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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 September 2010 (23.09.2010)

PCT

(10) International Publication Number
WO 2010/107315 A1

(51) International Patent Classification:
G10L 11/00 (2006.01) *G10L 21/06* (2006.01)
G10L 17/00 (2006.01) *G10L 15/02* (2006.01)

(21) International Application Number:
PCT/NL2010/050144

(22) International Filing Date:
19 March 2010 (19.03.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09155595.3 19 March 2009 (19.03.2009) EP
61/167,875 9 April 2009 (09.04.2009) US

(71) Applicant *(for all designated States except US)*: **RIJKSUNIVERSITEIT GRONINGEN** [NL/NL]; Broerstraat 5, NL-9712 CP Groningen (NL).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **ANDRINGA, Tjeerd, Catharinus** [NL/NL]; Burd 1, NL-9001 ZV Grou (NL). **KRIJNDERS, Dirk, Johannes** [NL/NL]; Soephuisstraatje 18-2, NL-9712 BZ Groningen (NL).

(74) Agent: **KETELAARS, Maarten**; J.W. Frisolaan 13, NL-2517 JS The Hague (NL).

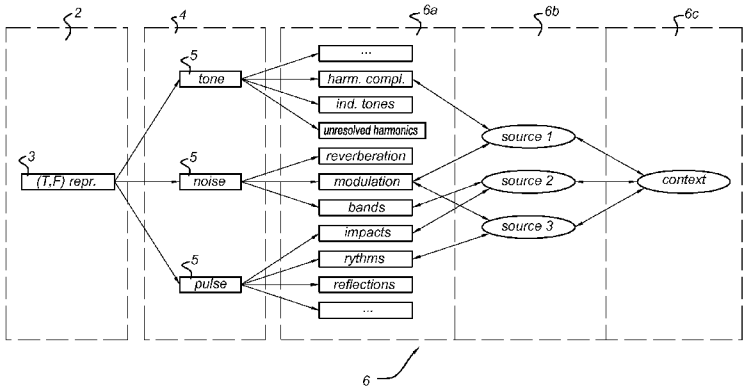
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG,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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) **Title**: TEXTURE BASED SIGNAL ANALYSIS AND RECOGNITION

Fig 2



(57) **Abstract**: System and method for analyzing an input signal. A preprocessing unit (2) is provided which is arranged to first process an input signal to obtain a representation (3) of the input signal in the time-frequency domain. A texture determination unit (4) connected to the preprocessing unit (2) is arranged to subsequently segment the representation in the time-frequency domain into a plurality of regions, in which each region is of a type corresponding to one of a plurality of basic signal textures, in which a basic signal texture is a coherent region in the time frequency domain representation with one or more characteristic properties. An input source classifier (6) is connected to the texture determination unit and arranged to analyze each of the plurality of regions according to the type of basic signal texture to obtain input signal source information.

WO 2010/107315 A1

Texture based signal analysis and recognition

Field of the invention

The present invention relates to a method and system for analyzing an input signal, comprising first processing the input signal to obtain a representation of the input signal in the time-frequency domain.

Prior art

International patent application WO2006/522023 describes a method in which feature vectors are used to classify sound signals. Features are being analyzed using mel-scale frequency cepstral coefficients, and together with output from a harmonic detector, a classifier classifies sound signals.

American patent publication US2008/0082323 discloses a classification system for sound signals. A feature extraction unit is used to determine a large number of possible audio features from an input signal using various descriptors in the temporal domain, frequency domain and a statistical value. These features are then processed and input to a classification unit.

The article by Krijnders et al., 'Demonstration of online auditory scene analysis', of 30 October 2008 discloses a method for signal analysis based on a transmission line model of the human ear, using a cochleogram. The article discloses that signal components are formed, and tonal signal components are grouped into harmonic complexes, which in its turn are used to group signal components likely to stem from a single source. However, no information is disclosed on specific implementations of details of this method.

In known signal processing systems it is attempted to derive sound features directly from the input signal, and neural networks or hidden Markov models are used to obtain a classification or recognition of a possible source. This usually requires training of the system with a large amount of data, and poses stringent requirements on the type of input signals that may be processed.

The problem of stringent limitations on the input is not normally acknowledged because it is not considered part of the domain of sound analysis and sound recognition research. This choice stands at the basis of the current state of the art in intelligent signal analysis and speech and sound recognition which considers the recognition

process as a within-class problem in a closed (and typically quite narrow) domain. This recognition approach considers only problems of the kind: "Which sound from some known and fixed subset is currently presented?" or "Which English words are spoken?" For practical applications this requires a user to guarantee that the input is suitable for the system. Implicitly this transferred the responsibility for the correct functioning of the system from the designer to the user. When optimal operating conditions cannot be guaranteed, these systems produce erratic output. Since the current approach functions by the grace of a suitably limited input it excludes uncontrolled, unconstrained and therefore completely unknown input characteristic of most operating environments. However, users of the prospected systems typically expect these systems to work in conditions they themselves function without difficulty. As such users are easily disappointed by these inbuilt constraints. The invention relaxes many of these constraints and is an enabling technology for a new level of intelligent signal processing.

To summarize. Most researchers and engineers aim to make the most of a suboptimal signal recognition approach. There does not seem to be much effort focused on the more general and more optimal approach as demonstrated by human audition, which functions adequately with almost any input it is exposed to. This generality is the focus of the invention and leads to new business opportunities.

Summary of the invention

The present invention provides a signal analysis method and system, allowing to analyze sources contributing to the signal processed, regardless of the type of input signal or environment thereof.

According to the present invention, a method as defined above is provided, further comprising subsequently segmenting the representation in the time-frequency domain into a plurality of regions, in which each region is of a type corresponding to one of a plurality of basic signal textures, in which a basic signal texture is a coherent region in the time frequency domain representation with one or more characteristic properties, and analyzing each of the plurality of regions according to the type of basic signal texture to obtain input signal source information. This allows to an estimate of information about the physical processes that shaped the signal and it allows a reliably estimate of the possible sources that produced the input signal.

In a further embodiment, segmenting the representation in the time-frequency domain into a plurality of regions comprises analyzing a plurality of time-frequency context areas in the representation, each time-frequency context area having dimensions determined by a time domain correlation length and a frequency domain correlation length as determined for a white noise signal processed as the input signal. As a white noise signal itself is uncorrelated, the calculation of the time-frequency context areas takes into account the correlations induced by the processing steps leading to the time frequency representation. The correlation length is defined herein as meaning a time lag or a distance in channel numbers, respectively, where the correlation drops below a threshold value (for the first time).

The method comprises in a further embodiment calculating for each time-frequency context area a pulse-like contribution and a tonal contribution, wherein the pulse-like contribution can be more than, equal to or less than a noise contribution expected from a white noise signal as input signal, and the tonal contribution can be more than, equal to or less than the noise contribution. According to this basic division in three classes in the time frequency domain for both the pulse-like contribution and the tonal contribution, it is possible to characterize the regions in the time-frequency representation, and hence provide a more reliable prediction of the source producing the input signal.

In a further embodiment, calculating the pulse-like contribution comprises a smoothing operation in the time direction, and calculating the tone-like contribution comprises a smoothing operation in the frequency direction.

In an even further embodiment the pulse-like contribution and tonal contribution are calculated as root-mean-square values. This allows to include both large negative and positive contributions.

In a further embodiment, basic signal textures are determined as one of the group of:

- Tone only, wherein the tonal contribution is larger than the noise contribution and the pulse-like contribution is lower than the noise contribution;
- Pulse only, wherein the pulse-like contribution is larger than the noise contribution and the tonal contribution is lower than the noise contribution;
- Noise, wherein the pulse-like contribution, tonal contribution and noise contribution are similar;

- Tonal noise, wherein the tonal contribution is larger than the noise contribution and the pulse-like contribution and noise contribution are similar;
- Pulse like noise, wherein the pulse-like contribution is larger than the noise contribution and the tonal contribution and noise contribution are similar;
- Unresolved harmonics, wherein the tonal contribution and noise contribution are similar, and the pulse-like contribution is lower than the noise contribution;
- Unresolved pulses, wherein the pulse-like contribution and noise contribution are similar, and the tonal contribution is lower than the noise contribution;
- No energy, wherein the tonal contribution and the pulse-like contribution are both lower than the noise contribution;
- Sweeps, wherein the tonal contribution and the pulse-like contribution are both larger than the noise contribution.

In an even more refined embodiment, the sweeps type of region is further segmented into:

- a shallow sweeps type, wherein the tonal contribution is larger than the pulse-like contribution; and
- a steep sweep type, wherein the pulse-like contribution is larger than the tonal contribution.

In a further embodiment, after segmentation in the plurality of regions, a characteristic property, such as an area, of each of the plurality of regions in the time frequency domain is determined, and a region is assigned to an adjacent (or even surrounding) region if the characteristic property of the region in the time frequency domain is lower than a threshold value. This eliminates spurious contributions and provides a more robust intermediate result in determining a source of an input signal.

Analyzing each of the plurality of regions includes determining a likelihood of a possible source based on context information in a further embodiment. This allows to use also a further source of information when determining the likelihood that a specific source has contributed to the input signal.

In a further aspect, the present invention relates to a system for analyzing an input signal, comprising a preprocessing unit arranged to first process an input signal to obtain a representation of the input signal in the time-frequency domain, a texture determination unit connected to the preprocessing unit arranged to subsequently segment the representation in the time-frequency domain into a plurality of regions, in

which each region is of a type corresponding to one of a plurality of basic signal textures, in which a basic signal texture is a coherent region in the time frequency domain representation with one or more characteristic properties, and an input source classifier connected to the texture determination unit and arranged to analyze each of the plurality of regions according to the type of basic signal texture to obtain input signal source information. The various elements of the system may be implemented in separate hardware units, or may be implemented as functional units (e.g. using software programming) in a single processing system. The texture determination unit and/or the input source classifier are further arranged to implement the present invention method embodiments.

In an even further aspect, a software program product is provided comprising computer executable code, which when loaded on a processing system receiving an input signal, provides the processing system with the functionality of the present invention method embodiments.

Short description of drawings

The present invention will be discussed in more detail below, using a number of exemplary embodiments, with reference to the attached drawings, in which

Fig. 1 shows a flow diagram according to an embodiment of the present invention;

Fig. 2 shows a schematic diagram depicting the various processing steps according to an embodiment of the present invention;

Fig. 3 shows a flow diagram showing in detail steps implemented in the texture determination unit of Fig. 3;

Fig. 4a-d show gray scale plots of parts of a time frequency domain representation of an input signal in the various steps of Fig. 3 for determining pulse like contributions;

Fig. 5a-d show gray scale plots of parts of a time frequency domain representation of an input signal in the various steps of Fig. 3 for determining tonal contributions; and

Fig. 6a-c show gray scale plots of a time frequency domain representation of an input signal in which qualitative basic texture regions are determined based on tone texture (a), pulse texture (b), and as basic signal textures (c).

Detailed description of exemplary embodiments

In Fig. 1 a schematic diagram is shown of (functional) units which execute parts of the present invention embodiments to analyze an input signal to obtain signal sources having formed a (complex) input signal. The functional units 1, 2, 4, 6 are e.g. implemented as software modules on a general purpose computer, which may comprise one or more (virtual) data processors. One or more of the functional units 1, 2, 4, 6 may also comprise dedicated signal processors, such as digital signal processors (DSP) and the like. Each of the units 1, 2, 4, 6 may also be equipped with memory units, used to store instructions as a software program, and also (intermediate) results of calculations. Furthermore, the data/signal processors may be part of a network, allowing decentralized data processing.

A signal input unit 1 is gathering and combining signals from one or more microphones, recordings, or other sources and creates an (audio) input signal, which is transferred to pre-processing unit 2. The input is here assumed to be a time-domain signal, but it can be any mapping from the real line into an arbitrary vector space. The pre-processing unit 2 is arranged to process the received input signal in a representation 3 of the input signal in the time-frequency domain (T, F), or any other domain constructed from applying frequency specific filters to the mapping from the real line to the vector space. In the frequency direction, a discrete number of channels are used, and in the time direction, a discrete time interval is used. The representation 3 of the input signal in the time frequency domain (T, F) may be plotted as the energy of the input signal as function of time and frequency. In order to match the (human) ear behavior the input signal can be processed to obtain a cochleogram, which is a special type of representation 3 in the time frequency domain. A cochleogram is a spectrogram like representation of the spectro-temporal energy distribution of the input signal. A cochleogram is based on an auditory model with the spatio-temporal continuity properties and the logarithmic place-frequency relation of the mammalian cochlea.

In known signal processing systems it is attempted to derive sound features directly from the input signal, and use neural networks or hidden Markov models to obtain a classification or recognition of a possible source. This usually requires training of the system, and stringent requirements on the type of input signals that may be processed. However in the present invention embodiments the analysis is based on the

time frequency representation of the input signal directly. More specifically, a texture determination unit 4 processes the (T, F) representation 3 to obtain a set of textures 5 each identifying a specific area in the (T, F) representation 3, as explained in more detail below. The set of textures 5 is then input to an input source classifier 6 to obtain the final classification result.

In general terms, the present method segments the representation 3 in the time frequency domain (T, F) directly into a plurality of regions, i.e. a number of qualitatively different, clearly defined and delineated coherent regions with one or more characteristic properties. Each region corresponds to one of a plurality of basic signal textures. For sounds and other 1-D time-series as input signal, these textures can be referred to as tonal, pulse-like, or noisy textures. The resulting data (i.e. each of the plurality of regions) are analyzed according to the type of basic signal texture to obtain input signal source information. Specific algorithms are associated with each of these basic signal textures, which can be used to elicit the source information that shaped the input signal in that area in the time frequency domain. The direct link between bottom-up estimated textures (basic signal textures) and the algorithms appropriate for detailed analysis of the regions is exploited in the system overview as depicted in Fig. 2.

The system is always driven by an input signal (e.g. sound) which is converted into a (T, F) representation 3 like a cochleogram in preprocessing unit 2. This (T, F) representation 3 is dissected in the texture determination unit 4 into different regions 5 of a specific type (tone, noise, or pulse indicated in Fig. 2), which are analyzed further by specific algorithms in an application part 6a of the input source classifier 6. In further embodiments, more types of textures may be present (see embodiments described below), which each are associated with one or more algorithms to be applied in the first part 6a. These detailed analyses may lead to evidence indicating the presence of specific sound sources, e.g. using statistical features of likelihood that a specific source (source 1, source 2, source 3 indicated in Fig. 2) has caused (part of) the input signal. This is stored and updated in a second part 6b of the input source classifier 6. Already the pattern of these sources (source 1, source 2, source 3) may indicate a more global context/environment/situation (e.g. airport, train area, football stadium), which is determined in a third part 6c of the input source classifier 6. This is a bottom-up or signal-driven approach.

Conversely, knowledge or expectations of being in some environment (context in third part 6c) will also indicate likely sources (source 1, source 2, source 3) to be present. This entails specific expectations and possibly a justified lowering of detection thresholds of the algorithms applied in part 6a (based on this knowledge or context). This is a top-down, or active and knowledge-driven, approach in which recognition results after checking a few hypotheses. In other words, analyzing each of the plurality of regions as implemented in second part 6b includes determining a likelihood of a possible source (source 1, source 2, source 3) based on context information available in third part 6c.

The flexible combination of these two approaches is characteristic of human audition, and is reflected in the present system as shown in Fig. 2. This embodiment makes it possible to analyze natural signals optimally through a combination of smart signal processing, artificial intelligence, source physics, and cognitive science.

As a result, the embodiments as described herein impose no restrictions whatsoever on the input signal. The input signal may be unknown and unconstrained. Furthermore, the input source classifier 6 can operate more efficiently, because specific types of algorithms are used depending on the type of basic signal texture being present.

In the present embodiments, the presence and properties of positive signal-to-noise (SNR) regions in the (T, F) representation 3 are estimated (in texture determination unit 4) since these regions correspond to how sources radiate most of their acoustic energy (pulse-like, tonal, or noisy). For that, method steps are provided in an embodiment of the present invention (as shown for an exemplary embodiment in Fig. 3) to determine the textures associated with the three classes (tonal, pulse-like and noise) for all points in the time-frequency plane. The texture determination measures the contributions of tonal, pulse like, and noisy patterns within a suitably chosen time-frequency context area.

The time-frequency context area (TFC-area) is determined by a time domain correlation length and a frequency domain correlation length. In other words, the correlation lengths in the frequency and temporal direction of the time-frequency representation of a white noise signal as processed in the preprocessing unit 2. Because a white noise signal is itself uncorrelated the correlations are induced by the processing

steps leading to the (T, F) representation 3, which are representation and channel dependent.

The correlation length in the frequency direction is computed by computing the correlation between segments when the cochlea-model is excited with a white noise signal. The frequency domain correlation length is the distance in channel numbers at which the correlation drops, below a given minimal correlation, e.g. a threshold value of 0.1, for the first time. The frequency domain correlation lengths may be different for the upward direction and the downward direction. In practice each channel is associated with a range of neighboring channels with which it correlates more than the threshold value.

The time domain correlation length or temporal correlation per channel can be calculated with an autocorrelation within each channel. The temporal correlation length is determined as the time lag where the correlation drops below the threshold value for the first time. In general, this is the same threshold value as used for determining the frequency domain correlation length, although it may be different.

The contribution of pulses and tones is measured in relation to what a white noise input signal would represent in the two kinds of excitations. The contribution of both pulse like and tonal information can be more, equal or less than a contribution to be expected of a white noise signal. Types of textures differ in the relative average (local) content of pulse like and tonal contributions according to the next table.

	Tone<noise	Tone ≈ noise	Tone > noise	Tone > pulse > noise
Pulse < noise	No energy	Unresolved harmonics	Tone only	-
Pulse ≈ noise	Unresolved pulses	Noise/ mix tones-pulses	Tonal noise	-
Pulse > noise	Pulse only	Pulse like noise	Sweeps/mixed	Sweeps shallow
Pulse > Tone > noise	-	-	Sweep steep	-

Texture computation (as e.g. implemented in the texture determination unit 4 of Fig. 1) is in an embodiment a three step process as shown in the flow diagram of Fig. 3. From the preprocessing unit 2, the time frequency domain representation 3 of the input signal is obtained (indicated as calculation step 11 in Fig.3):

Step 1. The computation of a local measure for each time-frequency point for both tonal contributions (calculation step 12a) and pulse-like contributions (calculation step 12b).

Step 2. The computation of the root-mean-square of the local measure in a suitably chosen environment for pulse-like and tonal information (calculation steps 13a and 13b, respectively) to obtain basic signal textures (calculation step 14).

Step 3. Combination of the results of step 2 into a single texture measure according to the table given above (calculation step 15).

The process of obtaining the set of single textures for a (T, F) representation 3, i.e. the steps 1-3 in Fig. 3, is visualized in Fig. 4 and 5, which show energy as function of time and frequency (or segment and channel in the cochleogram). Fig. 4a and 5a show a close-up of a region of the cochleogram (channels 31 to 65) containing both tonal (left) and pulsal (or pulse-like) (center) components. Fig. 4b and 5b show two smoothed versions of the cochleogram, e.g. using the correlation calculations over the time frequency context areas as described above.

Fig. 4b shows a version which is smoothed in the horizontal (time) direction with a scope depicted as the black line in Fig. 4a, which corresponds to the temporal correlation length. Fig. 5b shows a version which is smoothed in vertical (frequency) direction with a scope depicted as the vertical black line in Fig. 5a, which corresponds to the local channel correlation length (i.e. in frequency direction).

For pulses, smoothing in the vertical (frequency) direction does not lead to large differences. However smoothing in the horizontal direction leads to a much lower value around the energy peak of the pulses. For tones the situation is reversed: smoothing in the horizontal direction does not lead to lower values, but smoothing in the vertical direction does.

Fig 4c and 5c show the difference (subtraction) of the original cochleogram and the smoothed variants. In this picture the grey scale coding must be interpreted as gray for values around zero, black for negative, and light for positive values. The value corresponds to deviations in terms of standard deviations from the mean of the white

noise response. Fig. 4c shows high positive and negative values around the pulse-like vertical structure, and near zero values for the tonal components. Fig. 5c shows the reverse pattern: low values near the pulse and high values around the position of the tones. This is the result of step 1 of the method as described above: Fig. 4c depicts the local pulsality measure, Fig. 5c the local tonality measure.

Fig. 4d and 5d show the root-mean-square (RMS) of the local pulsality computed for the region depicted as the white rectangle in Fig. 4c and the RMS of the tonality computed for the region depicted as the white rectangle in Fig. 5c. The square operation in the RMS calculation treats large negative values the same as large positive values and returns a positive response. Therefore these values are normalized to a unit mean, unit standard deviation white noise response. This is the result of step 2. These values form the basic texture information and can be used in many ways.

As discussed above, in relation to the system overview of Fig. 2, the obtained measures for tonality and pulse-like contributions are used to obtain the basic signal textures. In Fig. 6, the resulting qualitative regions are shown in the time-frequency representations of the input signal. Coherent regions are shown in a corresponding grey scale.

In Fig. 6a, the qualitative regions based on tone textures are shown in one of five grey-scale values, indicating a region of no energy, less than noise evidence, noise-like evidence, tonal evidence, larger than pulse-like evidence, and pure tonal evidence.

In Fig. 6b a similar representation is shown based on pulse like textures. Again, five grey scales are used to indicate no energy, less than noise evidence, noise-like evidence, pulse-like larger than tonal evidence and pure pulse-like evidence.

In Fig. 6c a representation is shown in 8 grey scales, indicating one of eight basic signal textures (taken as a subset of the textures as indicated in the table shown above).

The grey scale values of the representation as shown in Fig. 6c each indicate a basic signal texture, which is the result of estimation of textures and segmentation of the time frequency domain. In Fig. 6c, e.g., the boxes labeled A and B (light grey) indicate regions which are dominated by pulses (23 Hz for box A, and 17 Hz for box B). The values 23 and 17 Hz indicate the average repetition frequency of the main constituents. The grey scale coding indicates which type of algorithm is appropriate for further analysis of that region in algorithm application part 6a (see Fig. 2). E.g. in this

case, repetitions in the time dimension can be searched for, neglecting details in the frequency direction.

In a further embodiment, the method is further refined to remove possible spurious contributions, making the eventual determination of a source more robust. For this, after segmentation in the plurality of regions, a characteristic property such as an area of each of the plurality of regions in the time frequency domain is determined, and a region is assigned to an adjacent region if the characteristic property of the region in the time frequency domain is lower than a threshold value. The adjacent region can also be a surrounding region.

This method embodiment describes a way to remove spurious contributions, or alternatively correct erroneous deletions from a binary mask, which e.g. resulted from a noisy signal. This mask can be the results of one of the earlier described embodiments, or any other binary operation on the input. During mask-forming in noisy conditions, the probability of (typically small) errors is non-zero. The number of these errors can be reduced by exploiting suitable noise-properties. For example when the size-distribution of the spurious areas favors certain sizes, it is possible to use this distribution to discard all areas that are likely to be too small. A threshold value of for example two standard deviations (2 std) above the mean size can be chosen to discard (or fill) selected regions (or holes in masks) that are smaller than the threshold value.

In a first step, a white noise signal is processed with the method embodiments described so far to arrive at a binary mask in the time frequency domain. The mask should include the complete TF plane, but mask-forming errors (e.g. noise not included in a noise-mask) will exist. Then for each channel the standard deviation of an area or other measure of these regions containing this channel is calculated

Then a threshold application procedure is executed. A signal is processed with the method embodiments described so far and masks are formed for the regions. All regions that have a area smaller then the threshold value of the corresponding channel are assigned to their surrounding (mask-)region.

In the following paragraphs, a theoretic description is given on how to implement the various calculations for determining the basic textures.

The uncertainty relation relates the temporal spread of a signal to its frequency spread by putting a lower limit on their product. If the spread Δx in a variable x is defined by:

$$\Delta x = \sqrt{\mu(x^2 - \mu^2(x))} \qquad (1)$$

Then the spread in two variables $t, f,$ which are related by a Fourier transformation and which corresponds to a transformation from the temporal to the spectral domain, the most general example being the Fourier transform, is given by the uncertainty relation

$$\Delta t \Delta f \geq C_1 \qquad (2)$$

The actual value of the constant C_1 depends on the details of the transformation relating the t and the f variables, but given such a transformation they are fixed. We define the localization as the logarithm of the inverse spread:

$$\begin{aligned} L_t &= \log(1/\Delta t) \\ L_f &= \log(1/\Delta f) \\ L &= L_t + L_f. \end{aligned} \qquad (3)$$

We can use these to express the uncertainty principle as an upper boundary on the sum of the localizations,

$$L = L_t + L_f \leq C_2. \qquad (4)$$

where $C_2 = log(1/C_1)$.

The texture analysis can be used with a time-frequency representation within which the white noise response has a bounded correlation length ε in both time and frequency. We define the correlation length around (t, f) with four values:

lower bound correlation in frequency direction: $\varepsilon_f(f)$

upper bound correlation in frequency direction: $\varepsilon^f(f)$

lower bound correlation in time direction: $\varepsilon_t(f)$

upper bound correlation in time direction: $\varepsilon^t(f)$

A small note on notation: In general TF-representations like the energy with no arguments E or two arguments $E(t, f)$ are used to denote the energy representation of the arbitrary signal being processed. In some cases when it is intended that the TF-representation refers to a reference signal like white noise w, an extra argument is added at the end of the argument list to denote the reference signal $E(w)$ and.

The correlation lengths are based on the correlations for white noise:

$$R_{\delta f}(E) = \frac{\mu((E(w)-\mu(E(w)))(E_{\delta f}(w)-\mu(E_{\delta f}(w))))}{\sigma(E(w))\sigma(E_{\delta f}(w))}$$
$$R_{\delta t}(E) = \frac{\mu((E(w)-\mu(E(w)))(E_{\delta t}(w)-\mu(E_{\delta t}(w))))}{\sigma(E(w))\sigma(E_{\delta t}(w))} \quad (5)$$

where $E_{\delta f}(w)$ stands for the frequency shifted $E(w)$ matrix $E(t, f + \delta f, w)$ and analogous $E_{\delta t}(w)$ stands for the time shifted $E(w)$ matrix $E(t+\delta t, f, w)$. μ and σ indicate the mean and the variance of the matrix they take as an argument.

Using the definition above the correlation lengths are defined as:

$$\epsilon_f(f) = \sup_{\delta f<0} : R_{\delta f}(E) < \theta$$
$$\epsilon^f(f) = \inf_{\delta f>0} : R_{\delta f}(E) < \theta$$
$$\epsilon_t(f) = \sup_{\delta t<0} : R_{\delta t}(E) < \theta$$
$$\epsilon^t(f) = \inf_{\delta t>0} : R_{\delta t}(E) < \theta \quad (6)$$

Where θ is a conveniently chosen threshold value

The sets $N_T(t,f)$, $N_F(t,f)$ and $N_N(t,f)$ contain different neighbourhoods of the point (t,f). They are defined by the following relations:

$$(t', f') \in N_T(t,f) \Leftrightarrow (\epsilon_t(f) < t' - t < \epsilon^t(f)) \wedge (f' = f)$$
$$(t', f') \in N_F(t,f) \Leftrightarrow (t' = t) \wedge (\epsilon_f(f) < f' - f < \epsilon^f(f))$$
$$(t', f') \in N_N(t,f) \Leftrightarrow (\epsilon_t(f) < t' - t < \epsilon^t(f)) \wedge (\epsilon_f(f) < f' - f < \epsilon^f(f)) \quad (7)$$

The N_T neighbourhood stretches out in the temporal direction and is limited to a single frequency channel. The N_F neighbourhood stretches out in the frequency direction and is limited to a single time point in time. The N_N neighbourhood stretches out in both directions. These neighbourhoods can be used to define smoothed (moving averaged) versions of the original time-frequency representation, for which we take here the energy development per frequency channel $E(t,f)$.

$$\overline{E}_T(t,f) = \mu_{N_T(t,f)}(E)$$
$$\overline{E}_F(t,f) = \mu_{N_F(t,f)}(E)$$
$$\overline{E}_N(t,f) = \mu_{N_N(t,f)}(E) \quad (8)$$

The symbol $\mu_{X(E)}$ with X a subset of the TF-plane is used to denote the average of the $E(t',f')$ for which $(t',f')\in X$. Similarly, we will use the symbol $\sigma_{X(E)}$ with X a subset of the TF-plane to denote the standard deviation of the $E(t',f')$ for which $(t',f')\in X$.

From these smoothed averages we can derive unnormalized Energy Peaks Above Surrounding values (EPAS-values):

$$\begin{aligned} \tilde{P}_T(t,f) &= E_T(t,f) - \overline{E}_T(t,f) \\ \tilde{P}_F(t,f) &= E_F(t,f) - \overline{E}_F(t,f) \\ \tilde{P}_N(t,f) &= E_N(t,f) - \overline{E}_N(t,f) \end{aligned} \qquad (9)$$

which are then used to define white noise normalized EPAS-values

$$\begin{aligned} P_T(t,f) &= \frac{\tilde{P}_T(t,f) - \mu_{\{w\}}(\tilde{P}_T(t,f,w))}{\sigma_{\{w\}}(\tilde{P}_T(t,f,w))} \\ P_F(t,f) &= \frac{\tilde{P}_F(t,f) - \mu_{\{w\}}(\tilde{P}_F(t,f,w))}{\sigma_{\{w\}}(\tilde{P}_F(t,f,w))} \\ P_N(t,f) &= \frac{\tilde{P}_N(t,f) - \mu_{\{w\}}(\tilde{P}_N(t,f,w))}{\sigma_{\{w\}}(\tilde{P}_N(t,f,w))} \end{aligned} \qquad (10)$$

where the $\{w\}$ in the expressions for the mean and the variance indicates that we are averaging over different realizations of white noise.

The P or EPAS-values we found here describe the peakedness of the energy development with respect to different directions in the TF-plane. In general, P_F will be high where tones are present and P_T will high where pulses are present. P_N will be highest for isolated peaks rising from the TF-plane.

We start with defining a mask to eliminate peaks which occur in relatively silent areas. As a threshold E_{min} we take a value of e.g. 60dB under the maximal energy found in the TF-plane. This energy we use to define the mask M_E,

$$(t',f') \in M_E \quad \Leftrightarrow \quad E(t,f) > E_{min}$$

The statistical properties of the $P_X(t, f)$ can be studied by studying the statistical properties in areas which are larger and oriented perpendicular than those used in the definitions of the area $N_T(t,f)$, $N_F(t,f)$ and $N_N(t,f)$. To achieve this we define new frequency dependent range boundaries,

$$c_t=a_t\varepsilon_t \qquad c^t=a^t\varepsilon^t$$

$$c_f^{\cdot}=a_f\varepsilon_f \qquad c^{f\cdot}=a^f\varepsilon^f$$

where the a_X and a^X are small constants larger than 1 which can be used to optimize the algorithm. Using these new range boundaries we can now define new neighbourhoods,

$$\begin{aligned}
(t',f') \in \Lambda_T(t,f) &\Leftrightarrow (c_t(f) < t'-t < c^t(f)) \wedge (f'=f) \\
(t',f') \in \Lambda_F(t,f) &\Leftrightarrow (t'=t) \wedge (c_f(f) < f'-f < c^f(f)) \\
(t',f') \in \Lambda_N(t,f) &\Leftrightarrow (c_t(f) < t'-t < c^t(f)) \wedge (c_f(f) < f'-f < c^f(f))
\end{aligned} \tag{11}$$

These areas are now used to define basic textures,

$$\begin{aligned}
\tilde{B}_T^{\cdot}(t,f) &= \mu_{\Lambda_T(t,f)\cap M_\infty}(P_T^{\cdot}) \\
\tilde{B}_F^{\cdot}(t,f) &= \mu_{\Lambda_F(t,f)\cap M_\infty}(P_F^{\cdot}) \\
\tilde{B}_N^{\cdot}(t,f) &= \mu_{\Lambda_N(t,f)\cap M_\infty}(P_N^{\cdot})
\end{aligned} \tag{12}$$

Here the dot in the exponent indicates that the exponentiation should be done element wise instead of matrix wise.

The textures are now defined by normalizing them by white noise values.

$$\begin{aligned}
B_T(t,f) &= \frac{\tilde{B}_T(t,f) - \mu_{[w]}(\tilde{B}_T(w))}{\sigma_{[w]}(\tilde{B}_T(w))} \\
B_F(t,f) &= \frac{\tilde{B}_F(t,f) - \mu_{[w]}(\tilde{B}_F(w))}{\sigma_{[w]}(\tilde{B}_F(w))} \\
B_N(t,f) &= \frac{\tilde{B}_N(t,f) - \mu_{[w]}(\tilde{B}_N(w))}{\sigma_{[w]}(\tilde{B}_N(w))}
\end{aligned} \tag{13}$$

The main application area of the present invention embodiments is autonomous intelligent signal analysis and signal recognition. It can be used for any system that receives an input signal (such as sounds) from a (possibly) unknown, and (possibly) unconstrained and (possibly) very rich environment. Current systems will function only when very limiting requirements are met: typically the signal should originate from the same type of environment as the training data and that this environment is sufficiently small. This is qualitatively different from human audition that functions adequately with almost any input it is exposed to. The same will hold for systems based on the present embodiments.

As such it is possible to develop intelligent autonomous signal analysis and recognition systems that function, just as human audition, under two very weak (generally valid) basic assumptions:

• The sound of a sound source consists of events defined by an onset, a continuous development (typically in the TF-plane), and an offset.

• Physical source constraints in signal allow and determine recognizability. These basic assumptions are satisfied for (virtually) all natural signals.

This entails that the application area of the present invention embodiments is huge. Some examples:

• Autonomous automatic signal analysis systems that have analyzed the input in full and that can tell the user about the specifics of the signal. This might revolutionize the way signal processing is taught on universities and it may even introduce quite sophisticated signal processing in secondary school projects (e.g. to analyze the sound of bees, speech, etc)

• Automatic monitoring systems that count and describe different events like the number and type of passing vehicles, when people are present and what kind of activities they do. Note that these monitoring systems are a generalization of for example an aggression detection system that is suitable for a single type of event.

• Condition monitoring systems that are able to determine the global state of a complex machine or system and that are able to track wear- and tear in machines, predict immanent failure, postpone unnecessary maintenance, and generate management information for industrial processes.

• Models of auditory perception to build systems that implement specific auditory functions. These can be used to determine the pleasantness of designer sounds, the acoustic quality of concert halls and rooms for mentally handicapped (who are often exquisitely sensitive to sound), and the mental intrusiveness of environmental sounds for use in schools, offices, houses, etc.

• Auditory monitoring for robots and other autonomous systems that need acoustic information to detect novelty and to detect and analyze relevant environmental activities and processes.

• Medical applications ranging from counting and analyzing the number of coughs to estimate medicine effectiveness, listening to sleepers with apneu to signal the activation

of a breathing assistant, to intelligent heart monitoring systems that function with a similar level of intelligence as the constant presence of a hart-specialist.

• Soundscape analysis for environmental monitoring or soundscape quality estimation, to measure noise pollution, to improve sound annoyances regulations, to measure the perceptual efficiency of noise mitigation measures (such as sound barriers).

• Automatic annotation systems for audiovisual libraries (tv, radio, movies, Youtube-content). Currently the content of these libraries is only minimally accessible and time-consuming to search through.

• Audio-coding will benefit from the invention because it becomes possible to approximate perceptual irrelevant textural detail by an efficient higher level description of the texture.

• Multisensor analysis of physiological signals, e.g. EEG, ECG, EMG and MEG and subdural electrode grids. At present there is strong need for systems able to predict seizures and other form of epileptic activity. And increasingly multisensor monitoring of physiological signals is used to create interfaces between physiological systems and machines, like e.g. prosthetic devices and computer interfaces.

• Automated movement, position and pressure and sound analysis in animal monitoring systems. Animal research is changing from low throughput to high throughput, for this purpose more and more animal screening is needed. Movement, position, pressure and sound give viable information about animal behavioral states, which with our invention can be automatically classified.

CLAIMS

1. Method for analyzing an input signal, comprising
first processing the input signal to obtain a representation of the input signal in the time-frequency domain, and subsequently
segmenting the representation in the time-frequency domain into a plurality of regions, in which each region is of a type corresponding to one of a plurality of basic signal textures, in which a basic signal texture is a coherent region in the time frequency domain representation with one or more characteristic properties, and
analyzing each of the plurality of regions according to the type of basic signal texture to obtain input signal source information.

2. Method according to claim 1, wherein segmenting the representation in the time-frequency domain into a plurality of regions comprises
analyzing a plurality of time-frequency context areas in the representation, each time-frequency context area having dimensions determined by a time domain correlation length and a frequency domain correlation length as determined for a white noise signal processed as the input signal.

3. Method according to claim 2, further comprising
calculating for each time-frequency context area a pulse-like contribution and a tonal contribution, wherein the pulse-like contribution can be more than, equal to or less than a noise contribution expected from a white noise signal as input signal, and the tonal contribution can be more than, equal to or less than the noise contribution.

4. Method according to claim 3, wherein calculating the pulse-like contribution comprises a smoothing operation in the time direction, and calculating the tone-like contribution comprises a smoothing operation in the frequency direction.

5. Method according to claim 3 or 4, wherein the pulse-like contribution and tonal contribution are calculated as root-mean-square values.

6. Method according to claim 3, 4 or 5, wherein the basic signal texture is determined as one of the group of:

- Tone only, wherein the tonal contribution is larger than the noise contribution and the pulse-like contribution is lower than the noise contribution;
- Pulse only, wherein the pulse-like contribution is larger than the noise contribution and the tonal contribution is lower than the noise contribution;
- Noise, wherein the pulse-like contribution, tonal contribution and noise contribution are similar;
- Tonal noise, wherein the tonal contribution is larger than the noise contribution and the pulse-like contribution and noise contribution are similar;
- Pulse like noise, wherein the pulse-like contribution is larger than the noise contribution and the tonal contribution and noise contribution are similar;
- Unresolved harmonics, wherein the tonal contribution and noise contribution are similar, and the pulse-like contribution is lower than the noise contribution;
- Unresolved pulses, wherein the pulse-like contribution and noise contribution are similar, and the tonal contribution is lower than the noise contribution;
- No energy, wherein the tonal contribution and the pulse-like contribution are both lower than the noise contribution;
- Sweeps, wherein the tonal contribution and the pulse-like contribution are both larger than the noise contribution.

7. Method according to claim 6, wherein the sweeps type of region is further segmented into:

- a shallow sweeps type, wherein the tonal contribution is larger than the pulse-like contribution; and
- a steep sweep type, wherein the pulse-like contribution is larger than the tonal contribution.

8 Method according to any one of claims 1-7, wherein after segmentation in the plurality of regions, a characteristic property of each of the plurality of regions in the time frequency domain is determined, and a region is assigned to an adjacent region if the characteristic property of the region in the time frequency domain is lower than a threshold value.

9. Method according to any one of claims 1-8, wherein analyzing each of the plurality of regions includes determining a likelihood of a possible source based on context information.

10. System for analyzing an input signal, comprising

a preprocessing unit (2) arranged to first process an input signal to obtain a representation of the input signal in the time-frequency domain,

a texture determination unit (4) connected to the preprocessing unit (2) arranged to subsequently segment the representation in the time-frequency domain into a plurality of regions, in which each region is of a type corresponding to one of a plurality of basic signal textures, in which a basic signal texture is a coherent region in the time frequency domain representation with one or more characteristic properties, and

an input source classifier (6) connected to the texture determination unit and arranged to analyze each of the plurality of regions according to the type of basic signal texture to obtain input signal source information.

11. System according to claim 10, wherein the texture determination unit (4) is further arranged to implement the method according to any one of claims 2-8.

12. System according to claim 10 or 11, wherein the input source classifier (6) is further arranged to implement the method according to claim 9.

13. Software program product comprising computer executable code, which when loaded on a processing system receiving an input signal, provides the processing system with the functionality of the method according to any one of claims 1-9.

****

Fig 1

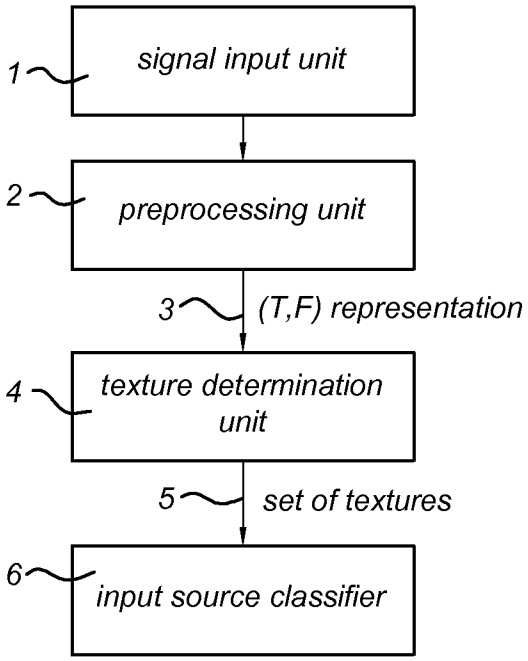

1
signal input unit
2
preprocessing unit
3
(T,F) representation
4
texture determination unit
5
set of textures
6
input source classifier

Fig 3

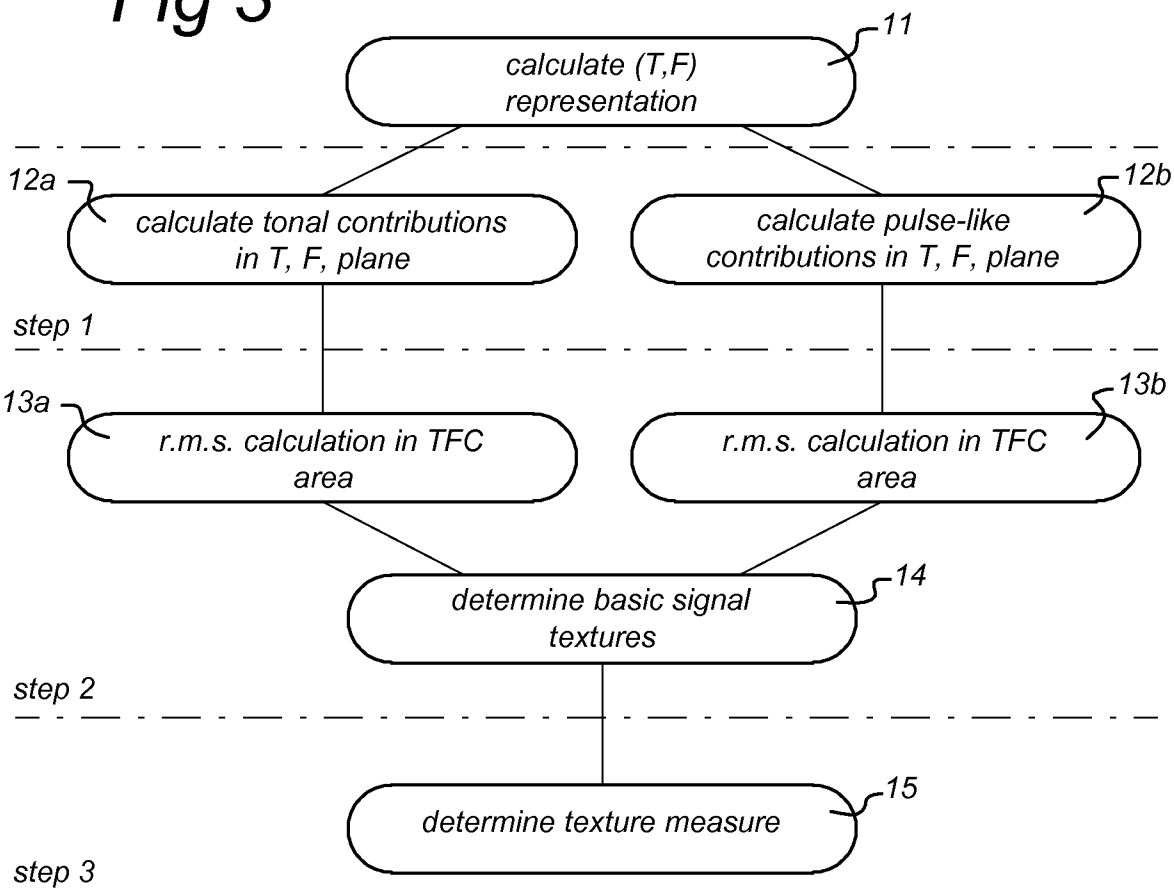
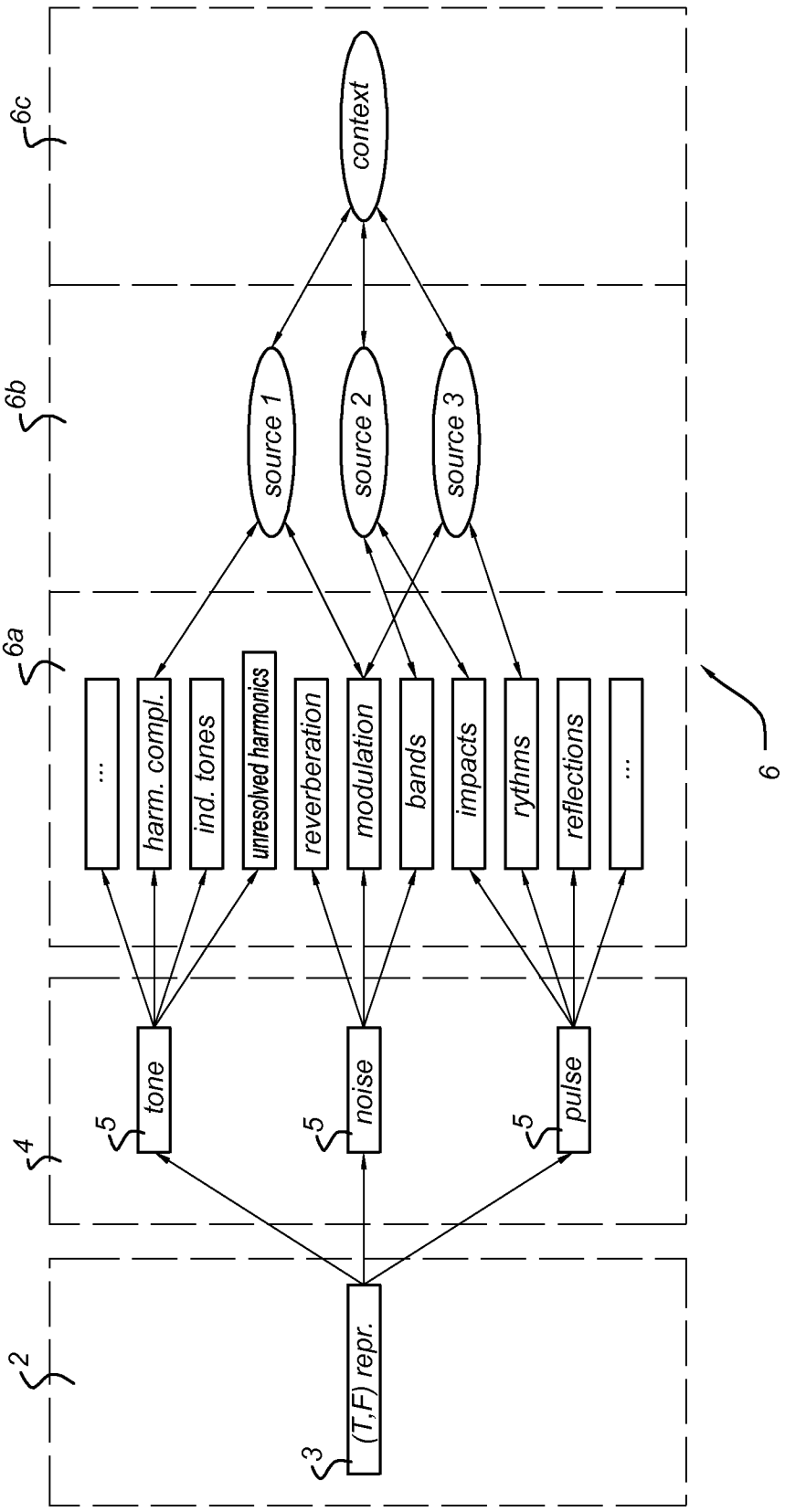

11
calculate (T,F) representation
12a
calculate tonal contributions in T, F, plane
12b
calculate pulse-like contributions in T, F, plane
step 1
13a
r.m.s. calculation in TFC area
13b
r.m.s. calculation in TFC area
14
determine basic signal textures
step 2
15
determine texture measure
step 3

Fig 2


6c
context
6b
source 1
source 2
source 3
6a
...
harm. compl.
ind. tones
unresolved harmonics
reverberation
modulation
bands
impacts
rythms
reflections
...
6
4
5
tone
5
noise
5
pulse
2
3
(T,F) repr.

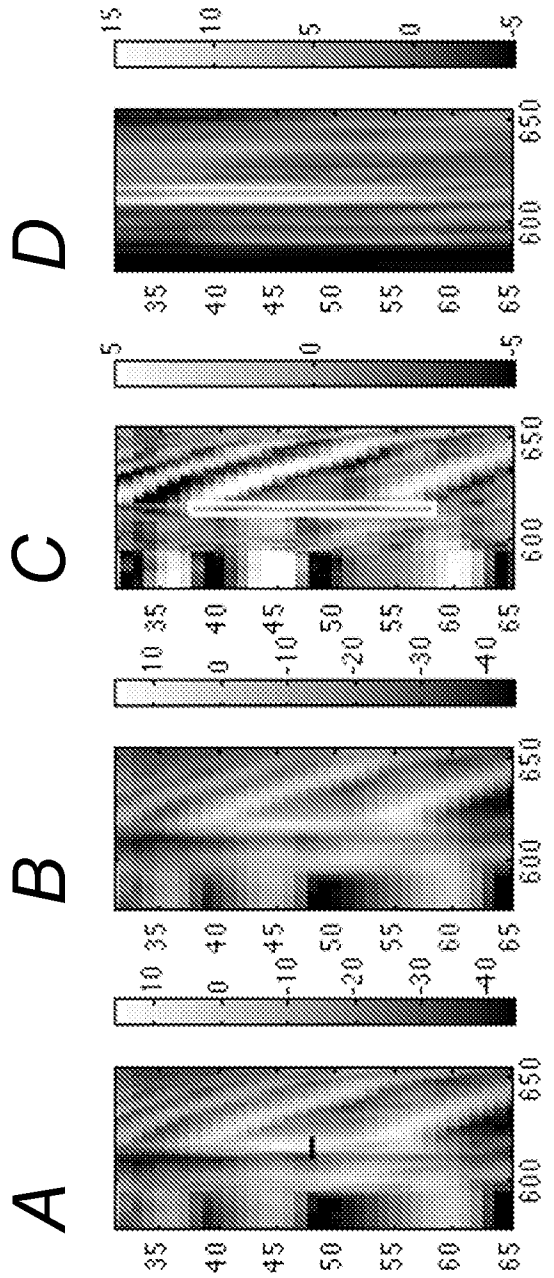

A
B
C
D
35 40 45 50 55 60 65
600 650
10 35 0 40 -10 45 -20 50 -30 55 -40 60 65
5 0 -5
15 10 5 0 -5

Fig 4

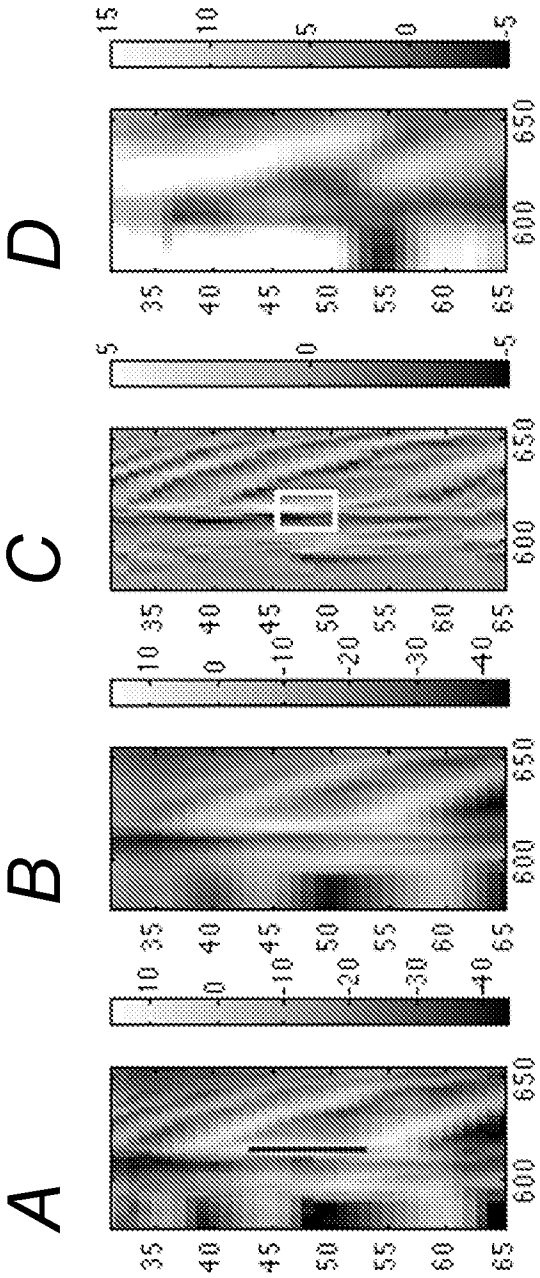

A
B
C
D
35 40 45 50 55 60 65
600 650
10 35 0 40 -10 45 -20 50 -30 55 -40 60 65
5 0 -5
15 10 5 0 -5

Fig 5

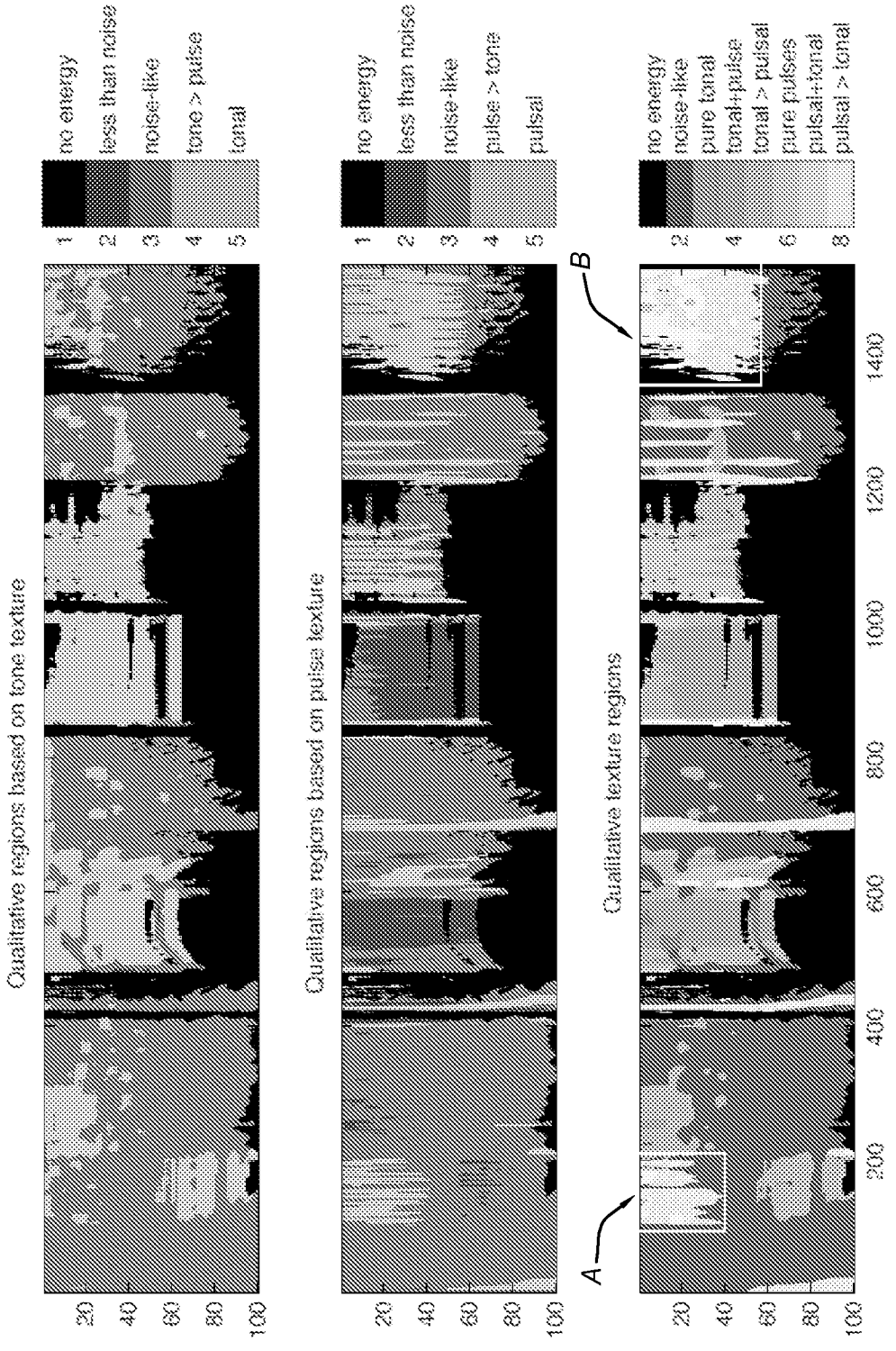


Fig 6a

Fig 6b

Fig 6c

INTERNATIONAL SEARCH REPORT

International application No
PCT/NL2010/050144

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L11/00 G10L17/00 G10L21/06 G10L15/02
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KRIJNDERS J D ET AL: "Demonstration of online auditory scene analysis" BELGIUM-NETHERLANDS CONFERENCE ON ARTIFICIAL INTELLIGENCE (BNAIC), 30 October 2008 (2008-10-30), - 31 October 2008 (2008-10-31) XP002527510 Retrieved from the Internet: URL:http://www.ai.rug.nl/%7Edirkjank/media-files/Publications/BNAIC2008-Abstract-demo.pdf> [retrieved on 2009-05-12] page 1 ----- -/--	1,3-5, 8-13

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

[X] See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 June 2010	24/06/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Fax: (+31–70) 340–3016	Ramos Sánchez, U

INTERNATIONAL SEARCH REPORT

International application No
PCT/NL2010/050144

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ANDRINGA T C ET AL: "Real-World Sound Recognition: A Recipe" PROC. OF 1ST WORKSHOP ON LEARNING THE SEMANTICS OF AUDIO SIGNALS (LSAS 2006), 6 December 2006 (2006-12-06), pages 106-118, XP002527511 Retrieved from the Internet: URL:http://www.ai.rug.nl/acg/pubs/Andringa 06.pdf> [retrieved on 2009-05-11] pages 109-111	1,10,13
X	US 2005/004797 A1 (AZENCOTT ROBERT [FR]) 6 January 2005 (2005-01-06) page 3, paragraph 79 - page 4, paragraph 79 page 1, paragraph 16 - paragraph 23	1,10,13
X	US 5 377 302 A (TSIANG ELAINE Y L [US]) 27 December 1994 (1994-12-27) column 13, line 26 - line 60 column 5, line 14 - line 45 column 2, line 19 - column 3, line 9	1,10,13

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/NL2010/050144

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005004797 A1	06-01-2005	EP 1473709 A1	03-11-2004
		FR 2854483 A1	05-11-2004
US 5377302 A	27-12-1994	WO 9616399 A1	30-05-1996