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

(11) **EP 2 139 000 A1**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
30.12.2009 Bulletin 2009/53

(21) Application number: **08159018.4**

(22) Date of filing: **25.06.2008**

(51) Int Cl.:
G10L 19/14 *(2006.01)* *G10L 19/02* *(2006.01)*
G10L 19/04 *(2006.01)* *G10L 11/02* *(2006.01)*

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MT NL NO PL PT RO SE SI SK TR
Designated Extension States:
AL BA MK RS

(71) Applicant: **Deutsche Thomson OHG**
30625 Hannover (DE)

(72) Inventors:
- **Wuebbolt, Oliver**
  30161 Hannover (DE)
- **Boehm, Johannes**
  37081 Göttingen (DE)

(74) Representative: **Hartnack, Wolfgang**
Deutsche Thomson OHG
European Patent Operations
Karl-Wiechert-Allee 74
30625 Hannover (DE)

(54) **Method and apparatus for encoding or decoding a speech and/or non-speech audio input signal**

(57) A disadvantage of known audio or speech codecs is a clear dependency of the coding quality on the types of content, i.e. music-like audio signals are best coded by audio codecs and speech-like audio signals are best coded by speech codecs. No known codec is holding a dominant position for mixed speech/music content. The inventive joined speech/audio codec uses speech coding processing as well as audio coding processing. Transform based audio coding processing is combined in an advantageous way with linear prediction based speech coding processing, using at the input a Modulated Lapped Transform the output spectrum of which is separated into frequency bins (low frequencies) assigned to the speech coding and the remaining frequency bins (high frequencies) are assigned to the transform-based audio coding. The invention achieves a uniform good codec quality for both speech-like and music-like audio signals, especially for very low bit rates but also for higher bit rates.

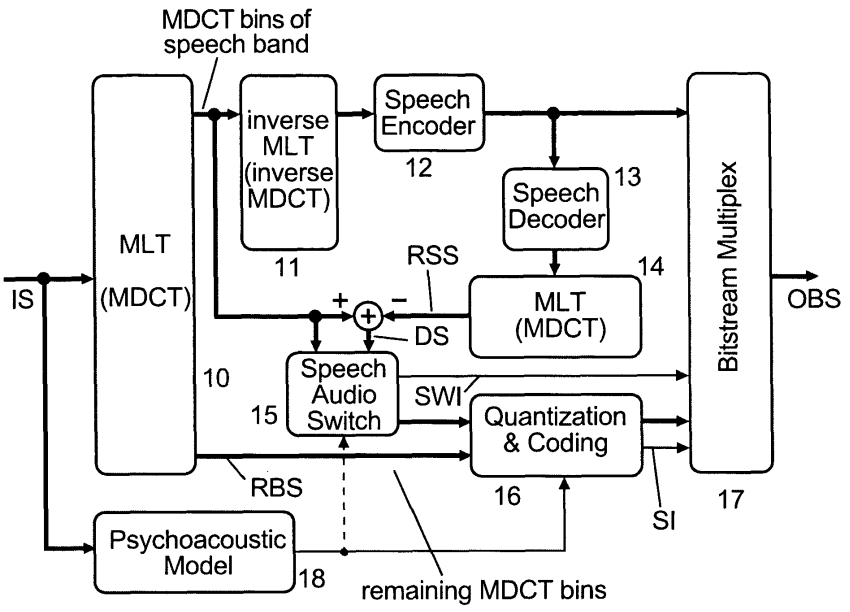

Fig. 1

EP 2 139 000 A1

Description

[0001] The invention relates to a method and to an apparatus for encoding or decoding a speech and/or non-speech audio input signal.

Background

[0002] Several wideband or speech/audio codecs are known, for example:

S. Ragot et al., "ITU-T G.729.1: An 8-32 Kbit/s scalable coder interoperable with G.729 for wideband telephony and voice over IP", IEEE International Conference on Acoustics, Speech and Signal Processing 2007, ICASSP 2007, vol.4, pp.IV-529 to IV-532.
This wideband speech coder includes an embedded G.729 speech coder, which is used permanently. Therefore the quality for music-like signals (non-speech) is not very good. Although this coder uses transform coding techniques it is a speech coder.

[0003] S.A. Ramprashad, "A two stage hybrid embedded speech/audio coding structure", Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing 1998, ICASSP 1998, vol.1, pp.337-340.
This coder uses a principle structure similar to that of the above-mentioned coder. The processing is based on time domain signals, which implies a difficult handling of the delay in the core encoder/decoder (speech coder). Therefore the processing is based on a common transform in order to reduce this problem. Again, the core coder (i.e. the speech coder) is used permanently, which results in a non-optimal quality for music like (non-speech) signals.

[0004] M. Purat, P. Noll, "A new orthonormal wavelet packet decomposition for audio coding using frequency-varying modulated lapped transforms", IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, 1995, pp.183-186.

[0005] M. Purat, P. Noll, "Audio coding with a dynamic wavelet packet decomposition based on frequency-varying modulated lapped transforms", IEEE International Conference on Acoustics, Speech, and Signal Processing 1996, ICASSP 1996, vol.2, pp.1021-1024.

Invention

[0006] A disadvantage of the known audio/speech codecs is a clear dependency of the coding quality on the types of content, i.e. music-like audio signals are best coded by audio codecs and speech-like audio signals are best coded by speech codecs. No known codec is holding a dominant position for mixed speech/music content.

[0007] A problem to be solved by the invention is to provide a good codec performance for both, speech and music, and to further improve the codec performance for such mixed signals. This problem is solved by the methods disclosed in claims 1 and 3. Apparatuses that utilise these methods are disclosed in claims 2 and 4.

[0008] The inventive joined speech/audio codec uses speech coding techniques as well as audio transform coding techniques. Known transform-based audio coding processing is combined in an advantageous way with linear prediction-based speech coding processing using one or more Modulated Lapped Transform (MLT) at the codec input and one or more inverse Modulated Lapped Transform (IMLT) at the codec output. The MLT output spectrum is separated into frequency bins (low frequencies) assigned to the speech coding section of the codec, and the remaining frequency bins (high frequencies) assigned to the transform-based coding section of the codec, wherein the transform length at the codec input and output can be switched signal adaptively.
As an alternative, in the transform-based coding/decoding sections the transform length can be switched input signal adaptively.

[0009] The invention achieves a uniform good codec quality for both speech-like and music-like audio signals, especially for very low bit rates but also for higher bit rates.

[0010] In principle, the inventive method is suited for encoding a speech and/or non-speech audio input signal, including the steps:

- transforming successive and possibly overlapping sections of said input signal by at least one initial MLT transform and splitting the resulting output frequency bins into a low band signal and a remaining band signal;
- passing said low band signal to a speech/audio switching and through a speech coding/decoding loop including at least one short first-type MLT transform, a speech encoding, a corresponding speech decoding, and at least one short second-type MLT transform having a type opposite than that of said first-type short MLT transform;
- quantising and encoding said remaining band signal, controlled by a psycho-acoustic model that receives as its input said audio input signal;

- combining the output signal of said quantising and encoding, a switching information signal of said switching, possibly the output signal of said speech encoding, and optionally other encoding side information, in order to form for said current section of said input signal an output bit stream,

wherein said speech/audio switching receives said low band signal and a second input signal derived from the output of said short second-type MLT transform and decides, whether said second input signal bypasses said quantising and encoding step or said low band signal is coded together with said remaining band signal in said quantising and encoding step, and wherein in the latter case said output signal of said speech encoding is not included in the current section of said output bit stream.

[0011] In principle the inventive apparatus is suited for encoding a speech and/or non-speech audio input signal, said apparatus including means being adapted for:

- transforming successive and possibly overlapping sections of said input signal by at least one initial MLT transform and splitting the resulting output frequency bins into a low band signal and a remaining band signal;
- passing said low band signal to a speech/audio switching and through a speech coding/decoding loop including at least one short first-type MLT transform, a speech encoding, a corresponding speech decoding, and at least one short second-type MLT transform having a type opposite than that of said first-type short MLT transform;
- quantising and encoding said remaining band signal, controlled by a psycho-acoustic model that receives as its input said audio input signal;
- combining the output signal of said quantising and encoding, a switching information signal of said switching, possibly the output signal of said speech encoding, and optionally other encoding side information, in order to form for said current section of said input signal an output bit stream,

wherein said speech/audio switching receives said low band signal and a second input signal derived from the output of said short second-type MLT transform and decides, whether said second input signal bypasses said quantising and encoding step or said low band signal is coded together with said remaining band signal in said quantising and encoding step, and wherein in the latter case said output signal of said speech encoding is not included in the current section of said output bit stream.

[0012] In principle, the inventive method is suited for decoding a bit stream representing an encoded speech and/or non-speech audio input signal that was encoded according to the above method, said decoding method including the steps:

- demultiplexing successive sections of said bitstream to regain the output signal of said quantising and encoding, said switching information signal, possibly the output signal of said speech encoding, and said encoding side information if present;
- if present in a current section of said bitstream, passing said output signal of said speech encoding through a speech decoding and said short second-type MLT transform;
- decoding said output signal of said quantising and encoding, controlled by said encoding side information if present, in order to provide for said current section a reconstructed remaining band signal and a reconstructed low band signal;
- providing a speech/audio switching with said reconstructed low band signal and a second input signal derived from the output of said second-type MLT transform, and passing according to said switching information signal either said reconstructed low band signal or said second input signal;
- inversely MLT transforming the output signal of said switching combined with said reconstructed remaining band signal, and possibly overlapping successive sections, in order to form a current section of the reconstructed output signal.

[0013] In principle the inventive apparatus is suited for decoding a bit stream representing an encoded speech and/or non-speech audio input signal that was encoded according to the above encoding method, said apparatus including means being adapted for:

- demultiplexing successive sections of said bitstream to regain the output signal of said quantising and encoding, said switching information signal, possibly the output signal of said speech encoding, and said encoding side information if present;
- if present in a current section of said bitstream, passing said output signal of said speech encoding through a speech decoding and said short second-type MLT transform;
- decoding said output signal of said quantising and encoding, controlled by said encoding side information if present, in order to provide for said current section a reconstructed remaining band signal and a reconstructed low band signal;
- providing a speech/audio switching with said reconstructed low band signal and a second input signal derived from

the output of said second-type MLT transform, and passing according to said switching information signal either said reconstructed low band signal or said second input signal;

- inversely MLT transforming the output signal of said switching combined with said reconstructed remaining band signal, and possibly overlapping successive sections, in order to form a current section of the reconstructed output signal.

[0014] Advantageous additional embodiments of the invention are disclosed in the respective dependent claims.

Drawings

[0015] Exemplary embodiments of the invention are described with reference to the accompanying drawings, which show in:

Fig. 1 Block diagram of the inventive joint speech and audio coder;
Fig. 2 Higher time resolution processing in the 'quantisation&coding' step/stage (short block coding);
Fig. 3 Block diagram of the inventive joint speech and audio decoder;
Fig. 4 Higher time resolution processing in the 'decoding' step/stage (short block decoding);
Fig. 5 Block diagram of an other embodiment of the inventive joint speech and audio coder;
Fig. 6 Higher time resolution processing in the 'quantisation&coding' step/stage (short block coding) of the other embodiment;
Fig. 7 Block diagram of the inventive joint speech and audio decoder of the other embodiment;
Fig. 8 Higher time resolution processing in the 'decoding' step/stage (short block decoding) of the other embodiment;
Fig. 9 Block diagram of a further embodiment of the inventive joint speech and audio coder (short block coding) .

Exemplary embodiments

[0016] In the inventive joint speech and audio codec according to Fig. 1, known coding processing for speech-like signals (linear prediction based speech coding processing, e.g. CELP, ACELP, cf. ISO/IEC 14496-3, Subparts 2 and 3, and MPEG4-CELP) is combined with state-of-the-art coding processing for general audio or music-like signals based on a time-frequency transform, e.g. MDCT. The PCM audio input signal IS is transformed by a Modulated Lapped Transform MLT having a pre-determined length in step/stage 10. As a special processing of an MLT, e.g. a Modified Discrete Cosine Transform MDCT, is appropriate for audio coding applications. The MDCT was first called by Princen and Bradley "Oddly-stacked Time Domain Alias Cancellation Transform" and was published in John P. Princen and Alan B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation", IEEE Transactions on Acoustics Speech Sigal Processing ASSP-34 (5), pp.1153-1161, 1986. H.S.Malvar, "Signal processing with lapped transform", Artech House Inc., Norwood, 1992, and M.Temerinac, B.Edler, "A unified approach to lapped or-thogonal transforms", IEEE Transactions on Image Processing, Vol.1, No.1, pp.111-116, Januar 1992, called it Modulated Lapped Transform (MLT) and showed its relations to Lapped orthogonal Transforms in general and also proved it to be a special case of a QMF Filter bank. The Modified Discrete Cosine Transformation (MDCT) and the inverse MDCT (iMDCT) can be regarded as a critically sampled filter-bank with perfect reconstruction properties. The MDCT is calculated by:

$$X(k) = \sqrt{\frac{4}{N}} \sum_{n=0}^{N-1} h(n) \cdot x(n) \cdot \cos\left[\frac{\pi}{K} \cdot (n + \frac{K+1}{2}) \cdot (k + \frac{1}{2})\right] \quad , k = 0,1..,K-1; K = N/2$$

$$x(n) = \sqrt{\frac{4}{N}} \sum_{k=0}^{K-1} h(n) \cdot X(k) \cdot \cos\left[\frac{\pi}{K} \cdot (n + \frac{K+1}{2}) \cdot (k + \frac{1}{2})\right] \quad , n = 0,1..,N-1$$

[0017] At the MLT output the obtained spectrum is separated into frequency bins belonging to the speech band (representing a low band signal) and the remaining bins (high frequencies) representing a remaining band signal RBS. In step/stage 11 the speech band bins are transformed back into time domain using the inverse MLT, e.g. an inverse MDCT, with a short transform length with respect to the pre-determined length used in step/stage 10. The resulting time signal has a lower sampling frequency than the input time signal and contains only the corresponding frequencies of the speech band bins. The theory behind using only a subset of the MLT bins in an inverse MLT is described in the above-cited 1995 and 1996 Purat articles.

[0018] The generated time domain signal is then used as input signal for a speech encoding step/stage 12. The output of the speech encoding can be transmitted in the output bit stream OBS, depending on a decision made by a below-described speech/audio switch 15. The encoded 'speech' signal is decoded in a related speech decoding step/stage 13, and the decoded 'speech' signal is transformed back into frequency domain in step/stage 14 using the MLT corresponding to the inverse MLT of step/stage 11 (i.e. an 'opposite type' MLT having the short length) in order to re-generate the speech band signal, i.e. a reconstructed speech signal RSS. The difference signal DS between these frequency bins and the original low frequency bins, as well as the original low frequency bins signal, serve as input to the speech/audio switch 15. In that switch it is decided, whether the original low frequency bins are coded together with the remaining high frequency bins (this indicates that the coded 'speech' signal is not transmitted in bit stream OBS), or the difference signal DS is coded together with the remaining high frequency bins in a following quantisation&coding step/stage 16 (this indicates that the coded 'speech' signal is transmitted in bit stream OBS). That switch may be operated by using a rate-distortion optimisation. An information item SWI about the decision of switch 15 is included in bit stream OBS for use in the decoding. In this switch, but also in the other steps/stages, the different delays introduced by the cascaded transforms are to be taken into account. The different delays can be balanced using corresponding buffering for these steps/stages.

It is possible to use a mixture of original frequency bins and difference signal frequency bins in the low frequency band as input to step/stage 16. In such case, information about how that mixture is composed is conveyed to the decoding side. In any case, the remaining frequency bins output by step/stage 10 (i.e. the high frequencies) are processed in quantisation&coding step/stage 16.

In step/stage 16 an appropriate quantisation is used (e.g. like the quantisation techniques used in AAC), and subsequently the quantised frequency bins are coded using e.g. Huffman coding or arithmetic coding.

[0019] In case the speech/audio switch 15 decides that a music-like input signal is present and therefore the speech coder/decoder or its output is not used at all, the original frequency bins corresponding to the speech band are to be encoded (together with the remaining frequency bins) in the quantisation&coding step/stage 16.

The quantisation&coding step/stage 16 is controlled by a psycho-acoustic model calculation 18 that exploits masking properties of the input signal IS for the quantisation. Therefore side information SI can be transmitted in the bit stream multiplex to the decoder.

Switch 15 can also receive suitable control information (e.g. degree of tonality or spectral flatness, or how noise-like the signal is) from psycho-acoustic model step/stage 18.

A bit stream multiplexer step/stage 17 combines the output code (if present) of the speech encoder 12, the switch information of switch 15, the output code of the quantisation&coding step/stage 16, and optionally side information code SI, and provides the output bit stream OBS.

[0020] As shown in Fig. 2, to achieve a higher time resolution in the transform-based coding, at the input of the quantisation&coding step/stage 16 several small inverse MLT (matching the type of MLT 10) can be used (e.g. inverse MDCT, iMDCT) for transforming 22 the long output spectrum of the initial MLT 10 having high frequency resolution into several shorter spectra with lower frequency resolution but higher time resolution. The inverse MLT steps/stages 22 are arranged between a first grouping step/stage 21 and a second grouping step/stage 23 and provide a doubled number of output values. Again the theory behind this processing is described in the above-cited 1995 and 1996 Purat articles. In the first grouping 21 several neighbouring MLT bins are combined and used as input for the inverse MLTs 22. The number of combined MLT bins, which means the transform length of the inverse MLT, defines the resulting time and frequency resolution, wherein a longer inverse MLTs delivers a higher time resolution. In the following grouping 23, overlap/add is performed (optionally involving application of window functions) and the output of the inverse MLTs applied on the same input spectrum is sorted such that it results in several (the quantity depends on the size of the inverse MLTs) temporally successive 'short block' spectra which are quantised and coded in step/stage 16.

The information about this 'short block coding' mode being used is included in the side information SI. Optionally, multiple 'short block coding' modes with different inverse MLT transform lengths can be used and signalled in SI. Thereby a non-uniform time-frequency resolution over the short block spectra is facilitated, e.g. a higher time resolution for high frequencies and a higher frequency resolution for low frequencies. For instance, for the lowest frequencies the inverse MLT can get a length of 2 successive frequency bins and for the highest frequencies the inverse MLT can get a length of 16 successive frequency bins. In case a non-uniform frequency resolution is chosen, it is not possible to group e.g. 8 short block spectra. A different order of coding the resulting frequency bins can be used, for example one 'spectrum' may contain not only different frequency bins at a time, but also the same frequency bins at different points in time may be included.

The input signal IS adaptive switching between the processing according to Fig. 1 and the processing according to Fig. 2 is controlled by psycho-acoustic model step/stage 18. For example, if from one frame to the following frame the signal energy in input signal IS rises above a threshold (i.e. there is a transient in the input signal), the processing according to Fig. 2 is carried out. In case the signal energy is below that threshold, the processing according to Fig. 1 is carried out. This switching information, too, is included in output bitstream OBS for a corresponding switching in the decoding.

The transform block sections can be weighted by a window function, in particular in an overlapping manner, wherein the length of a window function corresponds to the current transform length.
Analysis and synthesis windows can be identical, but need not. The functions of the analysis an synthesis windows h_A (n) and h_S(n) must fulfil some constraints for the overlapping regions of successive blocks *i* and *i*+1 in order to enable a perfect reconstruction:

$$h_A(i+1,n)\cdot h_S(i+1,n)+h_A(i,n+N/2)\cdot h_S(i,n+N/2)=1\,,$$

$$h_A(i+1,n)=h_S(i,N-1-n),\quad h_S(i+1,n)=h_A(i,N-1-n)\quad,\quad n = 0\ldots N/2-1$$

[0021] A known window function type is the sine window:

$$h_{\sin}(n)=\sin(\pi\cdot\frac{n+0.5}{N})\,,\quad n = 0\ldots N-1$$

[0022] A window with an improved far away rejection, but a broader main lobe is the OGG-window, which is very similar to the Kaiser-Bessel derived window:

$$h_{OGG}(n)=\sin(\sin(\frac{\pi}{N}\cdot(n+\frac{1}{2}))^2\cdot\frac{\pi}{2})\quad,\quad n = 0\ldots N-1$$

[0023] A further window function is disclosed in table 7.33. of the AC-3 audio coding standard.
In case of switching the transform length, transition window functions are used, e.g. as described in B.Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen", FREQUENZ, vol.43, pp. 252-256, 1989, or as used in mp3 and described in the MPEG1 standard ISO/IEC 11172-3 in particular section 2.4.3.4.10.3, or as in AAC (e.g. as described in the MPEG4 standard ISO/IEC 14496-3, Subpart 4).

[0024] In the inventive decoder in Fig. 3, the received or replayed bit stream OBS is demultiplexed in a corresponding step/stage 37, thereby providing code (if present) for the speech decoder 33, the switch information SWI for switch 35, the code and the switching information for the decoding step/stage 36, and optionally side information code SI. In case the speech subcoder 11,12,13,14 was used at encoding side for a current data frame, in that current frame the corresponding encoded speech band frequency bins are correspondingly reconstructed by the speech decoding step/stage 33 and the downstream MLT step/stage 34, thereby providing the reconstructed speech signal RSS. The remaining encoded frequency bins are correspondingly decoded in decoding step/stage 36, whereby the encoder-side quantisation operation is reversed correspondingly. The speech/audio switch 35 operates corresponding to its operation at encoding side, controlled by switch information SWI. In case the switch signal SWI indicates that a music-like input signal is present in the current frame and therefore the speech coding/decoding was not used, the frequency bins corresponding to the low band are decoded together with the remaining frequency bins in the decoding step/stage 36, thereby providing the reconstructed remaining band signal RRBS and the reconstructed low band signal RLBS.
The output signal or signals of step/stage 36 and of switch 35 are correspondingly combined in inverse MLT (e.g. iMDCT) step/stage 30 and are synthesised in order to provide the decoded output signal OS. In switch 35 and in the other steps/stages, the different delays introduced by the cascaded transforms are to be taken into account. The different delays can be balanced using corresponding buffering for these steps/stages.
In case the corresponding option was used at encoding side, not the frequency bins of the combined signal CS, but the frequency bins of the reconstructed speech signal RSS are used for the corresponding processing in switch 35 and in step/stage 30, i.e. in step/stages 16 and 36, respectively, there is no coding/decoding at all of the low band spectrum.
In case at encoding side the 'short block mode' encoding was used to achieve a higher time resolution in the transform-based coding, the decoding in step/stage 36 of the 'short block mode' is illustrated in Fig. 4. According to the encoding process, several temporally successive 'short block' spectra are to be decoded in step/stage 36 and collected in a first grouping step/stage 43. Overlap/add is performed (optionally involving application of window functions). Thereafter each set of temporally successive spectral coefficients is transformed using the corresponding MLT steps/stages 42, and

provides a halved number of output values. The generated spectral coefficients are then grouped in a second grouping step/stage 41 to one MLT spectrum with the initial high frequency resolution and transform length. Optionally, multiple 'short block decoding' modes with different MLT transform lengths can be used as signalled in SI, whereby a non-uniform time-frequency resolution over the short block spectra is facilitated, e.g. a higher time resolution for high frequencies and a higher frequency resolution for low frequencies.

[0025] As an alternative embodiment, a different cascading of the MLTs can be used wherein the order of the inner MLT/inverse MLT pair in the speech encoder is switched. In Fig. 5 a block diagram of a corresponding encoding is depicted, wherein Fig. 1 reference signs mean the same operation as in Fig. 1.

The inverse MLT 11 is replaced by an MLT step/stage 51, and the MLT 14 is replaced by an inverse MLT step/stage 54 (i.e. an 'opposite type' MLT). Due to the exchanged order of these MLTs the speech encoder input signal has different properties compared to those in Fig. 1. Therefore the speech coder 52 and the speech decoder 53 are adapted to these different properties (e.g. such that aliasing components are cancelled out).

[0026] Like in Fig. 2 for the Fig. 1 embodiment, in decoding step/stage 36 for the Fig. 5 embodiment a 'short block mode' processing can be used as shown in Fig. 6, wherein MLT steps/stages 62 corresponding to that in Fig. 4 replace the inverse MLT steps/stages 22 in Fig. 2.

[0027] In the alternative embodiment decoder shown in Fig. 7, the speech decoding step/stage 33 in Fig. 3 is replaced by a correspondingly adapted speech decoding step/stage 73 and the MLT step/stage 34 in Fig. 3 is replaced by a corresponding inverse MLT step/stage 74.

[0028] Like in Fig. 4 for the Fig. 3 embodiment, for the Fig. 7 embodiment a 'short block mode' processing can be used as shown in Fig. 8, wherein corresponding inverse MLT steps/stages 82 corresponding to that in Fig. 1 replace the MLT steps/stages 42 in Fig. 4.

[0029] Instead of achieving a higher time resolution by the processing described in connection with Fig. 2 and Fig. 6 (block switching in the quantisation&coding step/stage 16 and in the decoding step/stage 36), in the further embodiment of Fig. 9 a different way of block switching is carried out. Instead of using a fixed large MLT 10 (e.g. an MDCT) before the separation into speech and audio bands, several short MLTs (or MDCTs) 90 can be switched on. For example, instead of using one MDCT with a transform length of 2048 samples, 8 short MDCTs with a transform length of 256 samples can be used. However, it is not mandatory that the sum of the lengths of the short transforms is equal to the long transform length (although it makes buffer handling even more easier).

[0030] Correspondingly, several short inverse MLTs 91 are used in front of speech encoder 12 and several short MLTs 94 are used following speech decoder 13. Advantageously, for this Fig. 9 long/short block mode switching the internal buffer handling is easier than for the long/short block mode switching according to figures 1 to 8, at the cost of a less sharp band separation between the speech frequency band and the remaining frequency band. The reason for the internal buffer handling being easier is as follows: at least for each inverse MLT operation an additional buffer is required, which leads in case of an inner transform to the necessity of an additional buffer also in the parallel high frequency path. Therefore the switching at the outmost transform has the least side effects concerning buffers.

On the other hand, because the short blocks are used only for encoding transient input signals, the sharp separation in time domain is more important.

[0031] In Fig. 9, the Fig. 1 reference signs do mean the same operation as in Fig. 1. The MLT 10 is input signal IS adaptively replaced by short MLT steps/stages 90, the inverse MLT 11 is replaced by shorter inverse MLT steps/stages 91, and the MLT 14 is replaced by shorter MLT steps/stages 94. Due to this kind of blocks switching, the lengths of the first transform 90, 30 and the second transform 11, 34, 51, 74 (iMDCT to reconstruct the speech band) and the third transform 14, 54 are coordinated. Furthermore, several short blocks of the speech band signal can be buffered after the iMDCT 91 in Fig. 9 in order to collect enough samples for a complete input frame for the speech coder.

The encoding of Fig. 9 can also be adapted correspondingly to the encoding described for Fig. 5.

[0032] Based on the Fig. 9 embodiment, the decoding according to Fig. 3, or the decoding according to Fig. 7, is adapted correspondingly, i.e. the inverse MLTs 34 and 30 are each replaced by corresponding adaptively switched shorter inverse MLTs.

Based on the Fig. 9 embodiment, the transform block sections are weighted at encoding side in MLT 90 and at decoding side in inverse MLT 30 by window functions, in particular in an overlapping manner, wherein the length of a window function corresponds to the current transform length. In case of switching the transform length, to achieve a smooth transition between long and short blocks, especially shaped long windows (the start and stop windows, or transition windows) are used.

Claims

1. Method for encoding a speech and/or non-speech audio input signal (IS), said method including the steps:

- transforming (10, 90) successive and possibly overlapping sections of said input signal (IS) by at least one initial MLT transform and splitting the resulting output frequency bins into a low band signal and a remaining band signal (RBS);
- passing said low band signal to a speech/audio switching (15) and through a speech coding/decoding loop including at least one short first-type MLT transform (11, 51, 91), a speech encoding (12, 52), a corresponding speech decoding (13, 53), and at least one short second-type MLT transform (14, 54, 94) having a type opposite than that of said first-type short MLT transform;
- quantising and encoding (16) said remaining band signal (RBS), controlled by a psycho-acoustic model that receives as its input said audio input signal (IS);
- combining (17) the output signal of said quantising and encoding (16), a switching information signal (SWI) of said switching (15), possibly the output signal of said speech encoding (12, 52), and optionally other encoding side information (SI), in order to form for said current section of said input signal (IS) an output bit stream (OBS),

wherein said speech/audio switching (15) receives said low band signal and a second input signal (DS) derived from the output of said short second-type MLT transform (14, 54, 94) and decides, whether said second input signal bypasses said quantising and encoding (16) step or said low band signal is coded together with said remaining band signal (RBS) in said quantising and encoding (16) step,
and wherein in the latter case said output signal of said speech encoding (12, 52) is not included in the current section of said output bit stream (OBS).

2. Apparatus for encoding a speech and/or non-speech audio input signal (IS), said apparatus including means being adapted for:

- transforming (10, 90) successive and possibly overlapping sections of said input signal (IS) by at least one initial MLT transform and splitting the resulting output frequency bins into a low band signal and a remaining band signal (RBS);
- passing said low band signal to a speech/audio switching (15) and through a speech coding/decoding loop including at least one short first-type MLT transform (11, 51, 91), a speech encoding (12, 52), a corresponding speech decoding (13, 53), and at least one short second-type MLT transform (14, 54, 94) having a type opposite than that of said first-type short MLT transform;
- quantising and encoding (16) said remaining band signal (RBS), controlled by a psycho-acoustic model that receives as its input said audio input signal (IS);
- combining (17) the output signal of said quantising and encoding (16), a switching information signal (SWI) of said switching (15), possibly the output signal of said speech encoding (12, 52), and optionally other encoding side information (SI), in order to form for said current section of said input signal (IS) an output bit stream (OBS), wherein said speech/audio switching (15) receives said low band signal and a second input signal (DS) derived from the output of said short second-type MLT transform (14, 54, 94) and decides, whether said second input signal bypasses said quantising and encoding (16) step or said low band signal is coded together with said remaining band signal (RBS) in said quantising and encoding (16) step,
and wherein in the latter case said output signal of said speech encoding (12, 52) is not included in the current section of said output bit stream (OBS).

3. Method for decoding a bit stream (OBS) representing an encoded speech and/or non-speech audio input signal (IS) that was encoded according to the method of claim 1, said decoding method including the steps:

- demultiplexing (37) successive sections of said bitstream (OBS) to regain the output signal of said quantising and encoding (16), said switching information signal (SWI), possibly the output signal of said speech encoding (12, 52), and said encoding side information (SI) if present;
- if present in a current section of said bitstream (OBS), passing said output signal of said speech encoding through a speech decoding (33, 73) and said short second-type MLT transform (34, 74);
- decoding (36) said output signal of said quantising and encoding (16), controlled by said encoding side information (SI) if present, in order to provide for said current section a reconstructed remaining band signal (RRBS) and a reconstructed low band signal (RLBS);
- providing a speech/audio switching (15) with said reconstructed low band signal and a second input signal (CS) derived from the output of said second-type MLT transform (34, 74), and passing according to said switching information signal (SWI) either said reconstructed low band signal (RLBS) or said second input signal (CS);
- inversely MLT transforming (30) the output signal of said switching (15) combined with said reconstructed remaining band signal (RRBS), and possibly overlapping successive sections, in order to form a current section

of the reconstructed output signal (OS).

4. Apparatus for decoding a bit stream (OBS) representing an encoded speech and/or non-speech audio input signal (IS) that was encoded according to the method of claim 1, said apparatus including means being adapted for:

- demultiplexing (37) successive sections of said bitstream (OBS) to regain the output signal of said quantising and encoding (16), said switching information signal (SWI), possibly the output signal of said speech encoding (12, 52), and said encoding side information (SI) if present;
- if present in a current section of said bitstream (OBS), passing said output signal of said speech encoding through a speech decoding (33, 73) and said short second-type MLT transform (34, 74);
- decoding (36) said output signal of said quantising and encoding (16), controlled by said encoding side information (SI) if present, in order to provide for said current section a reconstructed remaining band signal (RRBS) and a reconstructed low band signal (RLBS);
- providing a speech/audio switching (15) with said reconstructed low band signal and a second input signal (CS) derived from the output of said second-type MLT transform (34, 74), and passing according to said switching information signal (SWI) either said reconstructed low band signal (RLBS) or said second input signal (CS);
- inversely MLT transforming (30) the output signal of said switching (15) combined with said reconstructed remaining band signal (RRBS), and possibly overlapping successive sections, in order to form a current section of the reconstructed output signal (OS).

5. Method according to claim 1 or 3, or apparatus according to claim 2 or 4, wherein in case a single MLT transform (10) is used at the input of the encoding and a single inverse MLT transform (30) is used at the output of the decoding, input signal (IS) adaptively at the input of said quantisation&coding (16) and at the output of said decoding (36) several short MLT transforms each having a length smaller than the length of said single MLT transform (10) and said single inverse MLT transform (30), respectively, are carried out:

either short inverse MLT transforms (22) at the input of said quantisation&coding (16) and short MLT transforms (22) at the output of said decoding (36),
or short MLT transforms (62) at the input of said quantisation&coding (16) and short inverse MLT transforms (82) at the output of said decoding (36).

6. Method or apparatus according to claim 5, wherein said short MLT transforms and said short inverse MLT transforms, respectively, are carried out if the signal energy in a current section of said input signal (IS) exceeds a threshold level.

7. Method according to claim 1 or 3, or apparatus according to claim 2 or 4, wherein at the input of the encoding it is switched input signal (IS) adaptively from a single MLT transform (10) to multiple shorter MLT transforms (90), and at the output of said decoding (36) correspondingly from a single inverse MLT transform (30) to multiple shorter inverse MLT transforms.

8. Method or apparatus according to claim 7, wherein said multiple shorter MLT transforms and said multiple shorter inverse MLT transforms, respectively, are carried out if the signal energy in a current section of said input signal (IS) exceeds a threshold level.

9. Method according to one of claims 1, 3 and 5 to 8, or apparatus according to one of claims 2 and 4 to 8, wherein said second input signal (DS) is the difference signal between said low band signal and the output signal (RSS) of said second-type MLT transform (14, 54, 94).

10. Method according to one of claims 1, 3 and 5 to 8, or apparatus according to one of claims 2 and 4 to 8, wherein said second input signal (DS) said output signal (RSS) of said second-type MLT transform (14, 54, 94).

11. Method according to one of claims 1, 3 and 5 to 10, or apparatus according to one of claims 2 and 4 to 10, wherein said switching (15) is controlled by information received from said psycho-acoustic model (18).

12. Method according to one of claims 1, 3 and 5 to 11, or apparatus according to one of claims 2 and 4 to 11, wherein said switching (15) is operated by using a rate-distortion optimisation.

13. Method according to one of claims 1, 3 and 5 to 12, or apparatus according to one of claims 2 and 4 to 12, wherein successive sections of said input signal (IS) and successive sections for said output signal (OS) are weighted by a

window function having a length corresponding to the related transform length, in particular in an overlapping manner, and wherein, if the transform length is switched, corresponding transition window functions are used.

14. Digital audio signal that is encoded according to the method of one of claims 1, 3 and to 5 to 13.

15. Storage medium, for example an optical disc, that contains or stores, or has recorded on it, a digital audio signal according to claim 14.

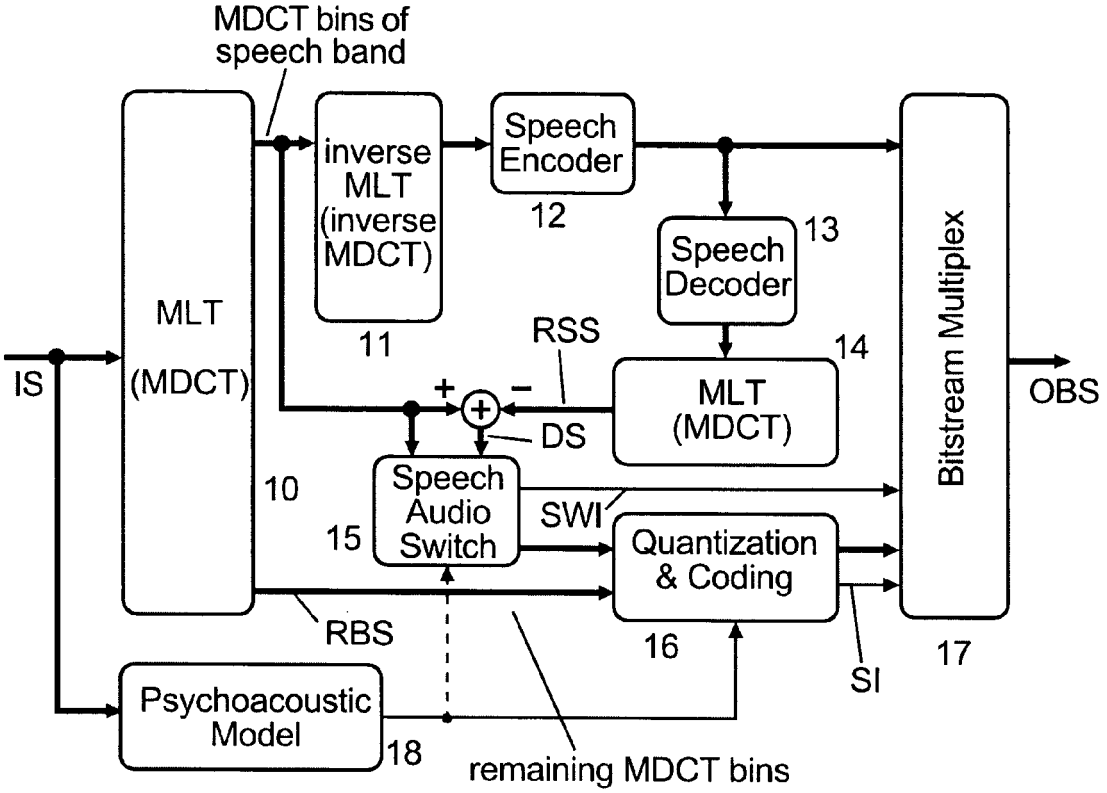
MDCT bins of speech band
inverse MLT (inverse MDCT)
11
Speech Encoder
12
Speech Decoder
13
MLT (MDCT)
14
Bitstream Multiplex
OBS
IS
MLT (MDCT)
10
RSS
DS
Speech Audio Switch
15
SWI
Quantization & Coding
16
17
SI
RBS
Psychoacoustic Model
18
remaining MDCT bins

Fig. 1

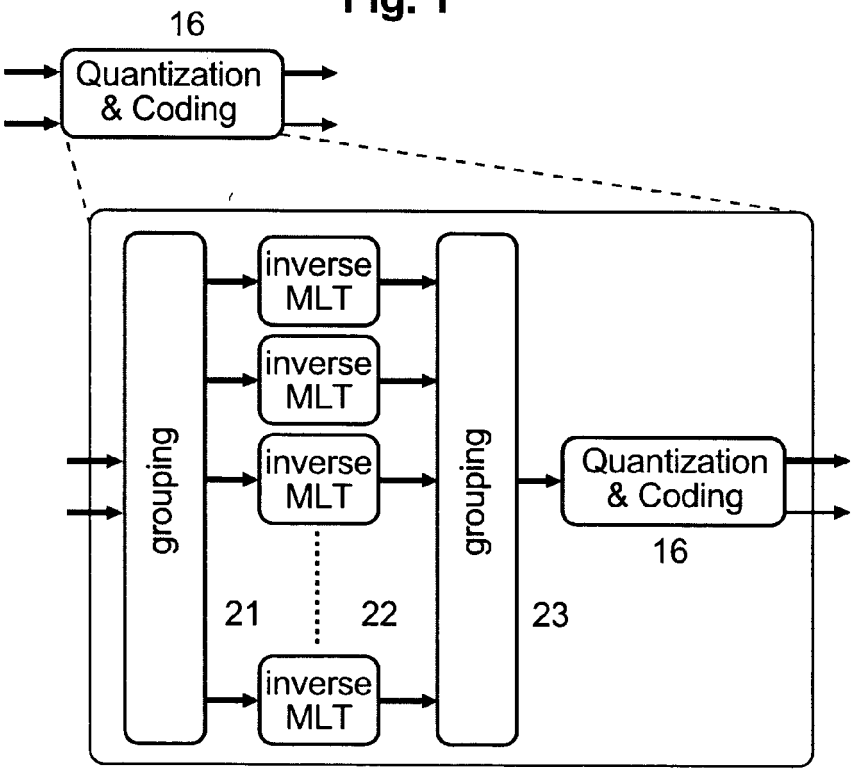
16
Quantization & Coding
grouping
inverse MLT
inverse MLT
inverse MLT
inverse MLT
21
22
grouping
23
Quantization & Coding
16

Fig. 2

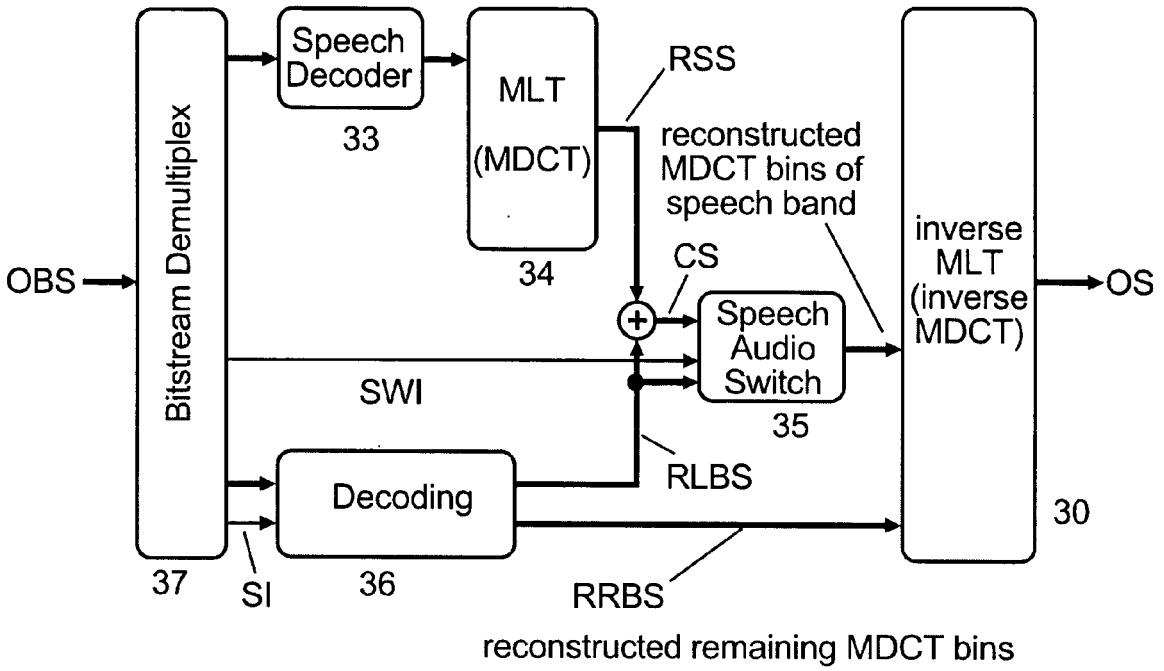
OBS
Bitstream Demultiplex
37
Speech Decoder
33
MLT (MDCT)
34
RSS
reconstructed MDCT bins of speech band
CS
Speech Audio Switch
35
SWI
SI
Decoding
36
RLBS
RRBS
reconstructed remaining MDCT bins
inverse MLT (inverse MDCT)
30
OS

Fig. 3

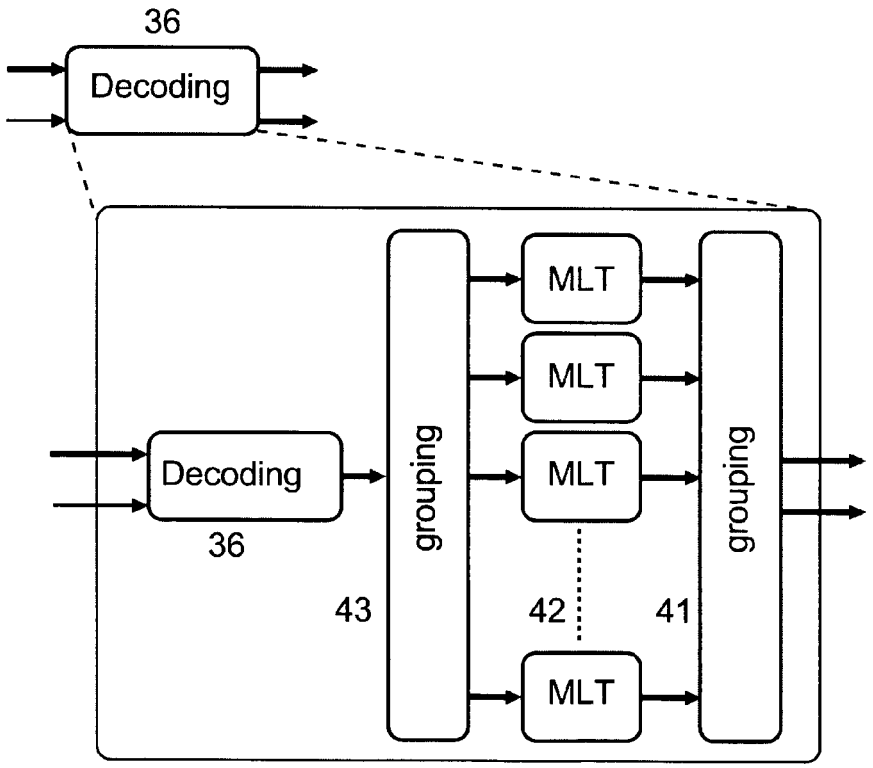
36
Decoding
Decoding
36
grouping
MLT
MLT
MLT
MLT
43
42
41
grouping

Fig. 4

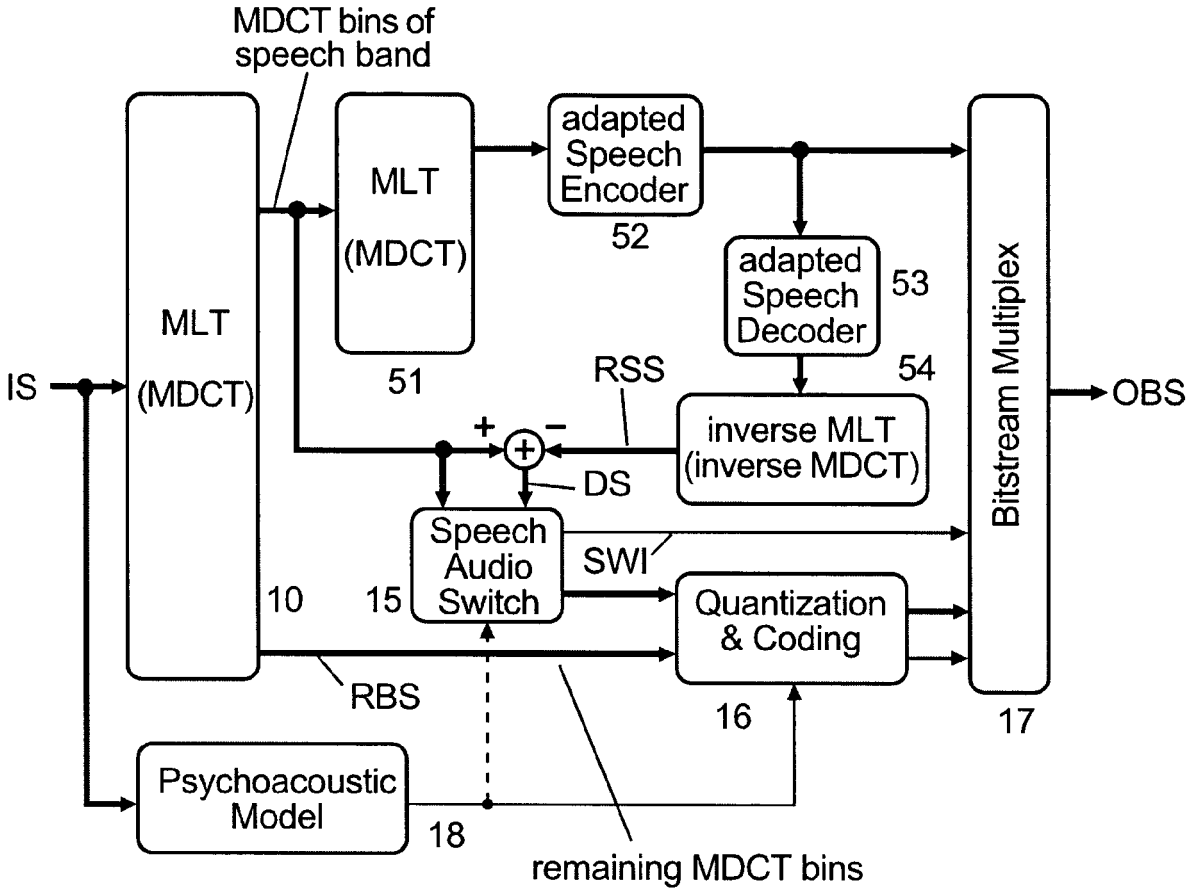
MDCT bins of speech band
MLT (MDCT)
51
adapted Speech Encoder
52
adapted Speech Decoder
53
54
Bitstream Multiplex
OBS
IS
MLT (MDCT)
RSS
inverse MLT (inverse MDCT)
DS
Speech Audio Switch
15
SWI
10
Quantization & Coding
16
17
RBS
Psychoacoustic Model
18
remaining MDCT bins

Fig. 5

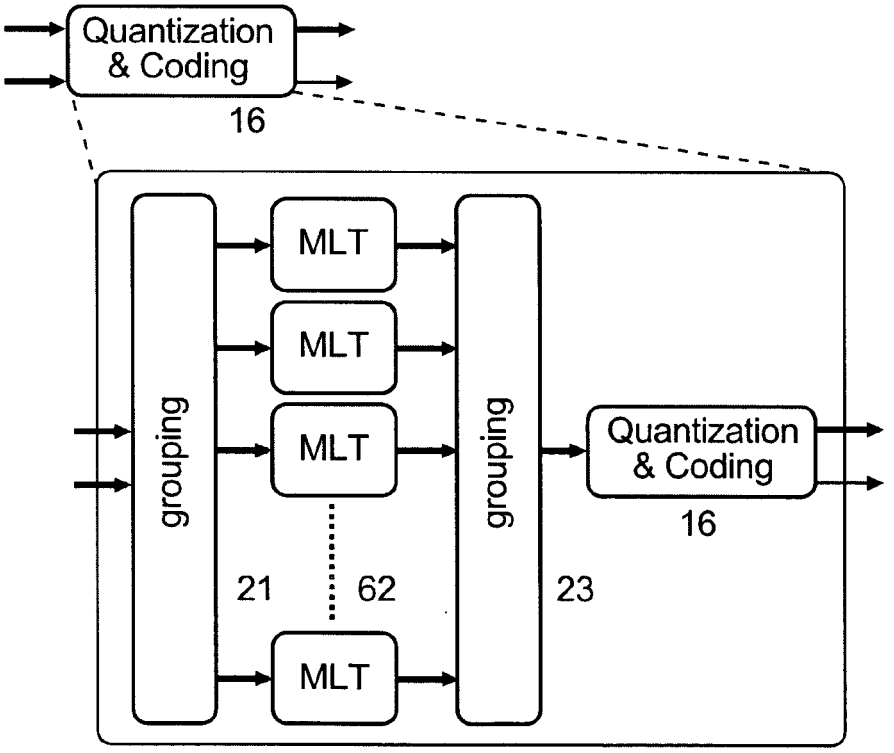
Quantization & Coding
16
grouping
MLT
MLT
MLT
MLT
21
62
grouping
23
Quantization & Coding
16

Fig. 6

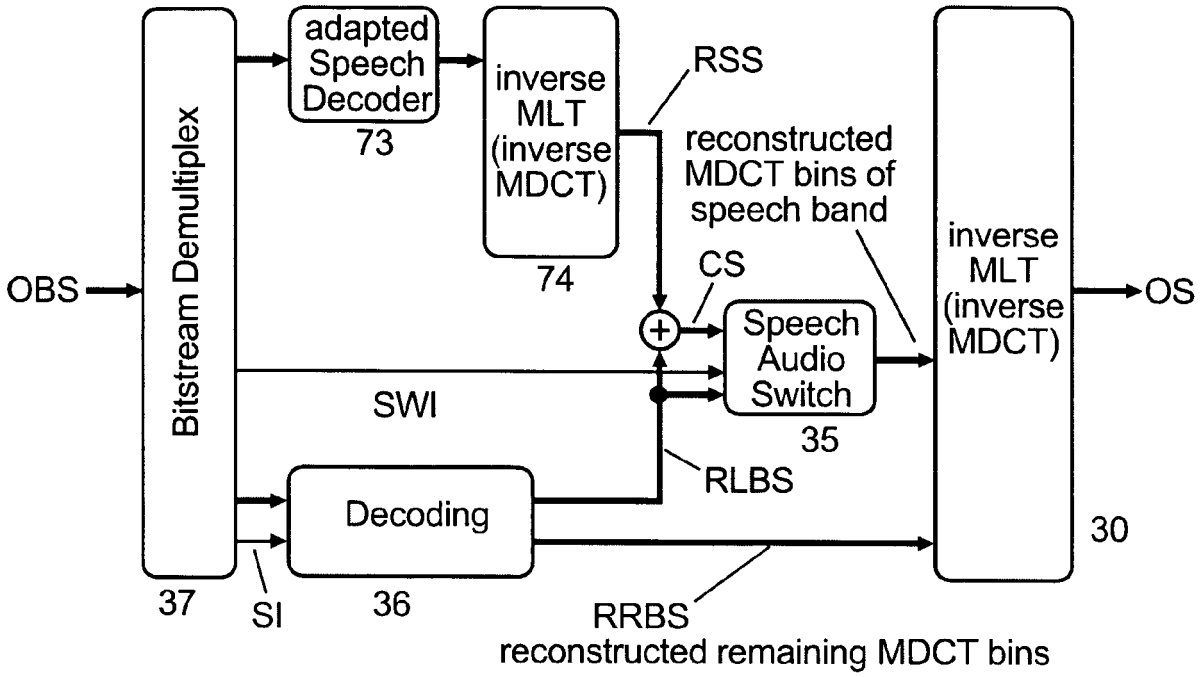
OBS
Bitstream Demultiplex
37
adapted Speech Decoder
73
inverse MLT (inverse MDCT)
74
RSS
reconstructed MDCT bins of speech band
CS
Speech Audio Switch
35
SWI
SI
Decoding
36
RLBS
RRBS
reconstructed remaining MDCT bins
inverse MLT (inverse MDCT)
30
OS

Fig. 7

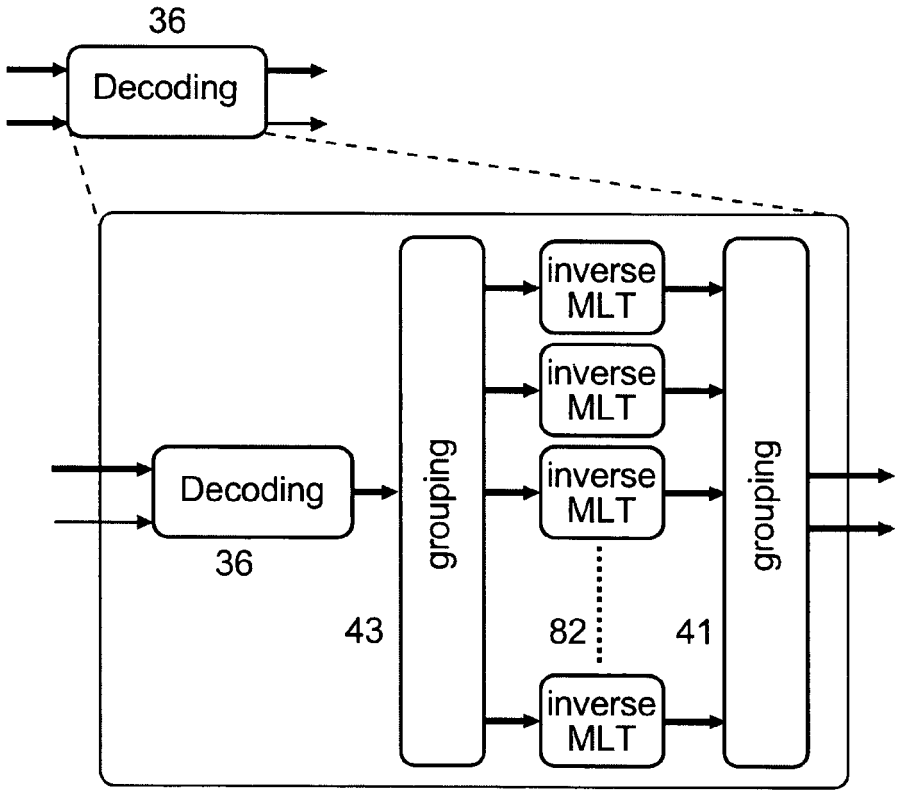
36
Decoding
Decoding
36
grouping
43
inverse MLT
inverse MLT
inverse MLT
82
inverse MLT
41
grouping

Fig. 8

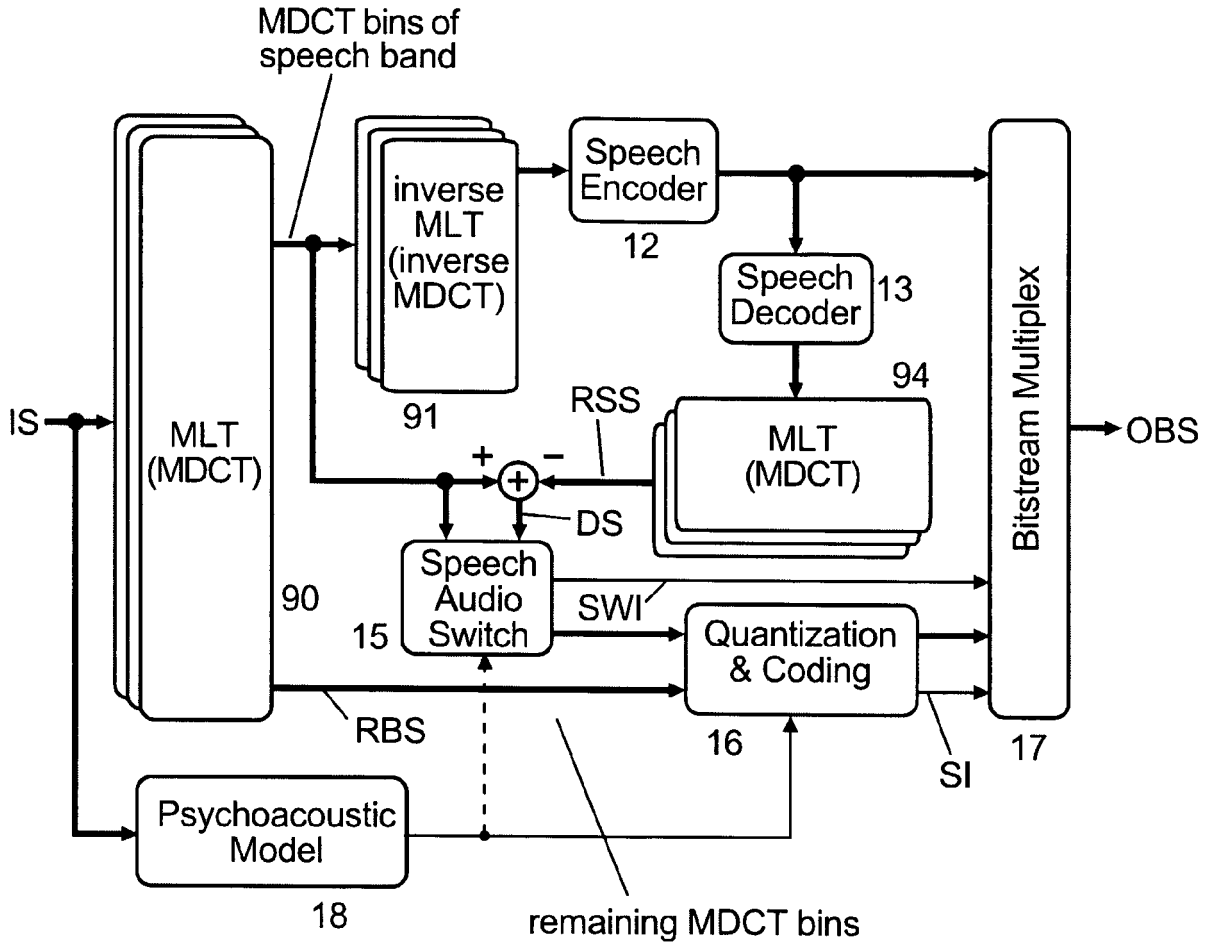
MDCT bins of speech band
IS
MLT (MDCT)
90
inverse MLT (inverse MDCT)
91
Speech Encoder
12
Speech Decoder
13
94
RSS
MLT (MDCT)
+
−
DS
Speech Audio Switch
15
SWI
Quantization & Coding
16
RBS
Psychoacoustic Model
18
remaining MDCT bins
SI
17
Bitstream Multiplex
OBS

Fig. 9

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 08 15 9018

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A	EP 1 278 184 A (MICROSOFT CORP [US]) 22 January 2003 (2003-01-22) * paragraph [0012] * * figures 2a,2b * -----	1-4,14, 15	INV. G10L19/14 ADD. G10L19/02 G10L19/04 G10L11/02
A,D	RAMPRASHAD S A: "A two stage hybrid embedded speech/audio coding structure" ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1998. PROCEEDINGS OF THE 1998 IEEE INTERNATIONAL CONFERENCE ON SEATTLE, WA, USA 12-15 MAY 1998, NEW YORK, NY, USA,IEEE, US, vol. 1, 12 May 1998 (1998-05-12), pages 337-340, XP010279163 ISBN: 978-0-7803-4428-0 * figures 1,2 * * paragraph [05.1] * -----	1-4,14, 15	

TECHNICAL FIELDS SEARCHED (IPC)

G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	5 September 2008	Krembel, Luc

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 08 15 9018

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

05-09-2008

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 1278184 A	22-01-2003	AT 388465 T	15-03-2008
		JP 2003044097 A	14-02-2003
		US 2003004711 A1	02-01-2003

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

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

- **S. Ragot et al.** ITU-T G.729.1: An 8-32 Kbit/s scalable coder interoperable with G.729 for wideband telephony and voice over IP. *IEEE International Conference on Acoustics, Speech and Signal Processing 2007, ICASSP 2007,* 2007, vol. 4, IV-529-IV-532 **[0002]**
- **S.A. Ramprashad.** A two stage hybrid embedded speech/audio coding structure. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing 1998, ICASSP 1998,* 1998, vol. 1, 337-340 **[0003]**
- **M. Purat ; P. Noll.** A new orthonormal wavelet packet decomposition for audio coding using frequency-varying modulated lapped transforms. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics,* 1995, 183-186 **[0004]**
- **M. Purat ; P. Noll.** Audio coding with a dynamic wavelet packet decomposition based on frequency-varying modulated lapped transforms. *IEEE International Conference on Acoustics, Speech, and Signal Processing 1996, ICASSP 1996,* 1996, vol. 2, 1021-1024 **[0005]**
- **John P. Princen ; Alan B. Bradley.** Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Transactions on Acoustics Speech Sigal Processing ASSP-34,* 1986, 1153-1161 **[0016]**
- **H.S.Malvar.** Signal processing with lapped transform. Artech House Inc, 1992 **[0016]**
- **M.Temerinac ; B.Edler.** A unified approach to lapped or-thogonal transforms. *IEEE Transactions on Image Processing,* January 1992, vol. 1 (1), 111-116 **[0016]**
- **B.Edler.** Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen. *FREQUENZ,* 1989, vol. 43, 252-256 **[0023]**