(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 777 698 B1

(12) EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention of the grant of the patent:
03.09.2008 Bulletin 2008/36

(51) Int Cl.:
G10L 19/02 *(2006.01)*

(21) Application number: **07002179.5**

(22) Date of filing: **27.08.2003**

(54) **Bit rate reduction in audio encoders by exploiting auditory temporal masking**

Bitraten-Reduktion in Audiokodierern unter Ausnutzung zeitlicher Maskierung

Réduction de débit dans un codeur audio utilisant un effet de masquage temporaire

(84) Designated Contracting States:
DE FR GB

(30) Priority: **27.08.2002 US 406055 P**

(43) Date of publication of application:
25.04.2007 Bulletin 2007/17

(62) Document number(s) of the earlier application(s) in accordance with Art. 76 EPC:
03405620.0 / 1 398 761

(73) Proprietor: **Her Majesty in Right of Canada as Represented by the Minister of Industry**
Ottawa,
Ontario K2H 8S2 (CA)

(72) Inventors:
- **Najaf-Zadeh, Hossein**
  Stittsville
  Ontario K2S 2G2 (CA)
- **Lahdili, Hassan**
  Hull
  Quebec J8Y 2B7 (CA)
- **Thibault, Louis**
  Hull
  Quebec J8Z 2G8 (CA)
- **Treurniet, William**
  Ajax
  Ontario L1Z 1N1 (CA)

(74) Representative: **Frei Patent Attorneys**
Frei Patentanwaltsbüro
Postfach 1771
8032 Zürich (CH)

(56) References cited:
- **YUAN-HAO HUANG ET AL: "A New Audio Coding Scheme Using a Forward Masking Model and Perceptually Weighted Vector Quantization" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, US, vol. 10, no. 5, July 2002 (2002-07), XP011079654 ISSN: 1063-6676**
- **TED PAINTER ET AL: "Perceptual Coding of Digital Audio" PROCEEDINGS OF THE IEEE, IEEE. NEW YORK, US, vol. 88, no. 4, April 2000 (2000-04), XP011044355 ISSN: 0018-9219**

EP 1 777 698 B1

Description

Field of the Invention

[0001] The present invention relates generally to the field of perceptual audio coding and more particularly to a method for determining masking thresholds using a psychoacoustic model.

Background of the Invention

[0002] In present state of the art audio coders, perceptual models based on characteristics of a human ear are typically employed to reduce the number of bits required to code a given input audio signal. The perceptual models are based on the fact that a considerable portion of an acoustic signal provided to the human ear is discarded - masked - due to the characteristics of the human hearing process. For example, if a loud sound is presented to the human ear along with a softer sound, the ear will likely hear only the louder sound. Whether the human ear will hear both, the loud and soft sound, depends on the frequency and intensity of each of the signals. As a result, audio coding techniques are able to effectively ignore the softer sound and not assign any bits to its transmission and reproduction under the assumption that a human listener is not capable of hearing the softer sound even if it is faithfully transmitted and reproduced. Therefore, psychoacoustic models for calculating a masking threshold play an essential role in state of the art audio coding. An audio component whose energy is less than the masking threshold is not perceptible and is, therefore, removed by the encoder. For the audible components, the masking threshold determines the acceptable level of quantization noise during the coding process.

[0003] However, it is a well-known fact that the psychoacoustic models for calculating a masking threshold in state of the art audio coders are based on simple models of the human auditory system resulting in unacceptable levels of quantization noise or reduced compression. Hence, it is desirable to improve the state of the art audio coding by employing better - more realistic - psychoacoustic models for calculating a masking threshold.

[0004] Furthermore, the MPEG-1 Layer 2 audio encoder is widely used in Digital Audio Broadcasting (DAB) and digital receivers based on this standard have been massively manufactured making it impossible to change the decoder in order to improve sound quality. Therefore, enhancing the psychoacoustic model is an option for improving sound quality without requiring a new standard.

Summary of the Invention

[0005] It is, therefore, an object of the present invention to provide a method for determining temporal masking thresholds as claimed in the appended claims.

Brief Description of the Drawings

[0006] Exemplary embodiments of the invention will now be described in conjunction with the drawings in which:

[0007] Fig. 1 is a simplified flow diagram of a first embodiment of a method for encoding an audio signal according to the present invention;

[0008] Fig. 2 is a diagram illustrating reduction in SMR due to temporal masking;

[0009] Figs. 3a and 3b are diagrams illustrating an example of a harmonic and an inharmonic signal, respectively;

[0010] Fig. 4 is a simplified flow diagram illustrating a process for determining inharmonicity of an audio signal according to the invention;

[0011] Figs. 5a and 5b are diagrams illustrating the outputs of a gammatone filterbank for a harmonic and an inharmonic signal, respectively;

[0012] Figs. 6a and 6b are diagrams illustrating the envelope autocorrelation for a harmonic and an inharmonic signal, respectively; and,

[0013] Fig. 7 is a simplified flow diagram of a second example of a method for encoding an audio signal.

Detailed Description of the Invention

[0014] Most psychoacoustic models are based on the auditory "simultaneous masking" phenomenon where a louder sound renders a weaker sound occurring at a same time instance inaudible. Another less prominent masking effect is "temporal masking". Temporal masking occurs when a masker - louder sound - and a maskee - weaker sound - are presented to the hearing system at different time instances. Detailed information about the temporal masking is disclosed in the following references:

B. Moore, "An Introduction to the Psychology of Hearing", Academic Press, 1997;

E. Zwicker, and T. Zwicker, "Audio Engineering and Psychoacoustics, Matching Signals to the Final Receiver, the Human Auditory System", J. Audio Eng. Soc., Vol. 39, No. 3, pp 115-126, Mar. 1991; and,

E. Zwicker and H. Fastl, "Psychoacoustics Facts and Models", Springer Verlag, Berlin, 1990.

[0015] The temporal masking characteristic of the human hearing system is asymmetric, i.e. "backward masking" is effective approximately 5 msec before occurrence of a masker, whereas "forward masking" lasts up to 200 msec after the end of the masker. Different phenomena contributing to temporal auditory masking effects include temporal overlap of basilar membrane responses to different stimuli, short term neural fatigue at higher neural levels and persistence of the neural activity caused by a masker, disclosed in B. Moore, "An Introduction to the Psychology of Hearing", Academic Press, 1997; and A. Harma, "Psychoacoustic Temporal Masking Effects with Artificial and Real Signals", Hearing Seminar, Espoo, Finland, pp. 665-668, 1999.

[0016] Since psychoacoustic models are used for adaptive bit allocation, the accuracy of those models greatly affects the quality of encoded audio signals. Since digital receivers have been massively manufactured and are now readily available, it is not desirable to change the decoder requirements by introducing a new standard. However, enhancing the psychoacoustic model employed within the encoders allows for improved sound quality of an encoded audio signal without modifying the decoder hardware. Incorporating non-linear masking effects such as temporal masking and inharmonicity into the MPEG-1 psychoacoustic model 2 significantly reduces the bit rate for transparent coding or equivalently, improves the sound quality of an encoded audio signal at a same bit rate.

[0017] In a first embodiment of a method for encoding an audio signal according to the invention a temporal masking index is determined in a non-linear fashion in time domain and implemented into a psychoacoustic model for calculating a masking threshold. In particular, a combined masking threshold considering temporal and simultaneous masking is calculated using the MPEG-1 psychoacoustic model 2. Listening tests have been performed with MPEG-1 Layer 2 audio encoder using the combined masking threshold. In the following it will become apparent to those of skill in the art that the method for encoding an audio signal according to the invention has been implemented into the MPEG-1 psychoacoustic model 2 in order to use a standard state of the art implementation but is not limited thereto.

[0018] Since the temporal masking method according to the invention is implemented in the MPEG-1 Layer 2 encoder, the relation between some of the encoder parameters and the temporal masking method will be discussed in the following. In the MPEG-1 psychoacoustic model 32 Signal-to-Mask-Ratios (SMR) corresponding to 32 subbands are calculated for each block of 1152 input audio samples. Since the time-to-frequency mapping in the encoder is critically sampled, the filterbank produces a matrix - frame - of 1152 subband samples, i.e. 36 subband samples in each of the 32 subbands. Accordingly, the temporal masking method according to the invention as implemented in the MPEG-1 psychoacoustic model acquires 72 subband samples - 36 samples belonging to a current frame and 36 samples belonging to a previous frame - in each subband and provides 32 temporal masking thresholds.

[0019] Referring to Fig. 1 a simplified flow diagram of the first embodiment of a method for encoding an audio signal is shown. The temporal masking method has been implemented using the following model suggested by W. Jesteadt, S. Bacon, and J. Lehman, "Forward masking as a function of frequency, masker level, and signal delay", J. Acoust. Soc. Am., Vol. 71, No. 4, pp. 950-962, April 1982:

$$M = a(b - \log_{10} t)(L_m - c)$$

where M is the amount of masking in dB, t is the time distance between the masker and the maskee in msec, L_m is the masker level in dB, and a, b, and c are parameters found from psychoacoustic data.

[0020] For determining the parameters in the above model the fact that forward temporal masking lasts for up to 200 msec whereas backward temporal masking decays in less than 5 msec has been considered. Furthermore, temporal masking at any time index is taken into account if the masker level is greater than 20 dB. Considering the above mentioned assumptions and based on listening tests of numerous audio materials the following forward and backward temporal masking functions have been determined, respectively. For forward masking

$$FTM(j,i) = 0.2(2.3 - \log_{10}(\tau(j-i)))(L_f(i) - 20),$$

where $j = i + 1,...,36$ is the subband sample index, τ is the time distance between successive subband samples - in msec, and $L_f(i)$ is the forward masker level in dB. For backward masking

$$BTM(j,i) = 0.2(0.7 - \log_{10}(\tau(i-j)))(L_b(i) - 20),$$

where $j = 1,...,i$-1 is the subband sample index, τ is the time distance between successive subband samples - in msec, and $L_b(i)$ is the backward masker level in dB. For the backward temporal masking function the time axis is reversed.

[0021] The time distance τ between successive subband samples is a function of the sampling frequency. Since the filterbank in the MPEG audio encoder is critically sampled - box 10 - one subband sample in each subband is produced for 32 input time samples. Therefore, the time distance τ between successive subband samples is $32/f_s$ msec, where f_s is the sampling frequency in kHz.

[0022] The masker level in forward masking at time index *i* is given by

$$L_f(i) = 10\log_{10}\frac{\sum_{k=-36}^{i} s^2(k)}{36+i},\ i = 1,...,35,$$

where $s(k)$ denotes the subband sample at time index *k* - box 12. At any time index *i* the masker level is calculated as the average energy of the 36 subband samples in the corresponding subband in the previous frame and the subband samples in the current frame up to time index *i.*

[0023] Similarly, the masker level in backward masking - box 14 - at time index *i* is given by

$$L_b(i) = 10\log_{10}\frac{\sum_{k=i}^{36} s^2(k)}{36-(i-1)},\ i = 2,...,36.$$

The above equation gives the backward masker level at any time as the average energy of the current and future subband samples.

[0024] The forward temporal masking level at time index *j* is then calculated - box 16 - as follows,

$$M_f(j) = \max\{FTM(j,i)\}.$$

[0025] Similarly, the backward temporal masking level at time index *j* is then calculated - box 18 - as,

$$M_b(j) = \max\{BTM(j,i)\}.$$

[0026] The total temporal masking energy at time index *j* is the sum of the two components - box 20,

$$E_T(j) = 10^{\frac{M_f(j)}{10}} + 10^{\frac{M_b(j)}{10}},$$

where M_f and M_b are the forward and the backward temporal masking level in dB at time index *j*, respectively.

[0027] The SMR at each subband sample is then calculated - box 22 - as,

$$\mathrm{SMR}(j)=\frac{s^2(j)}{E_T(j)},\ j=1,\ldots,36,$$

where $s(j)$ is the j-th subband sample.

[0028] Since in the MPEG audio encoder all the subband samples in each frame are quantized with the same number of bits, the maximum value of the 36 SMRs in each subband is taken to determine the required precision in the quantization process - box 24,

$$\mathrm{SMR}^{(n)}=\max\{\mathrm{SMR}(j)\},\ n=1,\ldots,32,$$

where $\mathrm{SMR}^{(n)}$ is the required Signal-to-Mask-Ratio in subband n.

[0029] A combined masking threshold is then calculated considering the effect of both temporal and simultaneous masking. First the SMRs due to temporal masking are translated into allowable noise levels within a frequency domain. In order to achieve a same SMR in each subband in the frequency domain, the noise level in a corresponding subband in the frequency domain is calculated - box 26 - as,

$$N_{TM}^{(n)}=\frac{E_{sb}^{(n)}}{\mathrm{SMR}^{(n)}},$$

where $N_{TM}^{(n)}$ is the allowable noise level due to temporal masking - temporal masking index - in subband n in the frequency domain, and $E_{sb}^{(n)}$ is the energy of the DFT components in subband n in the frequency domain. Alternatively, Parseval's theorem is used to calculate the equivalent noise level in the frequency domain.

[0030] In the following step, the noise levels due to temporal and simultaneous masking are combined - box 28. One possibility is to linearly sum the masking energies. However, according to psychoacoustic experiments the linear combination results in an under-estimation of the net masking threshold. Instead, a "power law" method is used for combining the noise levels,

$$N_{net}=\left(N_{TM}^{p}+N_{SM}^{p}\right)^{1/p},$$

where N_{TM} and N_{SM} are the allowable noise due to temporal and simultaneous masking, respectively, and N_{net} is the net masking energy. For the parameter p, a value of 0.4 has been found to provide an accurate combined masking threshold.

[0031] The net masking energy is used in the MPEG-1 psychoacoustic model 2 to calculate the corresponding SMR - masking threshold - in each subband - box 30,

$$\mathrm{SMR}_{net}^{(n)}=\frac{E_{sb}^{(n)}}{N_{net}^{(n)}}.$$

[0032] Finally, the acoustic signal is encoded using the masking threshold determined above - box 32.

[0033] Figure 2 shows an amount of reduction in SMR due to temporal masking in a frame of 1152 subband samples - 36 samples in each of 32 subbands.

[0034] Numerous audio materials have been encoded and decoded with the MPEG-1 Layer 2 audio encoder using psychoacoustic model 2 based on simultaneous masking and the method for encoding an audio signal according to the invention based on the improved psychoacoustic model including temporal masking. Bit allocation has been varied adaptively to lower the quantization noise below the masking threshold in each frame. Use of the combined masking model resulted in a bit-rate reduction of 5-12%.

Table 1

Audio Material	*Average Bit Rate Without TM*	*Average Bit Rate With TM*
Susan Vega	153.8	138.1
Tracy Chapman	167.2	157.7
Sax+Double Bass	191.2	177.4
Castanets	150.2	132.0
Male Speech	120.1	112.4
Electric Bass	145.6	129.9

[0035] Table 1 shows the average bit rate for a few test files coded with a MPEG-1 Layer 2 encoder using the standard psychoacoustic model 2 and using the modified psychoacoustic model. The test files were 2-channel stereo audio signals sampled at 48 kHz with 16-bit resolution.

[0036] In order to compare the subjective quality of the compressed audio materials semiformal listening tests involving six subjects have been conducted. The listening tests showed that using the method for encoding an audio signal according to the invention the subjective high quality of the decoded compressed sounds has been maintained while the bit rate was reduced by approximately 10%.

[0037] Since psychoacoustic models are used for adaptive bit allocation, the accuracy of those models greatly affects the quality of encoded audio signals. For instance, the MPEG-1 Layer 2 audio encoder is used in Digital Audio Broadcasting (DAB) in Europe and in Canada. Since digital receivers have been massively manufactured and are now readily available, it is not possible to change the decoder without introducing a new standard. However, enhancing the psychoacoustic model allows improving the sound quality of an encoded audio signal without modifying the decoder. Incorporating temporal masking into the MPEG-1 psychoacoustic model 2 significantly reduces the bit rate for transparent coding or equivalently, improves the sound quality of an encoded audio signal at a same bit rate.

[0038] W.C. Treumiet, and D.R. Boucher have shown in "A masking level difference due to harmonicity", J. Acoust. Soc. Am., 109(1), pp. 306-320, 2001, that the harmonic structure of a complex - multi-tonal - masker has an impact on the masking pattern. It has been found that if the partials in a multi-tonal signal are not harmonically related the resulting masking threshold increases by up to 10 dB. The amount of the increase depends on the frequency of the maskee and the frequency separation between the partials and the level of masker inharmonicity. For example, it has been found that for two different multi-tonal maskers having the same power, the one with a harmonic structure produces a lower masking threshold. This finding has been incorporated into a second example of an audio encoder comprising a modified MPEG-1 psychoacoustic model 2.

[0039] A sound is harmonic if its energy is concentrated in equally spaced frequency bins, i.e. harmonic partials. The distance between successive harmonic partials is known as the fundamental frequency whose inverse is called pitch. Many natural sounds such as harpsichord or clarinet consist of partials that are harmonically related. Contrary to harmonic sounds, inharmonic signals consist of individual sinusoids, which are not equally separated in the frequency domain.

[0040] A model developed to measure inharmonicity recognizes that an auditory filter output envelope is modulated when the filter passes two or more sinusoids as shown in Appendix A. since a harmonic masker has constant frequency differences between its adjacent partials, most auditory filters will have the same dominant modulation rate. On the other hand, for an inharmonic masker, the envelope modulation rate varies across auditory filters because the frequency differences are not constant.

[0041] When the signal is a complex masker comprising a plurality of partials, interaction of neighboring partials causes local variations of the basilar membrane vibration pattern. The output signal from an auditory filter centered at the corresponding frequency has an amplitude modulation corresponding to that location. To a first approximation, the modulation rate of a given filter is the difference between the adjacent frequencies processed by that filter. Therefore, the dominant output modulation rate is constant across filters for a harmonic signal because this frequency difference is constant. However, for inharmonic maskers, the modulation rate varies across filters. Consequently, in the case of a harmonic masker the modulation rate for each filter output signal is the fundamental frequency. When inharmonicity is introduced by perturbing the frequencies of the partials, a variation of the modulation rate across filters is noticeable. The variation increases with increasing inharmonicity. In general, the harmonicity nature of a complex masker is characterized by the variance calculated from the envelope modulation rates across a plurality of auditory filters.

[0042] Since a harmonic signal is characterized by particular relationships among sharp peaks in the spectrum, an appropriate starting point for measuring the effect of harmonicity is a masker having a similar distribution of energy across filters, but with small perturbations in the relationships among the spectral peaks. Fig. 3a shows an example of a harmonic signal comprising a fundamental frequency of 88 Hz, and a total of 45 equally spaced partials covering a range from 88 Hz to 3960 Hz. Fig. 3b shows an inharmonic signal generated by slightly perturbing the frequencies and randomizing the phases of the harmonic signal partials.

[0043] A process for estimating the harmonicity is illustrated in the flow chart of Fig. 4. The signal is analyzed using a "gammatone" filterbank based on the concept of critical bands disclosed in E. Zwicker, and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", J. Acoust. Soc. Am., 68(5), pp. 1523-1525, 1980. The output of each filter is processed with a Hilbert transform to extract the envelope. An autocorrelation is then applied to the envelope to estimate its period. Finally, the harmonicity measure is related to the variance of the modulation rates, i.e. envelope periods. This variance is negligible for a harmonic masker. However, for an inharmonic masker the variance is expected to be very large since the modulation rates vary across filters. For example, the two signals shown in Figs. 3a and 3b have been analyzed to verify the process. Figs. 5a, 5b, 6a, and 6b illustrate the output signals of the gammatone filterbank - channels 7-12 - and the corresponding autocorrelation functions for the harmonic - Figs. 5a and 6a - and inharmonic inputs- Figs. 5b and 6b. As shown in Figs. 6a and 6b, there is a notable difference between the autocorrelation functions. In the case of the harmonic signal all the peaks related to the dominant modulation rate are coincident. Consequently, the variance of the modulation rates is negligible. On the other hand, for the inharmonic signal, the peaks are not coincident. Therefore, the variance is much larger. A harmonicity estimation model based on the variability of envelope modulation rates differentiates harmonic from inharmonic maskers. The variance of the modulation rate measures the degree to which an audio signal departs from harmonicity, i.e. a near zero value implies a harmonic signal while a large value - a few hundreds - corresponds to a noise-like signal.

[0044] In the MPEG-1 Layer 2 psychoacoustic model 2, in order to achieve transparent coding, the minimum SMRs are computed for 32 subbands as follows. A block of 1056 input samples is taken from the input signal. The first 1024 samples are windowed using a Hanning window and transformed into the frequency domain using a 1024-point FFT. The tonality of each spectral line is determined by predicting its magnitude and phase from the two corresponding values in the previous transforms. The difference of each DFT coefficient and its predicted value is used to calculate the unpredictability measure. The unpredictability measure is converted to the "tonality" factor using an empirical factor with a larger value indicating a tonal signal. The required SNR for transparent coding is computed from the tonality using the following empirical formula

$$\mathbf{SNR}_j = t_j \mathbf{TMN}_j + (1 - t_j)\mathbf{NMT}_j ,$$

where t_j is the tonality factor, $\mathbf{TMN}_j$ and $\mathbf{NMT}_j$ are the value for tone-masking-noise and noise-masking-tone in subband j, respectively. $\mathbf{NMT}_j$ is set to 5.5 dB and $\mathbf{TMN}_j$ is given in a table provided in the MPEG audio standard. In order to take into account stereo unmasking effects $\mathbf{SNR}_j$ is determined to be larger than the minimum SNR $minval_j$ given in the standard. The SMR is calculated for each of the 32 subbands from the corresponding SNR. The above process is repeated for the next block of 1056 time samples - 480 old and 576 new samples - and another set of 32 SMR values is computed. The two sets of SMR values are compared and the larger value for each subband is taken as the required SMR.

[0045] Since the masking threshold due to a tonal and a noise-like signal is different, a tonality factor is calculated for each spectral line. The tonality factor is based on the unpredictability of the spectral components, meaning that higher unpredictability indicates a more noise-like signal. However, this measure does not distinguish between harmonic and inharmonic input signals as it is possible that they are equally predictable. In the second example of a method for encoding an audio signal, the MPEG-1 psychoacoustic model 2 has been modified considering imperfect harmonic structures of complex tonal sounds. It will become apparent to those skilled in the art that the method considering imperfect harmonic structures is not limited to the implementation in the MPEG-1 psychoacoustic model 2 but is also implementable into other psychoacoustic models. The example shown hereinbelow has been chosen because the MPEG-1 Layer 2 encoding is a widely used state of the art standard encoding process. The inharmonicity of an audio signal raises the masking threshold and, therefore, incorporating this effect into the encoding process of inharmonic input signals substantially reduces the bit rate.

[0046] In the MPEG-1 psychoacoustic model 2 the TMN parameter is given in a table. The values for the TMNs are based on psychoacoustic experiments in which a pure tone is used to mask a narrowband noise. In these experiments the masker is periodic, which is the case with an inharmonic masker. In fact, a noise probe is detected at a lower level when the masker is harmonic. This is likely caused by a disruption of the pitch sensation due to the periodic structure of the masker’s temporal envelope, as taught in W.C. Treumiet, and D.R. Boucher, "A masking level difference due to harmonicity", J. Acoust. Soc. Am., 109(1), pp. 306-320, 2001. In the second example of a method for encoding an audio signal, the TMN parameter is modified in dependence upon the input signal inharmonicity, as shown in the flow diagram of Fig. 7. Since in the MPEG-1 Layer 2 psychoacoustic model 2 a set of 32 SMRs is calculated for each 1152 time samples, the same time samples are analyzed for measuring the level of input signal inharmonicity. After determining the input signal inharmonicity, an inharmonicity index is calculated and subtracted from the TMN values. The inharmonicity index as a function of the periodic structure of the input signal is calculated as follows. The input block of 1632 time samples is decomposed using a gammatone filterbank - box 100. The envelope of each bandpass auditory filter output is detected using the Hilbert transform - box 102. The pitch of each envelope is calculated based on the autocorrelation of the envelope - box 104. Each pitch value is then compared with the other pitch values and an average error is determined and the variance of the average errors is calculated - box 106. According to W.C. Treumiet, and D.R. Boucher inharmonicity causes an increase of up to 10 dB in the masking threshold. Therefore, the inharmonicity index δ_{ih} as a function of the pitch variance V_p has been defined by the inventors to cover a range of 10 dB - box 108,

$$\delta_{ih} = 3\log_{10}\left(V_p + 1\right).$$

The above equation produces a zero value for a perfect harmonic signal and up to 10 dB for noise-like input signals. The new inharmonicity index is incorporated into the MPEG-1 psychoacoustic model 2 for calculating the masking threshold as

$$\mathbf{SNR}_j = \max\left\{\min val_j t_j\left(\mathbf{TMN}_j - \delta_{ih}\right) + \left(1 - t_j\right)\mathbf{NMT}_j\right\},$$

and the acoustic signal is encoded using the masking threshold determined above - box 110.

[0047] As shown above, the level of inharmonicity is defined as the variance of the periods of the envelopes of auditory filters outputs. The period of each envelope is found using the autocorrelation function. The location of the second peak of the autocorrelation function - ignoring the largest peak at the origin - determines the period. Since the autocorrelation function of a periodic signal has a plurality of peaks, the second largest peak sometimes does not correspond to the correct period. In order to overcome this problem in calculating the difference between two periods the smaller period is compared to a submultiple of the larger period if the difference becomes smaller. A MATLAB script for calculating the pitch variance is presented in Appendix B. Another problem occurs when there is no peak in the autocorrelation function. This situation implies an aperiodic envelope. In this case the period is set to an arbitrary or random value.

[0048] As shown in Appendix A, if at least two harmonics pass through an auditory filter the envelope of the output signal is periodic. Therefore, in order to correctly analyze an audio signal the lowest frequency of the gammatone filterbank is chosen such that the auditory filter centered at this frequency passes at least two harmonics. Therefore, the corresponding critical bandwidth centered at this frequency is chosen to be greater than twice the fundamental frequency of the input signal. The fundamental frequency is determined by analyzing the input signal either in the time domain or the frequency domain. However, in order to avoid extra computation for determining the fundamental frequency the median of the calculated pitch values is assumed to be the period of the input signal. The fundamental frequency of the input signal is then simply the inverse of the pitch value. Therefore, the lower bound for the analysis frequency range is set to twice the inverse of the pitch value.

[0049] In order to compare the subjective quality of the compressed audio materials informal listening tests have been conducted. Several audio files have been encoded and decoded using the standard MPEG-1 psychoacoustic model 2 and the modified version according to the invention. The bit allocation has been varied adaptively on a frame by frame basis. When the inharmonicity model was included the bit rate was reduced without adverse effects on the sound quality. The informal listening tests have shown that for multi-tonal audio-material the required bit rate decreases by approximately 10%.

[0050] As disclosed above a single value has been used to adjust the masking threshold for the entire frequency range of the input signal based on the complete frequency spectrum of the input signal. Alternatively, the masking threshold is modified based on the local harmonic structure of the input signal based on a local wideband frequency spectrum of the input signal.

[0051] Optionally, a combination of both non-linear masking effects indicated by the temporal masking index and the inharmonicity index are implemented into the MPEG-1 psychoacoustic model 2.

[0052] Of course, numerous other examples will be apparent to persons skilled in the art without departing from the scope of the invention as defined in the appended claims.

Appendix A

[0053] In the following it is shown that the envelope of the following signal is periodic with a period of either multiple or submultiple of P_0, i.e. the inverse of the fundamental frequency f_0.

$$y(t) = a_m \cos(m\omega_0 t + \phi_m) + a_n \cos(n\omega_0 t + \phi_n) \quad \text{(A1)}$$

Rewriting equation (A1) yiels

$$y(t) = a_m \cos(m\omega_0 t + \phi_m) + a_m \cos(n\omega_0 t + \phi_n) + (a_n - a_m)\cos(n\omega_0 t + \phi_n) \quad \text{(A2)}$$

$$y(t) = 2a_m \cos\left(\frac{(m-n)\omega_0 t + \phi_m - \phi_n}{2}\right) \times \cos\left(\frac{(m+n)\omega_0 t + \phi_m + \phi_n}{2}\right) + (a_n - a_m)\cos(n\omega_0 t + \phi_n) \quad \text{(A3)}$$

If ($m + n$) is much greater than ($m - n$), the first term in the above equation (A3) implies amplitude modulation. The lowpass signal is then expressed as

$$\xi(t) = a \cos\left(\frac{(m-n)\omega_0 t + \phi_m - \phi_n}{2}\right) \quad \text{(A4)}$$

The period of the envelope $\xi(t)$ is $\frac{2P_0}{(m-n)}$ which is a (sub)multiple of P_0. The second term in equation (A3) has no effect on the envelope due to being filtered out by the demodulator.

Appendix B

[0054] The pitch variance is calculated using the following MATLAB routine:

```
for i =1 : N
     s = 0;
     for j=1:N
        if (j ~= i)
             pmax = max ( P (i) , P (j) ) ;
             pmin = min ( P (i), P (j) ) ;
             a = round (pmax / pmin) ;
             s = s + abs ( pmin - pmax/a) ;
         end
      end
      d(i) = s/(N-1);
end
Vp = var (d)
```

[0055] In this routine, N is the number of auditory filters and P (.) is the pitch value.

Claims

1. A method for determining temporal masking thresholds comprising the steps of:

   receiving digital data indicative of samples of an analog audio signal (10);
   partitioning the digital data into overlapping blocks, each block comprising a predetermined number of samples (10);
   transforming the overlapping blocks into frequency domain using a filterbank, each transformed overlapping block comprising digital data indicative of a predetermined number of frequency subbands (10); and,
   for each frequency subband determining a temporal masking threshold calculated as an average energy in dependence upon the samples of the transformed overlapping blocks (12, 14).

2. A method for determining temporal masking thresholds as defined in claim 1 comprising providing the temporal masking threshold for each frequency subband for incorporation into a psychoacoustic model (32).

3. A method for determining temporal masking thresholds as defined in any one of claims 1 and 2, **characterized in that** the psychoacoustic model is the MPEG-1 psychoacoustic model 2 (32).

4. A method for determining temporal masking thresholds as defined in any one of claims 1 to 3 comprising combining for each frequency subband the temporal masking threshold with a simultaneous masking threshold (28).

5. A method for determining temporal masking thresholds as defined in any one of claims 1 to 4 comprising determining a forward temporal masker level for each current subband sample using past subband samples (16).

6. A method for determining temporal masking thresholds as defined in claim 5 **characterized in that** the forward temporal masker level is determined using a forward temporal masking function (16).

7. A method for determining temporal masking thresholds as defined in claim 5 or 6 comprising determining a backward temporal masker level for each current subband sample using future subband samples (18).

8. A method for determining temporal masking thresholds as defined in claim 7 **characterized in that** the backward temporal masker level is determined using a backward temporal masking function (18).

9. A method for determining temporal masking thresholds as defined in claim 8 comprising determining a total temporal masking level for each subband sample using the forward temporal masker level and the backward temporal masker level (20).

10. A method for determining temporal masking thresholds as defined in claim 9 comprising determining a signal-to-mask-ratio for each subband sample using the total temporal masking level of the corresponding subband sample (22).

11. A method for determining temporal masking thresholds as defined in claim 10 comprising determining a required signal-to-mask-ratio for each frequency subband, the required signal-to-mask-ratio being a maximum of the signal-to-mask-ratios of the subband samples of the corresponding frequency subband (24).

12. A method for determining temporal masking thresholds as defined in claim 11 comprising determining for each frequency subband an allowable noise level due to temporal masking using the required signal-to-mask-ratio for the corresponding frequency subband (26).

13. A method for determining temporal masking thresholds as defined in claim 12 comprising determining for each frequency subband a combined allowable noise level using the allowable noise level due to temporal masking and an allowable noise level due to simultaneous masking for the corresponding frequency subband (28).

14. A method for determining temporal masking thresholds as defined in claim 13 comprising determining for each frequency subband a combined signal-to-mask-ratio using the combined allowable noise level for the corresponding frequency subband (30).

Patentansprüche

1. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, welches die Schritte aufweist:

   • Empfangen von digitalen Daten, welche Abtastwerte eines analogen Tonsignals (10) beschreiben;
   • Aufteilen der digitalen Daten in sich überlappende Blöcke, wobei jeder Block eine vorbestimmte Anzahl von Abtastwerten (10) aufweist;
   • Transformation der sich überlappenden Blöcke in den Frequenzbereich mittels einer Filterreihe, wobei jeder transformierte überlappende Block digitale Daten entsprechend einer vorbestimmten Anzahl von Unterfrequenzbändern (Frequenz-Unterbändern) (10) aufweist; und
   • des Bestimmen, für jedes Frequenz-Unterband, eines als eine durchschnittliche Energie berechneten zeitlichen Maskierungs-Schwellenwertes in Abhängigkeit von den Abtastwerten der transformierten sich überlappenden Blöcke (12, 14).

2. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 3 definiert, welches das Vorsehen des Schwellwertes zur zeitlichen Maskierung für jedes Frequenz-Unterband zur Einfügung in ein psycho-akustisches Modell (32) umfasst.

3. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie in einem der Ansprüche 1 und 2 definiert, **dadurch gekennzeichnet, dass** das psycho-akustische Modell das MPEG-1 psycho-akustische Modell 2 (32) ist.

4. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie in einem der Ansprüche 1 bis 3 definiert, aufweisend das Kombinieren, für jedes Frequenz-Unterband, des Schwellwertes zur zeitlichen Maskierung mit einem Schwellwert zur simultanen Maskierung (28).

5. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie in einem der Ansprüche 1 bis 4 definiert, aufweisend das Bestimmen eines Niveaus zur vorwertsgerichteten zeitlichen Maskierung für jeden aktuellen Unterband-Abtastwerten unter Verwendung von vergangenen Unterband-Abtastwerten (16).

6. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 5 definiert, **dadurch gekennzeichnet, dass** das Niveau zur vorwäftsgerichteten zeitlichen Maskierung mittels einer vorwärtsgerichteten zeitlichen Maskierungsfunktion (16) bestimmt wird.

7. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 5 oder 6 definiert, aufweisend das Bestimmen eines Niveaus zur rückwärtsgerichteten zeitlichen Maskierung für jeden aktuellen Unterband-Abtastwert unter Verwendung von zukünftigen Unterband-Abtastwerten (18).

8. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 7 definiert, **dadurch gekennzeichnet, dass** das Niveau zur rückwärtsgerichteten zeitlichen Maskierung, mittels einer rückwärtsgerichteten zeitlichen Maskierungsfunktion (18) bestimmt wird.

9. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 8 defmiert, aufweisend das Bestimmen eines Niveaus zur gesamthaften zeitlichen Maskierung für jeden Unterband-Abtastwert unter Verwendung des Niveaus zur vorwärtsgerichteten zeitlichen Maskierung und des Niveaus zur rückwärtsgerichteten zeitlichen Maskierung (20).

10. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 9 definiert, aufweisend das Bestimmen eines Signal-Maskierungsverhältnisses für jeden Unterband-Abtastwert unter Verwendung des Niveaus zur gesamthaften zeitlichen Maskierung des entsprechenden Unterband-Abtastwerts (22).

11. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 10 definiert, aufweisend das Bestimmen eines erforderlichen Signal-Maskierungs-Verhältnisses für jedes Unterfrequenzband, wobei das erforderliche Signal-Maskierungsverhältnis ein Maximum der Signal-Maskierungs-Verhältnisse der Unterband-Abtastwerte des entsprechenden Unterfrequenzbandes (24) ist.

12. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 11 definiert, aufweisend das Bestimmen, für jedes Unterfrequenzband, eines zulässigen Rauschniveaus als Folge der zeitlichen

Maskierung, unter Verwendung des erforderlichen Signal-Maskierungsverhältnisses für das entsprechende Unterfrequenzband (26).

13. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 12 definiert, aufweisend das Bestimmen, für jedes Unterfrequenzband, eines kombinierten zulässigen Rauschniveaus als Folge der zeitlichen Maskierung und eines zulässiges Rauschniveau in Folge der simultanen Maskierung für das entsprechende Unterfrequenzband (28).

14. Ein Verfahren für das Bestimmen von Schwellwerten zur zeitlichen Maskierung, wie im Anspruch 13 definiert, aufweisend das Bestimmen, für jedes Unterfrequenzband, eines kombinierten Signal-Maskierungsverhältnisses, unter Verwendung des kombinierten zulässige Rauschniveaus für das entsprechende Unterfrequenzband (30).

Revendications

1. Procédé de détermination de seuils de masquage temporel, lequel procédé comprend les étapes qui consistent à :

recevoir des données numériques indicatives d'échantillons d'un signal audio analogique (10),
diviser les données numériques en blocs superposés, chaque bloc comprenant un nombre prédéterminé d'échantillons (10),
transformer les blocs superposés en le domaine de fréquences en utilisant une batterie de filtres, chaque bloc superposé et transformé contenant des données numériques indicatives d'un nombre prédéterminé des sous-bandes de fréquences (10) et
pour chaque sous-bande de fréquences, déterminer un seuil de masquage temporel calculé sous la forme d'une énergie moyenne dépendante des échantillons de blocs superposés transformés (12, 14).

2. Procédé de détermination de seuils de masquage temporel selon la revendication 1, qui comprend l'étape qui consiste à délivrer le seuil de masquage temporel de chaque sous-bande de fréquences pour l'incorporer dans un modèle psychoacoustique (32).

3. Procédé de détermination de seuils de masquage temporel selon l'une quelconque des revendications 1 et 2, **caractérisé en ce que** le modèle psychoacoustique est le modèle psychoacoustique 2 MPEG-1 (32).

4. Procédé de détermination de seuils de masquage temporel selon l'une quelconque des revendications 1 à 3, qui comprend l'étape qui consiste à combiner pour chaque sous-bande de fréquences le seuil de masquage temporel à un seuil de masquage simultané (28).

5. Procédé de détermination de seuils de masquage temporel selon l'une quelconque des revendications 1 à 4, qui comprend l'étape qui consiste à déterminer un niveau de masquage temporel avant pour chaque échantillon de sous-bande actuel en utilisant des échantillons passés de sous-bande (16).

6. Procédé de détermination de seuils de masquage temporel selon la revendication 5, **caractérisé en ce que** le niveau de masquage temporel avant est déterminé en utilisant une fonction (16) de masquage temporel avant.

7. Procédé de détermination de seuils de masquage temporel selon les revendications 5 ou 6, qui comprend l'étape qui consiste à déterminer un niveau de masquage temporel arrière pour chaque échantillon de sous-bande actuel en utilisant des échantillons futurs de sous-bande (18) .

8. Procédé de détermination de seuils de masquage temporel selon la revendication 7, **caractérisé en ce que** le niveau de masquage temporel arrière est déterminé en utilisant une fonction de masquage temporel arrière (18).

9. Procédé de détermination de seuils de masquage temporel selon la revendication 8, qui comprend l'étape qui consiste à déterminer un niveau de masquage temporel total pour chaque échantillon de sous-bande en utilisant le niveau de masquage temporel avant et le niveau de masquage temporel arrière (20).

10. Procédé de détermination de seuils de masquage temporel selon la revendication 9, qui comprend l'étape qui consiste à déterminer un rapport signal-masque pour chaque échantillon de sous-bande en utilisant le niveau de masquage temporel total de l'échantillon de sous-bande (22) correspondant.

11. Procédé de détermination de seuils de masquage temporel selon la revendication 10, qui comprend l'étape qui consiste à déterminer un rapport signal-masque requis pour chaque sous-bande de fréquences, le rapport signal-masque requis étant le plus grand des rapports signal-masque des échantillons de sous-bande de la sous-bande de fréquence correspondante (24).

12. Procédé de détermination de seuils de masquage temporel selon la revendication 11, qui comprend l'étape qui consiste à déterminer pour chaque sous-bande de fréquences le niveau admissible de bruit provoqué par le masquage temporel en utilisant le rapport signal-masque requis sur la sous-bande de fréquences (26) correspondante.

13. Procédé de détermination de seuils de masquage temporel selon la revendication 12, qui comprend l'étape qui consiste à déterminer pour chaque sous-bande de fréquences un niveau de bruit combiné admissible en utilisant le niveau de bruit admissible dû au masquage temporel et le niveau de bruit admissible dû au masquage simultané de la sous-bande de fréquences (28) correspondante.

14. Procédé de détermination de seuils de masquage temporel selon la revendication 13, qui comprend la détermination pour chaque sous-bande de fréquences un rapport signal-masque combiné en utilisant le niveau de bruit combiné admissible sur la sous-bande de fréquences (30) correspondante.

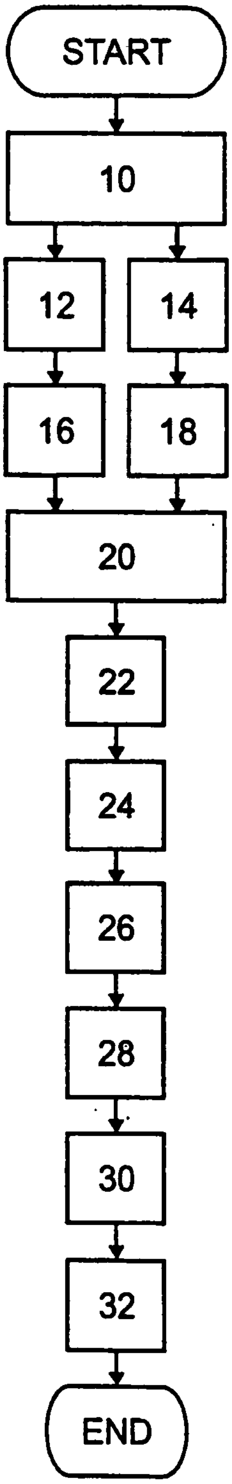

Fig. 1

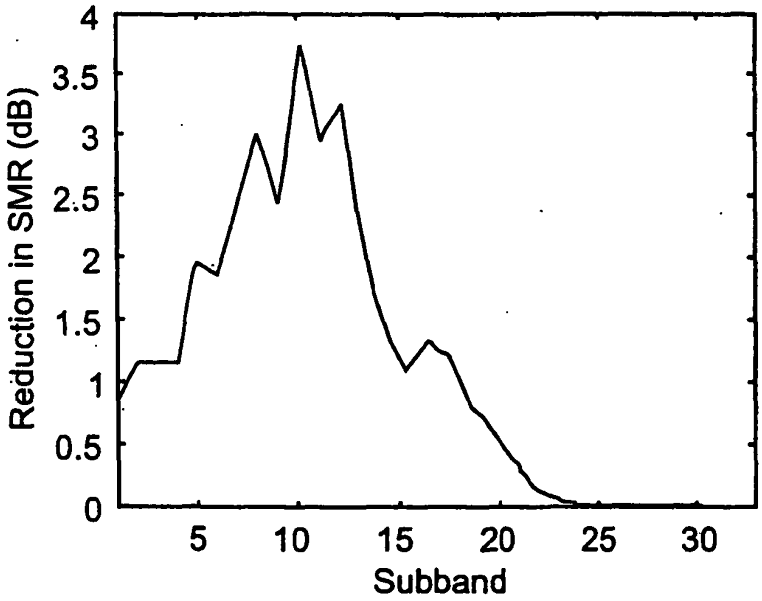

Fig. 2

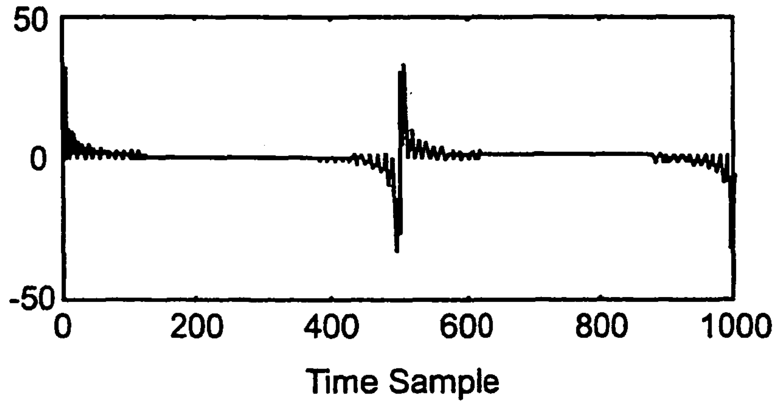

Fig. 3a

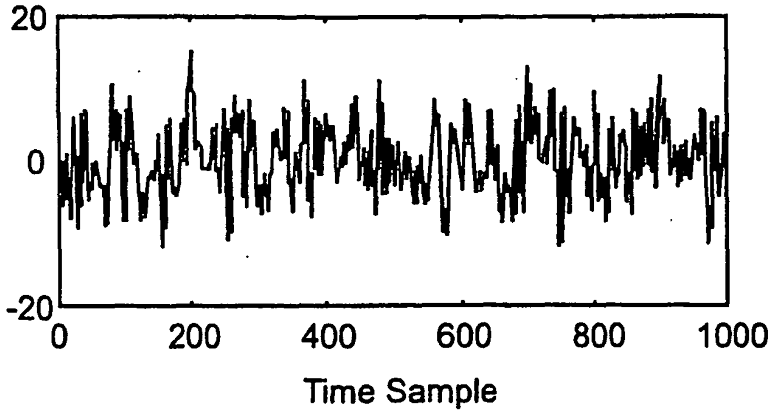

Fig. 3b

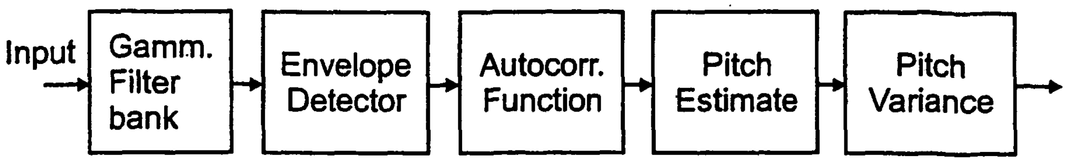
Input
Gamm. Filter bank
Envelope Detector
Autocorr. Function
Pitch Estimate
Pitch Variance

Fig. 4

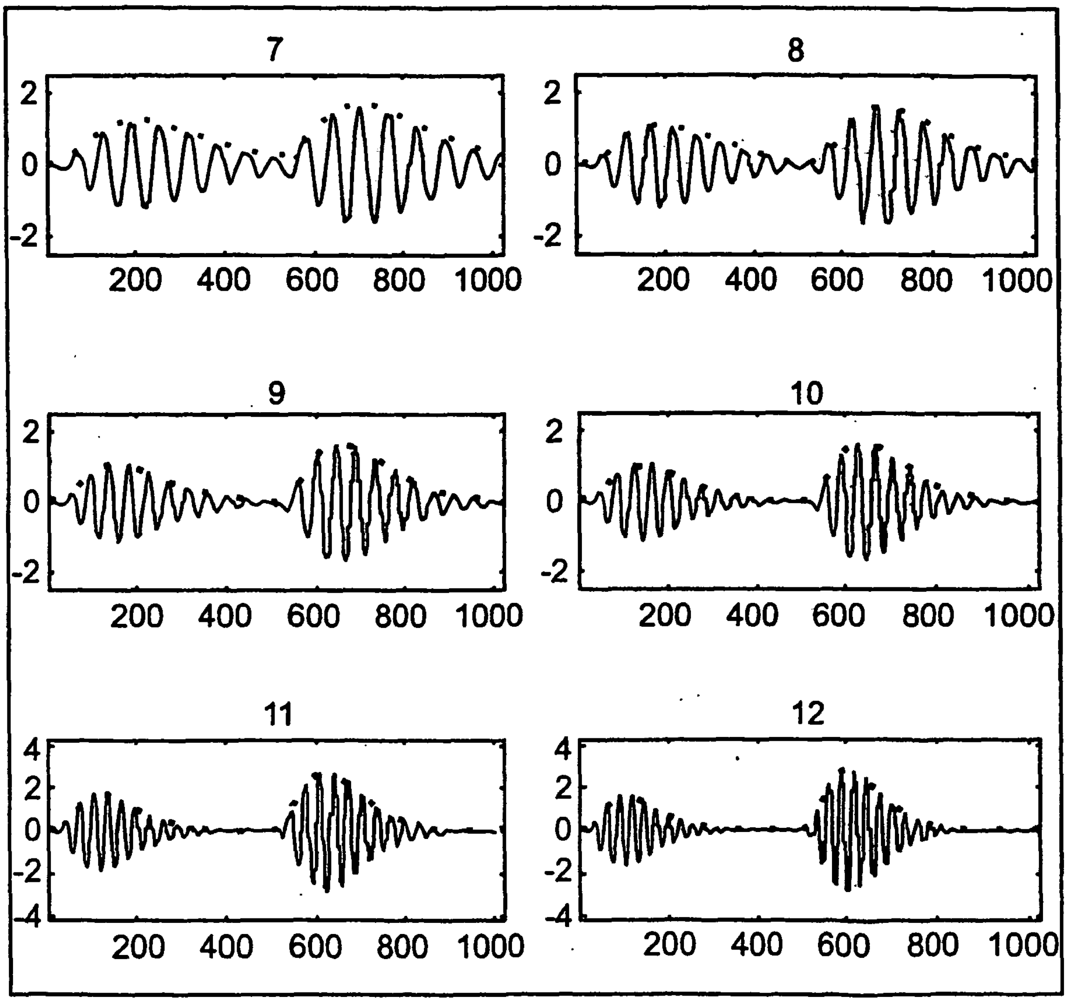
7
8
9
10
11
12
200 400 600 800 1000

Fig. 5a

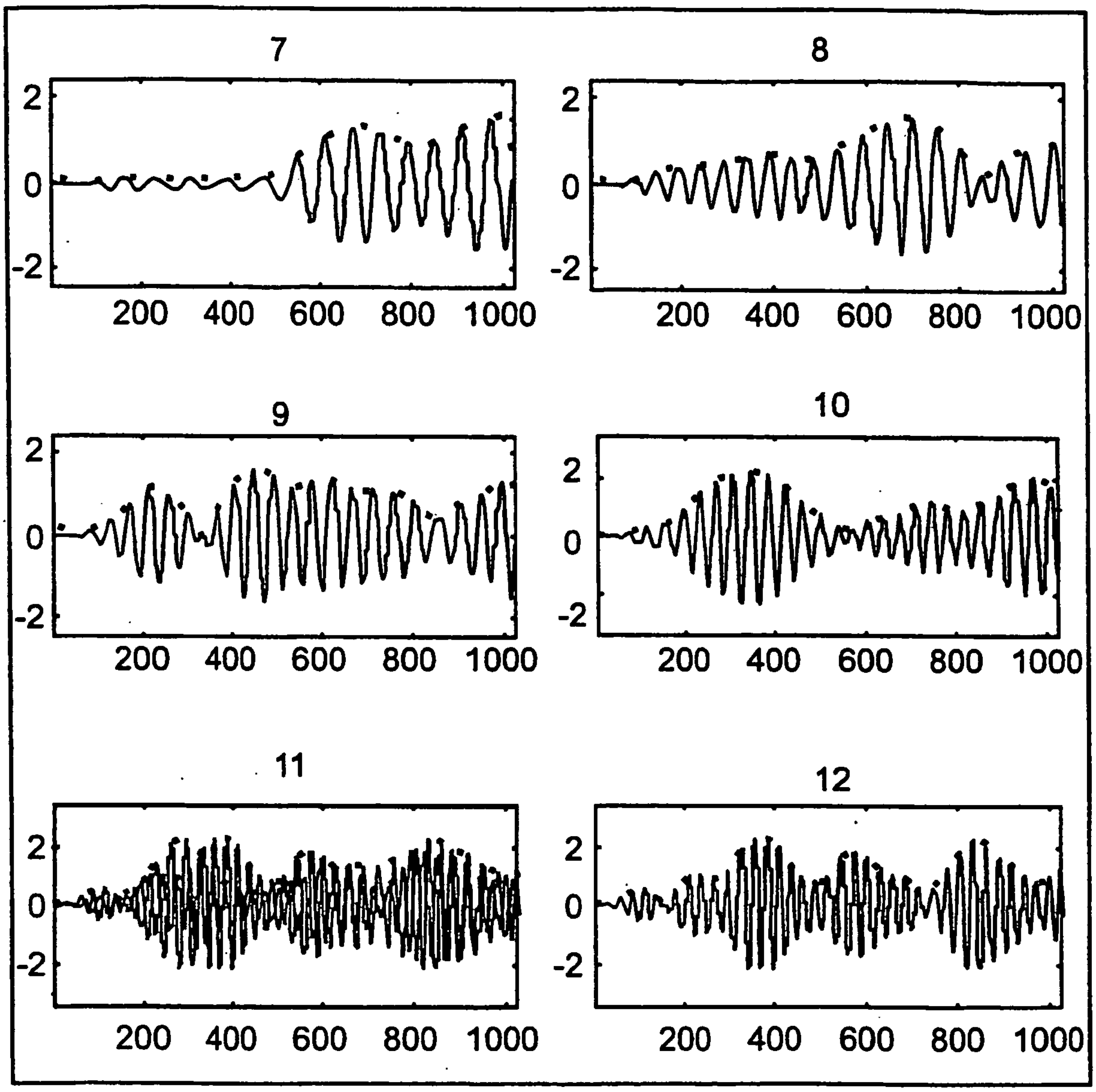

Fig. 5b

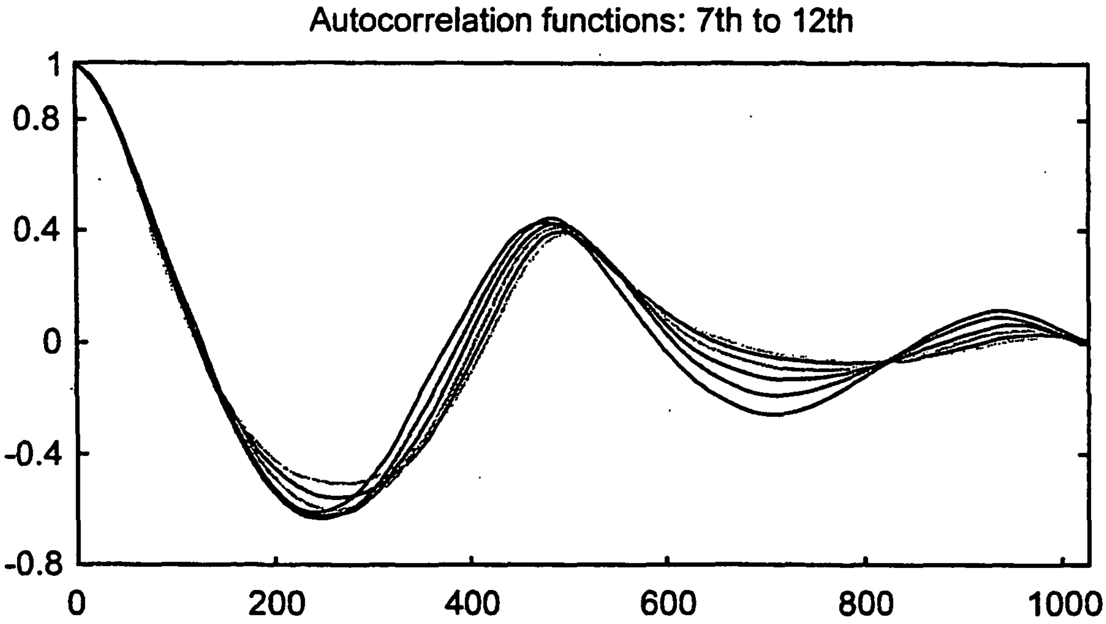
Autocorrelation functions: 7th to 12th
1
0.8
0.4
0
-0.4
-0.8
0
200
400
600
800
1000

Fig. 6a

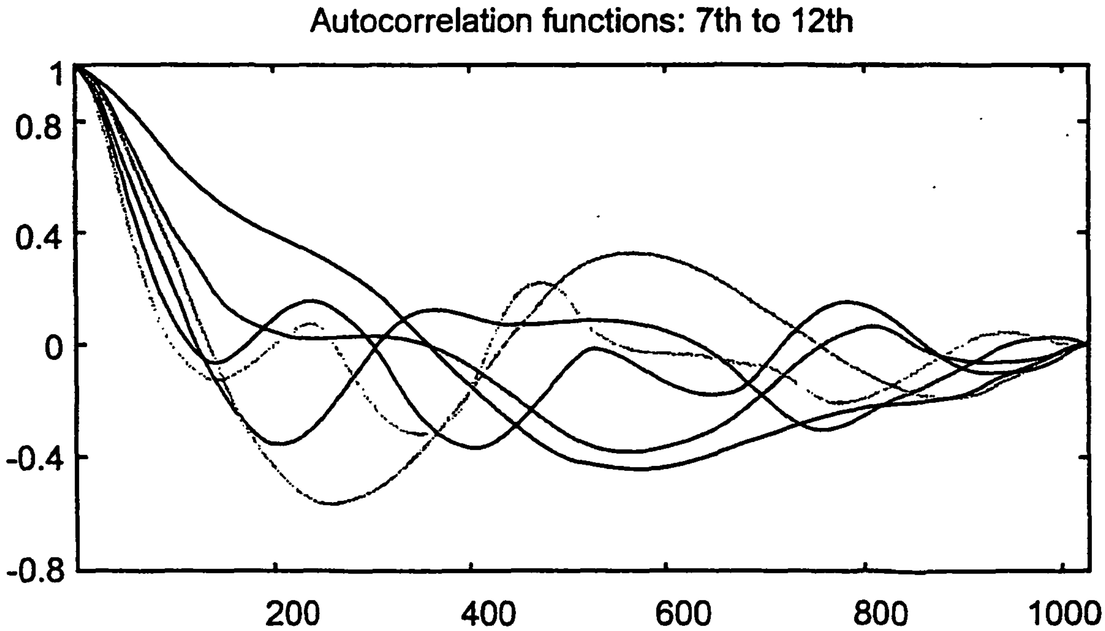
Autocorrelation functions: 7th to 12th
1
0.8
0.4
0
-0.4
-0.8
200
400
600
800
1000

Fig. 6b

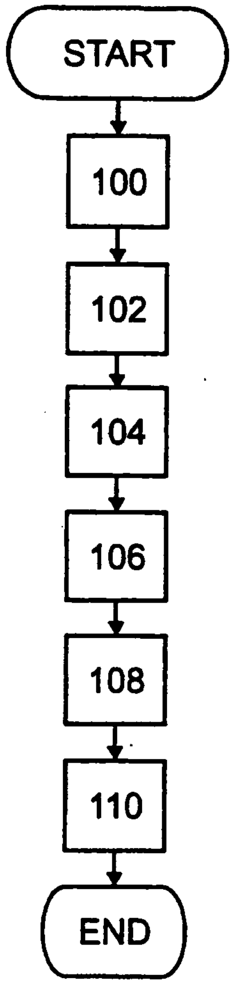

Fig. 7

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **B. MOORE.** An Introduction to the Psychology of Hearing. Academic Press, 1997 **[0014] [0015]**
- **E. ZWICKER ; T. ZWICKER.** Audio Engineering and Psychoacoustics, Matching Signals to the Final Receiver, the Human Auditory System. *J. Audio Eng. Soc.,* March 1991, vol. 39 (3), 115-126 **[0014]**
- **E. ZWICKER ; H. FASTL.** Psychoacoustics Facts and Models. Springer Verlag, 1990 **[0014]**
- **A. HARMA.** Psychoacoustic Temporal Masking Effects with Artificial and Real Signals. *Hearing Seminar,* 1999, 665-668 **[0015]**
- **W. JESTEADT ; S. BACON ; J. LEHMAN.** Forward masking as a function of frequency, masker level, and signal delay. *J. Acoust. Soc. Am.,* April 1982, vol. 71 (4), 950-962 **[0019]**
- **W.C. TREUMIET ; D.R. BOUCHER.** A masking level difference due to harmonicity. *J. Acoust. Soc. Am.,* 2001, vol. 109 (1), 306-320 **[0038] [0046]**
- **E. ZWICKER ; E. TERHARDT.** Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.,* 1980, vol. 68 (5), 1523-1525 **[0043]**