Apparatus and Method for Estimating an Inter-Channel Time Difference

Specification

The present application is related to stereo processing or, generally, multi-channel processing, where a multi-channel signal has two channels such as a left channel and a right channel in the case of a stereo signal or more than two channels, such as three, four, five or any other number of channels.

Stereo speech and particularly conversational stereo speech has received much less scientific attention than storage and broadcasting of stereophonic music. Indeed in speech communications monophonic transmission is still nowadays mostly used. However with the increase of network bandwidth and capacity, it is envisioned that communications based on stereophonic technologies will become more popular and bring a better listening experience.

Efficient coding of stereophonic audio material has been for a long time studied in perceptual audio coding of music for efficient storage or broadcasting. At high bitrates, where waveform preserving is crucial, sum-difference stereo, known as mid/side (M/S) stereo, has been employed for a long time. For low bit-rates, intensity stereo and more recently parametric stereo coding has been introduced. The latest technique was adopted in different standards as HeAACv2 and Mpeg USAC. It generates a down-mix of the two-channel signal and associates compact spatial side information.

Joint stereo coding are usually built over a high frequency resolution, i.e. low time resolution, time-frequency transformation of the signal and is then not compatible to low delay and time domain processing performed in most speech coders. Moreover the engendered bit-rate is usually high.

On the other hand, parametric stereo employs an extra filter-bank positioned in the front-end of the encoder as pre-processor and in the back-end of the decoder as post-processor. Therefore, parametric stereo can be used with conventional speech coders like ACELP as it is done in MPEG USAC. Moreover, the parametrization of the auditory scene can be achieved with minimum amount of side information, which is suitable for low bit-rates. However, parametric stereo is as for example in MPEG USAC not specifically de-

signed for low delay and does not deliver consistent quality for different conversational scenarios. In conventional parametric representation of the spatial scene, the width of the stereo image is artificially reproduced by a decorrelator applied on the two synthesized channels and controlled by Inter-channel Coherence (ICs) parameters computed and transmitted by the encoder. For most stereo speech, this way of widening the stereo image is not appropriate for recreating the natural ambience of speech which is a pretty direct sound since it is produced by a single source located at a specific position in the space (with sometimes some reverberation from the room). By contrast, music instruments have much more natural width than speech, which can be better imitated by decorrelating the channels.

Problems also occur when speech is recorded with non-coincident microphones, like in A-B configuration when microphones are distant from each other or for binaural recording or rendering. Those scenarios can be envisioned for capturing speech in teleconferences or for creating a virtually auditory scene with distant speakers in the multipoint control unit (MCU). The time of arrival of the signal is then different from one channel to the other unlike recordings done on coincident microphones like X-Y (intensity recording) or M-S (Mid-Side recording). The computation of the coherence of such non time-aligned two channels can then be wrongly estimated which makes fail the artificial ambience synthesis.

Prior art references related to stereo processing are US Patent 5,434,948 or US Patent 8,811,621.

Document WO 2006/089570 A1 discloses a near-transparent or transparent multi-channel encoder/decoder scheme. A multi-channel encoder/decoder scheme additionally generates a waveform-type residual signal. This residual signal is transmitted together with one or more multi-channel parameters to a decoder. In contrast to a purely parametric multi-channel decoder, the enhanced decoder generates a multi-channel output signal having an improved output quality because of the additional residual signal. On the encoder-side, a left channel and a right channel are both filtered by an analysis filterbank. Then, for each subband signal, an alignment value and a gain value are calculated for a subband. Such an alignment is then performed before further processing. On the decoder-side, a de-alignment and a gain processing is performed and the corresponding signals are then synthesized by a synthesis filterbank in order to generate a decoded left signal and a decoded right signal.

In such stereo processing applications, the calculation of an inter-channel or inter channel time difference between a first channel signal and a second channel signal is useful in order to typically perform a broadband time alignment procedure. However, other applications do exist for the usage of an inter-channel time difference between a first channel and a second channel, where these applications are in storage or transmission of parametric data, stereo/multi-channel processing comprising a time alignment of two channels, a time difference of arrival estimation for a determination of a speaker position in a room, beamforming spatial filtering, foreground/background decomposition or the location of a sound source by, for example, acoustic triangulation in order to only name a few.

For all such applications, an efficient, accurate and robust determination of an inter-channel time difference between a first and a second channel signal is necessary.

There do already exist such determinations known under the term "GCC-PHAT" or, stated differently, generalized cross-correlation phase transform. Typically, a cross-correlation spectrum is calculated between the two channel signals and, then, a weighting function is applied to the cross-correlation spectrum for obtaining a so-called generalized cross-correlation spectrum before performing an inverse spectral transform such as an inverse DFT to the generalized cross-correlation spectrum in order to find a time-domain representation. This time-domain representation represents values for certain time lags and the highest peak of the time-domain representation then typically corresponds to the time delay or time difference, i.e., the inter-channel time delay of difference between the two channel signals.

However, it has been shown that, particularly in signals that are different from, for example, clean speech without any reverberation or background noise, the robustness of this general technique is not optimum.

It is, therefore, an object of the present invention to provide an improved concept for estimating an inter-channel time difference between two channel signals.

This object is achieved by an apparatus for estimating an inter-channel time difference in accordance with claim 1, or a method for estimating an inter-channel time difference in accordance with claim 15 or a computer program in accordance with claim 16.

The present invention is based on the finding that a smoothing of the cross-correlation spectrum over time that is controlled by a spectral characteristic of the spectrum of the first channel signal or the second channel signal significantly improves the robustness and accuracy of the inter-channel time difference determination.

In preferred embodiments, a tonality/noisiness characteristic of the spectrum is determined, and in case of tone-like signal, a smoothing is stronger while, in case of a noisiness signal, a smoothing is made less stronger.

Preferably, a spectral flatness measure is used and, in case of tone-like signals, the spectral flatness measure will be low and the smoothing will become stronger, and in case of noise-like signals, the spectral flatness measure will be high such as about 1 or close to 1 and the smoothing will be weak.

Thus, in accordance with the present invention, an apparatus for estimating an inter-channel time difference between a first channel signal and a second channel signal comprises a calculator for calculating a cross-correlation spectrum for a time block for the first channel signal in the time block and the second channel signal in the time block. The apparatus further comprises a spectral characteristic estimator for estimating a characteristic of a spectrum of the first channel signal and the second channel signal for the time block and, additionally, a smoothing filter for smoothing the cross-correlation spectrum over time using the spectral characteristic to obtain a smoothed cross-correlation spectrum. Then, the smoothed cross-correlation spectrum is further processed by a processor in order to obtain the inter-channel time difference parameter.

For preferred embodiments related to the further processing of the smoothed cross-correlation spectrum, an adaptive thresholding operation is performed, in which the time-domain representation of the smoothed generalized cross-correlation spectrum is analyzed in order to determine a variable threshold, that depends on the time-domain representation and a peak of the time-domain representation is compared to the variable threshold, wherein an inter-channel time difference is determined as a time lag associated with a peak being in a predetermined relation to the threshold such as being greater than the threshold.

In one embodiment, the variable threshold is determined as a value being equal to an integer multiple of a value among the largest, for example ten percents of the values of

the time domain representation or, alternatively, in a further embodiment for the variable determination, the variable threshold is calculated by a multiplication of the variable threshold and the value, where the value depends on a signal-to-noise ratio characteristic of the first and the second channel signals, where the value becomes higher for a higher signal-to-noise ratio and becomes lower for a lower signal-to-noise ratio.

As stated before, the inter-channel time difference calculation can be used in many different applications such as the storage or transmission of parametric data, a stereo/multi-channel processing/encoding, a time alignment of two channels, a time difference of arrival estimation for the determination of a speaker position in a room with two microphones and a known microphone setup, for the purpose of beamforming, spatial filtering, foreground/background decomposition or a location determination of a sound source, for example by acoustic triangulation based on time differences of two or three signals.

In the following, however, a preferred implementation and usage of the inter-channel time difference calculation is described for the purpose of broadband time alignment of two stereo signals in a process of encoding a multi-channel signal having the at least two channels.

An apparatus for encoding a multi-channel signal having at least two channels comprises a parameter determiner to determine a broadband alignment parameter on the one hand and a plurality of narrowband alignment parameters on the other hand. These parameters are used by a signal aligner for aligning the at least two channels using these parameters to obtain aligned channels. Then, a signal processor calculates a mid-signal and a side signal using the aligned channels and the mid-signal and the side signal are subsequently encoded and forwarded into an encoded output signal that additionally has, as parametric side information, the broadband alignment parameter and the plurality of narrowband alignment parameters.

On the decoder-side, a signal decoder decodes the encoded mid-signal and the encoded side signal to obtain decoded mid and side signals. These signals are then processed by a signal processor for calculating a decoded first channel and a decoded second channel. These decoded channels are then de-aligned using the information on the broadband alignment parameter and the information on the plurality of narrowband parameters included in an encoded multi-channel signal to obtain the decoded multi-channel signal.

In a specific implementation, the broadband alignment parameter is an inter-channel time difference parameter and the plurality of narrowband alignment parameters are inter channel phase differences.

The present invention is based on the finding that specifically for speech signals where there is more than one speaker, but also for other audio signals where there are several audio sources, the different places of the audio sources that both map into two channels of the multi-channel signal can be accounted for using a broadband alignment parameter such as an inter-channel time difference parameter that is applied to the whole spectrum of either one or both channels. In addition to this broadband alignment parameter, it has been found that several narrowband alignment parameters that differ from subband to subband additionally result in a better alignment of the signal in both channels.

Thus, a broadband alignment corresponding to the same time delay in each subband together with a phase alignment corresponding to different phase rotations for different subbands results in an optimum alignment of both channels before these two channels are then converted into a mid/side representation which is then further encoded. Due to the fact that an optimum alignment has been obtained, the energy in the mid-signal is as high as possible on the one hand and the energy in the side signal is as small as possible on the other hand so that an optimum coding result with a lowest possible bitrate or a highest possible audio quality for a certain bitrate can be obtained.

Specifically for conversional speech material, it appears that there are typically speakers being active at two different places. Additionally, the situation is such that, normally, only one speaker is speaking from the first place and then the second speaker is speaking from the second place or location. The influence of the different locations on the two channels such as a first or left channel and a second or right channel is reflected by different time of arrivals and, therefore, a certain time delay between both channels due to the different locations, and this time delay is changing from time to time. Generally, this influence is reflected in the two channel signals as a broadband de-alignment that can be addressed by the broadband alignment parameter.

On the other hand, other effects, particularly coming from reverberation or further noise sources can be accounted for by individual phase alignment parameters for individual bands that are superposed on the broadband different arrival times or broadband de-alignment of both channels.

In view of that, the usage of both, a broadband alignment parameter and a plurality of narrowband alignment parameters on top of the broadband alignment parameter result in an optimum channel alignment on the encoder-side for obtaining a good and very compact mid/side representation while, on the other hand, a corresponding de-alignment subsequent to a decoding on the decoder side results in a good audio quality for a certain bitrate or in a small bitrate for a certain required audio quality.

An advantage of the present invention is that it provides a new stereo coding scheme much more suitable for a conversion of stereo speech than the existing stereo coding schemes. In accordance with the invention, parametric stereo technologies and joint stereo coding technologies are combined particularly by exploiting the inter-channel time difference occurring in channels of a multi-channel signal specifically in the case of speech sources but also in the case of other audio sources.

Several embodiments provide useful advantages as discussed later on.

The new method is a hybrid approach mixing elements from a conventional M/S stereo and parametric stereo. In a conventional M/S, the channels are passively downmixed to generate a Mid and a Side signal. The process can be further extended by rotating the channel using a Karhunen-Loeve transform (KLT), also known as Principal Component Analysis (PCA) before summing and differentiating the channels. The Mid signal is coded in a primary code coding while the Side is conveyed to a secondary coder. Evolved M/S stereo can further use prediction of the Side signal by the Mid Channel coded in the present or the previous frame. The main goal of rotation and prediction is to maximize the energy of the Mid signal while minimizing the energy of the Side. M/S stereo is waveform preserving and is in this aspect very robust to any stereo scenarios, but can be very expensive in terms of bit consumption.

For highest efficiency at low bit-rates, parametric stereo computes and codes parameters, like Inter-channel Level differences (ILDs), Inter-channel Phase differences (IPDs), Inter-channel Time differences (ITDs) and Inter-channel Coherence (ICs). They compactly represent the stereo image and are cues of the auditory scene (source localization, panning, width of the stereo…). The aim is then to parametrize the stereo scene and to code only a downmix signal which can be at the decoder and with the help of the transmitted stereo cues be once again spatialized.

Our approach mixed the two concepts. First, stereo cues ITD and IPD are computed and applied on the two channels. The goal is to represent the time difference in broadband and the phase in different frequency bands. The two channels are then aligned in time and phase and M/S coding is then performed. ITD and IPD were found to be useful for modeling stereo speech and are a good replacement of KLT based rotation in M/S. Unlike a pure parametric coding, the ambience is not more modeled by the ICs but directly by the Side signal which is coded and/or predicted. It was found that this approach is more robust especially when handling speech signals.

The computation and processing of ITDs is a crucial part of the invention. ITDs were already exploited in the prior art Binaural Cue Coding (BCC), but in a way that it was inefficient once ITDs change over time. For avoiding this shortcoming, specific windowing was designed for smoothing the transitions between two different ITDs and being able to seamlessly switch from one speaker to another positioned at different places.

Further embodiments are related to the procedure that, on the encoder-side, the parameter determination for determining the plurality of narrowband alignment parameters is performed using channels that have already been aligned with the earlier determined broadband alignment parameter.

Correspondingly, the narrowband de-alignment on the decoder-side is performed before the broadband de-alignment is performed using the typically single broadband alignment parameter.

In further embodiments, it is preferred that, either on the encoder-side but even more importantly on the decoder-side, some kind of windowing and overlap-add operation or any kind of crossfading from one block to the next one is performed subsequent to all alignments and, specifically, subsequent to a time-alignment using the broadband alignment parameter. This avoids any audible artifacts such as clicks when the time or broadband alignment parameter changes from block to block.

In other embodiments, different spectral resolutions are applied. Particularly, the channel signals are subjected to a time-spectral conversion having a high frequency resolution such as a DFT spectrum while the parameters such as the narrowband alignment parameters are determined for parameter bands having a lower spectral resolution. Typically, a

parameter band has more than one spectral line than the signal spectrum and typically has a set of spectral lines from the DFT spectrum. Furthermore, the parameter bands increase from low frequencies to high frequencies in order to account for psychoacoustic issues.

Further embodiments relate to an additional usage of a level parameter such as an inter-level difference or other procedures for processing the side signal such as stereo filling parameters, etc. The encoded side signal can represented by the actual side signal itself, or by a prediction residual signal being performed using the mid signal of the current frame or any other frame, or by a side signal or a side prediction residual signal in only a subset of bands and prediction parameters only for the remaining bands, or even by prediction parameters for all bands without any high frequency resolution side signal information. Hence, in the last alternative above, the encoded side signal is only represented by a prediction parameter for each parameter band or only a subset of parameter bands so that for the remaining parameter bands there does not exist any information on the original side signal.

Furthermore, it is preferred to have the plurality of narrowband alignment parameters not for all parameter bands reflecting the whole bandwidth of the broadband signal but only for a set of lower bands such as the lower 50 percents of the parameter bands. On the other hand, stereo filling parameters are not used for the couple of lower bands, since, for these bands, the side signal itself or a prediction residual signal is transmitted in order to make sure that, at least for the lower bands, a waveform-correct representation is available. On the other hand, the side signal is not transmitted in a waveform-exact representation for the higher bands in order to further decrease the bitrate, but the side signal is typically represented by stereo filling parameters.

Furthermore, it is preferred to perform the entire parameter analysis and alignment within one and the same frequency domain based on the same DFT spectrum. To this end, it is furthermore preferred to use the generalized cross correlation with phase transform (GCC-PHAT) technology for the purpose of inter-channel time difference determination. In a preferred embodiment of this procedure, a smoothing of a correlation spectrum based on an information on a spectral shape, the information preferably being a spectral flatness measure is performed in such a way that a smoothing will be weak in the case of noise-like signals and a smoothing will become stronger in the case of tone-like signals.

Furthermore, it is preferred to perform a special phase rotation, where the channel amplitudes are accounted for. Particularly, the phase rotation is distributed between the two channels for the purpose of alignment on the encoder-side and, of course, for the purpose of de-alignment on the decoder-side where a channel having a higher amplitude is considered as a leading channel and will be less affected by the phase rotation, i.e., will be less rotated than a channel with a lower amplitude.

Furthermore, the sum-difference calculation is performed using an energy scaling with a scaling factor that is derived from energies of both channels and is, additionally, bounded to a certain range in order to make sure that the mid/side calculation is not affecting the energy too much. On the other hand, however, it is to be noted that, for the purpose of the present invention, this kind of energy conservation is not as critical as in prior art procedures, since time and phase were aligned beforehand. Therefore, the energy fluctuations due to the calculation of a mid-signal and a side signal from left and right (on the encoder side) or due to the calculation of a left and a right signal from mid and side (on the decoder-side) are not as significant as in the prior art.

Subsequently, preferred embodiments of the present invention are discussed with respect to the accompanying drawings in which:

Fig. 1 is a block diagram of a preferred implementation of an apparatus for encoding a multi-channel signal;

Fig. 2 is a preferred embodiment of an apparatus for decoding an encoded multi-channel signal;

Fig. 3 is an illustration of different frequency resolutions and other frequency-related aspects for certain embodiments;

Fig. 4a illustrates a flowchart of procedures performed in the apparatus for encoding for the purpose of aligning the channels;

Fig. 4b illustrates a preferred embodiment of procedures performed in the frequency domain;

Fig. 4c illustrates a preferred embodiment of procedures performed in the apparatus for encoding using an analysis window with zero padding portions and overlap ranges;

Fig. 4d illustrates a flowchart for further procedures performed within the apparatus for encoding;

Fig. 4e illustrates a flowchart for showing a preferred implementation of an inter-channel time difference estimation;

Fig. 5 illustrates a flowchart illustrating a further embodiment of procedures performed in the apparatus for encoding;

Fig. 6a illustrates a block chart of an embodiment of an encoder;

Fig. 6b illustrates a flowchart of a corresponding embodiment of a decoder;

Fig. 7 illustrates a preferred window scenario with low-overlapping sine windows with zero padding for a stereo time-frequency analysis and synthesis;

Fig. 8 illustrates a table showing the bit consumption of different parameter values;

Fig. 9a illustrates procedures performed by an apparatus for decoding an encoded multi-channel signal in a preferred embodiment;

Fig. 9b illustrates a preferred implementation of the apparatus for decoding an encoded multi-channel signal;

Fig. 9c illustrates a procedure performed in the context of a broadband dealignment in the context of the decoding of an encoded multi-channel signal;

Fig. 10a illustrates an embodiment of an apparatus for estimating an inter-channel time difference;

Fig. 10b illustrates a schematic representation of a signal further processing where the inter-channel time difference is applied;

Fig. 11a illustrates procedures performed by the processor of Fig. 10a;

Fig. 11b illustrates further procedures performed by the processor in Fig. 10a;

Fig. 11c illustrates a further implementation of the calculation of a variable threshold and the usage of the variable threshold in the analysis of the time-domain representation;

Fig. 11d illustrates a first embodiment for the determination of the variable threshold;

Fig. 11e illustrates a further implementation of the determination of the threshold;

Fig. 12 illustrates a time-domain representation for a smoothed cross-correlation spectrum for a clean speech signal;

Fig. 13 illustrates a time-domain representation of a smoothed cross-correlation spectrum for a speech signal having noise and ambiance.

Fig. 10a illustrates an embodiment of an apparatus for estimating an inter-channel time difference between a first channel signal such as a left channel and a second channel signal such as a right channel. These channels are input into a time-spectral converter 150 that is additionally illustrated, with respect to Fig. 4e as item 451.

Furthermore, the time-domain representations of the left and the right channel signals are input into a calculator 1020 for calculating a cross-correlation spectrum for a time block from the first channel signal in the time block and the second channel signal in the time block. Furthermore, the apparatus comprises a spectral characteristic estimator 1010 for estimating a characteristic of a spectrum of the first channel signal or the second channel signal for the time block. The apparatus further comprises a smoothing filter 1030 for smoothing the cross-correlation spectrum over time using the spectral characteristic to obtain a smoothed cross-correlation spectrum. The apparatus further comprises a processor 1040 for processing the smoothed correlation spectrum to obtain the inter-channel time difference.

Particularly, the functionalities of the spectral characteristic estimator are also reflected by Fig. 4e, items 453, 454 in a preferred embodiment.

Furthermore, the functionalities of the cross-correlation spectrum calculator 1020 are also reflected by item 452 in Fig. 4e described later on in a preferred embodiment.

Correspondingly, the functionalities of the smoothing filter 1030 are also reflected by item 453 in the context of Fig. 4e to be described later on. Additionally, the functionalities of the processor 1040 are also described in the context of Fig. 4e in a preferred embodiment as items 456 to 459.

Preferably, the spectral characteristic estimation calculates a noisiness or a tonality of the spectrum where a preferred implementation is the calculation of a spectral flatness measure being close to 0 in the case of tonal or non-noisy signals and being close to 1 in the case of noisy or noise-like signals.

Particularly, the smoothing filter is then configured to apply a stronger smoothing with a first smoothing degree over time in case of a first less noisy characteristic or a first more tonal characteristic, or to apply a weaker smoothing with a second smoothing degree over time in case of a second more noisy or second less tonal characteristic.

Particularly, the first smoothing is greater than the second smoothing degree, where the first noisy characteristic is less noisy than the second noisy characteristic or the first tonal characteristic is more tonal than the second tonal characteristic. The preferred implementation is the spectral flatness measure.

Furthermore, as illustrated in Fig. 11a, the processor is preferably implemented to normalize the smoothed cross-correlation spectrum as illustrated at 456 in Fig. 4e and 11a before performing the calculation of the time-domain representation in step 1031 corresponding to steps 457 and 458 in the embodiment of Fig. 4e. However, as also outlined in Fig. 11a, the processor can also operate without the normalization in step 456 in Fig. 4e. Then, the processor is configured to analyze the time-domain representation as illustrated in block 1032 of Fig. 11a in order to find the inter-channel time difference. This analysis can be performed in any known way and will already result in an improved robustness,

since the analysis is performed based on the cross-correlation spectrum being smoothed in accordance with the spectral characteristic.

As illustrated in Fig. 11b, a preferred implementation of the time-domain analysis 1032 is a low-pass filtering of the time-domain representation as illustrated at 458 in Fig. 11b corresponding to item 458 of Fig. 4e and a subsequent further processing 1033 using a peak searching/peak picking operation within the low-pass filtered time-domain representation.

As illustrated in Fig. 11c, the preferred implementation of the peak picking or peak searching operation is to perform this operation using a variable threshold. Particularly, the processor is configured to perform the peak searching/peak picking operation within the time-domain representation derived from the smoothed cross-correlation spectrum by determining 1034 a variable threshold from the time-domain representation and by comparing a peak or several peaks of the time-domain representation (obtained with or without spectral normalization) to the variable threshold, wherein the inter-channel time difference is determined as a time lag associated with a peak being in a predetermined relation to the threshold such as being greater than the variable threshold.

As illustrated in Fig. 11d, one preferred embodiment illustrated in the pseudo code related to Fig. 4e-b described later on consists in the sorting 1034a of values in accordance with their magnitude. Then, as illustrated in item 1034b in Fig. 11d, the highest for example 10 or 5 % of the values are determined.

Then, as illustrated in step 1034c, a number such as the number 3 is multiplied to the lowest value of the highest 10 or 5 % in order to obtain the variable threshold.

As stated, preferably, the highest 10 or 5 % are determined, but it can also be useful to determine the lowest number of the highest 50 % of the values and to use a higher multiplication number such as 10. Naturally, even a smaller amount such as the highest 3 % of the values are determined and the lowest value among these highest 3 % of the values is then multiplied by a number which is, for example, equal to 2.5 or 2, i.e., lower than 3. Thus, different combinations of numbers and percentages can be used in the embodiment illustrated in Fig. 11d. Apart from the percentages, the numbers can also vary, and numbers greater than 1,5 are preferred.

In a further embodiment illustrated in Fig. 11e, the time-domain representation is divided into subblocks as illustrated by block 1101, and these subblocks are indicated in Fig. 13 at 1300. Here, about 16 subblocks are used for the valid range so that each subblock has a time lag span of 20. However, the number of subblocks can be greater than this value or lower and preferably greater than 3 and lower than 50.

In step 1102 of Fig. 11e, the peak in each subblock is determined, and in step 1103, the average peak in all the subblocks is determined. Then, in step 1104, a multiplication value a is determined that depends on a signal-to-noise ratio on the one hand and, in a further embodiment, depends on the difference between the threshold and the maximum peak as indicated to the left of block 1104. Depending on these input values, one of preferably three different multiplication values is determined where the multiplication value can be equal to a_{low}, a_{high} and a_{lowest}.

Then, in step 1105, the multiplication value a determined in block 1104 is multiplied by the average threshold in order to obtain the variable threshold that is then used in the comparison operation in block 1106. For the comparison operation, once again the time-domain representation input into block 1101 can be used or the already determined peaks in each subblock as outlined in block 1102 can be used.

Subsequently, further embodiments regarding the evaluation and detection of a peak within the time-domain cross-correlation function is outlined.

The evaluation and detection of a peak within the time-domain cross correlation function resulting from the generalized cross-correlation (GCC-PHAT) method in order to estimate the Inter-channel Time Difference (ITD) is not always straightforward due to different input scenarios. Clean speech input can result to a low deviation cross-correlation function with a strong peak, while speech in a noisy reverberant environment can produce a vector with high deviation and peaks with lower but still outstanding magnitude indicating the existence of ITD. A peak detection algorithm that is adaptive and flexible to accommodate different input scenarios is described.

Due to delay constraints, the overall system can handle channel time alignment up to a certain limit, namely ITD_MAX. The proposed algorithm is designed to detect whether a valid ITD exists in the following cases:

- **Valid ITD due to outstanding peak**. An outstanding peak within the [-ITD_MAX, ITD_MAX] bounds of the cross-correlation function is present.

- **No correlation**. When there is no correlation between the two channels, there is no outstanding peak. A threshold should be defined, above which the peak is strong enough to be considered as a valid ITD value. Otherwise , no ITD handling should be signaled, meaning ITD is set to zero and no time alignment is performed.

- **Out of bounds ITD**. Strong peaks of the cross-correlation function outside the region [-ITD_MAX, ITD_MAX] should be evaluated in order to determine whether ITDs that lie outside the handling capacity of the system exist. In this case no ITD handling should be signaled and thus no time alignment is performed.

To determine whether the magnitude of a peak is high enough to be considered as a time difference value, a suitable threshold needs to be defined. For different input scenarios, the cross-correlation function output varies depending on different parameters, e.g. the environment (noise, reverberation etc.), the microphone setup (AB, M/S, etc.). Therefore, to adaptively define the threshold is essential.

In the proposed algorithm, the threshold is defined by first calculating the mean of a rough computation of the envelope of the magnitude of the cross-correlation function within the [-ITD_MAX, ITD_MAX] region (Fig. 13), the average is then weighted accordingly depending on the SNR estimation.

The step-by-step description of the algorithm is described below.

The output of the inverse DFT of the GCC-PHAT, which represents the time-domain cross-correlation, is rearranged from negative to positive time lags (Fig. 12).

The cross-correlation vector is divided in three main areas: the area of interest namely [-ITD_MAX, ITD_MAX] and the area outside the ITD_MAX bounds, namely time lags smaller than –ITD_MAX (max_low) and higher than ITD_MAX (max_high). The maximum peaks of the "out of bound" areas are detected and saved to be compared to the maximum peak detected in the area of interest.

In order to determine whether a valid ITD is present, the sub-vector area [-ITD_MAX, ITD_MAX] of the cross-correlation function is considered. The sub-vector is divided into N sub-blocks (Fig. 13).

For each sub-block the maximum peak magnitude peak_sub and the equivalent time lag position index_sub is found and saved.

The maximum of the local maxima peak_max is determined and will be compared to the threshold to determine the existence of a valid ITD value.

The maximum value peak_max is compared to max_low and max_high. If peak_max is lower than either of the two than no itd handling is signaled and no time alignment is performed. Because of the ITD handling limit of the system, the magnitudes of the out of bound peaks do not need to be evaluated.

The mean of the magnitudes of the peaks is calculated:

$$peak_{mean} = \frac{\Sigma_N\, peak_sub}{N}$$

The threshold $thres$ is then computed by weighting $peak_{mean}$ with an SNR depended weighting factor a_w:

$$thres = a_w peak_{mean}, \text{ where } a_w = \begin{cases} a_{low}, & SNR \leq SNR_{threshold} \\ a_{high}, & SNR > SNR_{threshold} \end{cases}$$

In cases where $SNR \ll SNR_{threshold}$ and $|thres - peak_max| < \varepsilon$, the peak magnitude is also compared to a slightly more relaxed threshold ($a_w = a_{lowest}$), in order to avoid rejecting an outstanding peak with high neighboring peaks. The weighting factors could be for example a_{high} = 3, a_{low} = 2.5 and a_{lowest} = 2, while the $SNR_{threshold}$ could be for example 20dB and the bound ε = 0.05.

Preferred ranges are 2.5 to 5 for a_{high}; 1.5 to 4 for a_{low}; 1.0 to 3 for a_{lowest}; 10 to 30 dB for $SNR_{threshold}$; and 0.01 to 0.5 for ε, where a_{high} is greater than a_{low} that is greater than a_{lowest}.

If *peak_max* > *thres* the equivalent time lag is returned as the estimated ITD, elsewise no itd handling is signaled (ITD=0).

Further embodiments are described later on with respect to Fig. 4e.

Subsequently, a preferred implementation of the present invention within block 1050 of Fig. 10b for the purpose of a signal further processor is discussed with respect to Figs. 1 to 9e, i.e., in the context of a stereo/multi-channel processing/encoding and time alignment of two channels.

However, as stated and as illustrated in Fig. 10b, many other fields exist, where a signal further processing using the determined inter-channel time difference can be performed as well.

Fig. 1 illustrates an apparatus for encoding a multi-channel signal having at least two channels. The multi-channel signal 10 is input into a parameter determiner 100 on the one hand and a signal aligner 200 on the other hand. The parameter determiner 100 determines, on the one hand, a broadband alignment parameter and, on the other hand, a plurality of narrowband alignment parameters from the multi-channel signal. These parameters are output via a parameter line 12. Furthermore, these parameters are also output via a further parameter line 14 to an output interface 500 as illustrated. On the parameter line 14, additional parameters such as the level parameters are forwarded from the parameter determiner 100 to the output interface 500. The signal aligner 200 is configured for aligning the at least two channels of the multi-channel signal 10 using the broadband alignment parameter and the plurality of narrowband alignment parameters received via parameter line 10 to obtain aligned channels 20 at the output of the signal aligner 200. These aligned channels 20 are forwarded to a signal processor 300 which is configured for calculating a mid-signal 31 and a side signal 32 from the aligned channels received via line 20. The apparatus for encoding further comprises a signal encoder 400 for encoding the mid-signal from line 31 and the side signal from line 32 to obtain an encoded mid-signal on line 41 and an encoded side signal on line 42. Both these signals are forwarded to the output interface 500 for generating an encoded multi-channel signal at output line 50. The encoded signal at output line 50 comprises the encoded mid-signal from line 41, the encoded side signal from line 42, the narrowband alignment parameters and the broadband alignment parameters from line 14 and, optionally, a level parameter from line 14 and, additionally optionally, a stereo filling parameter generated by the signal encoder 400 and forwarded to the output interface 500 via parameter line 43.

Preferably, the signal aligner is configured to align the channels from the multi-channel signal using the broadband alignment parameter, before the parameter determiner 100

actually calculates the narrowband parameters. Therefore, in this embodiment, the signal aligner 200 sends the broadband aligned channels back to the parameter determiner 100 via a connection line 15. Then, the parameter determiner 100 determines the plurality of narrowband alignment parameters from an already with respect to the broadband characteristic aligned multi-channel signal. In other embodiments, however, the parameters are determined without this specific sequence of procedures.

Fig. 4a illustrates a preferred implementation, where the specific sequence of steps that incurs connection line 15 is performed. In the step 16, the broadband alignment parameter is determined using the two channels and the broadband alignment parameter such as an inter-channel time difference or ITD parameter is obtained. Then, in step 21, the two channels are aligned by the signal aligner 200 of Fig. 1 using the broadband alignment parameter. Then, in step 17, the narrowband parameters are determined using the aligned channels within the parameter determiner 100 to determine a plurality of narrowband alignment parameters such as a plurality of inter-channel phase difference parameters for different bands of the multi-channel signal. Then, in step 22, the spectral values in each parameter band are aligned using the corresponding narrowband alignment parameter for this specific band. When this procedure in step 22 is performed for each band, for which a narrowband alignment parameter is available, then aligned first and second or left/right channels are available for further signal processing by the signal processor 300 of Fig. 1.

Fig. 4b illustrates a further implementation of the multi-channel encoder of Fig. 1 where several procedures are performed in the frequency domain.

Specifically, the multi-channel encoder further comprises a time-spectrum converter 150 for converting a time domain multi-channel signal into a spectral representation of the at least two channels within the frequency domain.

Furthermore, as illustrated at 152, the parameter determiner, the signal aligner and the signal processor illustrated at 100, 200 and 300 in Fig. 1 all operate in the frequency domain.

Furthermore, the multi-channel encoder and, specifically, the signal processor further comprises a spectrum-time converter 154 for generating a time domain representation of the mid-signal at least.

Preferably, the spectrum time converter additionally converts a spectral representation of the side signal also determined by the procedures represented by block 152 into a time domain representation, and the signal encoder 400 of Fig. 1 is then configured to further encode the mid-signal and/or the side signal as time domain signals depending on the specific implementation of the signal encoder 400 of Fig. 1.

Preferably, the time-spectrum converter 150 of Fig. 4b is configured to implement steps 155, 156 and 157 of Fig. 4c. Specifically, step 155 comprises providing an analysis window with at least one zero padding portion at one end thereof and, specifically, a zero padding portion at the initial window portion and a zero padding portion at the terminating window portion as illustrated, for example, in Fig. 7 later on. Furthermore, the analysis window additionally has overlap ranges or overlap portions at a first half of the window and at a second half of the window and, additionally, preferably a middle part being a non-overlap range as the case may be.

In step 156, each channel is windowed using the analysis window with overlap ranges. Specifically, each channel is widowed using the analysis window in such a way that a first block of the channel is obtained. Subsequently, a second block of the same channel is obtained that has a certain overlap range with the first block and so on, such that subsequent to, for example, five windowing operations, five blocks of windowed samples of each channel are available that are then individually transformed into a spectral representation as illustrated at 157 in Fig. 4c. The same procedure is performed for the other channel as well so that, at the end of step 157, a sequence of blocks of spectral values and, specifically, complex spectral values such as DFT spectral values or complex subband samples is available.

In step 158, which is performed by the parameter determiner 100 of Fig. 1, a broadband alignment parameter is determined and in step 159, which is performed by the signal alignment 200 of Fig. 1, a circular shift is performed using the broadband alignment parameter. In step 160, again performed by the parameter determiner 100 of Fig. 1, narrowband alignment parameters are determined for individual bands/subbands and in step 161, aligned spectral values are rotated for each band using corresponding narrowband alignment parameters determined for the specific bands.

Fig. 4d illustrates further procedures performed by the signal processor 300. Specifically, the signal processor 300 is configured to calculate a mid-signal and a side signal as illustrated at step 301. In step 302, some kind of further processing of the side signal can be performed and then, in step 303, each block of the mid-signal and the side signal is transformed back into the time domain and, in step 304, a synthesis window is applied to each block obtained by step 303 and, in step 305, an overlap add operation for the mid-signal on the one hand and an overlap add operation for the side signal on the other hand is performed to finally obtain the time domain mid/side signals.

Specifically, the operations of the steps 304 and 305 result in a kind of cross fading from one block of the mid-signal or the side signal in the next block of the mid signal and the side signal is performed so that, even when any parameter changes occur such as the inter-channel time difference parameter or the inter-channel phase difference parameter occur, this will nevertheless be not audible in the time domain mid/side signals obtained by step 305 in Fig. 4d.

The new low-delay stereo coding is a joint Mid/Side (M/S) stereo coding exploiting some spatial cues, where the Mid-channel is coded by a primary mono core coder, and the Side-channel is coded in a secondary core coder. The encoder and decoder principles are depicted in Figs. 6a, 6b.

The stereo processing is performed mainly in Frequency Domain (FD). Optionally some stereo processing can be performed in Time Domain (TD) before the frequency analysis. It is the case for the ITD computation, which can be computed and applied before the frequency analysis for aligning the channels in time before pursuing the stereo analysis and processing. Alternatively, ITD processing can be done directly in frequency domain. Since usual speech coders like ACELP do not contain any internal time-frequency decomposition, the stereo coding adds an extra complex modulated filter-bank by means of an analysis and synthesis filter-bank before the core encoder and another stage of analysis-synthesis filter-bank after the core decoder. In the preferred embodiment, an oversampled DFT with a low overlapping region is employed. However, in other embodiments, any complex valued time-frequency decomposition with similar temporal resolution can be used.

The stereo processing consists of computing the spatial cues: inter-channel Time Difference (ITD), the inter-channel Phase Differences (IPDs) and inter-channel Level Differences (ILDs). ITD and IPDs are used on the input stereo signal for aligning the two chan-

nels L and R in time and in phase. ITD is computed in broadband or in time domain while IPDs and ILDs are computed for each or a part of the parameter bands, corresponding to a non-uniform decomposition of the frequency space. Once the two channels are aligned a joint M/S stereo is applied, where the Side signal is then further predicted from the Mid signal. The prediction gain is derived from the ILDs.

The Mid signal is further coded by a primary core coder. In the preferred embodiment, the primary core coder is the 3GPP EVS standard, or a coding derived from it which can switch between a speech coding mode, ACELP, and a music mode based on a MDCT transformation. Preferably, ACELP and the MDCT-based coder are supported by a Time Domain BandWidth Extension (TD-BWE) and or Intelligent Gap Filling (IGF) modules respectively.

The Side signal is first predicted by the Mid channel using prediction gains derived from ILDs. The residual can be further predicted by a delayed version of the Mid signal or directly coded by a secondary core coder, performed in the preferred embodiment in MDCT domain. The stereo processing at encoder can be summarized by Fig. 5 as will be explained later on.

Fig. 2 illustrates a block diagram of an embodiment of an apparatus for decoding an encoded multi-channel signal received at input line 50.

In particular, the signal is received by an input interface 600. Connected to the input interface 600 are a signal decoder 700, and a signal de-aligner 900. Furthermore, a signal processor 800 is connected to a signal decoder 700 on the one hand and is connected to the signal de-aligner on the other hand.

In particular, the encoded multi-channel signal comprises an encoded mid-signal, an encoded side signal, information on the broadband alignment parameter and information on the plurality of narrowband parameters. Thus, the encoded multi-channel signal on line 50 can be exactly the same signal as output by the output interface of 500 of Fig. 1.

However, importantly, it is to be noted here that, in contrast to what is illustrated in Fig. 1, the broadband alignment parameter and the plurality of narrowband alignment parameters included in the encoded signal in a certain form can be exactly the alignment parameters as used by the signal aligner 200 in Fig. 1 but can, alternatively, also be the inverse val-

ues thereof, i.e., parameters that can be used by exactly the same operations performed by the signal aligner 200 but with inverse values so that the de-alignment is obtained.

Thus, the information on the alignment parameters can be the alignment parameters as used by the signal aligner 200 in Fig. 1 or can be inverse values, i.e., actual "de-alignment parameters". Additionally, these parameters will typically be quantized in a certain form as will be discussed later on with respect to Fig. 8.

The input interface 600 of Fig. 2 separates the information on the broadband alignment parameter and the plurality of narrowband alignment parameters from the encoded mid/side signals and forwards this information via parameter line 610 to the signal de-aligner 900. On the other hand, the encoded mid-signal is forwarded to the signal decoder 700 via line 601 and the encoded side signal is forwarded to the signal decoder 700 via signal line 602.

The signal decoder is configured for decoding the encoded mid-signal and for decoding the encoded side signal to obtain a decoded mid-signal on line 701 and a decoded side signal on line 702. These signals are used by the signal processor 800 for calculating a decoded first channel signal or decoded left signal and for calculating a decoded second channel or a decoded right channel signal from the decoded mid signal and the decoded side signal, and the decoded first channel and the decoded second channel are output on lines 801, 802, respectively. The signal de-aligner 900 is configured for de-aligning the decoded first channel on line 801 and the decoded right channel 802 using the information on the broadband alignment parameter and additionally using the information on the plurality of narrowband alignment parameters to obtain a decoded multi-channel signal, i.e., a decoded signal having at least two decoded and de-aligned channels on lines 901 and 902.

Fig. 9a illustrates a preferred sequence of steps performed by the signal de-aligner 900 from Fig. 2. Specifically, step 910 receives aligned left and right channels as available on lines 801, 802 from Fig. 2. In step 910, the signal de-aligner 900 de-aligns individual subbands using the information on the narrowband alignment parameters in order to obtain phase-de-aligned decoded first and second or left and right channels at 911a and 911b. In step 912, the channels are de-aligned using the broadband alignment parameter so that, at 913a and 913b, phase and time-de-aligned channels are obtained.

In step 914, any further processing is performed that comprises using a windowing or any overlap-add operation or, generally, any cross-fade operation in order to obtain, at 915a or 915b, an artifact-reduced or artifact-free decoded signal, i.e., to decoded channels that do not have any artifacts although there have been, typically, time-varying de-alignment parameters for the broadband on the one hand and for the plurality of narrowbands on the other hand.

Fig. 9b illustrates a preferred implementation of the multi-channel decoder illustrated in Fig. 2.

In particular, the signal processor 800 from Fig. 2 comprises a time-spectrum converter 810.

The signal processor furthermore comprises a mid/side to left/right converter 820 in order to calculate from a mid-signal M and a side signal S a left signal L and a right signal R.

However, importantly, in order to calculate L and R by the mid/side-left/right conversion in block 820, the side signal S is not necessarily to be used. Instead, as discussed later on, the left/right signals are initially calculated only using a gain parameter derived from an inter-channel level difference parameter ILD. Generally, the prediction gain can also be considered to be a form of an ILD. The gain can be derived from ILD but can also be directly computed. It is preferred to not compute ILD anymore, but to compute the prediction gain directly and to transmit and use the prediction gain in the decoder rather than the ILD parameter.

Therefore, in this implementation, the side signal S is only used in the channel updater 830 that operates in order to provide a better left/right signal using the transmitted side signal S as illustrated by bypass line 821.

Therefore, the converter 820 operates using a level parameter obtained via a level parameter input 822 and without actually using the side signal S but the channel updater 830 then operates using the side 821 and, depending on the specific implementation, using a stereo filling parameter received via line 831. The signal aligner 900 then comprises a phased-de-aligner and energy scaler 910. The energy scaling is controlled by a scaling factor derived by a scaling factor calculator 940. The scaling factor calculator 940 is fed by the output of the channel updater 830. Based on the narrowband alignment parameters

received via input 911, the phase de-alignment is performed and, in block 920, based on the broadband alignment parameter received via line 921, the time-de-alignment is performed. Finally, a spectrum-time conversion 930 is performed in order to finally obtain the decoded signal.

Fig. 9c illustrates a further sequence of steps typically performed within blocks 920 and 930 of Fig. 9b in a preferred embodiment.

Specifically, the narrowband de-aligned channels are input into the broadband de-alignment functionality corresponding to block 920 of Fig. 9b. A DFT or any other transform is performed in block 931. Subsequent to the actual calculation of the time domain samples, an optional synthesis windowing using a synthesis window is performed. The synthesis window is preferably exactly the same as the analysis window or is derived from the analysis window, for example interpolation or decimation but depends in a certain way from the analysis window. This dependence preferably is such that multiplication factors defined by two overlapping windows add up to one for each point in the overlap range. Thus, subsequent to the synthesis window in block 932, an overlap operation and a subsequent add operation is performed. Alternatively, instead of synthesis windowing and overlap/add operation, any cross fade between subsequent blocks for each channel is performed in order to obtain, as already discussed in the context of Fig. 9a, an artifact reduced decoded signal.

When Fig. 6b is considered, it becomes clear that the actual decoding operations for the mid-signal, i.e., the "EVS decoder" on the one hand and, for the side signal, the inverse vector quantization VQ^{-1} and the inverse MDCT operation (IMDCT) correspond to the signal decoder 700 of Fig. 2.

Furthermore, the DFT operations in blocks 810 correspond to element 810 in Fig. 9b and functionalities of the inverse stereo processing and the inverse time shift correspond to blocks 800, 900 of Fig. 2 and the inverse DFT operations 930 in Fig. 6b correspond to the corresponding operation in block 930 in Fig. 9b.

Subsequently, Fig. 3 is discussed in more detail. In particular, Fig. 3 illustrates a DFT spectrum having individual spectral lines. Preferably, the DFT spectrum or any other spectrum illustrated in Fig. 3 is a complex spectrum and each line is a complex spectral line having magnitude and phase or having a real part and an imaginary part.

Additionally, the spectrum is also divided into different parameter bands. Each parameter band has at least one and preferably more than one spectral lines. Additionally, the parameter bands increase from lower to higher frequencies. Typically, the broadband alignment parameter is a single broadband alignment parameter for the whole spectrum, i.e., for a spectrum comprising all the bands 1 to 6 in the exemplary embodiment in Fig. 3.

Furthermore, the plurality of narrowband alignment parameters are provided so that there is a single alignment parameter for each parameter band. This means that the alignment parameter for a band always applies to all the spectral values within the corresponding band.

Furthermore, in addition to the narrowband alignment parameters, level parameters are also provided for each parameter band.

In contrast to the level parameters that are provided for each and every parameter band from band 1 to band 6, it is preferred to provide the plurality of narrowband alignment parameters only for a limited number of lower bands such as bands 1, 2, 3 and 4.

Additionally, stereo filling parameters are provided for a certain number of bands excluding the lower bands such as, in the exemplary embodiment, for bands 4, 5 and 6, while there are side signal spectral values for the lower parameter bands 1, 2 and 3 and, consequently, no stereo filling parameters exist for these lower bands where wave form matching is obtained using either the side signal itself or a prediction residual signal representing the side signal.

As already stated, there exist more spectral lines in higher bands such as, in the embodiment in Fig. 3, seven spectral lines in parameter band 6 versus only three spectral lines in parameter band 2. Naturally, however, the number of parameter bands, the number of spectral lines and the number of spectral lines within a parameter band and also the different limits for certain parameters will be different.

Nevertheless, Fig. 8 illustrates a distribution of the parameters and the number of bands for which parameters are provided in a certain embodiment where there are, in contrast to Fig. 3, actually 12 bands.

As illustrated, the level parameter ILD is provided for each of 12 bands and is quantized to a quantization accuracy represented by five bits per band.

Furthermore, the narrowband alignment parameters IPD are only provided for the lower bands up to a boarder frequency of 2.5 kHz. Additionally, the inter-channel time difference or broadband alignment parameter is only provided as a single parameter for the whole spectrum but with a very high quantization accuracy represented by eight bits for the whole band.

Furthermore, quite roughly quantized stereo filling parameters are provided represented by three bits per band and not for the lower bands below 1 kHz since, for the lower bands, actually encoded side signal or side signal residual spectral values are included.

Subsequently, a preferred processing on the encoder side is summarized with respect to Fig. 5. In a first step, a DFT analysis of the left and the right channel is performed. This procedure corresponds to steps 155 to 157 of Fig. 4c. In step 158, the broadband alignment parameter is calculated and, particularly, the preferred broadband alignment parameter inter-channel time difference (ITD). As illustrated in 170, a time shift of L and R in the frequency domain is performed. Alternatively, this time shift can also be performed in the time domain. An inverse DFT is then performed, the time shift is performed in the time domain and an additional forward DFT is performed in order to once again have spectral representations subsequent to the alignment using the broadband alignment parameter.

ILD parameters, i.e., level parameters and phase parameters (IPD parameters), are calculated for each parameter band on the shifted L and R representations as illustrated at step 171. This step corresponds to step 160 of Fig. 4c, for example. Time shifted L and R representations are rotated as a function of the inter-channel phase difference parameters as illustrated in step 161 of Fig. 4c or Fig. 5. Subsequently, the mid and side signals are computed as illustrated in step 301 and, preferably, additionally with an energy conversation operation as discussed later on. In a subsequent step 174, a prediction of S with M as a function of ILD and optionally with a past M signal, i.e., a mid-signal of an earlier frame is performed. Subsequently, inverse DFT of the mid-signal and the side signal is performed that corresponds to steps 303, 304, 305 of Fig. 4d in the preferred embodiment.

In the final step 175, the time domain mid-signal m and, optionally, the residual signal are coded as illustrated in step 175. This procedure corresponds to what is performed by the signal encoder 400 in Fig. 1.

At the decoder in the inverse stereo processing, the *Side* signal is generated in the DFT domain and is first predicted from the *Mid* signal as:

$$\widetilde{Side} = g \cdot Mid$$

where *g* is a gain computed for each parameter band and is function of the transmitted Inter-channel Level Difference (ILDs).

The residual of the prediction $Side - g \cdot Mid$ can be then refined in two different ways:

- By a secondary coding of the residual signal:

$$\widehat{Side} = g \cdot Mid + g_{cod} \cdot (Side - \widetilde{g \cdot Mid})$$

where g_{cod} is a global gain transmitted for the whole spectrum

- By a residual prediction, known as stereo filling, predicting the residual side spectrum with the previous decoded *Mid* signal spectrum from the previous DFT frame:

$$\widehat{Side} = g \cdot Mid + g_{pred} \cdot Mid \cdot z^{-1}$$

where g_{pred} is a predictive gain transmitted per parameter band.

The two types of coding refinement can be mixed within the same DFT spectrum. In the preferred embodiment, the residual coding is applied on the lower parameter bands, while residual prediction is applied on the remaining bands. The residual coding is in the preferred embodiment as depict in Fig.1 performs in MDCT domain after synthesizing the residual Side signal in Time Domain and transforming it by a MDCT. Unlike DFT, MDCT is critical sampled and is more suitable for audio coding. The MDCT coefficients are directly vector quantized by a Lattice Vector Quantization but can be alternatively coded by a Scalar Quantizer followed by an entropy coder. Alternatively, the residual side signal can be also coded in Time Domain by a speech coding technique or directly in DFT domain.

1. Time-Frequency Analysis: DFT

It is important that the extra time-frequency decomposition from the stereo processing done by DFTs allows a good auditory scene analysis while not increasing significantly the

overall delay of the coding system. By default, a time resolution of 10 ms (twice the 20 ms framing of the core coder) is used. The analysis and synthesis windows are the same and are symmetric. The window is represented at 16 kHz of sampling rate in Fig. 7. It can be observed that the overlapping region is limited for reducing the engendered delay and that zero padding is also added to counter balance the circular shift when applying ITD in frequency domain as it will be explained hereafter.

2. Stereo parameters

Stereo parameters can be transmitted at maximum at the time resolution of the stereo DFT. At minimum it can be reduced to the framing resolution of the core coder, i.e. 20ms. By default, when no transients is detected, parameters are computed every 20ms over 2 DFT windows. The parameter bands constitute a non-uniform and non-overlapping decomposition of the spectrum following roughly 2 times or 4 times the Equivalent Rectangular Bandwidths (ERB). By default, a 4 times ERB scale is used for a total of 12 bands for a frequency bandwidth of 16kHz (32kbps sampling-rate, Super Wideband stereo). Fig. 8 summarized an example of configuration, for which the stereo side information is transmitted with about 5 kbps.

3. Computation of ITD and channel time alignment

The ITD are computed by estimating the Time Delay of Arrival (**TDOA**) using the Generalized Cross Correlation with Phase Transform (**GCC-PHAT**):

$$ITD = argmax(IDFT(\frac{L_i(f)R^*_i\ (k)}{|L_i(f)R^*_i\ (k)|}))$$

where L and R are the frequency spectra of the of the left and right channels respectively. The frequency analysis can be performed independently of the DFT used for the subsequent stereo processing or can be shared. The pseudo-code for computing the ITD is the following:

```
L =fft(window(l));
R =fft(window(r));
tmp = L .* conj( R );
sfm_L = prod(abs(L).^(1/length(L)))/(mean(abs(L))+eps);
sfm_R = prod(abs(R).^(1/length(R)))/(mean(abs(R))+eps);
sfm = max(sfm_L,sfm_R);
h.cross_corr_smooth = (1-sfm)*h.cross_corr_smooth+sfm*tmp;
tmp = h.cross_corr_smooth ./ abs( h.cross_corr_smooth+eps );
tmp = ifft( tmp );
tmp = tmp([length(tmp)/2+1:length(tmp) 1:length(tmp)/2+1]);
tmp_sort = sort( abs(tmp) );
thresh = 3 * tmp_sort( round(0.95*length(tmp_sort)) );
xcorr_time=abs(tmp(- ( h.stereo_itd_q_max - (length(tmp)-1)/2 - 1 ):- (
h.stereo_itd_q_min - (length(tmp)-1)/2 - 1 )));
%smooth output for better detection
xcorr_time=[xcorr_time 0];
xcorr_time2=filter([0.25 0.5 0.25],1,xcorr_time);
[m,i] = max(xcorr_time2(2:end));
if m > thresh
  itd = h.stereo_itd_q_max - i + 1;
else
  itd = 0;
end
```

Fig. 4e illustrates a flow chart for implementing the earlier illustrated pseudo code in order to obtain a robust and efficient calculation of an inter-channel time difference as an example for the broadband alignment parameter.

In block 451, a DFT analysis of the time domain signals for a first channel (l) and a second channel (r) is performed. This DFT analysis will typically be the same DFT analysis as has been discussed in the context of steps 155 to 157 in Fig. 5 or Fig. 4c, for example.

A cross-correlation is then performed for each frequency bin as illustrated in block 452.

Thus, a cross-correlation spectrum is obtained for the whole spectral range of the left and the right channels.

In step 453, a spectral flatness measure is then calculated from the magnitude spectra of L and R and, in step 454, the larger spectral flatness measure is selected. However, the selection in step 454 does not necessarily have to be the selection of the larger one but this determination of a single SFM from both channels can also be the selection and calculation of only the left channel or only the right channel or can be the calculation of weighted average of both SFM values.

In step 455, the cross-correlation spectrum is then smoothed over time depending on the spectral flatness measure.

Preferably, the spectral flatness measure is calculated by dividing the geometric mean of the magnitude spectrum by the arithmetic mean of the magnitude spectrum. Thus, the values for SFM are bounded between zero and one.

In step 456, the smoothed cross-correlation spectrum is then normalized by its magnitude and in step 457 an inverse DFT of the normalized and smoothed cross-correlation spectrum is calculated. In step 458, a certain time domain filter is preferably performed but this time domain filtering can also be left aside depending on the implementation but is preferred as will be outlined later on.

In step 459, an ITD estimation is performed by peak-picking of the filter generalized cross-correlation function and by performing a certain thresholding operation.

If no peak above the threshold is obtained, then ITD is set to zero and no time alignment is performed for this corresponding block.

The ITD computationcan also be summarized as follows. The cross-correlation is computed in frequency domain before being smoothed depending of the Spectral Flatness Measurement. SFM is bounded between 0 and 1. In case of noise-like signals, the SFM will be high (i.e. around 1) and the smoothing will be weak. In case of tone-like signal, SFM will be low and the smoothing will become stronger. The smoothed cross-correlation is then normalized by its amplitude before being transformed back to time domain. The normalization corresponds to the Phase –transform of the cross-correlation, and is known to show better performance than the normal cross-correlation in low noise and relatively high reverberation environments. The so-obtained time domain function is first filtered for

achieving a more robust peak peaking. The index corresponding to the maximum amplitude corresponds to an estimate of the time difference between the Left and Right Channel (ITD). If the amplitude of the maximum is lower than a given threshold, then the estimated of ITD is not considered as reliable and is set to zero.

If the time alignment is applied in Time Domain, the ITD is computed in a separate DFT analysis. The shift is done as follows:

$$\begin{cases} r(n) = r(n + ITD) \text{ if } ITD > 0 \\ l(n) = l(n - ITD) \text{ if } ITD < 0 \end{cases}$$

It requires an extra delay at encoder, which is equal at maximum to the maximum absolute ITD which can be handled. The variation of ITD over time is smoothed by the analysis windowing of DFT.

Alternatively the time alignment can be performed in frequency domain. In this case, the ITD computation and the circular shift are in the same DFT domain, domain shared with this other stereo processing. The circular shift is given by:

$$\begin{cases} L(f) = L(f)e^{-j2\pi f\frac{ITD}{2}} \\ R(f) = R(f)e^{+j2\pi f\frac{ITD}{2}} \end{cases}$$

Zero padding of the DFT windows is needed for simulating a time shift with a circular shift. The size of the zero padding corresponds to the maximum absolute ITD which can be handled. In the preferred embodiment, the zero padding is split uniformly on the both sides of the analysis windows, by adding 3.125ms of zeros on both ends. The maximum absolute possible ITD is then 6.25ms. In A-B microphones setup, it corresponds for the worst case to a maximum distance of about 2.15 meters between the two microphones. The variation in ITD over time is smoothed by synthesis windowing and overlap-add of the DFT.

It is important that the time shift is followed by a windowing of the shifted signal. It is a main distinction with the prior art Binaural Cue Coding (BCC), where the time shift is applied on a windowed signal but is not windowed further at the synthesis stage. As a consequence, any change in ITD over time produces an artificial transient/click in the decoded signal.

4. Computation of IPDs and channel rotation

The IPDs are computed after time aligning the two channels and this for each parameter band or at least up to a given ipd_max_band, dependent of the stereo configuration.

$$IPD[b] = angle\left(\sum_{k=band_{limits}[b]}^{band_{limits}[b+1]} L[k]\,R^*[k]\right)$$

IPDs is then applied to the two channels for aligning their phases:

$$\begin{cases} L'(k) = L(k)e^{-j\beta} \\ R'(k) = R(k)e^{j(IPD[b]-\beta)} \end{cases}$$

Where $\beta = \text{atan2}(\sin(\text{IPD}_i[b]), \cos(\text{IPD}_i[b]) + c)$, $c = 10^{ILD_i[b]/20}$ and b is the parameter band index to which belongs the frequency index k. The parameter β is responsible of distributing the amount of phase rotation between the two channels while making their phase aligned. β is dependent of IPD but also the relative amplitude level of the channels, ILD. If a channel has higher amplitude, it will be considered as leading channel and will be less affected by the phase rotation than the channel with lower amplitude.

5. Sum-difference and side signal coding

The sum difference transformation is performed on the time and phase aligned spectra of the two channels in a way that the energy is conserved in the Mid signal.

$$\begin{cases} M(f) = \left(L'(f) + R'(f)\right) \cdot a \cdot \sqrt{\frac{1}{2}} \\ S(f) = \left(L'(f) - R'(f)\right) \cdot a \cdot \sqrt{\frac{1}{2}} \end{cases}$$

where $a = \sqrt{\frac{L'^2 + R'^2}{(L'+R')^2}}$ is bounded between 1/1.2 and 1.2, i.e. -1.58 and +1.58 dB. The limitation avoids artefact when adjusting the energy of M and S. It is worth noting that this

energy conservation is less important when time and phase were beforehand aligned. Alternatively the bounds can be increased or decreased.

The side signal S is further predicted with M:

$$S'(f) = S(f) - g(ILD)M(f)$$

where $g(\mathrm{ILD}) = \frac{c-1}{c+1}$, where $c = 10^{ILD_i[b]/20}$. Alternatively the optimal prediction gain g can be found by minimizing the Mean Square Error (MSE) of the residual and ILDs deduced by the previous equation.

The residual signal $S'(f)$ can be modeled by two means: either by predicting it with the delayed spectrum of M or by coding it directly in the MDCT domain in the MDCT domain.

6. Stereo decoding

The Mid signal X and Side signal S are first converted to the left and right channels L and R as follows:

$$L_i[\mathrm{k}] = M_i[k] + gM_i[k], \text{for } band_limits[b] \leq k < band_limits[b+1],$$
$$R_i[\mathrm{k}] = M_i[k] - gM_i[k], \text{for } band_limits[b] \leq k < band_limits[b+1],$$

where the gain g per parameter band is derived from the ILD parameter:

$g = \frac{c-1}{c+1}$, where $c = 10^{ILD_i[b]/20}$.

For parameter bands below cod_max_band, the two channels are updated with the decoded Side signal:

$$L_i[\mathrm{k}] = L_i[k] + cod_gain_i \cdot S_i[k], \text{for } 0 \leq k < band_limits[cod_max_band],$$
$$R_i[\mathrm{k}] = R_i[k] - cod_gain_i \cdot S_i[k], \text{for } 0 \leq k < band_limits[cod_max_band],$$

For higher parameter bands, the side signal is predicted and the channels updated as:

$$L_i[\mathrm{k}] = L_i[k] + cod_pred_i[b] \cdot M_{i-1}[k], \text{for } band_limits[b] \leq k < band_limits[b+1],$$

$$R_i[\mathrm{k}] = R_i[k] - cod_pred_i[b] \cdot M_{i-1}[k], \text{for } band_limits[b] \leq k < band_limits[b+1],$$

Finally, the channels are multiplied by a complex value aiming to restore the original energy and the inter-channel phase of the stereo signal:

$$L_i[\mathrm{k}] = a \cdot e^{j2\pi\beta} \cdot L_i[k]$$

$$R_i[\mathrm{k}] = a \cdot e^{j2\pi\beta - \mathrm{IPD_i[b]}} \cdot R_i[k]$$

where

$$a = \sqrt{2 \cdot \frac{\sum_{k=band_limits[b]}^{band_limits[b+1]} M_i^2[k]}{\sum_{k=band_limits[b]}^{band_limits[b+1]-1} L_i^{\,2}[k] + \sum_{k=band_limits[b]}^{band_limits[b+1]-1} R_i^{\,2}[k]}}$$

where a is defined and bounded as defined previously, and where $\beta = \mathrm{atan2}(\sin(\mathrm{IPD_i[b]}), \cos(\mathrm{IPD_i[b]}) + c)$, and where atan2(x,y) is the four-quadrant inverse tangent of x over y.

Finally, the channels are time shifted either in time or in frequency domain depending of the transmitted ITDs. The time domain channels are synthesized by inverse DFTs and overlap-adding.

Specific features of the invention relate to the combination of spatial cues and sum-difference joint stereo coding. Specifically, the spatial cues IDT and IPD are computed and applied on the stereo channels (left and right). Furthermore, sum-difference (M/S signals) are calculated and preferably a prediction is applied of S with M.

On the decoder-side, the broadband and narrowband spatial cues are combined together with sum-different joint stereo coding. In particular, the side signal is predicted with the mid-signal using at least one spatial cue such as ILD and an inverse sum-difference is calculated for getting the left and right channels and, additionally, the broadband and the narrowband spatial cues are applied on the left and right channels.

Preferably, the encoder has a window and overlap-add with respect to the time aligned channels after processing using the ITD. Furthermore, the decoder additionally has a windowing and overlap-add operation of the shifted or de-aligned versions of the channels after applying the inter-channel time difference.

The computation of the inter-channel time difference with the GCC-Phat method is a specifically robust method.

The new procedure is advantageous prior art since is achieves bit-rate coding of stereo audio or multi-channel audio at low delay. It is specifically designed for being robust to different natures of input signals and different setups of the multichannel or stereo recording. In particular, the present invention provides a good quality for bit rate stereos speech coding.

The preferred procedures find use in the distribution of broadcasting of all types of stereo or multichannel audio content such as speech and music alike with constant perceptual quality at a given low bit rate. Such application areas are a digital radio, internet streaming or audio communication applications.

An inventively encoded audio signal can be stored on a digital storage medium or a non-transitory storage medium or can be transmitted on a transmission medium such as a wireless transmission medium or a wired transmission medium such as the Internet.

Although some aspects have been described in the context of an apparatus, it is clear that these aspects also represent a description of the corresponding method, where a block or device corresponds to a method step or a feature of a method step. Analogously, aspects described in the context of a method step also represent a description of a corresponding block or item or feature of a corresponding apparatus.

Depending on certain implementation requirements, embodiments of the invention can be implemented in hardware or in software. The implementation can be performed using a digital storage medium, for example a floppy disk, a DVD, a CD, a ROM, a PROM, an EPROM, an EEPROM or a FLASH memory, having electronically readable control signals stored thereon, which cooperate (or are capable of cooperating) with a programmable computer system such that the respective method is performed.

Some embodiments according to the invention comprise a data carrier having electronically readable control signals, which are capable of cooperating with a programmable computer system, such that one of the methods described herein is performed.

Generally, embodiments of the present invention can be implemented as a computer program product with a program code, the program code being operative for performing one of the methods when the computer program product runs on a computer. The program code may for example be stored on a machine readable carrier.

Other embodiments comprise the computer program for performing one of the methods described herein, stored on a machine readable carrier or a non-transitory storage medium.

In other words, an embodiment of the inventive method is, therefore, a computer program having a program code for performing one of the methods described herein, when the computer program runs on a computer.

A further embodiment of the inventive methods is, therefore, a data carrier (or a digital storage medium, or a computer-readable medium) comprising, recorded thereon, the computer program for performing one of the methods described herein.

A further embodiment of the inventive method is, therefore, a data stream or a sequence of signals representing the computer program for performing one of the methods described herein. The data stream or the sequence of signals may for example be configured to be transferred via a data communication connection, for example via the Internet.

A further embodiment comprises a processing means, for example a computer, or a programmable logic device, configured to or adapted to perform one of the methods described herein.

A further embodiment comprises a computer having installed thereon the computer program for performing one of the methods described herein.

In some embodiments, a programmable logic device (for example a field programmable gate array) may be used to perform some or all of the functionalities of the methods described herein. In some embodiments, a field programmable gate array may cooperate

with a microprocessor in order to perform one of the methods described herein. Generally, the methods are preferably performed by any hardware apparatus.

The above described embodiments are merely illustrative for the principles of the present invention. It is understood that modifications and variations of the arrangements and the details described herein will be apparent to others skilled in the art. It is the intent, therefore, to be limited only by the scope of the impending patent claims and not by the specific details presented by way of description and explanation of the embodiments herein.

Claims

1. Apparatus for estimating an inter-channel time difference between a first channel signal and a second channel signal, comprising:

a calculator for calculating a cross-correlation spectrum for a time block from the first channel signal in the time block and the second channel signal in the time block;

a spectral characteristic estimator for estimating a characteristic of a spectrum of the first channel signal or the second channel signal for the time block;

a smoothing filter for smoothing the cross-correlation spectrum over time using the spectral characteristic to obtain a smoothed cross-correlation spectrum; and

a processor for processing the smoothed cross-correlation spectrum to obtain the inter-channel time difference.

2. Apparatus of claim 1,

wherein the processor is configured to normalize the smoothed cross-correlation spectrum using a magnitude of the smoothed cross-correlation spectrum.

3. Apparatus of claim 1 or 2,

wherein the processor is configured

to calculate a time-domain representation of the smoothed cross-correlation spectrum or a normalized smoothed cross-correlation spectrum; and

to analyze the time-domain representation to determine the inter-channel time difference.

4. Apparatus of one of the preceding claims,

wherein the processor is configured to low-pass filter the time-domain representation and to further process a result of the low-pass filtering.

5. Apparatus of one of the preceding claims,

wherein the processor is configured to perform the inter-channel time difference determination by performing a peak searching or peak picking operation within a time-domain representation determined from the smoothed cross-correlation spectrum.

6. Apparatus of one of the preceding claims,

wherein the spectral characteristic estimator is configured to determine, as the spectral characteristic, a noisiness or a tonality of the spectrum; and

wherein the smoothing filter is configured to apply a stronger smoothing over time with a first smoothing degree in case of a first less noisy characteristic or a first more tonal characteristic, or to apply a weaker smoothing over time with a second smoothing degree in case of a second more noisy characteristic or a second less tonal characteristic,

wherein the first smoothing degree is greater than the second smoothing degree, and wherein the first noisy characteristic is less noisy than the second noisy characteristic, or the first tonal characteristic is more tonal than the second tonal characteristic.

7. Apparatus of one of the preceding claims,

wherein the spectral characteristics estimator is configured to calculate, as the characteristic, a first spectral flatness measure of a spectrum of the first channel signal and a second spectral flatness measure of a second spectrum of the second channel signal, and to determine the characteristic of the spectrum from the first and the second spectral flatness measure by selecting a maximum value, by determining a weighted average or an unweighted average between the spectral flatness measures, or by selecting a minimum value.

8. Apparatus of one of the preceding claims,

wherein the smoothing filter is configured to calculate a smoothed cross-correlation spectrum value for a frequency by a weighted combination of the cross-correlation spectrum value for the frequency from the time block and a cross-correlation spectral value for the frequency from at least one past time block, wherein weighting factors for the weighted combination are determined by the characteristic of the spectrum.

9. Apparatus of one of the preceding claims,

wherein the processor is configured to determine a valid range and an invalid range within a time-domain representation derived from the smoothed cross-correlation spectrum,

wherein at least one maximum peak within the invalid range is detected and compared to a maximum peak within the valid range, wherein the inter-channel time difference is only determined, when the maximum peak within the valid range is greater than at least one maximum peak within the invalid range.

10. Apparatus of one of the preceding claims,

wherein the processor is configured

to perform a peak search operation within a time-domain representation derived from the smoothed cross-correlation spectrum,

to determine a variable threshold from the time-domain representation; and

to compare a peak to the variable threshold, wherein the inter-channel time difference is determined as a time lag associated with a peak being in a predetermined relation to the variable threshold.

11. Apparatus of claim 10,

wherein the processor is configured to determine the variable threshold as a value being equal to an integer multiple of a value among the largest 10 % of values of the time-domain representation.

12. Apparatus of one of claims 1 to 9,

wherein the processor is configured to determine a maximum peak amplitude in each subblock of a plurality of subblocks of a time-domain representation derived from the smoothed cross-correlation spectrum,

wherein the processor is configured to calculate a variable threshold based on a mean peak magnitude derived from the maximum peak magnitudes of the plurality of subblocks, and

wherein the processor is configured to determine the inter-channel time difference as a time lag value corresponding to a maximum peak of the plurality of subblocks being greater than the variable threshold.

13. Apparatus of claim 12,

wherein the processor is configured to calculate the variable threshold by a multiplication of the mean threshold determined as an average peak among the peaks in the subblocks and a value,

wherein the value is determined by an SNR (signal to noise ratio) characteristic of the first and the second channel signal, wherein a first value is associated with a first SNR value and a second value is associated with a second SNR value, wherein the first value is greater than the second value, and wherein the first SNR value is greater than the second SNR value.

14. Apparatus of claim 13,

wherein the processor is configured to use a third value being lower than the second value in case of a third SNR value being lower than the second SNR value and when a difference between the threshold and a maximum peak is lower than a predetermined value.

15. Method for estimating an inter-channel time difference between a first channel signal and a second channel signal, comprising:

calculating a cross-correlation spectrum for a time block from the first channel signal in the time block and the second channel signal in the time block;

estimating a characteristic of a spectrum of the first channel signal or the second channel signal for the time block;

smoothing the cross-correlation spectrum over time using the spectral characteristic to obtain a smoothed cross-correlation spectrum; and

processing the smoothed cross-correlation spectrum to obtain the inter-channel time difference.

16. Computer program for performing, when running on a computer or a processor, the method of claim 15.