

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

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
5 December 2002 (05.12.2002)

PCT

(10) International Publication Number
WO 02/097790 A1

(51) International Patent Classification⁷: **G10L 11/00**

(21) International Application Number: **PCT/US02/05329**

(22) International Filing Date: 22 February 2002 (22.02.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/293,825	25 May 2001 (25.05.2001)	US
10/045,644	11 January 2002 (11.01.2002)	US
60/351,498	23 January 2002 (23.01.2002)	US
PCT/US02/	12 February 2002 (12.02.2002)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

US	60/293,825 (CIP)
Filed on	25 May 2001 (25.05.2001)
US	60/351,498 (CIP)
Filed on	23 January 2002 (23.01.2002)
US	10/045,644 (CIP)
Filed on	11 January 2002 (11.01.2002)
US	PCT/US02/ (CIP)
Filed on	12 February 2002 (12.02.2002)

(71) Applicant (for all designated States except US): **DOLBY LABORATORIES LICENSING CORPORATION**
[US/US]; 100 Potrero Avenue, San Francisco, CA 94103 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **CROCKETT,**

Brett, G. [US/US]; 100 Potrero Avenue, San Francisco, CA 94103 (US). **SMITHERS, Michael, J.** [AU/US]; 100 Potrero Avenue, San Francisco, CA 94103 (US).

(74) Agents: **GALLACHER, Thomas, A.** et al.; Gallagher & Lathrop, A Professional Corporation, 601 California Street, Suite 1111, San Francisco, CA 94108-2805 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT (utility model), AT, AU (petty patent), AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), CZ, DE (utility model), DE, DK (utility model), DK, DM, DZ, EC, EE (utility model), EE, ES, FI (utility model), FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

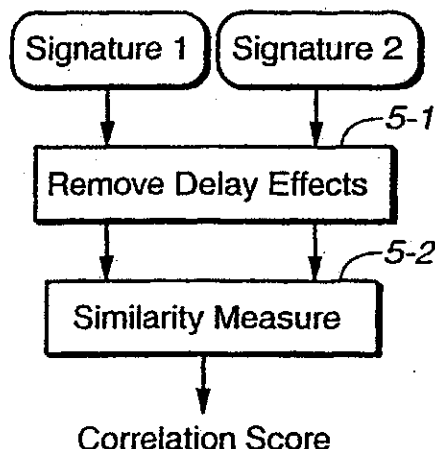
(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

- with international search report
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

[Continued on next page]

(54) Title: **COMPARING AUDIO USING CHARACTERIZATIONS BASED ON AUDITORY EVENTS**



(57) Abstract: A method for determining if one audio signal is derived from another audio signal or if two audio signals are derived from the same audio signal compares reduced-information characterizations of said audio signals, wherein said characterizations are based on auditory scene analysis. The comparison removes from the characterisations or minimizes in the characterisations the effect of temporal shift or delay on the audio signals(5-1), calculates a measure of similarity(5-2), and compares the measure of similarity against a threshold. In one alternative, the effect of temporal shift or delay is removed or minimized by cross-correlating the two characterizations. In another alternative, the effect of temporal shift or delay is removed or minimized by transforming the characterizations into a domain that is independent of temporal delay effects, such as the frequency domain. In both cases, a measure of similarity is calculated by calculating a coefficient of correlation.

WO 02/097790 A1

DESCRIPTION

Comparing Audio Using Characterizations Based on Auditory Events

TECHNICAL FIELD

5 The invention relates to audio signals. More particularly, the invention relates to characterizing audio signals and using characterizations to determine if one audio signal is derived from another audio signal or if two audio signals are derived from the same audio signal.

BACKGROUND ART

10 The division of sounds into units perceived as separate is sometimes referred to as "auditory event analysis" or "auditory scene analysis" ("ASA"). An extensive discussion of auditory scene analysis is set forth by Albert S. Bregman in his book *Auditory Scene Analysis - The Perceptual Organization of Sound*, Massachusetts
15 Institute of Technology, 1991, Fourth printing, 2001, Second MIT Press paperback edition. In addition, United States Patent 6,002,776 to Bhadkamkar, et al, December 14, 1999 cites publications dating back to 1976 as "prior art work related to sound separation by auditory scene analysis." However, the Bhadkamkar, et al patent discourages the practical use of auditory scene analysis, concluding that
20 "[t]echniques involving auditory scene analysis, although interesting from a scientific point of view as models of human auditory processing, are currently far too computationally demanding and specialized to be considered practical techniques for sound separation until fundamental progress is made."

 Bregman notes in one passage that "[w]e hear discrete units when the sound
25 changes abruptly in timbre, pitch, loudness, or (to a lesser extent) location in space." (*Auditory Scene Analysis - The Perceptual Organization of Sound*, supra at page 469). Bregman also discusses the perception of multiple simultaneous sound streams when, for example, they are separated in frequency.

 There are many different methods for extracting characteristics or features
30 from audio. Provided the features or characteristics are suitably defined, their extraction can be performed using automated processes. For example "ISO/IEC JTC

- 2 -

1/SC 29/WG 11" (MPEG) is currently standardizing a variety of audio descriptors as part of the MPEG-7 standard. A common shortcoming of such methods is that they ignore ASA. Such methods seek to measure, periodically, certain "classical" signal processing parameters such as pitch, amplitude, power, harmonic structure and spectral flatness. Such parameters, while providing useful information, do not analyze and characterize audio signals into elements perceived as separate according to human cognition.

Auditory scene analysis attempts to characterize audio signals in a manner similar to human perception by identifying elements that are separate according to human cognition. By developing such methods, one can implement automated processes that accurately perform tasks that heretofore would have required human assistance.

The identification of separately perceived elements would allow the unique identification of an audio signal using substantially less information than the full signal itself. Compact and unique identifications based on auditory events may be employed, for example, to identify a signal that is copied from another signal (or is copied from the same original signal as another signal).

DISCLOSURE OF THE INVENTION

According to an aspect of the present invention, there is provided a method for determining if one audio signal is derived from another audio signal or if two audio signals are derived from the same audio signal, comprising

comparing reduced-information characterizations of said audio signals, wherein said reduced-information characterizations are based on auditory scene analysis.

A method is described that generates a unique reduced-information characterization of an audio signal that may be used to identify the audio signal. The characterization may be considered a "signature" or "fingerprint" of the audio signal.

According to the present invention, an auditory scene analysis (ASA) is performed to identify auditory events as the basis for characterizing an audio signal. Ideally, the auditory scene analysis identifies auditory events that are most likely to be perceived by a human listener even after the audio has undergone processing, such as low bit rate coding or acoustic transmission through a loudspeaker. The audio signal may be characterized by the boundary locations of auditory events and, optionally, by the dominant frequency subband of each auditory event. The resulting information pattern constitutes a compact audio fingerprint or signature that may be compared to

one or more other such audio fingerprints or signatures. A determination that at least a portion of the respective signatures are the same (to a desired degree of confidence) indicates that the related portions of the audio signals from which the respective signatures were derived are the same or were derived from the same audio signal.

5 The auditory scene analysis method according to the present invention provides a fast and accurate method of comparing two audio signals, particularly music, by comparing signatures based on auditory event information. ASA extracts information or features underlying the perception of similarity, in contrast to traditional methods of feature extraction that extract features less fundamental to
10 perceiving similarities between audio signals (such as pitch amplitude, power, and harmonic structure). The use of ASA improves the chance of finding similarity in material that has undergone significant processing, such as low bit coding or acoustic transmission through a loudspeaker.

15 Although in principle the invention may be practiced either in the analog or digital domain (or some combination of the two), in practical embodiments of the invention, audio signals are represented by samples in blocks of data and processing is done in the digital domain.

20 Referring to FIG. 1A, auditory scene analysis 2 is applied to an audio signal in order to produce a "signature" or "fingerprint," related to that signal. In this case, there are two audio signals of interest. They may be similar in that one may be derived from the other or both may have been previously derived from the same original signal, but this is not known in advance. Thus, auditory scene analysis is applied to both signals. For simplicity, FIG. 1A shows only the application of ASA to one signal. As shown in FIG. 1B, the signatures for the two audio signals,
25 Signature 1 and Signature 2, are applied to a correlator or correlation function 4 that generates a correlation score. A user may set a minimum correlation score as providing a desired degree of confidence that at least a portion of the two signatures are the same. In practice, the two signatures may be stored data. In one practical application, one of the signatures may be derived, for example, from an unauthorized
30 copy of a musical work and the other signature may be one of a large number of

- 4 -

signatures in a database (each signature being derived from a copyright owner's musical work) against which the unauthorized copy signature is compared until a match, to a desired degree of confidence, if any, is obtained. This may be conducted automatically by a machine, the details of which are beyond the scope of the present invention.

Because the signatures are representative of the audio signals but are substantially shorter (*i.e.*, they are more compact or have fewer bits) than the audio signals from which they were derived, the similarity of the two signatures (or lack thereof) can be determined much faster than it would take to determine the similarity between the audio signals.

Further details of FIGS. 1A and 1B are set forth below.

In accordance with aspects of the present invention, a computationally efficient process for dividing audio into temporal segments or "auditory events" that tend to be perceived as separate is provided.

A powerful indicator of the beginning or end of a perceived auditory event is believed to be a change in spectral content. In order to detect changes in timbre and pitch (spectral content) and, as an ancillary result, certain changes in amplitude, the audio event detection process according to an aspect of the present invention detects changes in spectral composition with respect to time. Optionally, according to a further aspect of the present invention, the process may also detect changes in amplitude with respect to time that would not be detected by detecting changes in spectral composition with respect to time.

In its least computationally demanding implementation, the process divides audio into time segments by analyzing the entire frequency band of the audio signal (full bandwidth audio) or substantially the entire frequency band (in practical implementations, band limiting filtering at the ends of the spectrum are often employed) and giving the greatest weight to the loudest audio signal components. This approach takes advantage of a psychoacoustic phenomenon in which at smaller time scales (20 msec and less) the ear may tend to focus on a single auditory event at a given time. This implies that while multiple events may be occurring at the same

- 5 -

time, one component tends to be perceptually most prominent and may be processed individually as though it were the only event taking place. Taking advantage of this effect also allows the auditory event detection to scale with the complexity of the audio being processed. For example, if the input audio signal being processed is a solo instrument, the audio events that are identified will likely be the individual notes being played. Similarly for an input voice signal, the individual components of speech, the vowels and consonants for example, will likely be identified as individual audio elements. As the complexity of the audio increases, such as music with a drumbeat or multiple instruments and voice, the auditory event detection identifies the most prominent (*i.e.*, the loudest) audio element at any given moment. Alternatively, the "most prominent" audio element may be determined by taking hearing threshold and frequency response into consideration.

Optionally, according to further aspects of the present invention, at the expense of greater computational complexity, the process may also take into consideration changes in spectral composition with respect to time in discrete frequency bands (fixed or dynamically determined or both fixed and dynamically determined bands) rather than the full bandwidth. This alternative approach would take into account more than one audio stream in different frequency bands rather than assuming that only a single stream is perceptible at a particular time.

Even a simple and computationally efficient process according to an aspect of the present invention for segmenting audio has been found useful to identify auditory events.

An auditory event detecting process of the present invention may be implemented by dividing a time domain audio waveform into time intervals or blocks and then converting the data in each block to the frequency domain, using either a filter bank or a time-frequency transformation, such as a Discrete Fourier Transform (DFT) (implemented as a Fast Fourier Transform (FFT) for speed). The amplitude of the spectral content of each block may be normalized in order to eliminate or reduce the effect of amplitude changes. The resulting frequency domain representation provides an indication of the spectral content (amplitude as a function of frequency)

- 6 -

of the audio in the particular block. The spectral content of successive blocks is compared and a change greater than a threshold may be taken to indicate the temporal start or temporal end of an auditory event.

In order to minimize the computational complexity, only a single band of frequencies of the time domain audio waveform may be processed, preferably either the entire frequency band of the spectrum (which may be about 50 Hz to 15 kHz in the case of an average quality music system) or substantially the entire frequency band (for example, a band defining filter may exclude the high and low frequency extremes).

Preferably, the frequency domain data is normalized, as is described below. The degree to which the frequency domain data needs to be normalized gives an indication of amplitude. Hence, if a change in this degree exceeds a predetermined threshold, that too may be taken to indicate an event boundary. Event start and end points resulting from spectral changes and from amplitude changes may be ORed together so that event boundaries resulting from both types of change are identified.

In practical embodiments in which the audio is represented by samples divided into blocks, each auditory event temporal start and stop point boundary necessarily coincides with a boundary of the block into which the time domain audio waveform is divided. There is a trade off between real-time processing requirements (as larger blocks require less processing overhead) and resolution of event location (smaller blocks provide more detailed information on the location of auditory events).

As a further option, as suggested above, but at the expense of greater computational complexity, instead of processing the spectral content of the time domain waveform in a single band of frequencies, the spectrum of the time domain waveform prior to frequency domain conversion may be divided into two or more frequency bands. Each of the frequency bands may then be converted to the frequency domain and processed as though it were an independent channel. The resulting event boundaries may then be ORed together to define the event boundaries for that channel. The multiple frequency bands may be fixed, adaptive, or a combination of fixed and adaptive. Tracking filter techniques employed in audio

noise reduction and other arts, for example, may be employed to define adaptive frequency bands (e.g., dominant simultaneous sine waves at 800 Hz and 2 kHz could result in two adaptively-determined bands centered on those two frequencies).

Other techniques for providing auditory scene analysis may be employed to
5 identify auditory events in the present invention.

DESCRIPTION OF THE DRAWINGS

FIG. 1A is a flow chart showing the extraction of a signature from an audio signal in accordance with the present invention. The audio signal may, for example,
10 represent music (e.g., a musical composition or "song").

FIG. 1B is a flow chart illustrating the correlation of two signatures in accordance with the present invention.

FIG. 2 is a flow chart showing the extraction of audio event locations and the optional extraction of dominant subbands from an audio signal in accordance with
15 the present invention.

FIG. 3 is a conceptual schematic representation depicting the step of spectral analysis in accordance with the present invention.

FIGS. 4A and 4B are idealized audio waveforms showing a plurality of audio event locations or event borders in accordance with the present invention.

FIG. 5 is a flow chart showing in more detail the correlation of two signatures in accordance with the correlation 4 of FIG. 2 of the present invention.
20

FIGS. 6A-D are conceptual schematic representations of signals illustrating examples of signature alignment in accordance with the present invention. The figures are not to scale. In the case of a digital audio signal represented by samples,
25 the horizontal axis denotes the sequential order of discrete data stored in each signature array.

BEST MODE FOR CARRYING OUT THE INVENTION

In a practical embodiment of the invention, the audio signal is represented by
30 samples that are processed in blocks of 512 samples, which corresponds to about

- 8 -

11.6 msec of input audio at a sampling rate of 44.1 kHz. A block length having a time less than the duration of the shortest perceivable auditory event (about 20 msec) is desirable. It will be understood that the aspects of the invention are not limited to such a practical embodiment. The principles of the invention do not require
5 arranging the audio into sample blocks prior to determining auditory events, nor, if they are, of providing blocks of constant length. However, to minimize complexity, a fixed block length of 512 samples (or some other power of two number of samples) is useful for three primary reasons. First, it provides low enough latency to be acceptable for real-time processing applications. Second, it is a power-of-two
10 number of samples, which is useful for fast Fourier transform (FFT) analysis. Third, it provides a suitably large window size to perform useful auditory scene analysis.

In the following discussions, the input signals are assumed to be data with amplitude values in the range $[-1, +1]$.

Auditory Scene Analysis 2 (FIG. 1A)

15 Following audio input data blocking (not shown), the input audio signal is divided into auditory events, each of which tends to be perceived as separate, in process 2 ("Auditory Scene Analysis") of FIG. 1A. Auditory scene analysis may be accomplished by an auditory scene analysis (ASA) process discussed above. Although one suitable process for performing auditory scene analysis is described in
20 further detail below, the invention contemplates that other useful techniques for performing ASA may be employed.

FIG. 2 outlines a process in accordance with techniques of the present invention that may be used as the auditory scene analysis process of FIG. 1A. The ASA step or process 2 is composed of three general processing substeps. The first
25 substep 2-1 ("Perform Spectral Analysis") takes the audio signal, divides it into blocks and calculates a spectral profile or spectral content for each of the blocks. Spectral analysis transforms the audio signal into the short-term frequency domain. This can be performed using any filterbank; either based on transforms or banks of band-pass filters, and in either linear or warped frequency space (such as the Bark
30 scale or critical band, which better approximate the characteristics of the human ear).

- 9 -

With any filterbank there exists a tradeoff between time and frequency. Greater time resolution, and hence shorter time intervals, leads to lower frequency resolution. Greater frequency resolution, and hence narrower subbands, leads to longer time intervals.

5 The first substep 2-1 calculates the spectral content of successive time segments of the audio signal. In a practical embodiment, described below, the ASA block size is 512 samples of the input audio signal (FIG.3). In the second substep 2-2, the differences in spectral content from block to block are determined ("Perform spectral profile difference measurements"). Thus, the second substep calculates the
10 difference in spectral content between successive time segments of the audio signal. In the third substep 2-3 ("Identify location of auditory event boundaries"), when the spectral difference between one spectral-profile block and the next is greater than a threshold, the block boundary is taken to be an auditory event boundary. Thus, the third substep sets an auditory event boundary between successive time segments
15 when the difference in the spectral profile content between such successive time segments exceeds a threshold. As discussed above, a powerful indicator of the beginning or end of a perceived auditory event is believed to be a change in spectral content. The locations of event boundaries are stored as a signature. An optional process step 2-4 ("Identify dominant subband") uses the spectral analysis to identify
20 a dominant frequency subband that may also be stored as part of the signature.

In this embodiment, auditory event boundaries define auditory events having a length that is an integral multiple of spectral profile blocks with a minimum length of one spectral profile block (512 samples in this example). In principle, event boundaries need not be so limited.

25 Either overlapping or non-overlapping segments of the audio may be windowed and used to compute spectral profiles of the input audio. Overlap results in finer resolution as to the location of auditory events and, also, makes it less likely to miss an event, such as a transient. However, as time resolution increases, frequency resolution decreases. Overlap also increases computational complexity.
30 Thus, overlap may be omitted. FIG. 3 shows a conceptual representation of non-

- 10 -

overlapping 512 sample blocks being windowed and transformed into the frequency domain by the Discrete Fourier Transform (DFT). Each block may be windowed and transformed into the frequency domain, such as by using the DFT, preferably implemented as a Fast Fourier Transform (FFT) for speed.

5 The following variables may be used to compute the spectral profile of the input block:

N = number of samples in the input signal

M = number of windowed samples used to compute spectral profile

P = number of samples of spectral computation overlap

10 Q = number of spectral windows/regions computed

 In general, any integer numbers may be used for the variables above.

However, the implementation will be more efficient if M is set equal to a power of 2 so that standard FFTs may be used for the spectral profile calculations. In a practical embodiment of the auditory scene analysis process, the parameters listed may be set

15 to:

M = 512 samples (or 11.6 msec at 44.1 kHz)

P = 0 samples (no overlap)

 The above-listed values were determined experimentally and were found generally to identify with sufficient accuracy the location and duration of auditory
20 events. However, setting the value of P to 256 samples (50% overlap) has been found to be useful in identifying some hard-to-find events. While many different types of windows may be used to minimize spectral artifacts due to windowing, the window used in the spectral profile calculations is an M -point Hanning, Kaiser-Bessel or other suitable, preferably non-rectangular, window. The above-indicated
25 values and a Hanning window type were selected after extensive experimental analysis as they have shown to provide excellent results across a wide range of audio material. Non-rectangular windowing is preferred for the processing of audio signals with predominantly low frequency content. Rectangular windowing produces spectral artifacts that may cause incorrect detection of events. Unlike certain codec
30 applications where an overall overlap/add process must provide a constant level, such

- 11 -

a constraint does not apply here and the window may be chosen for characteristics such as its time/frequency resolution and stop-band rejection.

In substep 2-1 (FIG. 2), the spectrum of each M-sample block may be computed by windowing the data by an M-point Hanning, Kaiser-Bessel or other suitable window, converting to the frequency domain using an M-point Fast Fourier Transform, and calculating the magnitude of the FFT coefficients. The resultant data is normalized so that the largest magnitude is set to unity, and the normalized array of M numbers is converted to the log domain. The array need not be converted to the log domain, but the conversion simplifies the calculation of the difference measure in substep 2-2. Furthermore the log domain more closely matches the log domain amplitude nature of the human auditory system. The resulting log domain values have a range of minus infinity to zero. In a practical embodiment, a lower limit can be imposed on the range of values; the limit may be fixed, for example -60 dB, or be frequency-dependent to reflect the lower audibility of quiet sounds at low and very high frequencies. (Note that it would be possible to reduce the size of the array to M/2 in that the FFT represents negative as well as positive frequencies).

Substep 2-2 calculates a measure of the difference between the spectra of adjacent blocks. For each block, each of the M (log) spectral coefficients from substep 2-1 is subtracted from the corresponding coefficient for the preceding block, and the magnitude of the difference calculated (the sign is ignored). These M differences are then summed to one number. Hence, for the whole audio signal, the result is an array of Q positive numbers; the greater the number the more a block differs in spectrum from the preceding block. This difference measure could also be expressed as an average difference per spectral coefficient by dividing the difference measure by the number of spectral coefficients used in the sum (in this case M coefficients).

Substep 2-3 identifies the locations of auditory event boundaries by applying a threshold to the array of difference measures from substep 2-2 with a threshold value. When a difference measure exceeds a threshold, the change in spectrum is deemed sufficient to signal a new event and the block number of the change is recorded as an

- 12 -

event boundary. For the values of M and P given above and for log domain values (in substep 2-1) expressed in units of dB, the threshold may be set equal to 2500 if the whole magnitude FFT (including the mirrored part) is compared or 1250 if half the FFT is compared (as noted above, the FFT represents negative as well as positive frequencies — for the magnitude of the FFT, one is the mirror image of the other). This value was chosen experimentally and it provides good auditory event boundary detection. This parameter value may be changed to reduce (increase the threshold) or increase (decrease the threshold) the detection of events. The details of this practical embodiment are not critical. Other ways to calculate the spectral content of successive time segments of the audio signal, calculate the differences between successive time segments, and set auditory event boundaries at the respective boundaries between successive time segments when the difference in the spectral profile content between such successive time segments exceeds a threshold may be employed.

For an audio signal consisting of Q blocks (of size M samples), the output of the auditory scene analysis process of function 2 of FIG. 1A is an array $B(q)$ of information representing the location of auditory event boundaries where $q = 0, 1, \dots, Q-1$. For a block size of $M = 512$ samples, overlap of $P = 0$ samples and a signal-sampling rate of 44.1kHz, the auditory scene analysis function 2 outputs approximately 86 values a second. Preferably, the array $B(q)$ is stored as the signature, such that, in its basic form, without the optional dominant subband frequency information, the audio signal's signature is an array $B(q)$ representing a string of auditory event boundaries.

An example of the results of auditory scene analysis for two different signals is shown in FIGS. 4A and 4B. The top plot, FIG. 4A, shows the results of auditory scene processing where auditory event boundaries have been identified at samples 1024 and 1536. The bottom plot, FIG. 4B, shows the identification of event boundaries at samples 1024, 2048 and 3072.

Identify dominant subband (optional)

For each block, an optional additional step in the ASA processing (shown in FIG. 2) is to extract information from the audio signal denoting the dominant frequency "subband" of the block (conversion of the data in each block to the frequency domain results in information divided into frequency subbands). This block-based information may be converted to auditory-event based information, so that the dominant frequency subband is identified for every auditory event. This information for every auditory event provides the correlation processing (described below) with further information in addition to the auditory event boundary information.

The dominant (largest amplitude) subband may be chosen from a plurality of subbands, three or four, for example, that are within the range or band of frequencies where the human ear is most sensitive. Alternatively, other criteria may be used to select the subbands. The spectrum may be divided, for example, into three subbands.

The preferred frequency range of the subbands is:

Subband 1	301Hz to 560Hz
Subband 2	560Hz to 1938Hz
Subband 3	1938Hz to 9948Hz

To determine the dominant subband, the square of the magnitude spectrum (or the power magnitude spectrum) is summed for each subband. This resulting sum for each subband is calculated and the largest is chosen. The subbands may also be weighted prior to selecting the largest. The weighting may take the form of dividing the sum for each subband by the number of spectral values in the subband, or alternatively may take the form of an addition or multiplication to emphasize the importance of a band over another. This can be useful where some subbands have more energy on average than other subbands but are less perceptually important.

Considering an audio signal consisting of Q blocks, the output of the dominant subband processing is an array $DS(q)$ of information representing the dominant subband in each block ($q = 0, 1, \dots, Q-1$). Preferably, the array $DS(q)$ is stored in the signature along with the array $B(q)$. Thus, with the optional dominant subband

- 14 -

information, the audio signal's signature is two arrays $B(q)$ and $DS(q)$, representing, respectively, a string of auditory event boundaries and a dominant frequency subband within each block. Thus, in an idealized example, the two arrays could have the following values (for a case in which there are three possible dominant subbands).

5

1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	(Event Boundaries)
1	1	2	2	2	2	1	1	1	3	3	3	3	3	1	1	(Dominant Subbands)

In most cases, the dominant subband remains the same within each auditory event, as shown in this example, or has an average value if it is not uniform for all blocks within the event. Thus, a dominant subband may be determined for each auditory event and the array $DS(q)$ may be modified to provide that the same dominant subband is assigned to each block within an event.

Correlation

The determination of whether one signature is the same or similar to another stored signature may be accomplished by a correlation function or process. The correlation function or process compares two signatures to determine their similarity. This may be done in two steps as shown in FIG. 5: a step 5-1 that removes or minimizes the effect of temporal shift or delay on the signatures, followed by a step 5-2 that calculates a measure of similarity between the signatures.

The first-mentioned step 5-1 minimizes the effect of any delay between two signatures. Such delay may have been deliberately added to the audio signal or could be the result of signal processing and/or low bit rate audio coding. The output of this step is two modified signatures in a form suitable for calculation of a measure of their similarity.

The second-mentioned step 5-2 compares the two modified signatures to find a quantitative measure of their similarity (a correlation score). This measure of similarity can then be compared against a threshold to determine if the signatures are the same or different to a desired level of confidence. Two suitable correlation processes or functions are described. Either one of them or some other suitable correlation process or function may be employed as part of the present invention.

- 15 -

First Correlation Process or Function
Removal of Temporal Delay Effects

This correlation function or process isolates a single region or portion from each of the signatures such that these two regions are the most similar portions in the
5 respective signatures and have the same length. The isolated region could be the total overlapping regions between the two signatures, as shown in the examples in FIGS. 6A-D, or the isolated region could be smaller than the overlapping regions.

The preferred method uses the whole overlapping region from the two signatures. Some examples are shown in FIG. 6. The overlapping region for the two
10 signatures could be a portion from the end of one signature and the beginning of the other signature (FIGS. 6B and 6C). If one of the signatures is smaller than the other, then the overlapping region between the two signatures could be all of the smaller signature and a portion of the larger signature (FIG. 6A and 6D).

There are a number of different ways to isolate a common region from two
15 arrays of data. A standard mathematical technique involves using the cross correlation to find a lag or delay measure between the arrays of data. When the beginning of each of two arrays of data is aligned, the lag or delay is said to be zero. When the beginning of each of two arrays of data is not aligned, the lag or delay is non-zero. The cross correlation calculates a measure for each possible lag or delay
20 between the two arrays of data: this measure is stored as an array (the output of the cross correlation function). The lag or delay that represents the peak in the cross correlation array is considered to be the lag or delay of one array of data with respect to the other. The following paragraphs expresses such a correlation method in a mathematical form.

25 Let S_1 (length N_1) be an array from Signature 1 and S_2 (length N_2) an array from Signature 2. First calculate the cross-correlation array $R_{E_1 E_2}$ (see, for example, John G. Proakis, Dimitris G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Macmillan Publishing Company, 1992, ISBN 0-02-396815-X).

- 16 -

$$R_{E_1 E_2}(l) = \sum_{n=-\infty}^{\infty} S_1(n) S_2(n-l) \quad l = 0, \pm 1, \pm 2, \dots \quad (1)$$

Preferably, the cross-correlation is performed using standard FFT based techniques to reduce execution time.

Since both S_1 and S_2 are bounded, $R_{E_1 E_2}$ has length $N_1 + N_2 - 1$. Assuming S_1 and S_2 are similar, the lag l corresponding to the maximum element in $R_{E_1 E_2}$ represents the delay of S_2 relative to S_1 .

$$l_{peak} = l \text{ for } MAX(R_{E_1 E_2}(l)) \quad (2)$$

Since this lag represents the delay, the common spatial regions or spatially overlapping parts of signatures S_1 and S_2 are retained as S'_1 and S'_2 ; each having the same length, N_{12} .

Expressed as equations, the overlapping parts S'_1 and S'_2 of the signatures S_1 and S_2 are defined as:

$$S'_1(m) = S_1(n) \begin{cases} l_{peak} \leq n < l_{peak} + MIN(N_1 - l_{peak}, N_2) & m = n - l_{peak} & l_{peak} \geq 0 \\ 0 \leq n < MIN(N_1, N_2 + l_{peak}) & m = n & l_{peak} < 0 \end{cases}$$

$$S'_2(m) = S_2(n) \begin{cases} 0 \leq n < MIN(N_1 - l_{peak}, N_2) & m = n & l_{peak} \geq 0 \\ -l_{peak} \leq n < -l_{peak} + MIN(N_1, N_2 + l_{peak}) & m = n + l_{peak} & l_{peak} < 0 \end{cases}$$

(3)

The length of S'_1 and S'_2 is:

$$N_{12} = \begin{cases} MIN(N_1 - l_{peak}, N_2) & l_{peak} \geq 0 \\ MIN(N_1, N_2 + l_{peak}) & l_{peak} < 0 \end{cases} \quad (4)$$

First Correlation Process or Function

Similarity Measure

This step compares the two signatures to find a quantitative measure of their similarity. The preferred method uses the coefficient of correlation (Eqn. 5). This is

a standard textbook method (William Mendenhall, Dennis D. Wackerly, Richard L. Scheaffer, *Mathematical Statistics with Applications: Fourth Edition*, Duxbury Press, 1990, ISBN 0-534-92026-8).

$$\rho = \frac{\text{Cov}(S'_1, S'_2)}{\sigma_1 \sigma_2} \quad (5)$$

where σ_1 and σ_2 are the standard deviations of S'_1 and S'_2 respectively.

The covariance of S'_1 and S'_2 is defined as:

$$\text{Cov}(S'_1, S'_2) = \frac{\sum_{m=0}^{N_{12}-1} (S'_1(m) - \mu_1)(S'_2(m) - \mu_2)}{N_{12}} \quad (6)$$

where μ_1 and μ_2 are the means of S'_1 and S'_2 respectively.

The coefficient of correlation, ρ , is in the range $-1 \leq \rho \leq 1$ where -1 and 1 indicate perfect correlation. Preferably, a *threshold* is applied to the absolute value of this measure to indicate a correct match.

$$\text{Match} = \begin{cases} \text{TRUE} & \text{abs}(\rho) > \text{threshold} \\ \text{FALSE} & \text{abs}(\rho) \leq \text{threshold} \end{cases} \quad (7)$$

In practice, the value of the threshold may be tuned (on a large training set of signatures) to ensure acceptable false rejection and detection rates.

The first correlation process or function is preferred for signatures that have large misalignment or delay, and for signatures in which the length of one signature is significantly smaller than the length of the other signature.

Second Correlation Process or Function

Removal of Temporal Delay Effects

The second correlation process or function transforms the signatures from their current temporal domain into a domain that is independent of temporal delay effects.

- 5 The method results in two modified signatures that have the same length, such that they can be directly correlated or compared.

There are a number of ways to transform data in such a manner. The preferred method uses the Discrete Fourier Transform (DFT). The DFT of a signal can be separated into magnitude and phase. A spatial shift or time delay of the signal (input
10 to the DFT) alters the phase of the DFT but not the magnitude. Thus the magnitude of the DFT of a signal can be considered as a time-invariant representation of the signal.

This property of the DFT allows each of the two signatures to be transformed into a time-invariant representation. If both signatures have the same length, the
15 magnitude DFT can be directly computed for each of the signatures and the results stored as the modified signatures. If the length of each of the signatures is different, then prior to calculating the DFT, either the longer signature can be truncated to have the same length as the shorter signature, or the shorter signature can be zero padded or extended to have the same length as the longer signature. The following
20 paragraphs express the method in a mathematical form.

Let S_1 (length N_1) be an array from Signature 1 and S_2 (length N_2) an array from Signature 2. Firstly, the longer signature is truncated or the shorter signature zero padded such that both signatures have the same length, N_{12} . The transformed signature arrays, S'_1 and S'_2 , are created by taking the magnitude DFT as follows:

$$25 \quad S'_1(k) = \left| \sum_{n=0}^{N_{12}-1} S_1(n) \cdot e^{-j2\pi kn/N_{12}} \right| \quad k = 0, 1, 2, \dots, N_{12}-1 \quad (8)$$

$$S'_2(k) = \left| \sum_{n=0}^{N_{12}-1} S_2(n) \cdot e^{-j2\pi kn/N_{12}} \right| \quad k = 0, 1, 2, \dots, N_{12}-1 \quad (9)$$

In practice, for each signature it is beneficial to subtract its mean prior to calculating the DFT. Some windowing may also be applied to the S_1 and S_2

signatures prior to taking the Discrete Fourier Transform, however in practice no particular windowing has been found to produce the best results.

Second Correlation Process or Function

Similarity Measure

- 5 This similarity measure step compares the two signatures to find a quantitative measure of their similarity. The preferred method uses the coefficient of correlation (Eqn. 9). This is a standard textbook method (William Mendenhall, Dennis D. Wackerly, Richard L. Scheaffer, *Mathematical Statistics with Applications: Fourth Edition*, Duxbury Press, 1990, ISBN 0-534-92026-8).

10

$$\rho = \frac{\text{Cov}(S'_1, S'_2)}{\sigma_1 \sigma_2} \quad (9)$$

where σ_1 and σ_2 are the standard deviations of S'_1 and S'_2 respectively.

The covariance of S'_1 and S'_2 is defined as:

15

$$\text{Cov}(S'_1, S'_2) = \frac{\sum_{k=0}^{N_{12}-1} (S'_1(k) - \mu_1)(S'_2(k) - \mu_2)}{N_{12}} \quad (10)$$

where μ_1 and μ_2 are the means of S'_1 and S'_2 respectively.

- 20 The coefficient of correlation, ρ , is in the range $-1 \leq \rho \leq 1$ where -1 and 1 indicate perfect correlation. Preferably, a *threshold* is applied to the absolute value of this measure to indicate a correct match.

$$\text{Match} = \begin{cases} \text{TRUE} & \text{abs}(\rho) > \text{threshold} \\ \text{FALSE} & \text{abs}(\rho) \leq \text{threshold} \end{cases} \quad (11)$$

25

In practice, the value of the threshold may be tuned (on a large training set of signatures) to ensure acceptable false rejection and detection rates.

- 20 -

In practical applications, many signatures may be stored together to form a library of signatures representing "known" audio content. In this situation, the ability to discriminate between signatures can be improved by calculating a mean signature and subtracting this mean signature from each of two signatures under comparison.

5 For example, given a database containing W signatures, S'_0 to S'_{W-1} , the mean signature is calculated as follows.

$$S'_{MEAN}(k) = \frac{1}{W} \sum_{w=0}^{W-1} S'_w(k) \quad k = 0, 1, 2, \dots, N_{12} - 1 \quad (12)$$

10 When comparing two signatures (even if one of the signatures is not in the library) the mean signature is subtracted from both signatures prior to calculating the covariance (subsequently used in the coefficient of correlation). The covariance becomes:

$$15 \quad Cov(S'_1, S'_2) = \frac{\sum_{k=0}^{N_{12}-1} [(S'_1(k) - S'_{MEAN}(k)) - \mu_1] \cdot [(S'_2(k) - S'_{MEAN}(k)) - \mu_2]}{N_{12}} \quad (13)$$

where μ_1 and μ_2 are the means of $S'_1 - S'_{MEAN}$ and $S'_2 - S'_{MEAN}$ respectively.

The second correlation process or function is preferred for signatures that have small misalignment or delay, and for signatures where the lengths of the signatures are similar. It is also significantly faster than the first correlation process or function. However since some information is inherently lost (by discarding the phase of the DFTs), it results in a slightly less accurate measure of similarity.

Applications

As briefly mentioned earlier, an application of this invention is searchable audio databases; for example a record company's library of songs. Signatures could be created for all the songs from the library and the signatures stored in a database. This invention provides a means for taking a song of unknown origin, calculating its

signature and comparing its signature very quickly against all the signatures in the database to determine the identity of the unknown song.

In practice, the accuracy of (or confidence in) the similarity measure is proportional to the size of the signatures being compared. The greater the length of the signatures, the greater the amount of data being used in the comparison and hence the greater the confidence or accuracy in the similarity measure. It has been found that signatures generated from about 30 seconds of audio provide for good discrimination. However the larger the signatures, the longer the time required to perform a comparison.

Conclusion

It should be understood that implementation of other variations and modifications of the invention and its various aspects will be apparent to those skilled in the art, and that the invention is not limited by these specific embodiments described. It is therefore contemplated to cover by the present invention any and all modifications, variations, or equivalents that fall within the true spirit and scope of the basic underlying principles disclosed and claimed herein.

The present invention and its various aspects may be implemented as software functions performed in digital signal processors, programmed general-purpose digital computers, and/or special purpose digital computers. Interfaces between analog and digital signal streams may be performed in appropriate hardware and/or as functions in software and/or firmware.

- 22 -

CLAIMS

1. A method for determining if one audio signal is derived from another audio signal or if two audio signals are derived from the same audio signal, comprising comparing reduced-information characterizations of said audio signals,
5 wherein said reduced-information characterizations are based on auditory scene analysis.

2. The method of claim 1 wherein said comparing includes removing from the characterizations or minimizing in the characterizations the
10 effect of temporal shift or delay on the audio signals,
calculating a measure of similarity, and
comparing the measure of similarity against a threshold.

3. The method of claim 2 wherein said removing identifies a portion in each
15 of said characterizations, such that the respective portions are the most similar portions in the respective characterizations and the respective portions have the same length.

4. The method of claim 3 wherein said removing identifies a portion in each
20 of said characterizations by performing a cross-correlation.

5. The method of claim 4 wherein said calculating calculates a measure of similarity by calculating a coefficient of correlation of the identified portion in each of said characterizations.

25 6. The method of claim 2 wherein said removing transforms the characterizations into a domain that is independent of temporal delay effects.

7. The method of claim 6 wherein said removing transforms the
30 characterizations into the frequency domain.

8. The method of claim 7 wherein said calculating calculates a measure of similarity by calculating a coefficient of correlation of the identified portion in each of said characterizations.

5

9. The method of any one of claims 1-8 wherein one of said characterizations is a characterization from a library of characterizations representing known audio content.

10

10. The method of claim 9 further comprising subtracting a mean of the characterizations in said library from both characterizations after said removing and prior to said comparing.

15

11. The method of any one of claims 1-10 wherein said reduced-information characterizations based on auditory scene analysis are arrays of information representing at least the location of auditory event boundaries.

12. The method of claim 11 wherein said auditory event boundaries are determined by

20

calculating the spectral content of successive time segments of said audio signal,

calculating the difference in spectral content between successive time segments of said audio signal, and

25

identifying an auditory event boundary as the boundary between successive time segments when the difference in the spectral content between such successive time segments exceeds a threshold.

13. The method of claim 12 or claim 13 wherein said arrays of information also represent the dominant frequency subband of each of said auditory events.

30