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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 April 2010 (22.04.2010)

PCT

(10) International Publication Number
WO 2010/043258 A1

(51) International Patent Classification:
G10H 1/38 (2006.01)

(21) International Application Number:
PCT/EP2008/063911

(22) International Filing Date:
15 October 2008 (15.10.2008)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **MUSEEKA S.A.** [CH/CH]; P.O. Box 1568, 11, rue de Berne, CH-1211 Geneva 1 (CH).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **FÄRNSTRÖM, Lars** [DK/CH]; Route de Chesérex, 87, CH-1276 Gingins (CH). **LEONARDI, Riccardo** [CH/IT]; Via Trieste, 22, I-25060 Collebeato (BS) (IT). **SCARINGELLA, Nicolas** [FR/CH]; Chemin de Boston, 23, CH-1004 Lausanne (CH).

(74) Agents: **BONVICINI, Davide** et al.; Perani Mezzanotte & Partners, Piazza San Babila, 5, I-20122 Milano (MI) (IT).

(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, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI

[Continued on next page]

(54) Title: METHOD FOR ANALYZING A DIGITAL MUSIC AUDIO SIGNAL

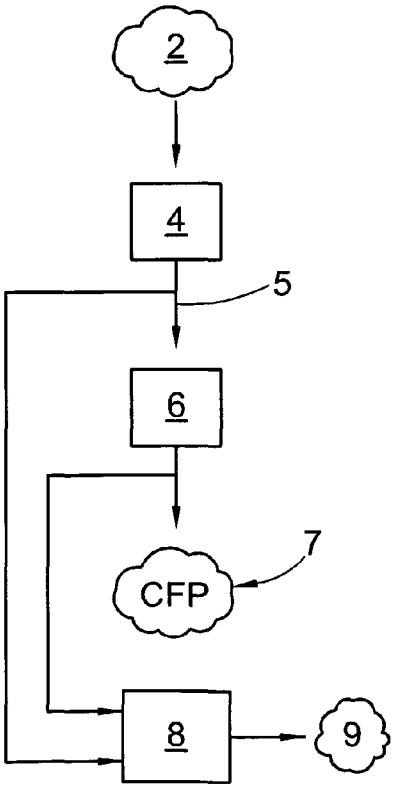


Figure 4

(57) Abstract: The present invention concerns a music audio representation method for analyzing a music audio signal (2) in order to extract a set of Chord Family Profiles (CFP) contained in the audio music signal (2), the method comprising the steps of: a) applying a first algorithm (4) to the music audio signal (2) in order to extract first data (5) representative of the tonality of music audio signal (2), and b) applying a second algorithm (6) to said first data (5) in order to provide second data (7) representative of the tonal centre contained in the first data (5).

WO 2010/043258 A1

(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

DESCRIPTION

Title: "Method for analyzing a digital music audio signal"

FIELD OF THE INVENTION

The invention relates to automatic analysis of music audio signal, preferably a digital audio music signal.

Particularly, the present invention relates to a music audio representation method and apparatus for analyzing a music audio signal in order to extract a set of characteristics representative of the informative content of the audio music signal, according to the preamble of claim 1 and 17, respectively.

DEFINITIONS

Several terms that are used in the description that follows are explained. Some of these terms are generally used in the field, and some were coined to communicate embodiments of the present invention.

As used herein, the following terms are intended to indicate:

Pitch - Perceived fundamental frequency of a sound. A pitch is associated to a single (possibly isolated) sound and is instantaneous (the percept is more or less as long as the sound itself, typically 200 to 500 ms duration in music signals). In the following Table 1, the pitches over the register of a piano have been associated to their corresponding fundamental frequencies (in Hertz) assuming a standard tuning, i.e. the pitch A3 corresponds to a fundamental frequency of 440Hz.

Pitch	Frequency (Hz)
A -1	27.50
Bb -1	29.13
B -1	30.86
C 0	32.70
C# 0	34.64
D 0	36.70
Eb 0	38.89
E 0	41.20
F 0	43.65
F# 0	46.24
G 0	48.99

G# 0	51.91
A 0	55.00
Bb 0	58.27
B 0	61.73
C 1	65.40
C# 1	69.29
D 1	73.41
Eb 1	77.78
E 1	82.40
F 1	87.30
F# 1	92.49
G 1	97.99
G# 1	103.82
A 1	110.00
Bb 1	116.54
B 1	123.47
C 2	130.81
C# 2	138.59
D 2	146.83
Eb 2	155.56
E 2	164.81
F 2	174.61
F# 2	184.99
G 2	195.99
G# 2	207.65
A 2	220.00
Bb 2	233.08
B 2	246.94
C 3	261.62
C# 3	277.18
D 3	293.66
Eb 3	311.12
E 3	329.62
F 3	349.22
F# 3	369.99
G 3	391.99

G# 3	415.30
A 3	440.00
Bb 3	466.16
B 3	493.88
C 4	523.25
C# 4	554.36
D 4	587.32
Eb 4	622.25
E 4	659.25
F 4	698.45
F# 4	739.98
G 4	783.99
G# 4	830.60
A 4	880.00
Bb 4	932.32
B 4	987.76
C 5	1046.50
C# 5	1108.73
D 5	1174.65
Eb 5	1244.50
E 5	1318.51
F 5	1396.91
F# 5	1479.97
G 5	1567.98
G# 5	1661.21
A 5	1760.00
Bb 5	1864.65
B 5	1975.53
C 6	2093.00
C# 6	2217.46
D 6	2349.31
Eb 6	2489.01
E 6	2637.02
F 6	2793.82
F# 6	2959.95
G 6	3135.96

G# 6	3322.43
A 6	3520.00
Bb 6	3729.31
B 6	3951.06
C 7	4186.01
C# 7	4434.92
D 7	4698.64
Eb 7	4978.03
E 7	5274.04
F 7	5587.65
F# 7	5919.91
G 7	6271.93
G# 7	6644.87
A 7	7040.00
Bb 7	7458.62
B 7	7902.13
C 8	8372.02
C# 8	8869.84
D 8	9397.27
Eb 8	9956.06
E 8	10548.08
F 8	11175.30
F# 8	11839.82
G 8	12543.85

Table 1

Interval – The difference in pitch between two pitched sounds.

Octave – An interval that corresponds to a doubling of fundamental frequency.

Pitch Class - A set of all pitches that are a whole number of octaves apart, e.g. the pitch class C consists of the Cs in all octaves.

Chord - In music theory, a chord is two or more different pitches that occur simultaneously; in this paper, single pitches may also be referred to as chords (see figure 1a and 1b for a sketch).

Chord Root – The note or pitch upon which a chord is perceived or labelled as being built or hierarchically centred upon (see figure 1a and 1b for a sketch).

Chord Family – A chord family is a set of chords that share a number of characteristics including (see figure 1a and 1b for an illustration):

- a number of pitch classes from which the chord takes its notes, (typically from 1 to 6 pitch classes per chord).
- a precise interval construction, sometimes called "chord quality", which defines the intervals between the pitch classes constituent of the chord.

Tonality - A system of music in which pitches are hierarchically organized (around a tonal centre) and tend to be perceived as referring to each other; notice that the percept of tonality is not instantaneous and requires a sufficiently long tonal context.

Tonal context - A combination of chords implying a particular tonality percept.

Key - Ordered set of pitch classes, i.e. the reunion of a tonic and a mode (see figure 2a and 2b for an illustration).

Tonal Centre or **Tonic** - The dominating pitch class in a particular tonal context upon which all other pitches are hierarchically referenced (see figure 2a and 2b for an illustration).

Mode – Ordered set of intervals (see figure 2a and 2b for an illustration).

Transposition – The process of moving a collection of pitches up or down in pitch by a constant interval.

Modulation – The process of changing from one tonal centre to another.

Chromatic scale – The set of all 12 pitch classes.

Meter - The underlying division of time in a musical piece which organises it into measures of stressed and unstressed beats (see figure 3 for a sketch).

Beat - Basic time unit of a piece of music (see figure 3 for an illustration).

Measure or **Bar** - Segment of time defined as a recurring sequence of stressed and unstressed beats; in figure 3 is shown an audio signal and detected onset positions, wherein the higher the amplitude associated to the onset, the higher its weight in the detected metrical hierarchy (i.e. musical bars have higher weights, bar have intermediate weights, unmetrical onsets have lower weights).

Frame of audio signal is a short slice of audio signal, typically 20 to 50 ms segments of audio signal.

BACKGROUND OF THE INVENTION

In the case of music audio signals, it is not possible to observe directly the various pitches present in the signal but rather a mixture of their harmonics. Consequently, most state-of-the-art algorithms rely on the use of Pitch Class Profiles (PCP) also called Chroma vectors as a basis for modelling the music audio signal (see for example [M.A. Bartsch and G.H. Wakefield, "Audio Thumbnailing of Popular Music Using Chroma-based Representa-tions", IEEE Transactions on Multimedia, 1996]).

The PCP/Chroma approach is a general low-level feature extraction method that measures the strength of pitch classes in the audio music signal.

A number of algorithms have been proposed in the art to infer the key or the chord progression of a music piece from its sequence of low-level PCPs.

For example, in a form of implementation of the PCPs algorithms, it is measured the intensity of each of the twelve semitones of the tonal scale. Such implementation consists in mapping some time/frequency representation to a time/pitch-class representation; in other words the spectrum peaks (or spectrum bins) are associated to the closest pitch of the chromatic scale.

In other embodiments of the PCPs algorithms it has been used a higher resolution for the PCP bins, i.e. PCPs algorithms of this type decrease the quantization level to less than a semitone.

Further, in some other implementations of the PCPs algorithms it is also considered the fact that a pitched instrument will not only exhibit an energy peak around a single frequency, but that it will also exhibit significant energy for some (more or less) harmonic frequencies.

As the number of notes and timbres increase (i.e. as the number of instruments playing simultaneously in a piece increases), partials of all composing notes overlap disorderly, making the PCPs extracted an improper representation of the actual content of a music piece.

A number of algorithms have been proposed in the art to infer high-level musical features such as e.g. the key or the chord progression of a music piece from its sequence of low-level PCPs (refer e.g. to O. Izmirli, "An Algorithm for Audio

Key Finding", Music Information Retrieval Evaluation eXchange (MIREX).

These algorithms typically rely on the use of "templates" that encode in the PCP space the musical object being searched for in the musical signal (see figure 1a and 1b for illustrations of chord templates and figure 2a and 2b for illustrations of key templates). By correlating such templates with actual PCP observations, it is possible to decide whether or not the musical objects corresponding to the templates are indeed hidden in the signal, i.e. the templates which maximally correlate with the PCPs correspond to musical objects being hidden in the signal.

The template based approach to high-level musical feature extraction is however restricted by the choice of templates. For example, in the case of key detection, state-of-the-art algorithms use templates for the Major key and for the Minor key (one such template for each of the 12 possible pitch class).

This restriction to a Major/Minor dichotomy comes from Western classical music. Popular music such as Rock may however not be properly described with Western classical ideas. Indeed, Rock music, and more generally Popular music, is a unique and diverse mix of cultural overlays that has produced new sets of rules for what is structurally acceptable in today's music.

This is even truer for the so-called World music, which comes from a totally different cultural background. As a matter of fact, there is a wider variety of musical colors and forms in the world of music than the Major/Minor dichotomy.

SUMMARY OF THE INVENTION

In view of the prior art as described above, the object of the present invention is to develop a feature extraction algorithm able to compute a musicologically valid description of the pitch content of the audio signal of a music piece.

Moreover, it is an object of the present invention to provide an algorithm for the detection of the tonal centre of a music piece in audio format and to provide a set of features that encode a transposition invariant representation of the distribution of pitches in a music piece and their correlations.

Moreover, it is an object of the present invention to propose an alternative low-level representation of the pitch content of a music piece robust to the variety of timbres and pitch combinations observable in real-world music signals. To achieve

this goal, it is notably proposed to use machine-learning algorithms so as to learn from data specificities of real-world music signals.

A further object of the present invention is to map directly spectral observations to a chord space without using an intermediate note identification unit.

It is another object of the present invention to allow for the following of the tonal centre along the course of a piece of music if a modulation occurs. It is a specificity of the tonal centre following algorithm to take into account a sufficiently long time scale to avoid tracking chord changes that occur at a faster rate than modulations.

It is an object of the present invention to take into account musical accentuation – more specifically, metrical accentuation - in the process of detecting the tonal centre of a music piece.

It is another object of the present invention to allow for an appropriate description of a larger variety of musical forms. To achieve this goal, it is notably proposed to use machine-learning algorithms so as to learn from data specificities of musical forms coming from different cultural backgrounds.

According to the present invention, these objects are fulfilled by a method for analyzing a music audio signal in order to extract a set of characteristics representative of the informative content of the audio music signal as defined in the features of claim 1.

Further, according to the present invention, these objects are fulfilled by an apparatus for analyzing a music audio signal in order to extract a set of characteristics representative of the informative content of the audio music signal as defined in the features of claim 17.

Thanks to the present invention it is possible to characterize the content of music pieces with an audio feature extraction method that generates compact descriptions of pieces that may be stored e.g. in a database or that may be embedded in audio files like e.g. ID3 tags.

Further, thanks to the present invention, it is possible to identify the tonal centre of a music piece and to allow for a transposition invariant selection of similar music pieces with features discriminating a large variety of musical forms as heard

notably in Popular and World music as well as in Classical Western music.

To this aim, a new set of features describing pitch distributions (Chord Family Profiles) is proposed and supervised machine learning approaches are used for both tonal centre detection and tonally similar music pieces selection to identify the patterns present in a large variety of musical forms.

It is a specificity of the present invention to extract Chord Family Profiles with a machine-learning algorithm trained in both supervised and unsupervised fashion.

DETAILED DESCRIPTION OF THE DRAWINGS

The characteristics and advantages of the invention will appear from the following detailed description of one practical embodiment, which is illustrated without limitation in the annexed drawings, in which:

- Figure 1a e 1b show graphical representations of Chord examples;
- Figure 2a e 2b show graphical representations of Key examples;
- Figure 3 shows a graphical representation of metrical levels;
- Figure 4 a block diagram of the music audio analysis method according to the present invention;
- Figure 5a shows a block diagram of a first algorithm of the music audio analysis method according to the present invention;
- Figure 5b shows the music audio signal and the plurality of vectors as result of the application to the audio music signal of the first algorithm;
- Figure 6a shows another block diagram of a first way for training of a step of the first algorithm according to the present invention;
- Figure 6b shows another block diagram of a second way for training of a step of the first algorithm according to the present invention;
- Figure 7 shows a block diagram of a second algorithm for the music audio analysis method according to the present invention;
- Figures 7A to 7D show graphically the way of working of the second algorithm;
- Figure 8 shows a block diagram of the music audio analysis apparatus according to the present invention;
- Figure 9 shows a graphical representation of a moving average when applied

to a power spectrum of the audio signal of Figure 3.

Referring to the accompanying figures 4 to 8, it is generally indicated with 1 a music audio analysis method for analyzing a digital music audio signal 2 in order to extract Chord Family Profiles (CFP).

It is to be noted that, the digital music audio signal 2 can be an extract of a signal audio representing a song or a complete version of a song.

Particularly the method 1 comprises the step of:

a) applying a first algorithm 4 to the music audio signal 2 in order to extract first data 5 representative of the tonal context of music audio signal 2, and

b) applying a second algorithm 6 to said first data 5 in order to provide second data 7 representative of the tonal centre contained in the first data 5.

It is to be noted having regard of the definition provided above, that with the term tonality it is encompassed a combination of chord roots and chord family hierarchically organized around a tonal centre, i.e. a combination of chord roots and chord family, which perceived significance is measured relatively to a tonal centre.

Therefore the step a) of the method 1, i.e. the first algorithm 4, is able to extract the first data 5 representing the combination of chord roots and chord families observed in the digital music audio signal 2, that is the first data 5 contains the tonal context of the digital music audio signal 2. Notice however that the step a) of the method 1, i.e. the first algorithm 4, does not aim explicitly at detecting chord roots and chord families contained in the digital music audio signal 2. On the contrary, it aims at obtaining an abstract, and possibly redundant, representation correlated with the chord roots and chord families observed in the digital music audio signal 2.

Moreover the step b) of the method 1, i.e. the second algorithm 6, is able to elaborate the first data 5 for providing second data 7 which represent the tonal centre Tc contained in said first data 5, that is in the second data 7 the dominating pitch class of a particular tonal context upon which all other pitches are hierarchically referenced (see figure 2a and 2b) are contained.

Therefore, once the tonal centre Tc of the digital music audio signal 2 by applying the first algorithm 4 and the second algorithm 6 has been found, the tonality of the digital music audio signal 2 is described thanks to the hierarchical position of

first data 5 in reference to second data 7.

Optionally, the method 1 further comprises the step of:

c) applying a third algorithm 8 to the first data 5 in function of the second data 7 in order to provide third data 9 which are the normalized version of the first data 5.

In the following, it is described in greater detail the way of working of the first algorithm 4, the second algorithm 6 and of the third algorithm 8.

FIRST ALGORITHM 4

Step a)

Referring to figure 5a and 5b, it is shown a block diagram of the first algorithm 4 that is suitable for the extraction of the first data 5 from the audio digital signal 2.

In particular, the first algorithm 4 comprises the steps of:

a1) identify 10 a sequence of note onsets in the music audio signal 2, in order to define the time position of a plurality of peaks p1, p2, p3, …, pi where "i" is an index that can vary between 1<i<N, being N the number of samples of the audio digital signal 2 and being in practice i << N;

a2) dividing the audio music signal 2 into a plurality of audio segments s-on-1, s-on-2, s-on-3, s-on-i, each audio segments containing a peak p1, p2, p3, …, pi,

a3) applying a frequency analysis to each audio segment s-on-1, s-on-2, s-on-3, …, s-on-i in order to obtain a plurality of spectrum segments sp-1, sp-2, sp-3, …, sp-i which represent the evolution in the time domain of the spectrum of the music audio signal 2, and

a4) processing said plurality of spectrum segments sp-1, sp-2, sp-3, …, sp-i by a computation network 12 in order to provide said first data 5.

The first data 5 comprise a plurality of vectors v1, v2, v3, …, vi, wherein each vector of the plurality of vectors v1, v2, v3, …, vi is associated to the respective audio segment s-on-1, s-on-2, s-on-3, s-on-i.

In particular, each vector v1, v2, v3, vi has a dimension equal to the twelve pitches (A to G#) times a predefined number "n" of chord type.

Advantageously the predefined number "n" of chord type can be set equal to five so as to represent, for example, "pitches", "major chords", "minor chords", "diminished chords" and "augmented chords".

Step a1)

The above mentioned step a1) of the first algorithm 4 is performed by an onset detection algorithm in order to detect the attacks of musical events of the audio signal 2.

In fact, each peak p1, p2, p3, …, pi represents an attack of musical event in the respective audio segments s-on-1, s-on-2, s-on-3, …, s-on-i.

The onset detection algorithm 10 can be implemented as described in [J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. Sandler, " A Tutorial on Onset Detection in Music Signals", in IEEE Transactions on Speech and Audio Processing, 2005].

Step a2)

The above mentioned step a2) of the first algorithm 4 divides the audio music signal 2 into the plurality of audio segments s-on-1, s-on-2, s-on-3, …, s-on-i each audio segments s-on-1, s-on-2, s-on-3, …, s-on-i having a duration "T".

The step a2) of the first algorithm 4 divides the audio music signal 2 into the audio segments s-on-1, s-on-2, s-on-3, …, s-on-i and each audio segments s-on-1, s-on-2, s-on-3, …, s-on-i has its own duration "T".

In other words the duration "T" of each audio segments s-on-1, s-on-2, s-on-3, s-on-i which can be different for each other.

Step a3)

The above mentioned step a3) of the first algorithm 4 applies, advantageously, the frequency analysis to each audio segment s-on-1, s-on-2, s-on-3, …, s-on-i only during a predetermined sub-duration "t", wherein the sub-duration "t" is less than the duration "T".

In other words the audio segments s-on-1, s-on-2, s-on-3, …, s-on-i are further analysed in frequency only during the sub-duration "t" even if they extend over such sub-duration "t".

It is to noted that the prefixed sub-duration "t" can be set manually by the user.

Preferably, the prefixed sub-duration "t" is within a range from 250 to 350 msec.

Therefore, if the duration "T" audio segment s-on-1, s-on-2, s-on-3, …, s-on-i

is longer than the pre-defined duration "t", i.e. more than 250-350 msec, only the data contained in the sub-duration "t" are considered while the rest of the segment is assumed to contain irrelevant data and therefore such remaining data are disregarded.

In case the duration T is less than the predetermined sub-duration "t" (the bounding peaks are less than "t" apart), zero samples are added to the audio segment so that its length equals the predetermined sub-duration "t". Therefore, the frequency analysis will be limited to the smallest time interval, i.e. the duration "T".

In the case which the duration T is equal to 50 msec and sub-duration "t" is equal to 200 msec, therefore, the frequency analysis of each audio segment s-on-1, s-on-2, s-on-3, …, s-on-i is performed only using the music samples occurring during the duration T, i.e. the smallest duration.

The frequency analysis, applied during step a3), is performed, in the preferred embodiment, by a D.F.T. (Discrete Fourier Transform).

It is to be noted during the step a3) can also be performed a further step during which is applied a function that reduces the uncertainty in the time-frequency representation of the audio signal 2.

To this aim, it is possible to apply an apodization function, such as a Hanning window.

In particular, in the case it is applied a Hanning window, the length of the Hanning window equals the length "T" of the audio segment s-on-1, s-on-2, s-on-3, s-on-i.

It is to be noted also that, the apodization function is applied to audio segment s-on-1, s-on-2, s-on-3, s-on-i. by multiplying on a sample by sample basis to the audio data of the corresponding segment prior to applying the frequency analysis performed by the D.F.T..

A further reason for which the apodization function is used is for the attenuation of the musical event attacks p1, p2, p3, …, pi, since they were located around the boundaries of the apodization window. In this way it is possible to create an attenuated version of the musical event attacks p1, p2, p3, …, pi.

Moreover the power spectrum is computed with the D.F.T. or any of its fast implementations, for example F.F.T. (Fast Fourier Transform).

It is to be noted that in the case of using the F.F.T., the choice of the sub-duration "t" allows for controlling the frequency resolution of the FFT (i.e. the longer the duration "t", the higher the frequency resolution) and normalizes the frequency resolution so that it remains constant even if the initial duration "T" of audio segments s-on-1, s-on-2, s-on-3, …, s-on-i is different for each segment.

In case of a radix-2 F.F.T. implementation, the choice of the sub-duration "t" is such that the length in samples of the resulting segment equals a power of two.

Step a4)

With reference to above mentioned step a4) and in connection with figures 6A e 6B, it is to be noted that the computation network 12 can be implemented, preferably, with a trained machine-learning algorithm.

In particular the trained machine-learning algorithm consists in a Multi-layer Perceptron (MLP).

The task of the Multi-layer Perceptron (MLP) is to estimate the posterior probabilities of each combination chord family (i.e a chord type) and chord root (i.e. a pitch class), given the spectrum segments sp-1, sp-2, sp-3, sp-i.

Particularly, the Multi-layer Perceptron (MLP) is trained in two steps:

1st step: training achieved in a supervised fashion by using a first set 13 of training data built upon a set of known isolated chords for which a first ground truth mapping can be established from the corresponding spectrum of said plurality of segments sp-1, sp-2, sp-3, sp-i to chord families and chord roots.

2nd step: in an unsupervised fashion by a second set 14 of training data comprising a large set of music pieces in order to refine the set of weights "ω" of the trained machine-learning algorithm obtained after the 1st step to the variety of mixtures of instruments encountered in real polyphonic music.

To recapitulate the trained machine-learning algorithm 12 is trained in two steps: a first supervised training with few hand labelled training data and a subsequent unsupervised training with a larger set of unlabelled training data.

More specifically the 1st step during which the machine-learning algorithm 12 is trained in a supervised fashion, the set of hand labelled training data consists of isolated chords saved as MIDI files. The set of chords should cover each considered

chord type (Major, Minor, Diminished, Augmented…), each pitch class (C, C#, D…) and should cover a number of octaves.

A large variety of audio training data is created from these MIDI files by using a variety of MIDI instruments. These audio examples together with their pitch class and chord type are used to train the machine-learning algorithm 12, which is set to produce from the ground truth a single output per "pitch class / chord type" pair.

The training of the various weights "ω" of the machine learning algorithm is performed thanks to a standard stochastic gradient descent. Once such training has been achieved, at the end of this 1st training step, a first preliminary mapping for any input spectral segment sp-1, sp-2, sp-3, sp-i to chord families can be produced.

It is to be noted that such a produced output vector of the machine learning algorithm 12 after this 1st training step will have components that determine the likelihood ratio for any "pitch class / chord type" pair. Yet, the machine-learning algorithm 12 does not yet successfully lead to a satisfactory correspondence for the variety of timbre encountered in real polyphonic music since it has been only trained so far from isolated chords produced by a variety of MIDI instruments.

Consequently, the training of the trained machine-learning algorithm 12 needs to be refined by using the data from a larger set of music pieces.

To this aim during the 2nd step the machine-learning algorithm 12 is trained in an unsupervised fashion. The initially trained machine-learning algorithm 12 after the 1st step is cascaded with a mirrored version of itself which uses as initial weights the same weights "ω" of the trained machine-learning network after the 1st step (so as to operate some sort of inversion of the corresponding operator, were it linear).

The machine-learning algorithm 12 (were it a linear operator) would achieve a projection of the high-dimensional input data (the spectral segments) into a low-dimensional space corresponding to the chord families. Its mirrored version attempts to go from the low-dimensional chord features back to the initial high dimensional spectral peak representation. For this purpose the initial setting of the cascaded algorithm adopts initially the transposed set of weights of the training engine algorithm.

Subsequently all the weights for the "machine-learning algorithm" and "its

initially mirrored version" are adjusted by stochastic gradient descent to minimize a distance between the input training patterns (i.e. spectral segments) and the reconstructed outputs using the complete set of available music pieces as training data.

This leads to a fine-tuning the weights of the network to learn a low-dimensional representation of the data that is steered to correspond to chord families because of the initial supervised training (performed during the 1st step).

This training approach is reminiscent of the training of auto-encoder networks.

In this case the initialisation of the network with a supervised strategy ensures finding an initial set of weights for the network which is consistent with the physical essence of a low level representation in terms of chord families.

Once the 2nd step training has been completed the "chord family – spectral segment" computation network can be removed, so as to retain only the first stage of processing elements which represents at this point the final trained machine learning algorithm 12.

With reference again to figure 5A, it is to be noted that the first algorithm 4 may comprise the further step a5) of filtering, after the D.F.T. step a3).

Such filtering step a5), also called peak detection 15, is an optional step of the method 1.

According to the way of working, the filtering step a5) is able to filter the plurality of spectrum segments sp-1, sp-2, sp-3, …, sp-i, generated by the block 11, by a moving average in order to emphasize the peak p1', p2', p3', …, pi' in each of said plurality of spectrum segments sp-1, sp-2, sp-3, sp-i.

Therefore at the output of the step a5), there are spectrum segments sp-1', sp-2', sp-3', …, sp-i' in which the peaks p1', p2', p3', …, pi' of the spectrum segments sp-1, sp-2, sp-3, …, sp-i are emphasized while the overall shape of the spectrum segments sp-1, sp-2, sp-3, …, sp-i is discarded.

In other words, also with reference to Figure 9, a moving average 20 typically operating over the power spectrum 21 as result from the step a4) is computed and the spectral components having power below this moving average are zeroed.

Moreover, after the filtering step 15, the music audio analysis method 1

comprises, before the computing step a4), a further step of decorrelating, also called whitening 16.

Also this decorrelating step is optional in the method 1.

Particularly, during the decorrelating step, the plurality of spectrum segments sp-1', sp-2', sp-3', ..., sp-i' is de-correlated with reference to a predetermined database 19 (Figure 8) of audio segment spectra in order to provide a plurality of decorrelated spectrum segments sp-1", sp-2", sp-3", ..., sp-i".

Therefore, once the plurality of spectrum segments sp-1, sp-2, sp-3, ..., sp-i are filtered in order to emphasize the peak p1', p2', p3', ..., pi' so as to obtain the plurality of spectrum segments sp-1', sp-2', sp-3', ..., sp-i', the latter are whitened with a whitening transformed obtained in a preferred embodiment of the invention through Principal Component Analysis (PCA) as computed on a large set of such audio segment spectra contained in the database.

In the case the optional step of filtering and decorrelating are implemented in the method 1, it is to be noted that the whitened spectrum segments sp-1", sp-2", sp-3", ..., sp-i" are hence fed into the computation network 12, i.e. the MLP.

SECOND ALGORITHM 6

Step b)

Referring now to figures 6 and 7, the second algorithm 6 of the music audio analysis method 1 comprises the steps of:

b1) providing a first window "w1" having a first prefixed duration T1 containing a first group "g1" of plurality of vectors composing the first data 5, and

b2) elaborating said first group "g1" of plurality of vectors contained in said first window "w1" for estimating a first tonal context Tc1 representative of the local tonal centre contained in said first window "w1".

It is to be noted that the first prefixed duration T1 of said first window "w1" is much longer than the sub-duration "t" of each plurality of audio segments s-on-1, s-on-2, s-on-3, ..., s-on-i.

Moreover, the second algorithm 6 comprises the further step of:

b3) providing a second window "w2", being a shifted window of said first window "w1", said second window "w2" having a second prefixed duration T2, said

second window "w2" comprising a second group "g2" of plurality of vectors;

b4) computing said second group "g2" of plurality of vectors contained in said second window "w2" for estimating a second tonal context Tc2 representative of the local tonal centre contained in said second window "w2";

b5) elaborating the tonal context Tc1 of said first window "w1" and the tonal context Tc2 of said second window "w2" in order to generate said second data 7 being representative of the evolution of the tonal centre of said first data 5.

In particular the second window "w2" is shifted by a prefixed duration Ts with respect to said temporal duration T1 of the first window "w".

It is to be noted that, the second prefixed duration T2 can vary in the range between T1-Ts and the first prefixed duration T1.

Therefore also the second prefixed duration T2 is much longer than the sub-period t.

Preferably the prefixed time Ts is considered to be less of the first prefixed duration T1, so that the first group g1 of vectors and the second group g2 of vectors overlap each other.

In fact, by choosing the prefixed time Ts less than first prefixed duration T1, it is advantageously possible to track in a more precise way the evolution of the tonal centre Tc of the data 5.

In fact, given a particular tonal context, some chords/pitches have to be more expected than others.

Since chords typically change with musical bars - or even faster at the beat level - tonality requires a longer time duration to be perceived.

Preferably the first prefixed duration T1 is typically set in the range of 25 - 35 sec, more preferably about 30 sec., whereas the prefixed time Ts, is typically set in the range of 10 - 20 sec, more preferably about 15 sec..

Alternatively, when the prefixed time Ts is equal to the first prefixed duration T1, the first group g1 of vectors is contiguous with the second group of vectors g2.

Moreover the second algorithm 6 of the music audio analysis method 1 comprises also the further step of:

b6) repeating the steps from b3) to b5) till to the end the plurality of audio

segments s-on-1, s-on-2, s-on-3, ..., s-on-i for defining further windows "wi" wherein each further window "wi" contains a group "gi" of vectors.

It is to be noted that, two consecutive windows, for example windows w3 and w4 (not shown in the drawings) have to be overlapping or at most consecutive without gaps but any subsequent window, i.e. windows w4, must not be contained in the previous windows, i.e. w1, w2 and w3.

Therefore the prefixed duration of the window w2, i.e. duration T2, could be equal to the prefixed duration T1 of the window w1 or could be greater than the prefixed duration T1, i.e. $T2 \geq 3/2\ T1$; T2 could also be adjusted locally to its associated window, so as to be tailored to local properties of the underlying audio signal, without however violating the principle of partial overlapping.

It is possible also to have multiple analysis windows overlapping, i.e. it could be possible to have e.g. 30 second long windows shifted by one onset at a time so that there is maximal overlap between windows.

Alternatively, the durations and positions of windows "w" may be tailored to the overall structure of the music signal, i.e. windows may be set so as to match sections like e.g. verse or chorus of a song. An automatic estimation of the temporal boundaries of these structural sections may be obtained by using a state-of-the-art music summarization algorithm such as well known to the skilled man in the art.

In this latter case, different windows may have different durations and may be contiguous instead of overlapping.

A first way to generate the second data 7 being representative of the tonal centre of said first data 5 is to elaborate a mean vector "m" of said first data 5 and choose the highest chord root value in such mean vector "m" in order to set the tonal centre.

A better way to capture the local temporal evolution of the tonal centre of said first data 5 is described in the following preferred embodiment according to the present invention and with reference to figure 6: Accordingly, the statistical estimates measured over time, such as mean, variance and first order covariance of the vectors contained in the first group g1 and the same statistical estimates for the others groups (i.e. g2, ..., gi) can be used to recover a better description of the local tonal context

of each audio segments s-on-1, s-on-2, s-on-3, …, s-on-i.

Such statistical estimates measured over time of data 5 can be calculated according to the following formulas in order to form the data 7A

$$\mu = \frac{1}{N}\sum_{1}^{N} X_i$$

$$\sigma^2 = \frac{1}{N-1}\sum_{1}^{N}(X_i - \mu)^2$$

$$\text{cov\_1} = \frac{1}{N-2}\sum_{1}^{N}(X_i - \mu) * (X_{i-1} - \mu)$$

where N is the number of vectors within the group "gi" of the window "wi", μ the mean, σ^2 the variance and cov_1 is first order covariance.

The data 8 output by the second algorithm 6 has a dimension equal to the:

$$D = 3 * 12 * F$$

Where D is the dimension, F is the number of considered chord families, is the number of semitones of the chromatic scale, i.e. the number of pitch class of the chromatic scale and 3 is the number of statistical estimates measured over time, i.e. mean, variance and first order covariance.

Optionally, it is possible to incorporate a weighting scheme during the extraction of data 7 to account for the fact that audio segments s-on-1, s-on-2, ..., s-on-i are perceived as being accentuated when synchronised with the underlying metrical grid.

Moreover, the most stable pitches producing the percept of tonality are typically played in synchrony with the metrical grid while less relevant pitches are more likely to be played on unmetrical time positions.

In a preferred embodiment, the incorporation of metrical information during the tonality estimation is as follows.

Each audio segment s-on-1, s-on-2, ..., s-on-i is associated to a particular metrical weight depending on its synchronisation with identified metrical events. For example, it is possible to assign a weight of 1.0 to the audio segment if a musical bar position has been detected at some time position covered by the corresponding audio segment. A lower weight of e.g. 0.5 may be used if a beat position has been detected

at some time position covered by the audio segment. Finally, the smallest weight of e.g. 0.25 may be used if no metrical event corresponds to the audio segment.

Given such weights, it is possible to re-evaluate data 7A as:

$$\mu_w = \frac{1}{N}\sum_{1}^{N} w_i X_i$$

$$\sigma_w{}^2 = \frac{1}{N-1}\sum_{1}^{N} (w_i X_i - \mu_w)^2$$

$$\text{cov\_1}_w = \frac{1}{N-2}\sum_{1}^{N} (w_i X_i - \mu_w) \ *(w_{i-1} X_{i-1} - \mu_w)$$

where N is the number of vectors within the group "gi" of the window "wi", μ_w the weighted mean, σ_w^2 the weighted variance and cov_1_w is first order weighted covariance.

Step b5)

In a preferred embodiment, the step b5) of the second algorithm 6 of the music audio analysis method 1, i.e. the extraction of data 7 being representative of the evolution of the tonal centre of the music piece given data 8, is implemented as follows.

Firstly, localized tonal centre estimates are computed by feeding independently each vector of data 7A into the Multi-Layer Perceptron (MLP).

The architecture of the MLP is such that its number of inputs matches the size of the vectors in data 7A.

In other words, the number of inputs of the MLP corresponds to the number of features describing the tonal context of window "w" (or generic window "wi").

In the preferred embodiment, there are D = 3*12*F such features.

The MLP may be built with an arbitrary number of hidden layers and hidden neurons.

The number of outputs is however fixed to 12 so that each output corresponds to one of the 12 possible pitches of the chromatic scale.

The parameters of the MLP are trained in a supervised fashion with stochastic gradient descent.

The training data consists of a large set of feature vectors describing the tonal

context of window "w" (or generic window "wi") for a variety of different music pieces.

To each such vector, a target tonal centre was manually associated by a number of expert musicologists.

The corresponding training data (i.e. the pairs feature vectors / tonal centre targets) can be enlarged by a factor 12 by considering all 12 possible transpositions of the CFP vectors (refer to third algorithm 8 for the transposition of CFPs hereinafter described).

The training consists in finding the set of parameters that maximises the output corresponding to the target tonal centre and that minimises the other outputs given the corresponding input data.

Using an appropriate choice of non-linearity functions (e.g. sigmoid function) and training cost function (e.g. cross-entropy cost function), the MLP outputs will estimate tonal centre posterior probabilities, i.e. each output will be bounded between 0 and 1 and they will sum to 1.

Once, localized tonal centre estimates have been computed thanks to the MLP, the corresponding localized posterior probabilities are smoothed over the course of the complete music piece assuming that tonal centre changes slowly and that if it does change, these changes follow some particular patterns.

In practice, it is assumed that the local estimate i only depends on the previous local estimate i-1, i.e. the process satisfies a 1st order Markov constraint.

This dependence between consecutive local estimates is modelled thanks to a transition matrix, which encodes the probability of going from the tonal centre estimate i-1 to the tonal centre estimate i.

Though these transition probabilities could be learnt from data, it is set manually according to some expert musicological knowledge (see table 2 for an example).

Moreover, it is assumed that initially all tonal centres are equally likely.

	C	**C#**	**D**	**Eb**	**E**	**F**	**F#**	**G**	**G#**	**A**	**Bb**	**B**
C	0.7	0.1	0.1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

C#	0.01	0.7	0.1	0.1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
D	0.01	0.01	0.7	0.1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Eb	0.01	0.01	0.01	0.7	0.1	0.1	0.01	0.01	0.01	0.01	0.01	0.01
E	0.01	0.01	0.01	0.01	0.7	0.1	0.1	0.01	0.01	0.01	0.01	0.01
F	0.01	0.01	0.01	0.01	0.01	0.7	0.1	0.01	0.01	0.01	0.01	0.01
F#	0.01	0.01	0.01	0.01	0.01	0.01	0.7	0.1	0.1	0.01	0.01	0.01
G	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.7	0.1	0.1	0.01	0.01
G#	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.7	0.1	0.1	0.01
A	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.7	0.1	0.1
Bb	0.1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.7	0.1
B	0.1	0.1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.7

Table 2

The problem of finding data 7, i.e. the optimal sequence of tonal centres over the course of the music piece, can be formulated as follows.

Let Tc1*, Tc2*,…, Tcn* be the optimal sequence of tonal centres and let Obs1, Obs2,…, Obsn be the sequence of feature vectors fed independently into the local tonal centre estimation MLP. Tc1*, Tc2*,…, Tcn* is such that:

$$Tc1^*,Tc2^*,\ldots,Tcn^* = \underset{Tc1,Tc2,\ldots,Tcn}{\operatorname{argmax}}\, p(Tc1,Tc2,\ldots,Tcn|Obs1,Obs2,\ldots,Obsn)$$

This is equivalent to find the most likely sequence of:

$$p(Tc1,Tc2,\ldots,Tcn,Obs1,Obs2,\ldots,Obsn) \approx \prod_t p(Tct|Obst)p(Tct|Tct\text{-}1)$$

where p(Tct|Obst) is the output of the local tonal centre estimation MLP corresponding to the local observation Obst and to the tonal centre Tct, p(Tct|Tct-1) is the entry of the transition probabilities matrix corresponding to the transition between Tct and Tct-1. Finally it is assumed initially that p(Tc0) = 1/12 (i.e. a uniform initial distribution for each tonal centre).

Given this formalisation, the most likely sequence of tonal centres Tc1*, Tc2*,…, Tcn* can be obtained thanks to the Viterbi algorithm. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states, in this case the most likely sequence of tonal centres, that results in a sequence of observed events, in this case the local tonal centre estimations of the MLP.

The modelling of the tonal context is implemented in practice by the computation of mean/variance/covariance 7A of the CFPs 7 in a generic window "wi" together with the MLP in charge of estimating the probability of each tonal centre Tci.

Figures 7A to 7D illustrate graphically the algorithm 6 once it has been applied on the first data 5.

In particular, Figure 7A shows a graphical representation of a sequence of CFP vectors of a music piece, i.e. first data 5, for F=2 chord families (i.e. CFP's of dimensionality 2*12 = 24) of a music audio signal 2, having in axis the vector for a generic audio segment s-on-i and in ordinate the dimension.

Figure 7B shows a graphical representation of a sequence of D dimensional vectors representative of the tonal content over window "wi", i.e. second data 7, having in axis the vector for a generic window "wi" and in ordinate the dimension. Particularly, Figure 7B shows the longer-term vectors corresponding to the mean/variance/covariance of the shorter-term CFP vectors over the windows "w".

Figure 7C shows a graphical representation of a sequence of local tonal centre estimates, i.e. the 12 dimensional outputs of the MLP, having in axis the vector for a generic window "wi" and in ordinate the pitch class.

Figure 7D finally shows a graphical representation of the corresponding optimal sequence of tonal centres obtained thanks to the Viterbi algorithm, i.e. the final tonal centre estimates for each window "wi", having in axis the vector for a generic window "wi" and in ordinate the pitch class.

THIRD ALGORITHM 8

Step c)

By referring again to figure 4, the third algorithm 8 comprises the step c1) of transposing to a reference pitch the first data 5 in function of second data 7 so as to generate the third data 9.

Thanks to the third algorithm 8, the third data 9 are made invariant with respect to the second data 7.

In fact, once an optimal tonal centre of the first data 5 has been identified using the second algorithm 6 heretofore described, each CFP vectors of the group g1 (or

g2, …, gi) is made invariant to transposition by transposing the vector values to a reference pitch.

For example, the reference pitch can be C.

In practice, this is implemented by a simple circular permutation:

$$TCFP_t(i,mod(j- Tt,12)) = CFPt(i,j)$$

where $TCFP_t$ is the transposed CFP vector at time t, i is the chord family index, j the pitch class and T_t the tonal centre pitch class at time t.

The step c1) of transposing to a reference pitch the first data 5 is a normalization operation, that allows to compare any kind of audio music signal based upon tonal considerations.

Referring now to figure 8, the apparatus able to perform the method heretofore described comprises:

- an input for receiving the digital music audio signal 2;
- a processor unit 18 for processing said digital music audio signal 2;
- and a database 19 in which are stored representatives of similar or dissimilar musical events (such events may correspond to known attacks of an original musical event), the database 19 being in signal communication with the processor unit 18.

Advantageously the processor unit 18 is configured to extract the CFP 7 representative of the tonal centre of the audio music signal 2.

Those skilled in the art will obviously appreciate that a number of changes and variants may be made to the embodiment as described hereinbefore to meet incidental and specific needs, without departure from the scope of the invention, as defined in the following claims.

CLAIMS

1. A music audio analysis method for analyzing a digital music audio signal (2) in order to extract a set of Chord Family Profiles (CFP) contained in said digital music audio signal (2), the method **comprising the steps of**:

a) applying a first algorithm (4) to the music audio signal (2) in order to extract first data (5) representative of the tonal context of music audio signal (2), and

b) applying a second algorithm (6) to said first data (5) in order to provide second data (7) representative of the tonal centre (Tc) contained in the first data (5).

2. A music audio analysis method according to claim 1, wherein the first algorithm comprises the steps of:

a1) identify (10) a sequence of note onsets in the music audio signal (2), in order to define the time position of a plurality of peaks (p1, p2, p3, ..., pi);

a2) dividing the audio music signal (2) into a plurality of audio segments (s-on-1, s-on-2, s-on-3, ..., s-on-i) having a duration (T), each audio segments containing one of said plurality of peaks (p1, p2, p3, ..., pi);

a3) applying a frequency analysis to each audio segment (s-on-1, s-on-2, s-on-3, s-on-i) during a predetermined sub-duration (t), wherein the length of the sub-duration (t) is less than the length of said duration (T), in order to obtain a plurality of spectrum segments (sp-1, sp-2, sp-3, sp-i).

3. A music audio analysis method according to claim 2, wherein the first algorithm comprises the steps of:

a4) processing said plurality of spectrum segments (sp-1, sp-2, sp-3, ..., sp-i) by a computation network (12) in order to provide said first data (5), the first data (5) comprising a plurality of vectors (v1, v2, v3, ..., vi) describing a "chord type/pitch class" pair, wherein each vector of the plurality of vectors (v1, v2, v3, ..., vi) corresponds to the respective audio segment (s-on-1, s-on-2, s-on-3, ..., s-on-i).

4. A music audio analysis method according to claim 3, wherein said computation network (12) is implemented with a trained machine-learning algorithm.

5. A music audio analysis method according to claim 4, wherein said trained machine-learning algorithm (12) is trained in two steps:

- first step in a supervised training with few hand labelled training data (13)

and

- second step in an unsupervised training with a larger set (14) of unlabelled training data.

6. A music audio analysis method according to claim 5, wherein the second step is performed in order to refine a set of weights (ω) of the trained machine-learning algorithm (12) obtained after the first step.

7. A music audio analysis method according to claim 3, wherein the first algorithm comprises, after the frequency analysis step a3), the further step of:

a5) filtering said plurality of spectrum segments (sp-1, sp-2, sp-3, ..., sp-i), by a moving average in order to emphasize the peak (p1', p2', p3' ..., pi') in each of said plurality of spectrum segments (sp-1, sp-2, sp-3, ..., sp-i).

8. A music audio analysis method according to claim 3, wherein said computing stage a4) is computed for each plurality of segments between two consecutive detected segments.

9. A music audio analysis method according to anyone of proceeding claims 2 to 8, wherein said frequency analysis is performed only during said sub-duration (t), said sub-duration (t) being in the range of 250-350 msec.

10. A music audio analysis method according to anyone of preceding claims, wherein the second algorithm comprises the steps of:

b1) providing a first window (w1) having a first prefixed duration (T1) containing a first group (g1) of plurality of vectors composing the first data (5);

b2) elaborating said first group (g1) of plurality of vectors contained in said window (w) for estimating a first tonal context (Tc1) representative of the local tonal centre contained in said first window (w1).

b3) providing a second window (w2) having a second prefixed duration (T2), said second window (w2) being a shifted window of a prefixed shifted time (Ts) of said first window (w1) so as said second window (w2) is overlapped with respect to said first window (w1), said second window (w2) comprising a second group (g2) of plurality of vectors;

b4) computing said second group (g2) of plurality of vectors contained in said second window (w2) for estimating a second tonal context (Tc2) representative of the

local tonal centre contained in said second window (w2);

b5) elaborating the tonal context (Tc1) of said first window (w1) and the tonal context (Tc2) of said second window (w2) in order to generate said second data (7), the latter being representative of the evolution of the tonal centre of said first data (5).

11. A music audio analysis method according to claim 10, wherein the second algorithm comprises the further step of:

b6) repeating the steps from b3) to b5) for defining further windows (wi) wherein each further windows (wi) contains a group (gi) of vectors for estimating the tonal context (Tc) contained in said first data (5).

12. A music audio analysis method according to claim 10, wherein the first prefixed duration (T1) is set in the range of 25 - 35 sec, more preferably about 30 sec.

13. A music audio analysis method according to claim 10, wherein the prefixed shifted time (Ts) is set in the range of 10 - 20 sec, more preferably about 15 sec., said second prefixed duration (T2) varying in the range between the difference of:

- the first prefixed duration (T1) and the prefixed shifted time (Ts) and
- the first prefixed duration (T1).

14. A music audio analysis method according to claim 10, wherein the step b5) is implemented by a Multi-Layer Perceptron (MLP).

15. A music audio analysis method according to anyone of preceding claims, wherein the method comprises the further step c) of applying a third algorithm (8) to the first data (5) in function of the second data (7) in order to provide said feature set (CFP) of characteristics music audio signal (2).

16. A music audio analysis method according to claim 15, wherein the third algorithm (8) comprises the step of transposing to a reference pitch said first data (5) in order to make invariant said first data (5).

17. A computer program product comprising a program for analyzing a music audio signal in order to extract at least a feature set representative of the content of the audio music signal, said computer program product **comprising the steps** of:

a) applying a first algorithm (4) to the music audio signal (2) in order to extract first data (5) representative of the tonality of music audio signal (2), and

b) applying a second algorithm (6) to said first data (5) in order to provide second data (7) representative of the tonal centre contained in the first data (5).

18. An apparatus for analyzing a music audio signal in order to extract at least a feature set representative of the content of the audio music signal, the apparatus comprising:

- an input for receiving a digital music audio signal (2);
- a processor unit (18) for processing said digital music audio signal (2);
- and a database (19) in which are stored representatives of similar or dissimilar musical events,

wherein the processor unit (18) is configured to extract the feature set representative of the content of the digital music audio signal (2) according to the music audio analysis method of any one of preceding claims from 1 to 16.

Chord Examples: family "Major"
- 3 pitches
- quality: intervals of 5 and 8 pitches between root pitch and each of the other pitches

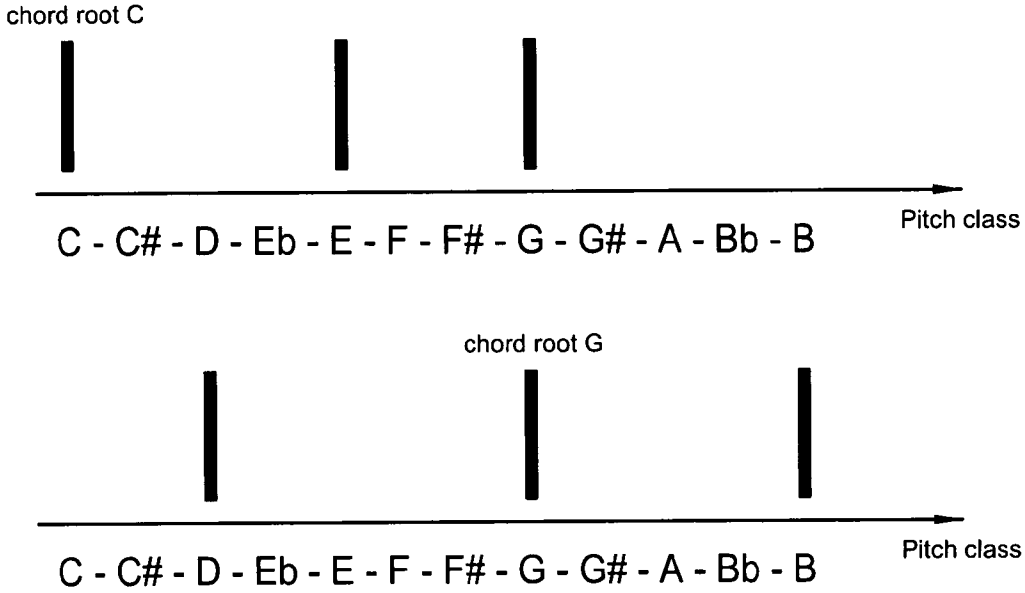


Figure 1a

Chord Examples: family "Minor"
- 3 pitches
- quality: intervals of 5 and 8 pitches between root pitch and each of the other pitches

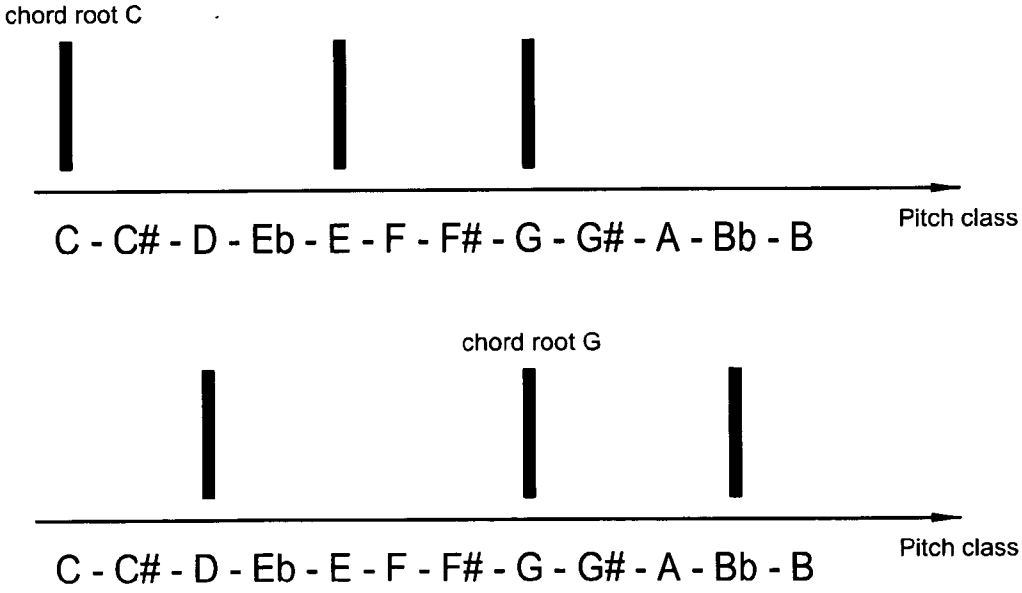
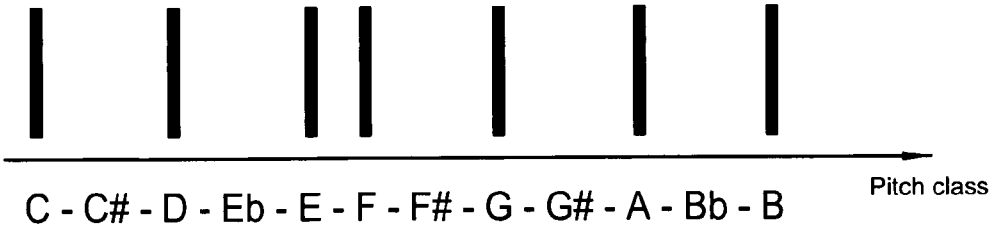


Figure 1b

key Examples: Mode "Major"
- 7 pitches
- intervals of 3, 5, 6, 8, 10 and 12 pitches between tonal center pitch and each of the other pitch

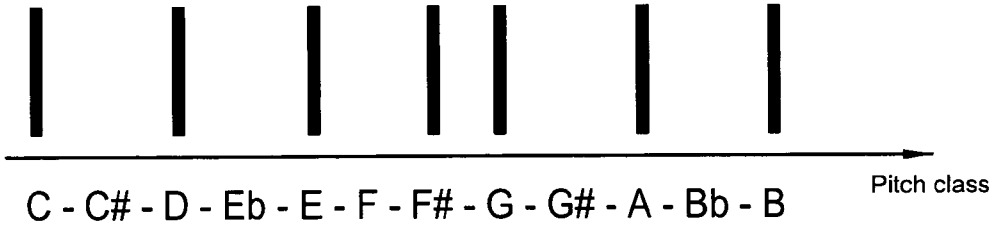
Tonal center C



C - C# - D - Eb - E - F - F# - G - G# - A - Bb - B

Pitch class

tonal center G



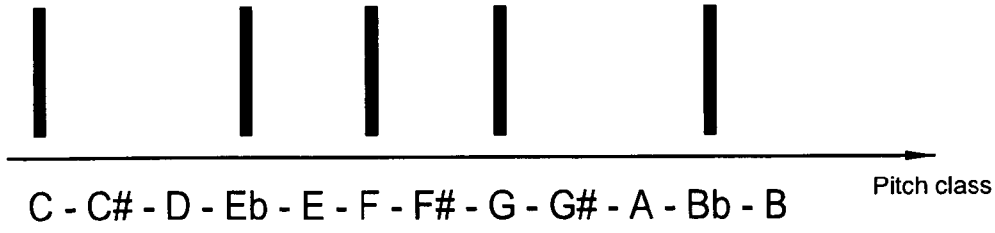
C - C# - D - Eb - E - F - F# - G - G# - A - Bb - B

Pitch class

Figure 2a

key Examples: Mode "Blues"
- 5 pitches
- intervals of 4, 6, 8, and 11 pitches between tonal center pitch and each of the other pitch

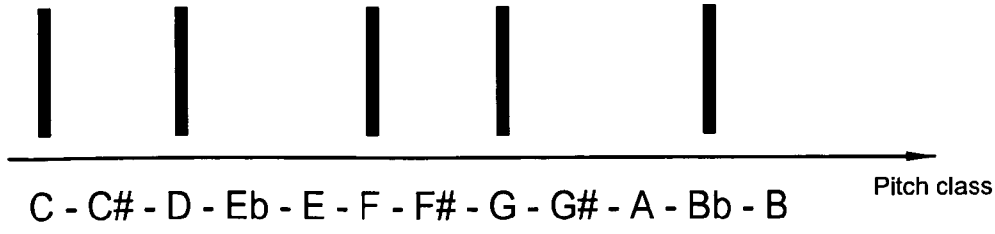
Tonal center C



C - C# - D - Eb - E - F - F# - G - G# - A - Bb - B

Pitch class

tonal center G



C - C# - D - Eb - E - F - F# - G - G# - A - Bb - B

Pitch class

Figure 2b

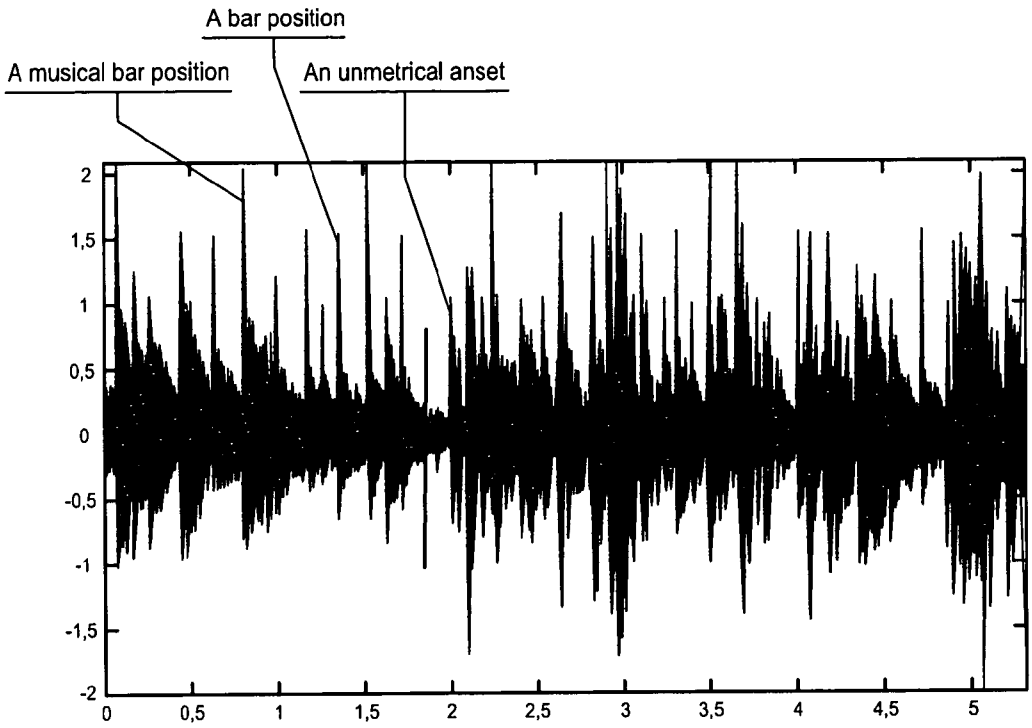


Figure 3

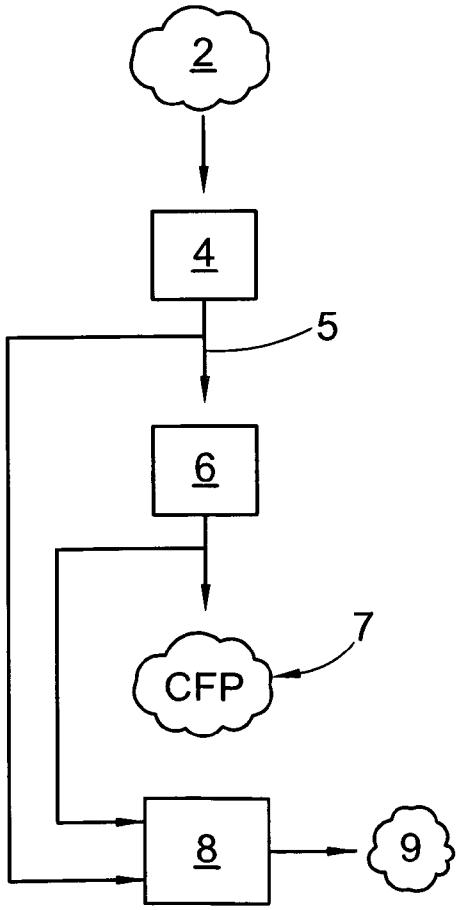


Figure 4

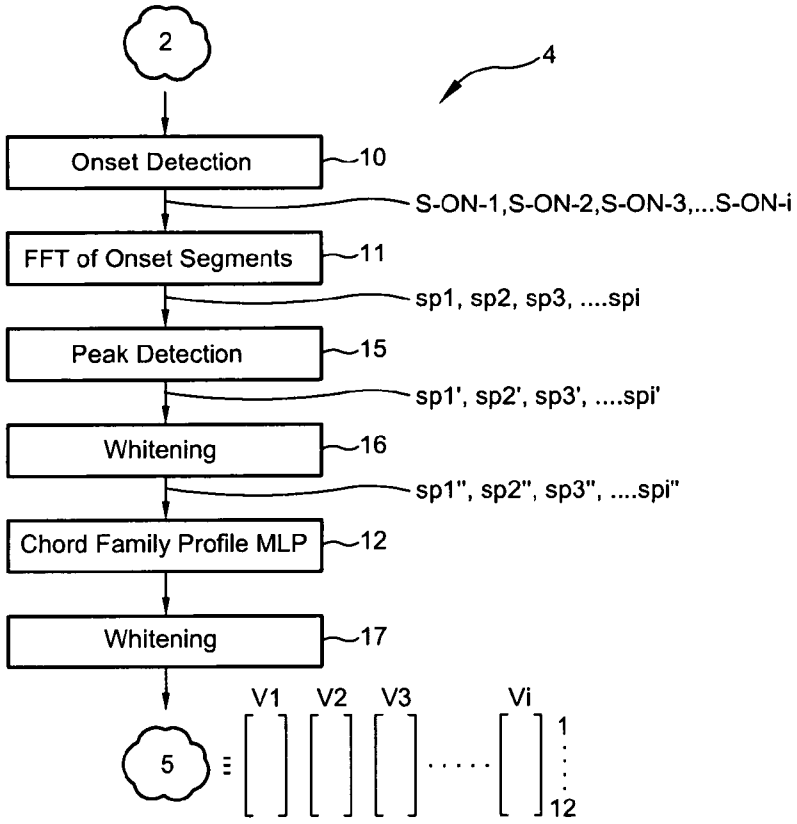


Figure 5a

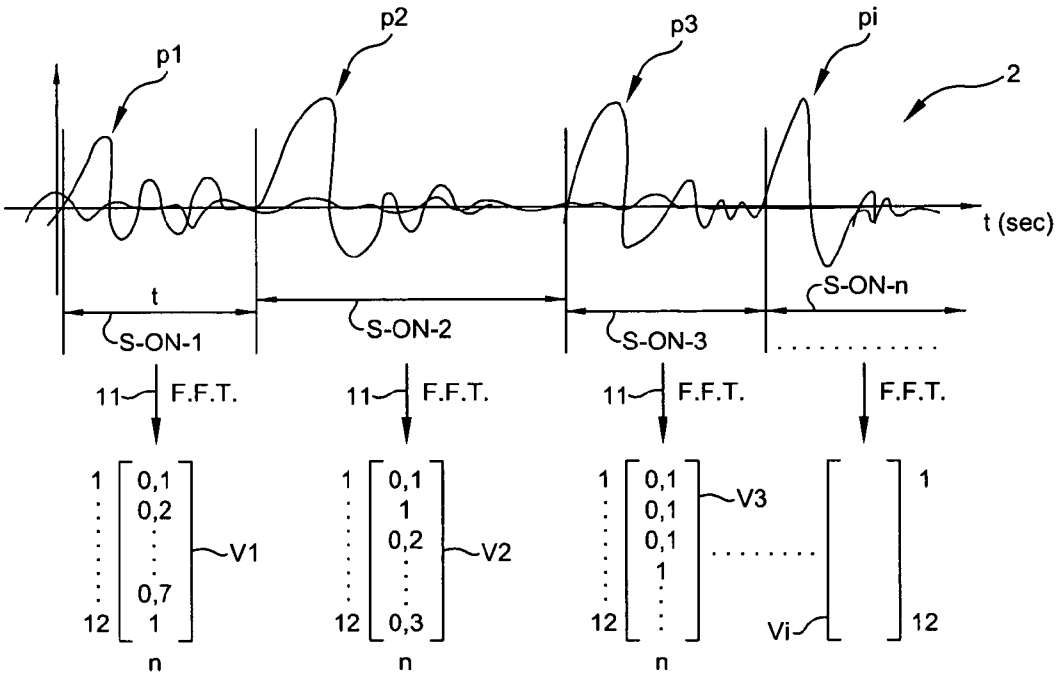


Figure 5b

Figure 6a

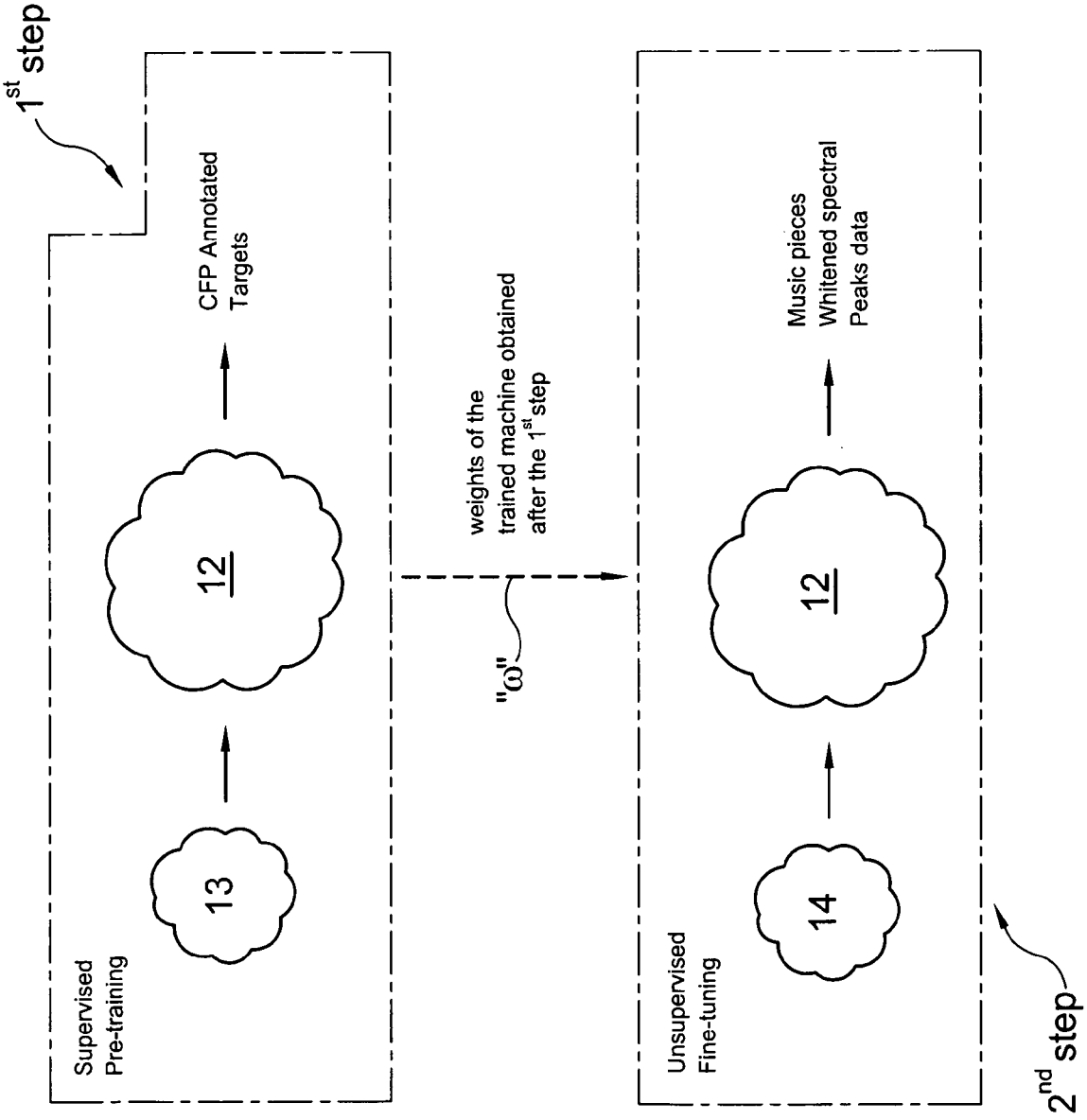


Figure 6b

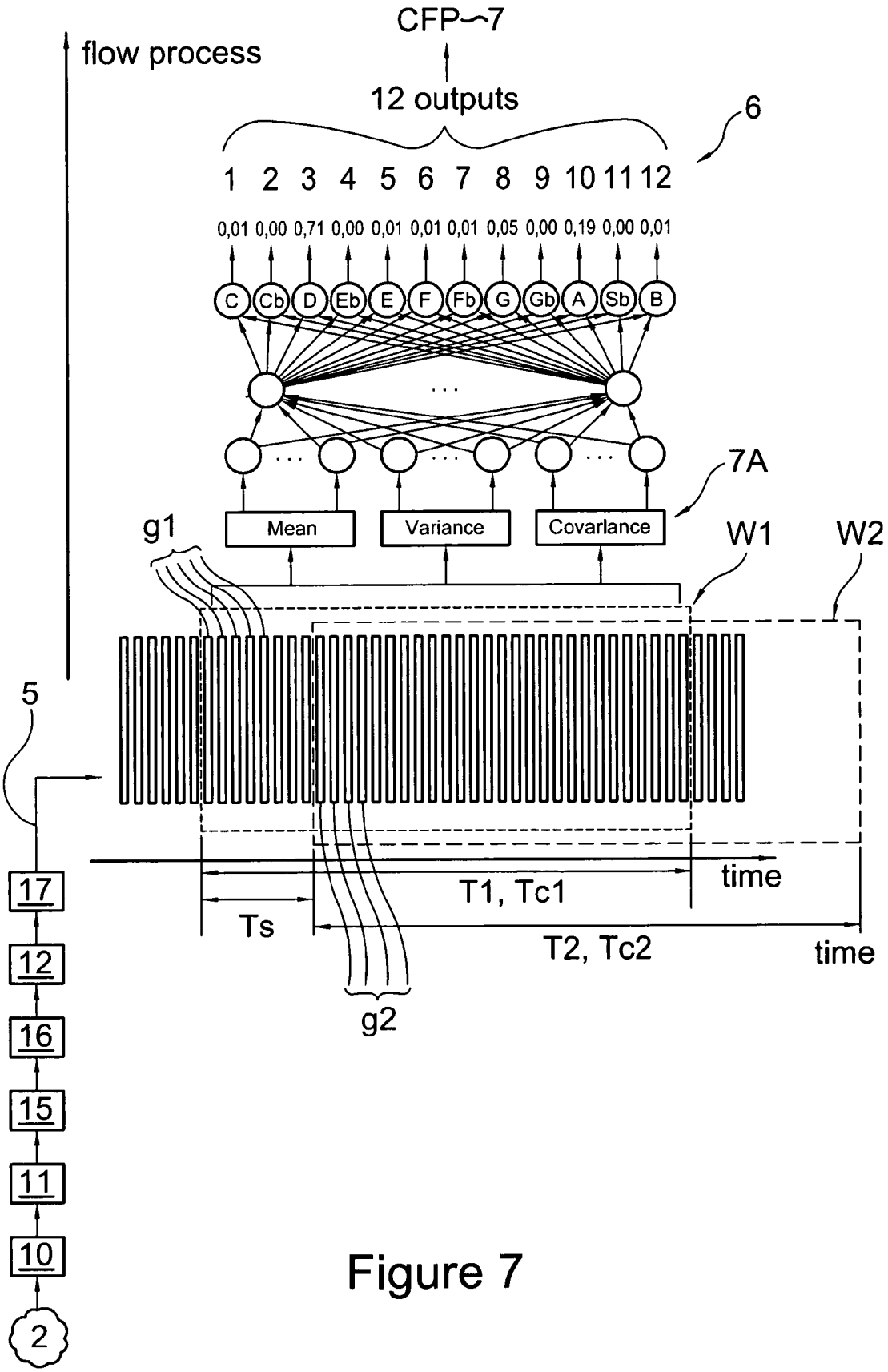

CFP
7
flow process
12 outputs
6
1 2 3 4 5 6 7 8 9 10 11 12
0,01 0,00 0,71 0,00 0,01 0,01 0,01 0,05 0,00 0,19 0,00 0,01
C Cb D Eb E F Fb G Gb A Sb B
7A
Mean
Variance
Covarlance
g1
W1
W2
5
17
12
16
15
11
10
2
time
T1, Tc1
Ts
T2, Tc2
time
g2

Figure 7

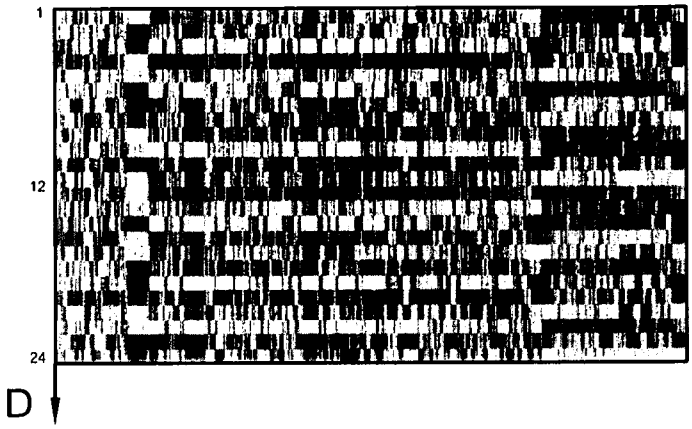


Figure 7a

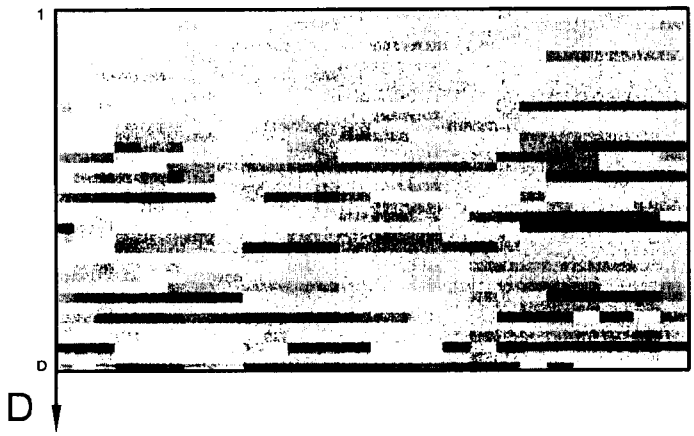


Figure 7b

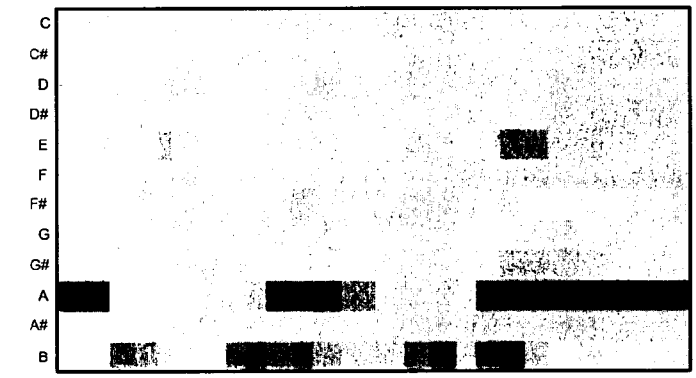


Figure 7c

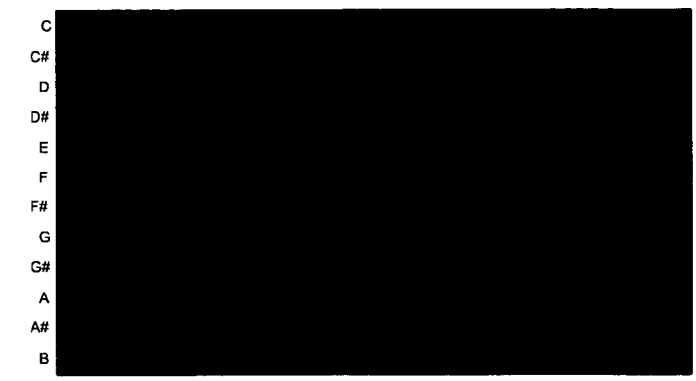


Figure 7d

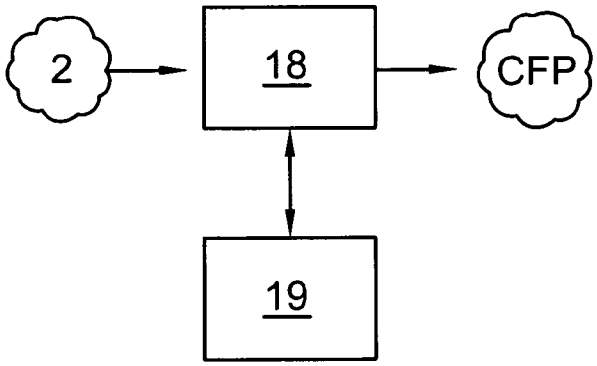


Figure 8

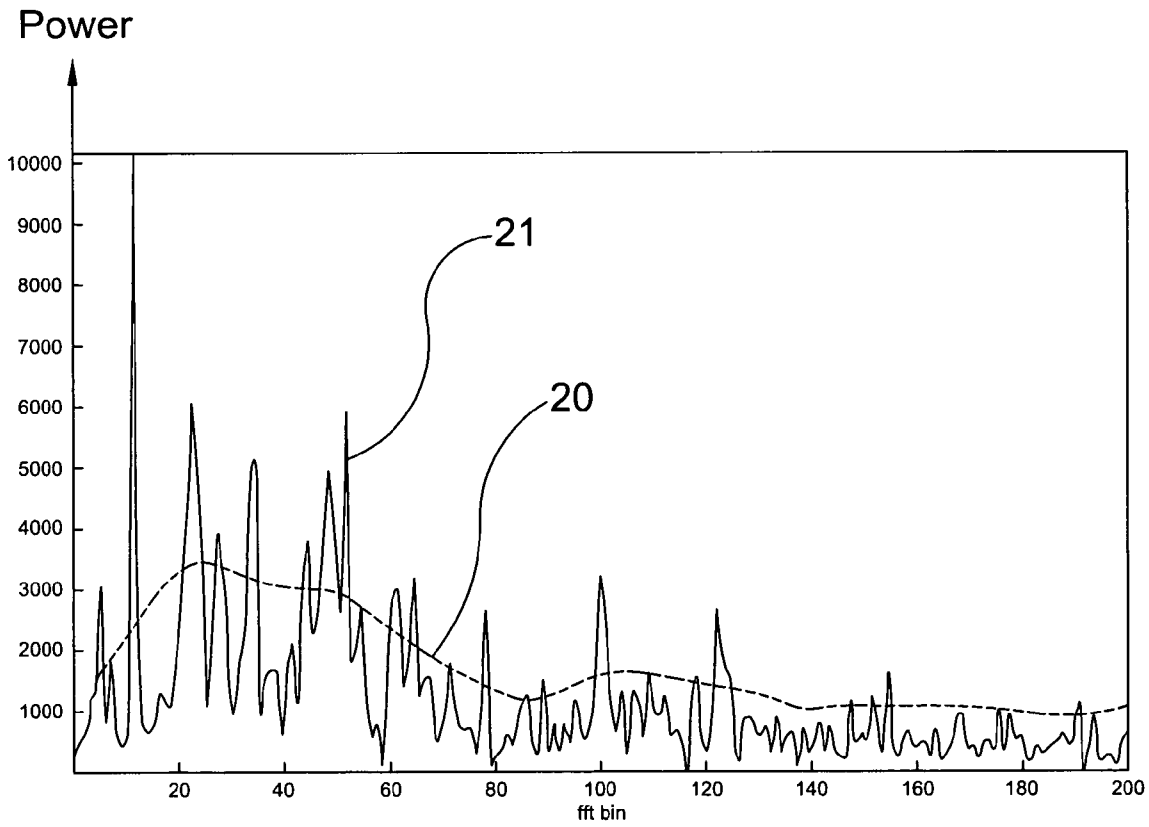


Figure 9

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2008/063911

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10H1/38

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)
G10H

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, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2008/245215 A1 (KOBAYASHI YOSHIYUKI [JP]) 9 October 2008 (2008-10-09) abstract; figures 1,12,14-28 paragraphs [0007] - [0023] paragraphs [0093] - [0107] paragraphs [0137] - [0147] paragraphs [0170] - [0201] paragraphs [0228] - [0234] paragraph [0294]	1-9,17, 18
X	US 6 057 502 A (FUJISHIMA TAKUYA [JP]) 2 May 2000 (2000-05-02) abstract; figures 1-16 column 1, lines 15-56 column 2, line 36 - column 4, line 37 column 6, line 42 - column 9, line 11 column 12, line 66 - column 14, line 58	1-3, 15-18

-/--

[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
20 February 2009	09/03/2009

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	Lecointe, Michael

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ZOIA G ET AL: "A multi-timbre chord/harmony analyzer based on signal processing and neural networks" MULTIMEDIA SIGNAL PROCESSING, 2004 IEEE 6TH WORKSHOP ON SIENA, ITALY SEPT. 29 - OCT. 1, 2004, PISCATAWAY, NJ, USA,IEEE, 29 September 2004 (2004-09-29), pages 219-222, XP010802125 ISBN: 978-0-7803-8578-8 abstract; figures 1-3 Sections III-IV	1,17,18
A		4-6,14
X	CHING-HUA CHUAN ET AL: "Polyphonic Audio Key Finding Using the Spiral Array CEG Algorithm" MULTIMEDIA AND EXPO, 2005. ICME 2005. IEEE INTERNATIONAL CONFERENCE ON AMSTERDAM, THE NETHERLANDS 06-06 JULY 2005, PISCATAWAY, NJ, USA,IEEE, 6 July 2005 (2005-07-06), pages 21-24, XP010843224 ISBN: 978-0-7803-9331-8 Sections 2.1-2.2	1,17,18
X	ÖZGÜR IZMIRLI: "An Algorithm for Audio Key Finding" INTERNET CITATION, [Online] XP002426658 Retrieved from the Internet: URL:http://www.music-ir.org/evaluation/mirex-results/articles/key_audio/izmirli.pdf> [retrieved on 2007-03-26] the whole document	1,10-13, 17,18
X	EMILIA GÓMEZ: "TONAL DESCRIPTION OF MUSIC AUDIO SIGNALS" 20060101, 1 January 2006 (2006-01-01), XP002501266 Sections 3.1-3.2 Sections 3.4-3.7	1,15-18
X	PEETERS G: "Chroma-based estimation of musical key from audio-signal analysis" PROCEEDINGS ANNUAL INTERNATIONAL SYMPOSIUM ON MUSIC INFORMATIONRETRIEVAL, XX, XX, 1 October 2006 (2006-10-01), pages 1-6, XP002447156 abstract; figures 1-4 Sections 1-2.4.2	1,10-13, 17,18

-/--

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	FUJISHIMA T: "REALTIME CHORD RECOGNITION OF MUSICAL SOUND: A SYSTEM USING COMMON LISP MUSIC" ICMC. INTERNATIONAL COMPUTER MUSIC CONFERENCE. PROCEEDINGS, XX, XX, 27 September 1999 (1999-09-27), pages 464-467, XP009053025 abstract; figures 1-6 Sections 1-4.1	1-4, 10-13, 17,18

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2008/063911

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2008245215 A1	09-10-2008	JP 2008102406 A	01-05-2008
US 6057502 A	02-05-2000	JP 3826660 B2	27-09-2006
		JP 2000298475 A	24-10-2000