US 20160247512A1

(19) United States

(12) Patent Application Publication

DUONG et al.

(10) Pub. No.: US 2016/0247512 A1

(43) Pub. Date: Aug. 25, 2016

(54) **METHOD AND APPARATUS FOR GENERATING FINGERPRINT OF AN AUDIO SIGNAL**

(71) Applicant: **THOMSON LICENSING**, Issy les Moulineaux (FR)

(72) Inventors: **Quang Khanh Ngoc DUONG**, Rennes (FR); **Alexey OZEROV**, Rennes (FR); **Frederic LEFEBVRE**, Rennes (FR)

(21) Appl. No.: **14/948,254**

(22) Filed: **Nov. 21, 2015**

(30) **Foreign Application Priority Data**

Nov. 21, 2014 (EP) 14306854.2

Publication Classification

(51) **Int. Cl.**

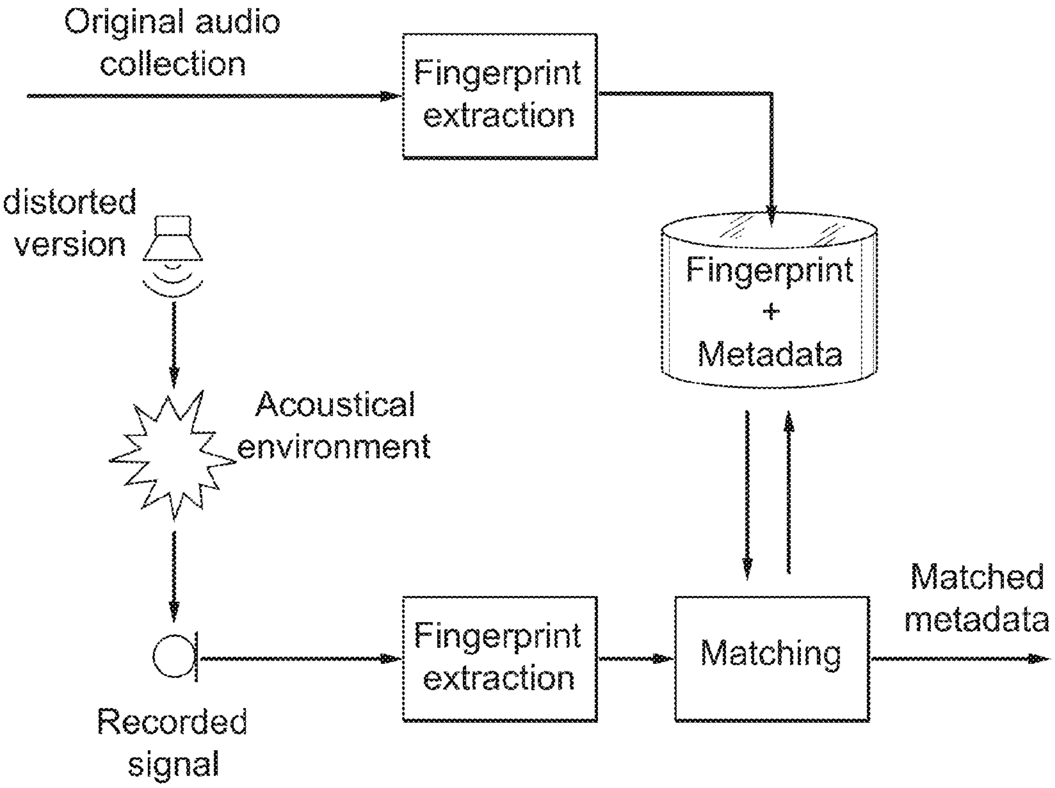
G10L 19/02	(2006.01)
G06F 17/30	(2006.01)
G10L 25/54	(2006.01)
G10L 19/038	(2006.01)
G10L 25/18	(2006.01)

(52) **U.S. Cl.**

CPC *G10L 19/02* (2013.01); *G10L 19/0212* (2013.01); *G10L 19/038* (2013.01); *G10L 25/18* (2013.01); *G10L 25/54* (2013.01); *G06F 17/30743* (2013.01)

(57) **ABSTRACT**

Methods and apparatus for generating a fingerprint of an audio signal are disclosed. The method comprises: detecting peaks in a representation of a temporal spectrum of frequencies of the audio signal, a peak being defined as a point in the representation which has a higher energy than its neighboring points; and generating the fingerprint of the audio signal as a function of a distribution of positions of the detected peaks along a frequency axis and a distribution of positions of the detected peaks along a time axis. The fingerprint of the disclosure is not only robust to many types of noise, but also robust against time scale modification and frequency shifting.



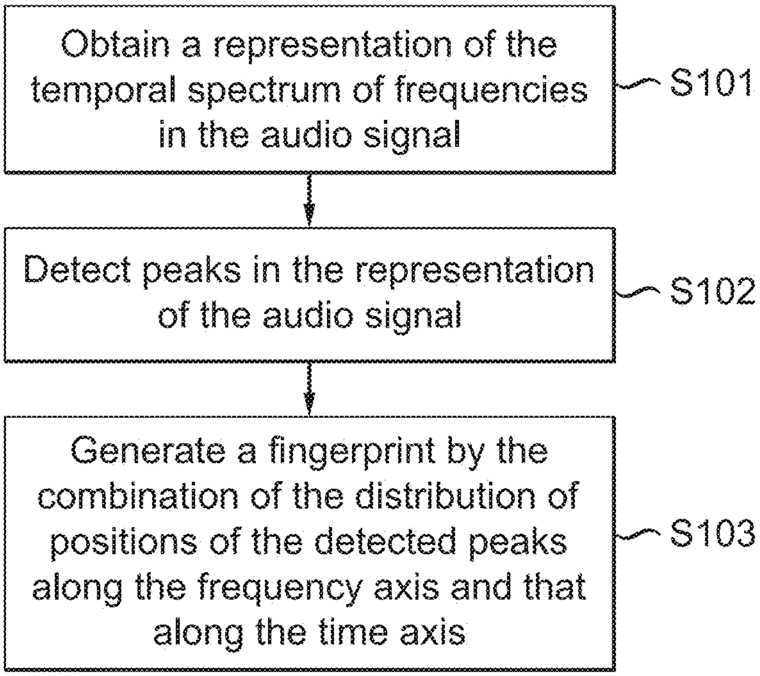

Obtain a representation of the temporal spectrum of frequencies in the audio signal
S101
Detect peaks in the representation of the audio signal
S102
Generate a fingerprint by the combination of the distribution of positions of the detected peaks along the frequency axis and that along the time axis
S103

FIG. 1

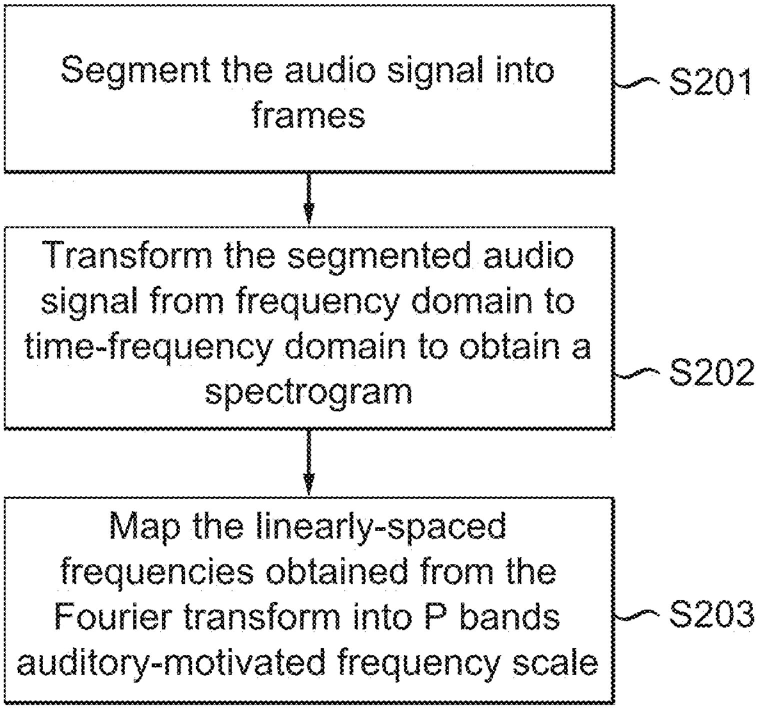

Segment the audio signal into frames
S201
Transform the segmented audio signal from frequency domain to time-frequency domain to obtain a spectrogram
S202
Map the linearly-spaced frequencies obtained from the Fourier transform into P bands auditory-motivated frequency scale
S203

FIG. 2

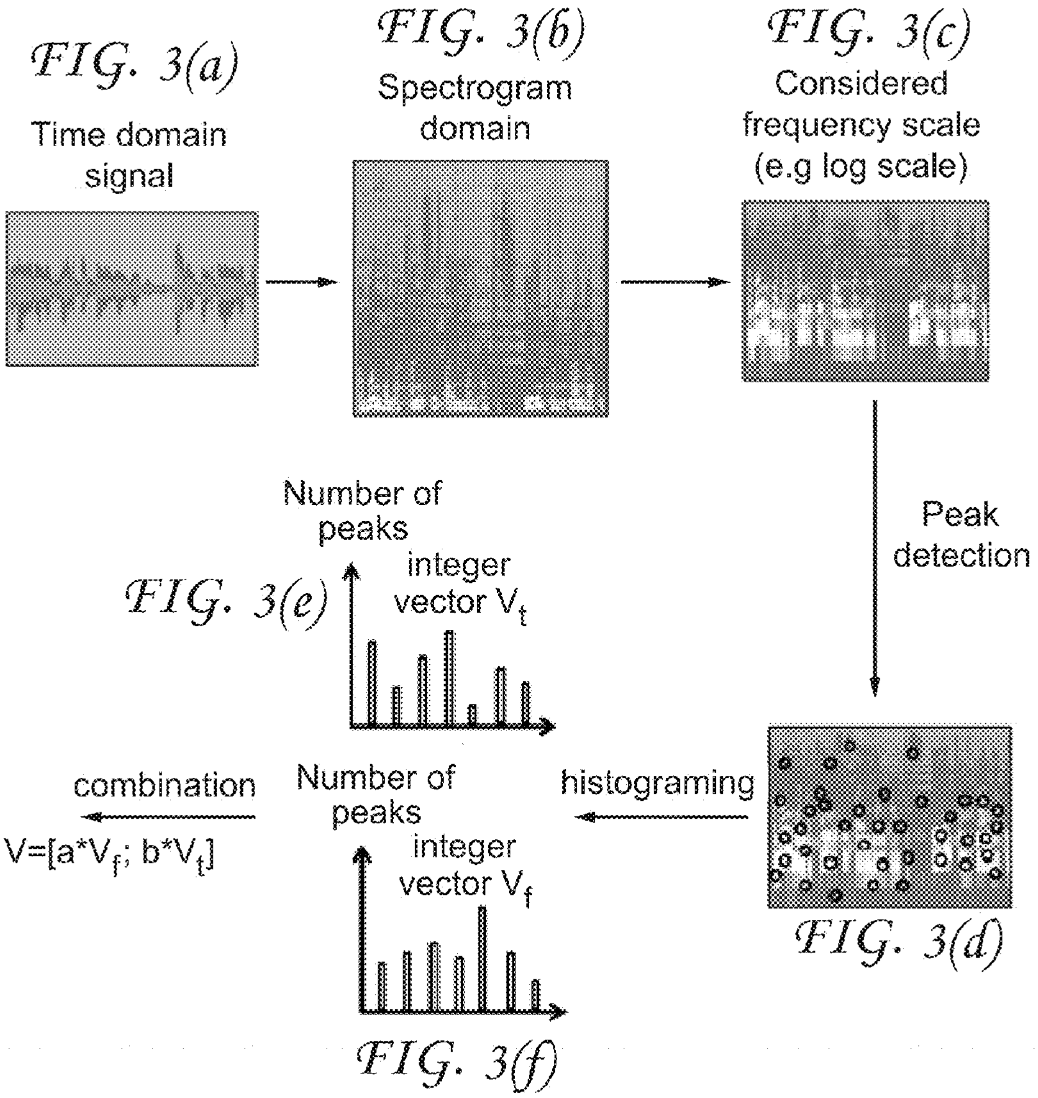


FIG. 3(a)

FIG. 3(b)

FIG. 3(c)

FIG. 3(d)

FIG. 3(e)

FIG. 3(f)

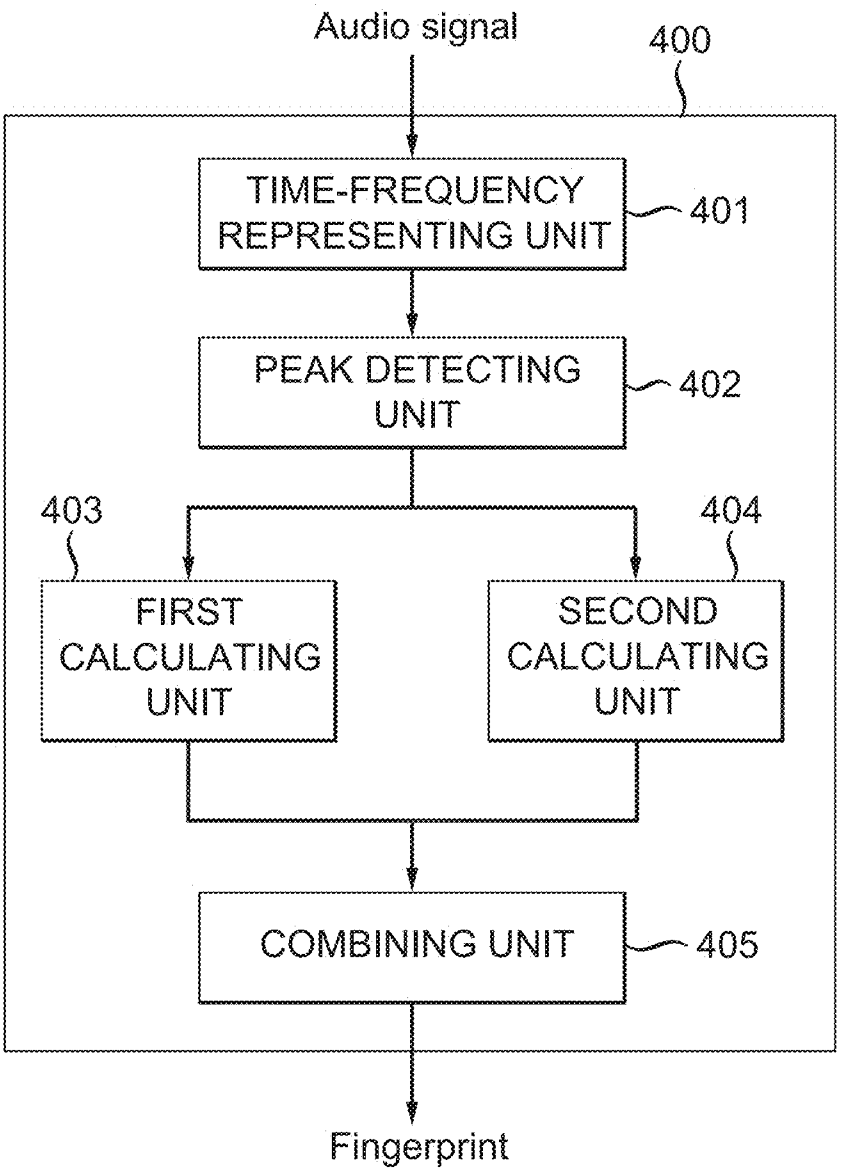

Audio signal
400
TIME-FREQUENCY REPRESENTING UNIT
401
PEAK DETECTING UNIT
402
403
404
FIRST CALCULATING UNIT
SECOND CALCULATING UNIT
COMBINING UNIT
405
Fingerprint

FIG. 4

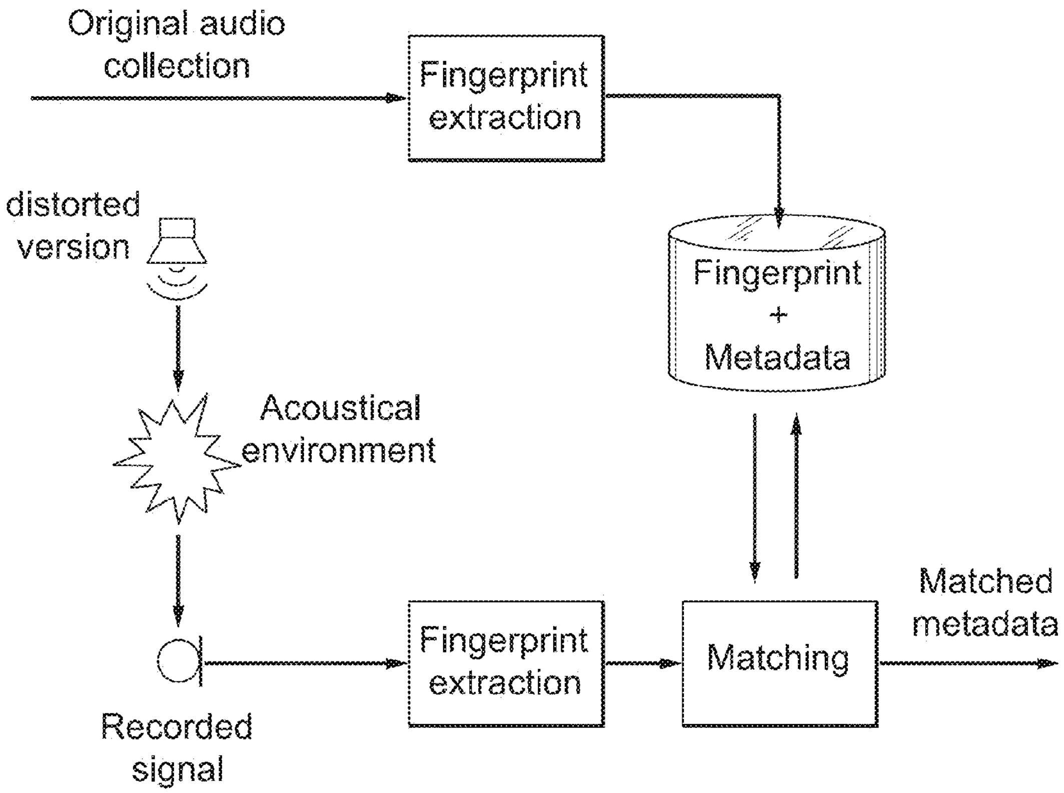

Original audio collection
Fingerprint extraction
distorted version
Fingerprint + Metadata
Acoustical environment
Recorded signal
Fingerprint extraction
Matching
Matched metadata

FIG. 5

METHOD AND APPARATUS FOR GENERATING FINGERPRINT OF AN AUDIO SIGNAL

TECHNICAL FIELD

[0001] The present disclosure relates to the digital audio technology, and in particular to a method and an apparatus for generating a fingerprint of an audio signal.

BACKGROUND

[0002] This section is intended to provide a background to the various embodiments of the technology described in this disclosure. The description in this section may include concepts that could be pursued, but are not necessarily ones that have been previously conceived or pursued. Therefore, unless otherwise indicated herein, what is described in this section is not prior art to the description and/or claims of this disclosure and is not admitted to be prior art by the mere inclusion in this section.

[0003] Audio fingerprinting technique can match distorted unlabeled audio snippets to corresponding labeled data. It has wide range of applications in digital audio technologies, such as audio classification, audio retrieval and content synchronization. As an example, a reference written by A. Wang, "An industrial-strength audio search algorithm", Proc. ISMIR 2003 (hereinafter referred to as reference 1) discusses an audio retrieval system, by which a person who is listening to a music (live, or on radio, . . .) and wants to know more about the singer, name of song, album of the music can simply record a short audio signal and uses it as a query to retrieve metadata information. Another example for the content synchronization is described in a reference written by N. Q. K Duong, C Howson, and Y Legallais, "Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation," IEEE International Conference in Consumer Electronics-Berlin (ICCE-Berlin), 2012 (hereinafter referred to as reference 2), where an audio fingerprint can assure fast and accurate synchronization of media components streamed over different networks to different rendering devices for the implementation of emerging second screen TV applications.

[0004] There are some known solutions for generating fingerprint in the art. In a reference written by Pedro Cano et al, "a review of audio fingerprinting", Journal of VLSI Signal Processing 41, 271-284, 2005 (hereinafter referred to as reference 3), several fingerprinting technologies were introduced. According to the reference 3, basically an audio signal will be subject to a preprocessing, a framing & overlap, a transform, a feature extract and a post-processing by a front end block and then the output is subject to a fingerprint modeling block to generate a fingerprint of the audio signal.

[0005] The above mentioned reference 1 also discusses the generation of an audio fingerprint. In the approach of the reference 1, locations of pairs of energy peaks in the audio spectrogram (i.e. the time-frequency representation of an audio signal) are encoded as fingerprint. In a reference written by J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system", Proc. ISMIR 2002 (hereinafter referred to as reference 4), energy differences between neighboring time-frequency point in the spectrogram are bit-quantized to generate signature.

[0006] Some known fingerprint approaches considered spectrogram as an image and applied computer vision techniques to this spectral image for designing fingerprint. For examples, a reference written by S. Baluja and M. Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," Patent recognition, 2008 (hereinafter referred to as reference 5) proposes to apply wavelet transform to the spectral images and designed Min-Hash signature based on sign of the top wavelet coefficients. In the algorithm provided by a reference written by K. Behun, "Image features in music style recognition", Proc. CESCG 2013 (hereinafter referred to as reference 6), the image based feature SIFT is computed and the histogram of SIFT (a.k.a. the bag-of-word (BoW) feature) is taken as signature. A reference written by M. Riley et al., "A text retrieval approach to content-based audio", Proc. ISMIR 2008 (hereinafter referred to as reference 7) provides an algorithm to use Bag-of-Audio-Word (BoA) for content-based audio retrieval. A reference written by S. Pancoast and M. Akbacak, " Bag-of-Audio-Words Approach for Multimedia Event Classification," Proc. Interspeech 2012 (hereinafter referred to as reference 8) proposes to use BoA for audio event classification.

[0007] However, most of the above known fingerprint solutions are not able to deal with the big time stretching (which for example happens in the process of changing speed or duration of an audio signal to fit the time in a TV or radio program) and the pitch variation (which happens for example in live concert, covered song) although they are robust against noise and distortions (such as A/D conversion, compression). Thus, the known solutions are not robust to some more challenging applications, such as in recognizing songs in live concert, where the recorded audio query is not exactly a distorted version of the original signal in the database (too much variation either time or frequency scale).

[0008] Therefore, there is a need for a method and an apparatus for generating a fingerprint of an audio signal, which is robust to time stretching and pitch variation in audio applications.

SUMMARY

[0009] The present invention disclosure is provided to solve at least one problem of the prior art. The present disclosure will be described in detail with reference to exemplary embodiments. However, the present disclosure is not limited to the exemplary embodiments.

[0010] According to a first aspect of the present invention disclosure, there is provided a method for generating a fingerprint of an audio signal. The method comprises detecting peaks in a representation of a temporal spectrum of frequencies of the audio signal, a peak being defined as a point in the representation which has a higher energy than its neighboring points; and generating the fingerprint of the audio signal as a function of a distribution of positions of the detected peaks along a frequency axis and a distribution of positions of the detected peaks along a time axis.

[0011] In an embodiment, the obtaining of the time-frequency representation of the audio signal comprises segmenting the audio signal into overlapping time frames; and transforming the segmented audio signal from a time domain to a time-frequency domain to generate a spectrogram of the audio signal comprising linearly-spaced frequencies.

[0012] In an embodiment, it further comprises mapping the linearly-spaced frequencies of the spectrogram into P bands of an auditory-motivated frequency scale.

[0013] In an embodiment, the distribution of positions of the detected peaks along the frequency axis is represented by a vector of integer numbers Vf=[V_{f1}, . . . , V_{fF}]T as a function of the number of peaks appearing at each frequency bin, wherein a parameter F is the number of frequency bins and T denote vector transpose; and the distribution of positions of the detected peaks along the time axis is represented by a vector of integer numbers Vt=[V_{t1}, . . . , V_{tN}]F as a function of the number of peaks appearing at each time frame bin, where a parameter N is the number of time frame bins.

[0014] In an embodiment, the function is a concatenation of the vector V_f=[V_{f1}, . . . , V_{fF}]^T and the vector Vt=[V_{t1}, . . . , V_{tN}]^T according to the equation below:

$$V=[a*V_f;b*V_t],$$

wherein a and b are constants.

[0015] In an embodiment, it further comprises adapting the parameters F and N according to a requirement on compactness and robustness of the fingerprint.

[0016] In an embodiment, it further comprises adapting the constants a and b according to a requirement on robustness to either frequency shifting or time scale shifting of the fingerprint.

[0017] In an embodiment, the segmented audio signal is transformed by a Fourier transform.

[0018] According to a second aspect of the present invention disclosure, there is provided an apparatus for generating a fingerprint of an audio signal. The apparatus comprises a time-frequency representing unit for obtaining a representation of the temporal spectrum of frequencies in the audio signal; a peak detecting unit for detecting peaks in the representation of the audio signal, a peak being defined as a point in the representation which has a higher energy than its neighboring points; a first calculating unit for obtaining a distribution of the positions of the detected peaks along a frequency axis; a second calculating unit for obtaining a distribution of positions of the detected peaks along a time axis; and a combining unit for combining the distribution of positions from the first calculating unit and the second calculating unit to generate the fingerprint of the audio signal.

[0019] In an embodiment, the time-frequency representing unit is adapted to segment the audio signal into overlapping time frames; and transform the segmented audio signal from time domain to time-frequency domain to generate a spectrogram of the audio signal comprising linearly-spaced frequencies.

[0020] In an embodiment, the time-frequency representing unit is further adapted to map the linearly-spaced frequencies of the spectrogram into P bands of an auditory-motivated frequency scale.

[0021] In an embodiment, the first calculating unit generates a vector of integer numbers V_f=[V_{f1}, . . . , V_{fF}]^T representing the distribution of positions of the detected peaks along the frequency axis as a function of the number of peaks appearing at each frequency bin, wherein a parameter F is the number of frequency bins and T denote vector transpose; and the second calculating unit generates a vector of integer numbers Vt=[V_{t1}, . . . , V_{tN}]^T to represent the distribution of positions of the detected peaks along the time axis as a function of the number of peaks appearing at each time frame bin, where a parameter N is the number of time frame bins.

[0022] In an embodiment, the combining unit combines the distribution of positions by a concatenation of the vector V_f=[V_{f1}, . . . , V_{fF}]^T and the vector Vt=[V_{t1}, . . . , V_{tN}]^T according to the equation below:

$$V=[a*V_f;b*V_t],$$

wherein a and b are constants.

[0023] According to a third aspect of the present disclosure, there is provided a computer program product downloadable from a communication network and/or recorded on a medium readable by computer and/or executable by a processor, comprising program code instructions for implementing the steps of a method according to the first aspect of the disclosure.

[0024] According to a fourth aspect of the present disclosure, there is provided a non-transitory computer-readable medium comprising a computer program product recorded thereon and capable of being run by a processor, including program code instructions for implementing the steps of a method according to the first aspect of the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0025] The above and other objects, features, and advantages of the present disclosure will become apparent from the following descriptions on embodiments of the present disclosure with reference to the drawings, in which:

[0026] FIG. **1** is a flowchart of a method for generating a fingerprint of an audio signal according to an embodiment of the present disclosure;

[0027] FIG. **2** is a flowchart of a process for obtaining a spectrogram of the audio signal according to an embodiment of the present disclosure;

[0028] FIGS. **3**(*a*)-**3**(*f*) are exemplary diagrams showing the objects resulting from the workflow of the generation of a fingerprint of an audio signal according to an embodiment of the present disclosure;

[0029] FIG. **4** is a block diagram of an apparatus for generating a fingerprint of an audio signal according to an embodiment of the present disclosure; and

[0030] FIG. **5** is illustrates an audio retrieval system which can use the fingerprint generated according to the embodiment of the disclosure for retrieving an audio signal.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0031] Hereinafter, the present disclosure is described with reference to embodiments shown in the attached drawings. However, it is to be understood that those descriptions are just provided for illustrative purpose, rather than limiting the present disclosure. Further, in the following, descriptions of known structures and techniques are omitted so as not to unnecessarily obscure the concept of the present disclosure.

[0032] FIG. **1** is a flowchart of a method for generating a fingerprint of an audio signal according to an embodiment of the present disclosure.

[0033] At step **S101**, it obtains a representation of a temporal spectrum of frequencies in the audio signal.

[0034] It can be appreciated that the representation can be called the spectrogram of the audio signal, which is a visual representation of the spectrum of frequencies in the audio signal varying with time. The spectrogram is actually the time-frequency representation of the audio signal which is normally viewed as a 2D image. In this case, normally the horizontal axis of the spectrogram represents time, and the vertical axis is frequency. There are known ways in the art to obtain the spectrogram of the audio signal, which can be used in the step S**101**. Hereinafter, a process for obtaining a spectrogram of the audio signal will be described with reference to FIG. **2**.

[0035] FIG. **2** is a flowchart of a process for obtaining a spectrogram of the audio signal according to an embodiment of the present disclosure.

[0036] As shown in FIG. **2**, at step **201**, it segments the audio signal into frames.

[0037] At step S**202**, it transforms the segmented audio signal from frequency domain to time-frequency domain to obtain a spectrogram of the audio signal.

[0038] The above steps S**201** and S**202** are for transforming the time domain audio signal into time-frequency domain representation known as spectrogram. In the step S**202**, a Fourier transform can be used for the transform. In this case, the steps S**201** and S**202** can be called a short time Fourier transform (STFT). The spectrogram obtained by the STFT comprises linearly-spaced frequencies varying with time. That is, the horizontal axis of the spectrogram is time, and the vertical axis represents linearly-spaced frequencies of the audio signal. The STFT is well-known in the art. No further details will be given in this respect.

[0039] As shown in FIG. **2**, the process for obtaining a spectrogram of the audio signal can further comprise a step S**203**, where it maps the linearly-spaced frequencies obtained from the STFT into P bands of an auditory-motivated frequency scale. The frequency scale can be Bark, Mel, log scale, or equivalent rectangular bandwidth (ERB) scale. These auditory-motivated frequency scales usually provide finer spectral resolution than the STFT at low frequencies and lower spectral resolution than STFT at high frequencies. Typically P=32, 64, . . .

[0040] The auditory-motivated frequency scales mentioned in the step S**203** are well-known in the art. No further details will be given in this respect.

[0041] Back to FIG. **1**, at the next step S**102**, it detects peaks in the representation, the spectrogram in this case, of the audio signal. Here, a peak is defined as a point in the spectrogram which has a higher energy than its neighboring points in a certain range. In this embodiment, it can be appreciated that the energy is defined as the square magnitude of the STFT coefficient used for the transform.

[0042] As an example, it can detect peaks in the spectrogram, which are points having higher energy than its neighboring points. Please note that the detection of peaks in a spectrogram of an audio signal is known in the art. For example, the reference 1 describes a detection method, which can be used for the step S**102**. No further details will be given in this respect.

[0043] At step S**103**, it generates a fingerprint of the audio signal as a function of the distribution of positions of the detected peaks along the frequency axis and those along the time axis.

[0044] In an example, the above-mentioned distribution can be represented by a histogram which is a graphical representation of the distribution of the peaks along two axes, each axis being divided into bins.

[0045] A detailed description of the generation of a histogram will be provided below.

[0046] Histogram of the positions of the detected peaks along the frequency axis can be obtained by counting the number of peaks appearing at each frequency bin f (denoted by V_f). This histogram feature can be denoted by a F-dimensional vector of integer numbers $V_f=[V_{f1}, \ldots, V_{fF}]^T$, where F is the number of frequency bins and T denote vector transpose. It provides the robustness to time scale modification because intrinsically when time is stretched, the number of peak in each frequency bin is not changed.

[0047] Histogram of the positions of the detected peaks along the time axis can be obtained by counting the number of peaks appearing at each time frame bin. This feature can be denoted by a N-dimensional vector of integer numbers Vt=$[V_{t1}, \ldots, V_{tN}]^T$, where N is the number of time frame bins. It provides the robustness to frequency shifting effect because intrinsically when pitch is shifted, the number of peak in each time frame bin is not changed.

[0048] Note that, the number N depends on both the signal length, and the number of frequency bins F. Given the fixed signal length, N will be higher if F is smaller and vice versa. Thus in a variant dealing mostly with frequency shifting, Vt is advantageously used as a robust fingerprint instead of Vf, and the smaller value of N, the more compact the fingerprint is. In another variant dealing mostly with time-scale distortion, Vf is advantageously used instead of Vt, and the smaller value of F, the more compact fingerprint is. Thus the fingerprint of the audio signal can be generated by a fu of the histogram along frequency axis and that along time axis of positions of the detected peaks. For example, the combination of both histograms can be built as below

$$V=[a*V_f;b*V_t] \qquad (1)$$

[0049] In this example, the generated fingerprint is the concatenation of V_f and V_t, which resulting in (F+N)-dimensional vector of integers. Note that the constant a and b in the equation (1) allow tuning the contribution (weight) between the two histogram in the final fingerprint signature. In applications where there is no scale shifting or the scale shifting is very small, it can set a=0 so as to reduce the fingerprint size, make the signature very robust to pitch variation, and fasten the matching process. Similarly, in applications where the frequency shifting is not concerned, it can set b=0 so that the signature is very robust to time stretch.

[0050] In an embodiment of the disclosure, a weighting scheme can be built for different peak locations, for example, based on prior knowledge about the important regions. In general case, one can set a=b=1, the number of frequency bins can be in the order of 128 (auditory-motivated scale), and the number of time frames N can be in the order of 100. Another way to balance the contribution of Vf and Vt is to set a=N/(F+N) and b=F/(N+F). For example, in case the query length is 4 seconds, the frameshift in the short time Fourier transform (STFT) is 20 ms.

[0051] FIGS. **3**(*a*)-**3**(*f*) are exemplary diagrams showing the objects resulting from the workflow of the generation of a fingerprint of an audio signal according to an embodiment of the present disclosure.

[0052] FIG. **3**(*a*) shows an audio signal in time domain. FIG. **3**(*b*) shows the spectrogram of the audio signal, which is obtained for example by the above-described steps S**201** and S**202**. FIG. **3**(*c*) illustrates the spectrogram after mapping the linearly-spaced frequencies obtained from the STFT into P bands auditory-motivated frequency scale.

[0053] FIG. **3**(*d*) is an exemplary diagram showing the detected peaks in the spectrogram of the audio signal.

[0054] FIG. **3**(*e*) illustrates an example of histogram of the positions of the detected peaks along the time axis, which is obtained by counting the number of peaks appearing at each time frame bin. The output is a vector of integer numbers V_t.

[0055] FIG. **3**(*f*) illustrates an example of histogram of the positions of the detected peaks along the frequency axis, which is obtained by counting the number of peaks appearing at each frequency bin. The output is a vector of integer numbers Vt. Finally as shown in FIG. **3**(*f*), a fingerprint of the audio signal can be generated by the concatenation of V_f and V_t. The generated fingerprint is represented by $V=[a*V_f; b*V_t]$.

[0056] FIG. **4** is a block diagram of an apparatus for generating a fingerprint of an audio signal according to an embodiment of the present disclosure.

[0057] As shown in FIG. **4**, the input of the apparatus **400** is an audio signal.

[0058] The apparatus **400** comprises a time-frequency representing unit **401** for obtaining a representation of the spectrum of frequencies in the audio signal varying with time. A spectrogram of the audio signal can be obtained according to the process described above.

[0059] The apparatus **400** further comprises a peak detecting unit **402** for detecting peaks in the representation of the audio signal.

[0060] The apparatus **400** further comprises a first calculating unit **403** for obtaining the distribution of the positions of the detected peaks along the frequency axis. As described above, the distribution can be represented by a histogram, which can be obtained by counting the number of peaks appearing at each frequency bin.

[0061] The apparatus **400** further comprises a second calculating unit **404** for obtaining the distribution of positions of the detected peaks along the time axis. As described above, the distribution can be represented by a histogram, which can be obtained by counting the number of peaks appearing at each time frame bin.

[0062] The apparatus **400** further comprises a combining unit **405** for combining the histograms from the first calculating unit **403** and the second calculating unit **404** to generate the fingerprint of the audio signal. The combination can be the concatenation of both histograms, which resulting in a vector of integers as the fingerprint of the audio signal.

[0063] The output of the apparatus **400** is a fingerprint of the audio signal. As described above, in an embodiment, it is a vector of integers.

[0064] According to the embodiments of the present disclosure, the peak locations, which are coordinates of peaks in time and frequency axes of the spectral image representation, are very robust to background noise due to the fact that background noise can only change the energy level in most cases, instead of the position of the local maximum energy point.

[0065] The fingerprint generated according to the embodiments of the disclosure is a vector of integer number. It can be applied to the application of similarity search, exhaustive search or Approximate Nearest Neighbor (ANN) search such as LSH, Hamming embedding, product quantization (PQ) code.

[0066] The fingerprint according to the embodiments of the disclosure is not only robust to many types of noise, but also robust against time scale modification and frequency shifting. The fingerprint is compact and therefore applicative for large-scale search. It therefore can bring wide range of applications in both audio retrieval and content synchronization.

[0067] FIG. **5** is illustrates an audio retrieval system which can use the fingerprint generated according to the embodiment of the disclosure for retrieving an audio signal.

[0068] As shown in FIG. **5**, the audio retrieval contains two major parts, one is fingerprint extraction and the other is fingerprint matching. In the fingerprint extraction, a fingerprint is extracted upon a query, for example, from a recorded signal. The fingerprint matching will compare the extracted fingerprint with a fingerprint of an available signal, for example, the original audio collection. The fingerprint of the available signal can be pre-computed and indexed in a database, as similarity search, for retrieval purpose. Detailed information of the matching and retrieval process will not be provided in the disclosure. It only needs to mention here that a robust and compact audio signature associated to each piece (segment) of audio signal is important for purpose of audio signal retrieval.

[0069] As described above, the fingerprint generated according to the embodiment of the disclosure is robust to time stretching and pitch variation in audio applications. In the known arts introduced in the background part, features used in the fingerprint of the reference 1 is robust to the background noise while the resulted fingerprint is not able to deal with the big time stretching and the pitch variation. The bag-of-word (BoW) feature used in the fingerprint of the references 6 and 7 can bring some benefits to those major distortions such as either time scale modification and or pitch shifting. The audio fingerprint according to the embodiment of the disclosure is proposed considering both features discussed in the references 1 and 6, 7. Therefore, the proposed fingerprint can be used in more challenging applications such as in recognizing songs in live concert where the recorded audio query is not exactly a distorted version of the original signal in the database (too much variation either time or frequency scale). In addition, since the fingerprint is a vector of integer numbers, it is very easily integrated to any search well-established engine.

[0070] It is to be understood that the present disclosure may be implemented in various forms of hardware, software, firmware, special purpose processors, or a combination thereof. Moreover, the software is preferably implemented as an application program tangibly embodied on a program storage device. The application program may be uploaded to, and executed by, a machine comprising any suitable architecture. Preferably, the machine is implemented on a computer platform having hardware such as one or more central processing units (CPU), a random access memory (RAM), and input/output (I/O) interface(s). The computer platform also includes an operating system and microinstruction code. The various processes and functions described herein may either be part of the microinstruction code or part of the application program (or a combination thereof), which is executed via the operating system. In addition, various other peripheral devices may be connected to the computer platform such as an additional data storage device and a printing device.

[0071] The present disclosure is described above with reference to the embodiments thereof. However, those embodiments are provided just for illustrative purpose, rather than limiting the present disclosure. The scope of the disclosure is defined by the attached claims as well as equivalents thereof. Those skilled in the art can make various alternations and modifications without departing from the scope of the disclosure, which all fall into the scope of the disclosure.

1. A method for generating a fingerprint of an audio signal, comprising:

detecting peaks in a representation of a temporal spectrum of frequencies of the audio signal, a peak being defined as a point in the representation which has a higher energy than its neighboring points; and

generating the fingerprint of the audio signal as a function of a distribution of positions of the detected peaks along a frequency axis and a distribution of positions of the detected peaks along a time axis.

2. The method according to claim **1**, wherein the obtaining of the representation of the spectrum of frequencies in the audio signal comprises:

segmenting the audio signal into overlapping time frames; and

transforming the segmented audio signal from a time domain to a time-frequency domain to generate a spectrogram of the audio signal comprising linearly-spaced frequencies.

3. The method according to claim **2**, further comprising:

mapping the linearly-spaced frequencies of the spectrogram into P bands of an auditory-motivated frequency scale.

4. The method according to claim **1**, wherein the distribution of positions of the detected peaks along the frequency axis is represented by a vector of integer numbers $V_f=[V_{f1}, \ldots, V_{fF}]^T$ as a function of the number of peaks appearing at each frequency bin, wherein a parameter F is the number of frequency bins and T denote vector transpose; and

the distribution of positions of the detected peaks along the time axis is represented by a vector of integer numbers $Vt=[V_{t1}, \ldots, V_{tN}]^T$ as a function of the number of peaks appearing at each time frame bin, where a parameter N is the number of time frame bins.

5. The method according to claim **4**, wherein the function is a concatenation of the vector $V_f=[V_{f1}, \ldots, V_{fF}]^T$ and the vector $Vt=[V_{t1}, \ldots, V_{tN}]^T$ according to the equation below:

$$V=[a*V_f;b*V_t]$$

wherein a and b are constants.

6. The method according to claim **4**, further comprising adapting the parameters F and N according to a requirement on compactness and robustness of the fingerprint.

7. The method according to claim **5**, further comprising adapting the constants a and b according to a requirement on robustness to either frequency shifting or time scale shifting of the fingerprint.

8. The method according to claim **2**, wherein the segmented audio signal is transformed by a Fourier transform.

9. An apparatus for generating a fingerprint of an audio signal, comprising:

a time-frequency representing unit for obtaining a representation of the temporal spectrum of frequencies in the audio signal;

a peak detecting unit for detecting peaks in the representation of the audio signal, a peak being defined as a point in the representation which has a higher energy than its neighboring points;

a first calculating unit for obtaining a distribution of the positions of the detected peaks along a frequency axis;

a second calculating unit for obtaining a distribution of positions of the detected peaks along a time axis; and

a combining unit for combining the distribution of positions from the first calculating unit and the second calculating unit to generate the fingerprint of the audio signal.

10. The apparatus according to claim **9**, wherein the time-frequency representing unit is adapted to:

segment the audio signal into overlapping time frames; and

transform the segmented audio signal from time domain to time-frequency domain to generate a spectrogram of the audio signal comprising linearly-spaced frequencies.

11. The apparatus according to claim **10**, wherein the time-frequency representing unit is further adapted to:

map the linearly-spaced frequencies of the spectrogram into P bands of an auditory-motivated frequency scale.

12. The apparatus according to claim **9**, wherein

the first calculating unit generates a vector of integer numbers $V_f=[V_{f1}, \ldots, V_{fF}]^T$ representing the distribution of positions of the detected peaks along the frequency axis as a function of the number of peaks appearing at each frequency bin, wherein a parameter F is the number of frequency bins and T denote vector transpose; and

the second calculating unit generates a vector of integer numbers $Vt=[V_{t1}, \ldots, V_{tN}]^T$ representing the distribution of positions of the detected peaks along the time axis as a function of the number of peaks appearing at each time frame bin, where a parameter N is the number of time frame bins.

13. The apparatus according to claim **12**, wherein.

wherein the combining unit combines the distribution of positions by a concatenation of the vector $V_f=[V_{f1}, \ldots, V_{fF}]^T$ and the vector $Vt=[V_{t1}, \ldots, V_{tN}]^T$ according to the equation below:

$$V=[a*V_f;b*V_t]$$

wherein a and b are constants.

14. Computer program comprising program code instructions executable by a processor for implementing the steps of a method according to claim **1**.

15. Computer program product which is stored on a non-transitory computer readable medium and comprises program code instructions executable by a processor for implementing the steps of a method according to claim **1**.

* * * * *