to use for assessing varying dominance of a particular musical instrument can be made using cross validation experiments on the T_2 training audio tracks. In such experiments, a subset of the T_2 training audio tracks are used to train regressors with different parameters and their accuracy in predicting the dominance of a particular musical instrument is evaluated using, for example the MAE criterion, on other ones of the T_2 training audio tracks that were not included in the subset. The regression model and parameters which provide the best prediction accuracy on the other T_2 training audio tracks may then be selected as the technique to be used for assessing varying dominance of that particular musical instrument.

The selection of the inputs to the varying dominance model is determined through univariate linear regression tests, in a similar manner to the selection of the Q features for the overall dominance model discussed above. In this particular example, the likelihoods L_{yes}, L_{no} of all the musical instruments to be evaluated are used as initial input, and the regressors are selected from these likelihoods.

The varying dominance model is then trained using the selected inputs, for example using Equation (11) or (12) above. For each instrument, the parameters and data used for the regression analysis are stored in the memory 206 for use in analysing further audio tracks. If linear regression is used, the stored parameters and data may include the number and indices of the selected inputs, together with the corresponding regression coefficients and intercept. If a support vector machine regression model is used, the parameters and data include support vectors, weights, the offset, and kernel parameters.

Figure 14 depicts an example method of determining the varying dominance 41 at s4.19, using the varying dominance model, starting at s14.0.

In this particular embodiment, the inputs to the varying dominance model include likelihoods L_{yes}, L_{no} that multiple segments of the audio track include a particular musical instrument. In embodiments where such likelihoods L_{yes}, L_{no} and, optionally, averages of those likelihoods are calculated when the probabilities p_{inst2}, are determined by the second classifiers 34 at s4.9, the first and second likelihoods L_{yes}, L_{no} and, where available, their averages as determined by the second classifiers 34 may be used in determining the varying dominance. However, for the sake of

completeness, a method of calculating the first and second likelihoods L_{yes}, L_{no} and their averages will now be described, with reference to s14.1 to s14.5.

Optionally, if the likelihoods L_{yes}, L_{no} are to be assessed over one or more temporal segments of the audio track, a segment length is set at s14.1. As discussed above in relation to s4.8, the duration of a segment may be set at a fixed value, such as 5 seconds. However, in some embodiments, the duration of a segment may be set to correspond to the tempo or bar times of the audio track. For example, the length of a bar may be determined, for example from tempo-related metadata for the audio track, and the segment length set to the duration of one bar. In other examples, the segment length may be set to a duration of multiple bars.

Acoustic features 31 are then extracted from the segment (s14.2), in a similar manner to that shown in Figure 5. In this example, the acoustic features are MFCCs and their first order time derivatives.

For a particular musical instrument corresponding to a tag 37 of the audio track, the number and indices of inputs to be selected for the varying dominance model, the corresponding regression coefficients and intercept are retrieved from the memory 206 (s14.3).

For each sample within the segment, a first likelihood L_{yes} that the sample includes the musical instrument is computed (s14.4) using the first GMM and the MFCCs and their first-order time-derivatives. A second likelihood L_{no} that the sample does not include the musical instrument is computed (s14.5) using the second GMM and the MFCCs and their first-order time-derivatives.

Respective averages and standard deviations for the first and second likelihoods L_{yes}, L_{no} over the duration of the segment are calculated or, if already available from the calculation of the second classifications, otherwise obtained (s14.6).

The varying dominance 41 for that instrument in that segment is then calculated using the varying dominance model and the inputs identified in s14.1 (s14.7), and then stored (s14.8). In this example, the varying dominance 41 is expressed as a value between 0 and 5.

If the dominance of another instrument is to be evaluated for that segment (s14.9), s14.3 to s14.8 are then repeated for the next instrument.

When the dominance of all of the instruments to be assessed for the segment has been determined (s14.10), the next segment is analysed by repeating s14.1 to s14.10 for the next segment.

Once all of the segments have been analysed (s14.10), the procedure ends (s14.11).

Figure 15 depicts varying dominance information for an example audio track. The solid line depicts the varying dominance 41 of electric guitar in the audio track. The dashed line depicts the varying dominance 41 of a vocals in the same audio track, while the dotted line shows an average of the dominance 41 of the other instruments, which include bass guitar and drums. In this example, the electric guitar is dominant in the beginning of the audio track. The vocals begin at around 30 seconds, which is reflected an increase in the vocals dominance value at that time point. As another example, at around 120 seconds, a section begins during which vocals dominate and the electric guitar is somewhat quieter in the background. This is reflected by an increase in the vocals dominance and a drop in the electric guitar dominance at that time point.

Returning to Figure 4, further features may, optionally, be calculated and stored based on the varying dominance 41 (s4.20). Such features may include dominance difference, based on the difference between the varying dominance 41 for a particular musical instrument and one or more other musical instruments. Figure 16 shows the difference between the dominance of the electric guitar and the average dominance of the other instruments in the example audio track discussed previously with reference to Figure 15. The change in dominance of the electric guitar at 30 seconds and 45 seconds, noted hereinabove, is reflected by the changes shown at those time points in Figure 16.

Other dominance-related features that may be calculated and stored at s4.20 instead of, or as well as, dominance difference include dominance change frequency and dominance section duration.

Dominance change frequency indicates how frequently dominance changes and may be calculated, for example, using a periodicity analysis in which a Fast-Fourier

Transform (FFT) is applied to the varying dominance 41 to determine a frequency and, optionally, amplitude, of a strongest dominance change frequency. Alternatively, the controller 202 may detect when the varying dominance 41 crosses an average dominance level, using a mean number of crossings in a time period and, optionally, derivatives of the varying dominance 41, to calculate a dominance change frequency. Instead of using the varying dominance 41, either of these methods may instead use the dominance difference. For example, such a periodicity analysis may be performed on the dominance difference, or the mean number of instances where the dominance difference crosses a zero level in a time period may be used to calculate a dominance change frequency.

Dominance section duration relates to the duration of sections of the audio track in which a particular musical instrument exhibits a strong dominance, for example, to the average dominance, or dominance difference, of that instrument over the duration of the audio track. To calculate the dominance section duration, the controller 202 detects the sections in which the particular musical instrument has a strong dominance or dominance difference, determines the average duration of those sections and, optionally, the variation in their durations.

While the above example relates to sections in which a particular musical instrument exhibits strong dominance, domination section duration may be based on sections in which the instrument exhibits a weak dominance. In other examples, the dominance of the particular musical instrument may be compared with a fixed threshold, or an adaptive threshold based on, for example, a running average, or with an average dominance of other instruments in the audio track, to determine whether its own dominance is strong or weak.

Optionally, if a recommendation of other audio tracks based on the analysed audio track is required (s4.21), for example if a request has been received indicating that the user wishes to find music that has a similarity to the analysed audio track, the controller 202 may then search for one or more similar audio tracks in an existing catalogue (s4.22). The catalogue may be a database stored in the memory of the analysis server 100 or accessible via the network 102 or other network.

Figure 17 shows an example user interface 60, that may be presented by the display 106, through which the user can provide input on which recommendations can be based. In this example, one or more sliders 61, 62 are provided to allow the user to

indicate additional preferences for the type of music tracks to be recommended. In this example, sliders 61 are provided for indicating instrument-based preferences and sliders 62 are provided for indicating music genre-based preferences. While Figure 17 depicts sliders 61, 62, in other embodiments, alternative input techniques for obtaining user preferences may be used, such as numerical values indicating relative importance or rankings for the preferences or input arranging the preferences in order of importance to the user. Where a user indicates a preference for a particular musical instrument, the controller 202 may search for music tracks in which that particular musical instrument has a strong overall dominance 40 and/or varying dominance 41.

In this particular embodiment, where the user indicates a preference for a particular musical instrument, a further slider 63 may be displayed to allow the user to indicate a preferred duration of dominance sections for a particular instrument. For example, the user may wish to search for music tracks with extended sections where electric guitar is dominant, and may select a minimum duration threshold of, perhaps, 90 seconds. In this particular example, a counter 64 is displayed beneath the slider 63 to indicate the currently indicated duration.

At s4.22, the controller 202 searches the catalogue for music tracks. The search may be based on similarity of tags and dominances between the analysed track, or another track that has been identified, played or ranked by the user and candidate music tracks from the catalogue, with any adjustments as indicated by the sliders 61, 62, 63.

Alternatively, the search may be based on tags and dominances selected based on user input as indicated by the sliders 61, 62, 63, without being based on a particular track.

Where the catalogue already includes tag and dominance information for its music tracks, the controller 202 may simply search for music tracks having particular tags and having instrument- and/or genre-based dominances that exceed thresholds set based on the user input. Where the catalogue does not include such information, the controller 202 may first compile a preliminary list of candidate tracks, for example, based on recommendations from user ratings in music databases, digital music stores or social media, and determine dominance information for the

candidate tracks in a similar manner to that described above in relation to s4.18 to s4.20.

A list of recommendations 43 is then compiled by the controller 202, based on the results of the search, and presented to the user (s4.23), for example, by being transmitted to the device 104 for presentation on display 106. The process ends at s4.24.

It will be appreciated that the above-described embodiments are not limiting on the scope of the invention, which is defined by the appended claims and their alternatives. Various alternative implementations will be envisaged by the skilled person, and all such alternatives are intended to be within the scope of the claims.

It is noted that the disclosure of the present application should be understood to include any novel features or any novel combination of features either explicitly or implicitly disclosed herein or any generalization thereof and during the prosecution of the present application or of any application derived therefrom, new claims may be formulated to cover any such features and/or combination of such features.

Embodiments of the present invention may be implemented in software, hardware, application logic or a combination of software, hardware and application logic. The software, application logic and/or hardware may reside on memory, or any computer media. In an example embodiment, the application logic, software or an instruction set is maintained on any one of various conventional computer-readable media. In the context of this document, a "computer-readable medium" may be any media or means that can contain, store, communicate, propagate or transport the instructions for use by or in connection with an instruction execution system, apparatus, or device, such as a computer.

A computer-readable medium may comprise a computer-readable storage medium that may be any tangible media or means that can contain or store the instructions for use by or in connection with an instruction execution system, apparatus, or device, such as a computer as defined previously. The computer-readable medium may be a volatile medium or non-volatile medium.

According to various embodiments of the previous aspect of the present invention, the computer program according to any of the above aspects, may be implemented

in a computer program product comprising a tangible computer-readable medium bearing computer program code embodied therein which can be used with the processor for the implementation of the functions described above.

Reference to "computer-readable storage medium", "computer program product", "tangibly embodied computer program" etc, or a "processor" or "processing circuit" etc. should be understood to encompass not only computers having differing architectures such as single/multi processor architectures and sequencers/parallel architectures, but also specialised circuits such as field programmable gate arrays FPGA, application specify circuits ASIC, signal processing devices and other devices. References to computer program, instructions, code etc. should be understood to express software for a programmable processor firmware such as the programmable content of a hardware device as instructions for a processor or configured or configuration settings for a fixed function device, gate array, programmable logic device, etc.

If desired, the different functions discussed herein may be performed in a different order and/or concurrently with each other. Furthermore, if desired, one or more of the above-described functions may be optional or may be combined.

Although various aspects of the invention are set out in the independent claims, other aspects of the invention comprise other combinations of features from the described embodiments and/or the dependent claims with the features of the independent claims, and not solely the combinations explicitly set out in the claims.

Claims

1. An apparatus comprising:

a controller; and

a memory in which is stored computer-readable instructions which, when executed by the controller, cause the controller to:

determine one or more acoustic features of an audio track;

determine dominance of an audible characteristic in the audio track based at least in part on said one or more acoustic features; and

store metadata for the audio track indicating said dominance of the audible characteristic.

2. An apparatus according to claim 1, wherein said computer-readable instructions, when executed by the controller, further cause the controller to:

select one or more tracks from a catalogue having a dominance of the audible characteristic within a range of dominance values defined at least in part based on the dominance of the audible characteristic of the audio track; and

output information identifying said one or more selected tracks.

3. An apparatus comprising:

a controller; and

a memory in which is stored computer-readable instructions which, when executed by the controller, cause the controller to:

select one or more tracks from a catalogue having a dominance of an audible characteristic within a range of dominance values; and

output information identifying said one or more selected tracks.

4. An apparatus according to claim 3, wherein the range of dominance values is based on a dominance for the audible characteristic in a first audio track.

5. An apparatus according to claim 2, 3 or 4, further comprising:

a user interface to receive input indicating a preferred dominance of the audible characteristic;

wherein said range of dominance values is further based on the input received via the user interface.

6. An apparatus according to any of the preceding claims, wherein said dominance includes an overall dominance indicating a level of audible distinguishability of a musical instrument in the audio track and/or an overall dominance indicating a degree of conformity to a musical genre of the audio track.

7. An apparatus according to any of the preceding claims, wherein said dominance includes a varying dominance indicating a level of audible distinguishability of a musical instrument in one or more temporal segments of the audio track.

8. An apparatus according to claim 7, said computer-readable instructions, when executed by the controller, further cause the controller to determine at least one of:

a difference in dominance of the musical instrument and an average of other musical instruments in the audio track;

a frequency of changes in dominance for the musical instrument; and

a duration of at least one section of the audio track for which the musical instrument is dominant.

9. A method comprising:

determining one or more acoustic features of an audio track;

determining dominance of an audible characteristic in the audio track based at least in part on said one or more acoustic features; and

storing metadata for the audio track indicating said dominance of the audible characteristic.

10. A method according to claim 9, further comprising:

selecting one or more tracks from a catalogue having a dominance of the audible characteristic within a range of dominance values defined at least in part based on the dominance of the audible characteristic of the audio track; and

outputting information identifying said one or more selected tracks.

11. A method comprising:

selecting one or more tracks from a catalogue having a dominance of an audible characteristic within a range of dominance values; and

outputting information identifying said one or more selected tracks.

12. A method according to claim 10 or 11, further comprising:

receiving input indicating a preferred dominance of the audible characteristic;

wherein said range of dominance values is further based on the received input.

13. A method according to any of claims 9 to 12, wherein said dominance includes at least one of:

an overall dominance indicating a level of audible distinguishability of a musical instrument in the audio track; and

an overall dominance indicating a degree of conformity to a musical genre of the audio track.

14. A method according to any of claims 9 to 13, wherein said dominance includes a varying dominance indicating a level of audible distinguishability of a musical instrument in one or more temporal segments of the audio track.

15. A method according to claim 14, further comprising determining at least one of:

a difference in dominance of the musical instrument and an average of other musical instruments in the audio track;

a frequency of changes in dominance for the musical instrument; and

a duration of at least one section of the audio track for which the musical instrument is dominant.

Intellectual Property Office

Application No:	GB1503467.1	**Examiner:**	Mr Tristan Ballard
Claims searched:	1, 2, 5-10, 12-15	**Date of search:**	6 August 2015

Patents Act 1977: Search Report under Section 17

Documents considered to be relevant:

Category	Relevant to claims	Identity of document and passage or figure of particular relevance
X,Y	X: 1, 6, 9, 13; Y: 2, 5, 10, 12	WO 2006/129274 A1 (PHILIPS ELECTRONICS) See especially pages 2 and 4
X	1, 9	US 2007/0174274 A1 (SAMSUNG ELECTRONICS) See especially paragraphs [0010], [0028], [0033], [0037] and [0038]
X	1, 9	JP 2008170991 A (SONY CORP) See especially paragraphs [0007]-[0009]
Y	2, 5, 10, 12	US 2002/0002899 A1 (GJERDINGEN ET AL) See especially paragraphs [0014], [0015], [0026], [0028], [0390]-[0446]

Categories:

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.

Field of Search:

Search of GB, EP, WO & US patent documents classified in the following areas of the UKC^{X} :

Worldwide search of patent documents classified in the following areas of the IPC

G06F; G10H

The following online and other databases have been used in the preparation of this search report

EPODOC, WPI

International Classification:

Subclass	Subgroup	Valid From
G06F	0017/30	01/01/2006