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

(11) EP 3 544 005 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
25.09.2019 Bulletin 2019/39

(21) Application number: **18187597.2**

(22) Date of filing: **06.08.2018**

(51) Int Cl.:
G10L 19/22 *(2013.01)* *G10L 19/032* *(2013.01)*
G10L 19/00 *(2013.01)*

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

(30) Priority: **22.03.2018 EP 18163459**

(71) Applicants:
- **Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.**
  80686 München (DE)
- **Friedrich-Alexander-Universität Erlangen-Nürnberg**
  91054 Erlangen (DE)

(72) Inventors:
- **FISCHER, Johannes**
  91330 Bammersdorf (Eggolsheim) (DE)
- **BÄCKSTRÖM, Tom**
  00100 Helsinki (FI)
- **DAS, Sneha**
  00076 Aalto (IN)

(74) Representative: **Schairer, Oliver**
Schoppe, Zimmermann, Stöckeler
Zinkler, Schenk & Partner mbB
Patentanwälte
Radlkoferstraße 2
81373 München (DE)

(54) AUDIO ENCODER, AUDIO DECODER, AUDIO ENCODING METHOD AND AUDIO DECODING METHOD FOR DITHERED QUANTIZATION FOR FREQUENCY-DOMAIN SPEECH AND AUDIO CODING

(57) An audio encoder for encoding an audio signal, wherein the audio signal is represented in a spectral domain, is provided. The audio encoder comprises a spectral envelope encoder (110) configured for determining a spectral envelope of the audio signal and for encoding the spectral envelope. Moreover, the audio encoder comprises a spectral sample encoder (120) configured for encoding a plurality of spectral samples of the audio signal. The spectral sample encoder (120) is configured to estimate an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the plurality of spectral samples depending on the spectral envelope. Moreover, the spectral sample encoder (120) is configured to encode each spectral sample of the plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

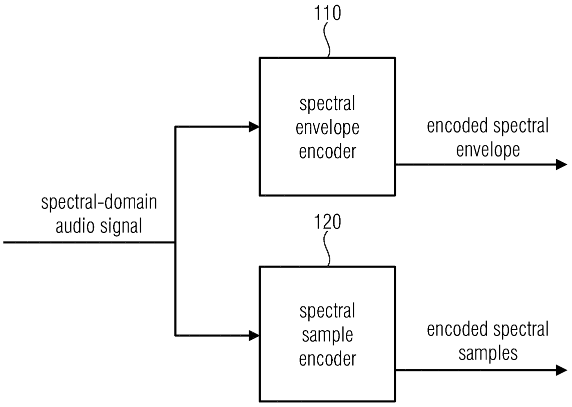

110
spectral envelope encoder
encoded spectral envelope
spectral-domain audio signal
120
spectral sample encoder
encoded spectral samples

Fig. 1a

Printed by Jouve, 75001 PARIS (FR)

Description

[0001] The present invention relates to audio encoding, audio processing and audio decoding, and in particular, to an audio encoder, an audio decoder, an audio encoding method and an audio decoding method for dithered quantization for frequency-domain speech and audio coding.

[0002] A common issue in coding speech and audio in the frequency domain, which appears with decreasing bitrate, is that quantization levels become increasingly sparse. With low accuracy, high-frequency components are typically quantized to zero, which leads to a muffled output signal and musical noise. Bandwidth extension and noise-filling methods attempt to treat the problem by inserting noise of similar energy as the original signal, at the cost of low signal to noise ratio.

[0003] State-of-the art codecs, such as 3GPP Enhanced Voice Services (EVS) and MPEG Unified speech and audio coding (USAC) use frequency domain coding in their intermediate and high bitrate ranges, but revert to time-domain coding at lower bitrates [1], [2], [4]. The reason is that the scalability of frequency-domain codecs in terms of coding efficiency at low bitrates remains a bottleneck even if they provide several other advantages such as low algorithmic complexity. A symptom of the issue is that frequency-domain codecs tend to quantize low-energy areas to zero, which further reduces their energy. This leads to a muffled character in the quantized output, since high-frequency components often have low energy and are thus zeroed (see Fig. 2).

[0004] Fig. 2 illustrates a mean energy of perceptually weighted and normalized MDCT-spectra over the TIMIT database, for original signal (thick line), conventional quantization (dotted), dithered (dashed), as well as dithered in combination with Wiener filtering (crosses) and energy matching (thin line). Quantization was scaled to match a bitrate of 13.2 kbit/s.

[0005] The problem of uniform quantization, in the conventional application, is that if the quantization bins are zero-centered, then the energy of the quantized signal decreases with decreasing accuracy. Alternatively, with off-center quantization the average energy can be retained, but limited in bit-rate to above 1 bit/sample, since the sign has to be transmitted. Moreover, at the extreme, at low bitrates, non-zero values can require so many bits to encode in the entropy coder, that one cannot ever afford to transmit them.

[0006] Entropy coding with uniform quantization therefore does not scale well to bitrates below 1 bit/sample.

[0007] This problem has been addressed in prior works primarily with two approaches. Firstly, one can encode high-frequency regions with bandwidth extension methods, where the objective is to retain the spectral magnitude envelope of the original signal, but sacrifice phase-information and other fine-structure, such that bitrate is limited [1]. Sometimes such methods also copy spectral structures from lower frequencies (copy-up) since the fine-structures are generally similar.

[0008] Secondly, with a method known as noise filling, one can insert noise in areas of the spectrum which have been quantized to zero such that absence of energy is avoided [64]. Both approaches thus aim to retain energy at a similar level as the original signal, but they do not optimize signal-to-noise ratio. A recent improvement, known as intelligent gap filling, combines these methods by using both noise filling and copy-up [65].

[0009] Classical dithering algorithms however also include methods which can retain the signal distribution without reduction in signal to noise ratio [66]. Common dithering methods such as Floyd-Steinberg, are based on error-diffusion or randomization of quantization levels, such that quantization errors can be diffused without loss in accuracy [67]. Alternatively, one can modify quantization bin locations to retain the probability distribution of the original signal even after quantization and coding [68] or use lattice quantization to pack quantization more densely [18].

[0010] These approaches however do not address the issue of very low bitrates, where one cannot afford to encode anything else than the most likely quantization bin. Algebraic coding can be used also at very low bitrates, but its output is also very sparse and it is not applicable on all bitrates [1], [69]. A further alternative would be vector coding, which provides optimal coding efficiency also at very low bitrates. However, vector coding approaches are not easily scalable across bitrates. Moreover, their computational complexity becomes prohibitively high at higher bitrates and if the vector length is high [1], [19]. Vector coding is thus also not a scalable approach.

[0011] It would therefore be highly be appreciated, if improved concepts to address the above-described problems would be provided.

[0012] The object of the present invention is to provide improved concepts for audio encoding and audio decoding. The object of the present invention is solved by an audio encoder according to claim 1, by an audio decoder according to claim 8, by a system according to claim 15, by a method according to claim 16, by a method according to claim 17 and by a computer program according to claim 18.

[0013] An audio encoder for encoding an audio signal, wherein the audio signal is represented in a spectral domain, is provided. The audio encoder comprises a spectral envelope encoder configured for determining a spectral envelope of the audio signal and for encoding the spectral envelope. Moreover, the audio encoder comprises a spectral sample encoder configured for encoding a plurality of spectral samples of the audio signal. The spectral sample encoder is configured to estimate an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the plurality of spectral samples depending on the spectral envelope. Moreover, the spectral sample encoder is configured to encode each spectral sample of the plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

[0014] Moreover, an audio decoder for decoding an encoded audio signal is provided. The audio decoder comprises an interface configured for receiving an encoded spectral envelope of the audio signal and configured for receiving an encoded plurality of spectral samples of the audio signal. Moreover, the audio decoder comprises a decoding unit configured for decoding the encoded audio signal by decoding the encoded spectral envelope and by decoding the encoded plurality of spectral samples. The decoding unit is configured to receive or to estimate an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the encoded plurality of spectral samples. Moreover, the decoding unit is configured to decode each spectral sample of the encoded plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples of the encoded plurality of spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

[0015] Furthermore, a method for encoding an audio signal, wherein the audio signal is represented in a spectral domain, is provided. The method comprises:

- Determining a spectral envelope of the audio signal and for encoding the spectral envelope. And:

- Encoding a plurality of spectral samples of the audio signal.

[0016] Encoding the plurality of spectral samples of the audio signal is conducted by estimating an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the plurality of spectral samples depending on the spectral envelope. Moreover, encoding the plurality of spectral samples of the audio signal is conducted by encoding each spectral sample of the plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

[0017] Furthermore, a method for decoding an encoded audio signal is provided. The method comprises:

- Receiving an encoded spectral envelope of the audio signal and for receiving an encoded plurality of spectral samples of the audio signal. And:

- Decoding the encoded audio signal by decoding the encoded spectral envelope and by decoding the encoded plurality of spectral samples.

[0018] Decoding the encoded audio signal is conducted by receiving or by estimating an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the encoded plurality of spectral samples. Moreover, decoding the encoded audio signal is conducted by decoding each spectral sample of the encoded plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples of the encoded plurality of spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

[0019] Moreover, computer programs are provided, wherein each of the computer programs is configured to implement one of the above-described methods when being executed on a computer or signal processor.

[0020] Embodiments are based on the following findings:

In the prior art, frequency-domain codecs use uniform quantization and arithmetic coding to efficiently encode spectral components, and such entropy codec is near-optimal at high bitrates.

[0021] The combination of uniform quantization and arithmetic coding saturates with decreasing bitrate, such that quality drops significantly below the theoretical accuracy/bitrate limit.

[0022] At low bitrates (<1bit/sample) encoding efficiency decreases; the quantized spectrum becomes increasingly sparse which causes a muffled character to the output sound as well as musical noise. Several methods have been introduced to treat this problem, including noise filling and bandwidth extension.

[0023] Concepts which come close to the theoretical maximum efficiency, but with low computational complexity are provided

Dithering methods have been used in other fields of signal processing, but not so much in speech and audio coding. Dithered coding concepts which use random rotations and 1-bit quantization are employed. In embodiments, such dithered coding concepts are applied in frequency-domain speech and audio coding.

[0024] The dithered coding concepts work well for low bitrates (<1bit/sample) whereas conventional entropy coding concepts work efficiently for high bitrates (>1bit/sample). In embodiments, a hybrid coder is provided, which may, for example, apply a threshold at 1bit/sample to choose between the two codecs. An estimate of the bit-consumption for individual samples may, for example, be extracted from the LPC-based arithmetic codec which is used in 3GPP EVS [48].

[0025] In embodiments, hybrid coding between uniform quantization with arithmetic coding (or other entropy codec, or other quantization) and dithered coding is conducted.

[0026] In particular, for dithered coding, randomization may, e.g., be based on orthonormal rotations and 1-bit coding (sign quantization), following [70].

[0027] In particular, 1-bit coding, where 1-bit coding (sign quantization) may e.g., be used on a select B samples of a vector length of N, to obtain a desired bit-consumption of B bits, which is equivalent with B/N bits per sample.

[0028] According to embodiments, hybrid coding is conducted such that the codec is chosen based on expected bit-rate of samples.

[0029] In particular, a bit-rate may, e.g., be estimated from an envelope model such as the LPC, following [48].

[0030] Optionally, one may, e.g., have multiple thresholds on the bit-rate to split the samples into categories of different accuracy, to allow for noise shaping.

[0031] In embodiments, dithered quantization and/or coding is conducted such that the output is post-filtered after reconstruction.

[0032] In particular, post-filtering based on a minimum mean square error minimization, or based on scaling output energy to match the (estimated) original energy or based on other scaling or optimization or a combination of these may, e.g., be conducted.

[0033] In particular, post-filtering may, e.g., be chosen to optimize a perceptual quality and/or a signal to noise ratio and/or another quality criteria.

[0034] According to embodiments, an audio encoder or a method of audio encoding or a computer program for implementing the method of audio coding using a hybrid coding as described in this application is provided.

[0035] An audio decoder or a method of audio decoding or a computer program for implementing the method of audio decoding using a hybrid decoding as described in this application is provided.

[0036] In particular embodiments, dithered quantization for frequency-domain speech and audio coding relating to a hybrid coding scheme for low-bitrate coding is employed, where low energy samples may, e.g., be quantized using dithering, instead of the conventional uniform quantizer. For dithering, one bit quantization in a randomized sub-space may, e.g., be applied.

[0037] In embodiments, a hybrid coder is provided, which applies a threshold at 1 bit/sample to choose between the two codecs. An estimate of the bit-consumption for individual samples may, e.g., be be extracted from the LPC-based arithmetic codec which is used in 3GPP EVS [48].

[0038] It is to be noted that described concepts equally relate to an apparatus for encoding and also apply for the apparatus for decoding where appropriate.

[0039] Some embodiments are based on using dithering instead of the conventional uniform quantizer. For dithering, one bit quantization in a randomized sub-space is applied.

[0040] In the following, embodiments of the present invention are described in more detail with reference to the figures, in which:

Fig. 1a illustrates an audio encoder according to an embodiment.

Fig. 1b illustrates an audio decoder according to an embodiment.

Fig. 1c illustrates a system according to an embodiment.

Fig. 2 illustrates a mean energy of perceptually weighted and normalized MDCT-spectra over the TIMIT database according to an embodiment.

Fig. 3 illustrates a diagram of the speech and audio encoder according to an embodiment.

Fig. 4 illustrates hybrid coding of spectral components according to an embodiment.

Fig. 5 illustrates collated histograms of K = 10000 vectors of unit variance Gaussian input, quantized by uniform quantization and entropy coding as well as the provided dithered coder (N = 32, _2), with 1 bit/sample according to an embodiment.

Fig. 6 depicts an illustration of performance for a typical speech spectrum at 13.2 kbits/s according to an embodiment.

Fig. 7 depicts results of a subjective MUSHRA listening test, comparing the provided 1 bit dithered quantizers with conventional arithmetic coding, as well as a synthetic dithering serving as an anchor according to an embodiment.

Fig. 8 illustrates differential scores of a subjective MUSHRA listening test, comparing the provided 1 bit dithered quantizers with conventional arithmetic coding, as well as a synthetic dithering serving as an anchor according to an embodiment.

Fig. 9 illustrates a histogram of relative differences in bit-consumptions when using the normal and Laplacian distributions.

Fig. 10 illustrates differential AB scores and their 95% confidence intervals of a comparison listening test.

Fig. 11 illustrates a flow diagram of the randomization process.

Fig. 12 illustrates a structure of an encoder and of a decoder of one node of the distributed speech and audio codec.

Fig. 13 illustrates an encoder/decoder structure of the distributed speech and audio codec with N encoder nodes and a single decoder node.

Fig. 14 depicts an illustration of histograms of the normalized covariance for different orthogonal matrices.

Fig. 15 depicts normalized histograms of the output samples after M consecutive randomizations.

Fig. 16 illustrates a convergence of distribution with increasing number of rotations to the normalized Gaussian and the normalized Laplacian.

Fig. 17 illustrates results of a MUSHRA test, given for different items and an average over items.

Fig. 18 illustrates the difference scores of the performed MUSHRA test of Fig. 17.

[0041] Embodiments of the present invention are based on particular combinations of first coding concepts and of second coding concepts.

[0042] For providing a better understanding of the embodiments of the present invention at first, the first coding concepts and the second coding concepts are described.

[0043] In the following, the first coding concepts are described. One concept or two or more concepts or all concepts of the first coding concepts may be referred to as a first coding rule.

[0044] The first coding concepts relate to arithmetic coding of speech and audio spectra using TCX based on linear predictive spectral envelopes.

[0045] Unified speech and audio codecs often use a frequency domain coding technique of the transform coded excitation (TCX) type. It is based on modeling the speech source with a linear predictor, spectral weighting by a perceptual model and entropy coding of the frequency components. While previous approaches have used neighbouring frequency components to form a probability model for the entropy coder of spectral components, it is proposed to use the magnitude of the linear predictor to estimate the variance of spectral components.

[0046] Since the linear predictor is transmitted in any case, this method does not require any additional side info. Subjective measurements show that the first coding concepts give a statistically significant improvement in perceptual quality when the bit-rate is held constant. Consequently, the first coding concepts have been adopted to the 3GPP Enhanced Voice Services speech coding standard.

[0047] At first, an introduction to the first coding concepts is provided.

[0048] Speech and audio coding technologies applied in modern standards such as MPEG USAC, G.718, AMR-WB+ and, importantly, the ETSI 3GPP Enhanced Voice Services, use multiple modes such as time-domain coding with ACELP and frequency-domain coding with TCX to gain efficient coding for as many signal types as possible [2], [57], [58], [59].

[0049] Generally, time-domain coding provides superior performance for signals with rapidly changing character and temporal events, such as spoken consonants, applause and percussive signals. Coding in the frequency-domain, on the other hand, is more effective for stationary signals such as harmonic music signals and sustained voiced speech sounds.

[0050] In the following, a focus is put on coding in the frequency-domain using models of the spectral envelope. Observe that there are two distinct types of spectral envelope models in classical literature and technologies; First of all, dedicated speech codecs are generally based on linear predictive models, which model the spectral energy envelope using an IIR filter. In contrast, classical audio codecs such as MP3 and the AAC family model the perceptual masking envelope [22], [60].

[0051] While these two envelopes do have many common features - their peaks and valleys are located at the same general frequency areas - the magnitude of peaks and valleys as well as the overall spectral tilt are very different. Roughly speaking, masking envelopes are much smoother and exhibit smaller variations in magnitude then the energy envelopes.

[0052] AAC-type codecs use the perceptual masking model to scale the spectrum such that the detrimental effect of quantization on spectral components has perceptually the same expected magnitude in every part of the spectrum [22]. To allow efficient coding of the perceptual spectrum, these codecs then apply entropy coding of the frequency components. For higher efficiency, the arithmetic coder can use the neighbouring spectral components to determine the probability distribution of the spectral components, such as in USAC [57], [61].

[0053] Speech codecs on the other hand use energy envelopes as a signal model and apply a perceptual weighting filter, much like the perceptual masking model, on top.

[0054] The assumptions rely on the fact that the spectral envelope, as described by the linear predictive model, provides information of the energy envelope of the spectrum. Since it thus describes the energy distribution of the spectral components, it can be used to describe their probability distributions. This distribution can, in turn, be used to design a highly efficient arithmetic coder for the spectral components. Since the linear predictive model is generally transmitted also for TCX frames, this spectral envelope information comes without additional side-information. In contrast to AAC-type codecs, an explicit source model may, e.g., be used in the form of the linear predictor, and in difference to TCX-type codecs, an adaptive probability distribution may, e.g., be used for the arithmetic codec derived from the magnitude of the linear predictor. A signal adaptive model of the probability distributions of spectral components is described based on the linear predictive model. The goal is to obtain a fixed bit-rate arithmetic coder applicable in speech and audio codecs which use linear predictive modeling. Moreover, the objective is to design a generic method which is efficient on a variety of bit-rates and bandwidths.

[0055] The encoder of the first coding concepts may, e.g., be configured to conduct three steps. First, the perceptually weighted linear predictor may, e.g., be used as a model for the shape of the perceptually weighted spectrum. Since this envelope does not contain information of the signal magnitude, the envelope is scaled such that the expected bit-consumption of a signal, whose variance follows the envelope, matches the desired bit-rate. Second, the actual perceptually weighted spectrum is scaled and quantized such that the bit-rate matches the desired bit-rate, when using the envelope model. Finally, one can then encode the spectrum with an arithmetic coder. The decoder can then apply the same scaling of the envelope to decode the spectral lines.

[0056] Now, a modeling of the probability distribution of the first coding concepts is described.

[0057] Let A_k^{-1} be the samples of the discrete Fourier transform of hat linear predictive model which describes the short-time temporal structure and thus the spectral envelope of a signal spectrum S_k. The filter residual can be obtained by multiplying S_k with A_k to obtain the residual $X_k = A_k S_k$. Given that A_k is an efficient model of S_k, then X_k will be the spectrum of a white-noise signal. It follows that the expected energy of every frequency component k is constant $\sigma_x^2 = \mathcal{E}\left[|X_k|^2\right]$.

[0058] Conversely, the expectation of the energy of the perceptual weighted signal S_k is

$$\sigma_{s,k}^2 = \mathcal{E}\left[|S_k|^2\right] = \sigma_x^2 \left|A_k^{-1}\right|^2 . \qquad (1.1)$$

[0059] For perceptual weighting of quantization errors, prior to quantization, the spectrum is weighted by a perceptual masking model W_k. It follows that the expected energy of the weighted spectrum $Y_k = W_k S_k$ is

$$\sigma_{y,k}^2 = \mathcal{E}\left[|W_k S_k|^2\right] = \sigma_x^2 \left|W_k A_k^{-1}\right|^2 . \qquad (1.2)$$

[0060] This relation quantifies the relative energy of spectral components and can be used as an estimate of their relative variance in the design of models of the probability distribution of weighted spectral components.

[0061] A probability distribution model for the individual spectral components can then be chosen. The most obvious candidates are either the normal or the Laplacian distribution, since they are both simple to implement and commonly known to be fairly accurate models of speech and audio signals. To determine which distribution fits the approach better, the following experiment has been conducted.

[0062] As material for the test, the 16 critical items of speech, mixed and music material used in the standardization of MPEG USAC [57] were used. The material was resampled to 12.8kHz and windowed with sine-windows, and transformed to the frequency domain using the MDCT, with 256 frequency bands and full-overlap. Linear predictive models of order 16 were estimated for each frame using a Hamming window of length 30ms, centered to align with the MDCT windows. The spectrum S_k of each frame was whitened in the frequency domain with the corresponding envelope model A_k, and perceptually weighted with W_k, to obtain the perceptually white spectrum $\tilde{X}_k = A_k W_k S_l$. The spectrum was then scaled by the standard deviation, to remove the effect of signal gain, whereby variance of the scaled signal was unity. The bit-consumption of both distributions was then estimated in each frame by

$$b = -\sum_k \log_2 p(\tilde{X}_k), \tag{1.3}$$

where the probability with the distributions is estimated

$$p(\tilde{X}_k) = \sqrt{2}\exp\left(-\frac{|\tilde{X}_k|}{\sqrt{2}}\right)$$

for the Laplacian distribution and

$$p(\tilde{X}_k) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{|\tilde{X}_k|^2}{2}\right)$$

for the normal distribution.

[0063] Finally, the bit-consumption of all frames in the data for both the normal and Laplacian distributions were calculated.

[0064] The histogram of the relative difference in bit-consumption between the distributions is shown in Fig. 9.

[0065] Fig. 9 illustrates a histogram of relative differences in bit-consumptions bn and bl when using the normal and Laplacian distributions, respectively, as measured by r = bn□bl (bn+bl)=2 .

[0066] It can be observed that in a majority of cases, modeling with a normal distribution requires more bits than with a Laplacian. Moreover, importantly, encoding with the Laplacian never yields a large increase in bit-consumption, whereas it sometimes gives a big reduction in bit-rate. On average, the Laplace distribution gave a bit-consumption 6% smaller than the normal distribution. In addition, the highest gains where found for frames with stationary harmonic content, where the spectrum is sparse, which is also exactly the frames where TCX should be used. It is thus concluded that in an application, a Laplacian is better than a normal distribution.

[0067] Since the target is to design a codec with a fixed bit-rate, it is then necessary to scale the weighted envelope such that its expected bit-consumption matches the target bit-rate. A Laplacian random variable with variance σ^2 has a bit-consumption expectation of 1 + $\log_2 (e\sigma)$, where e is Euler's number. The variance for the magnitude of the weighted spectrum is then

$$\sigma_{y,k}^2(\gamma) = \gamma^{-2}|W_k A_k^{-1}|^2, \tag{1.4}$$

where γ is the scaling coefficient. It follows that the expectation of bit-consumption of a spectrum with N coefficients is

$$B = N + \sum_{k=0}^{N-1} \log_2 \left(e\gamma\sigma_{y,k} \right) . \tag{1.5}$$

[0068] The above equation (1.5) can then be solved for:

$$\gamma = \frac{2^{\frac{B-1}{N}}}{e} \left(\prod_{k=0}^{N-1} \sigma_{y,k} \right)^{-1/N} . \tag{1.6}$$

[0069] It follows that when the spectral envelope is scaled by the given by equation (1.6), then the expected bit-rate is B for signals whose magnitude follows the envelope.

[0070] In the following, coding of spectral components of the first coding concepts is described.

[0071] The arithmetic coder of the first coding concepts is based on a Laplacian distribution, which is equivalent with a signed exponential distribution. To simplify the process, one can thus first encode the absolute magnitude of spectral lines and for non-zero values then also encode their sign with one bit.

[0072] The absolute magnitude can thus be encoded using an arithmetic coder with an exponential distribution. If a spectral component is quantized to an integer value *q*, then the original value has been in the interval $\left[q - \frac{1}{2}, q + \frac{1}{2}\right]$, whose probability is given by

$$\begin{aligned} p(|\widehat{Y}_k|) &= e^{-\lambda(q-\frac{1}{2})} - e^{-\lambda(q+\frac{1}{2})} \\ &= \left[e^{+\frac{\lambda}{2}} - e^{-\frac{\lambda}{2}} \right] e^{-\lambda q} = Ce^{-\lambda q}, \end{aligned} \tag{1.7}$$

where $\lambda = \sqrt{2}/\sigma_k$, the scalar σ_k is the standard deviation of the kth frequency component and $C = e^{+\frac{\lambda}{2}} - e^{-\frac{\lambda}{2}}$ is a constant. Using this probability, a standard arithmetic coder for the spectral lines [62] can then be designed. With this arithmetic coder each frequency component can then be consecutively encoded from the low to high frequencies.

[0073] Since speech and audio spectra are often dominated by low-frequency content, at low bit-rates it is a common that large sections of the high-frequency spectra is sparse or zero.

[0074] To improve efficiency, trailing zeros can therefore be discarded. The decoder may, e.g., be set to decode the spectra until the maximal bit-consumption is reached and set any remaining spectral components to zero. The encoder thus encodes the spectrum up to the last non-zero frequency. Informal experiments showed that with this approach the number of encoded frequency components is often reduced by 30 to 40%.

[0075] To use the allocated bits efficiently, the spectrum must be scaled to obtain the highest accuracy which can be encoded with the available bits. As an initial guess for the scaling of the input spectrum Y_k, one can scale it such that its energy matches the energy of the envelope that was scaled to the desired bit-rate. The optimal scaling can then be determined in a rate-loop implemented as a binomial search. Informal experiments showed that the best scaling can usually be found within 5 iterations.

[0076] In the following, implementation details of the first coding concepts are described.

[0077] The objective of the arithmetic coder is to encode the spectrum into the bit-stream as efficiently as possible. To encode and decode the bit-stream, the numerical operations must be implemented with fixed-point operations such that differences in numerical accuracy across different platforms will not change the outcome. A 14 bit integer representation stored in a 16 bit integer variable can be chosen, which allows for a sign bit and avoids problems in the last bit due to differences in rounding.

[0078] A related problem in implementation of the arithmetic coder is that when the standard deviation σ_k. is small compared to the quantized magnitude $|\hat{Y}_k|$, then the probability $p(|\hat{Y}_k|$ will become so small that it is rounded to zero in the 14 bit representation. In other words, the numerical accuracy is exhausted and one would be unable to encode such values. To overcome this problem, the memory-lessness property of exponential distributions can be used. This principle states that

$$p(|\widehat{Y}_k| > r + s \mid |\widehat{Y}_k| > r) = p(|\widehat{Y}_k| > s). \qquad (1.8)$$

[0079] Here, if r is the threshold above which the numerical representation saturates, then the probability $p(|\hat{Y}_k| > r)$ can first be encoded and then, one can continue by encoding the probability $p(|\hat{Y}_k| - r)$. By successively increasing r, it can be guaranteed that numerical saturation is avoided.

[0080] In the implementation of the rate-loop, the bit-stream should be output only for the final scaling coefficient. To reduce computational complexity, it is then sufficient to only estimate bit-consumption in the rate-loop and invoke the arithmetic coder only with the optimal scaling.

[0081] The bit-consumption of the spectrum for a given scaling coefficient γ can be efficiently estimated as follows. Let the quantized value be $\hat{Y}_k = \text{round}(\gamma Y_k)$ whereby the total bit-consumption $B(\gamma)$ can be calculated by

$$B(\gamma) = \sum_k - \log_2 p(\widehat{Y}_k) = - \log_2 \prod_k p(\widehat{Y}_k). \qquad (1.9)$$

[0082] Instead of calculating the logarithm for every component, a logarithm of the product $\Pi_k p(\hat{Y}_k)$ can thus be calculated, and thereby, computational complexity can be reduced. It should be noted, however, that the product can become a very large number, whereby the calculation of the product shall be implemented in a representation where it can be guaranteed that an overflow cannot occur.

[0083] Once the best possible scaling for quantization has been achieved, one can encode the spectral lines. However, since the rate-loop described above only approximates actual arithmetic coding, it may sometimes happen that the bit-budget is slightly exceeded. A safe-guard against this improbable event is to reduce the magnitude of the last encoded line until the actual bit-consumption remains within the budget.

[0084] Now, conducted experiments regarding the first coding concepts are described.

[0085] To determine the performance of the first coding concepts, a subjective AB comparison listening test [63] was performed. For this test, the codec was implemented according to the first coding concepts in a candidate version of the 3GPP Enhanced Voice Services speech and audio codec [2]. As comparison, an arithmetic coder derived from the MPEG Unified Speech and Audio Coder was used, which forms a probability model of spectral lines using neighbouring lines [57], [61]. The codecs were operating at a fixed bit-rate of 8 kbps in the wide-band mode. Bandwidth extension was encoded normally but its output was disabled to concentrate on differences of the core bandwidth. The double-blind AB comparison test was performed by 7 expert listeners in a silent room with high-quality headphones.

[0086] The test included 19 samples of mono speech, music, and mixed material. The differential results of the compare test are illustrated in Fig. 10, where a positive score indicates a subjective improvement of the codec of the first coding concepts over the reference codec using the arithmetic coder of MPEG USAC.

[0087] Fig. 10 illustrates differential AB scores and their 95% confidence intervals of a comparison listening test measuring the performance of the arithmetic coder in comparison to the coder from MPEG USAC at 8 kbps for wide-band signals.

[0088] These results show that the mean AB score indicates a statistically significant improvement of 0.48 points with 95% confidence. Moreover, no item had a statistically significant reduction in perceptual quality, but 6 out of 19 items had a statistically significant improvement. Note that it is remarkable that the improvement is statistically highly significant even with a limited number of listeners.

[0089] Now, a discussion and conclusions regarding the first coding concepts follow:

The first coding concepts provide concepts for modeling the probability distribution of perceptually weighted frequency components of speech and audio signals using a model of the spectral envelope and the perceptual weighting function. Frequency-domain codecs based on the TCX concept model the spectral envelope using linear prediction, from which an estimate of the perceptual masking curve can be obtained. Since the linear predictor is transmitted in any case, the first coding concepts can be applied without transmission of additional side-information.

[0090] The first coding concepts use the spectral envelope model as a model of the speech source for construction of a probability model for the entropy coder. In contrast, conventional methods have used preceding frequency components to predict the magnitude of the current component [57], [61]. The conventional methods thus use an implicit source model, whereas the first coding concepts model the source explicitly.

[0091] It should be noted that in estimation of the bit-consumption of the spectral envelope, the theoretical entropy of Laplacian distributions may, e.g., be used, which is accurate only when the quantization accuracy is very high. The bias at lower bit-rates is due to fact that when a spectral line is quantized to zero, its sign does not need to be transmitted, whereby 1 bit is saved. When a significant part of the spectrum is quantized to zero, a rather large number of bits is saved, whereby the bit-consumption estimates are too high. Informal experiments, however, show that a more accurate estimate of bit-consumption increases complexity significantly, but that the impact on overall bit-consumption was marginal. Such more accurate and complex estimates of the bit-consumption of spectral envelopes could thus be avoided. Whereas here, only the spectral envelope associated with the linear predictor may, e.g., be used, it should be observed that speech codecs regularly use also other information which can be used to estimate spectral magnitude. Namely, long term prediction (LTP) is regularly used to model the fundamental frequency. The long term predictor can be used to model the comb-structure of spectra with a dominant fundamental frequency.

[0092] The presented results demonstrate that the first coding concepts improve perceptual quality at low bit-rates when the bit-consumption is kept constant. Specifically, subjective measurements with an AB test showed a statistically significant improvement. The coding scheme of the first coding concepts can thus be used to either increase quality at a given fixed bit-rate or decrease the bit-rate without losing perceptual quality. The presented approach is applicable in all speech and audio codecs which employ a frequency-domain coding of the TCX type, where a model of the spectral envelope is transmitted to the decoder. Such codecs include standards such as MPEG USAC, G.718 and AMR-WB+ [57], [58], [59]. In fact, the method has already been included in the ETSI 3GPP Enhanced Voice Services standard [2].

[0093] In the following, the second coding concepts are described. One concept or two or more concepts or all concepts of the second coding concepts may be referred to as a second coding rule.

[0094] The second coding concepts relate to fast randomization for distributed low-bitrate coding of speech and audio.

[0095] Efficient coding of speech and audio in a distributed system requires that quantization errors across nodes are uncorrelated. Yet, with conventional methods at low bitrates, quantization levels become increasingly sparse, which does not correspond to the distribution of the input signal and, importantly, also reduces coding efficiency in a distributed system.

[0096] The second coding concepts provide a distributed speech and audio codec design, which applies quantization in a randomized domain such that quantization errors are randomly rotated in the output domain. Similar to dithering, this ensures that quantization errors across nodes are uncorrelated and coding efficiency is retained.

[0097] The second coding concepts achieve fast randomization with a computational complexity of $O(N \log N)$. The presented experiments demonstrate that the randomizations of the second coding concepts yield uncorrelated signals, that perceptual quality is competitive, and that the complexity of the second coding concepts is feasible for practical applications.

[0098] Now, an introduction to the second coding concepts is provided.

[0099] The second coding concepts describe an alternative method using dithered quantization, where the input signal is multiplied with a random rotation before quantization such that the quantization levels are obscured when the rotation is inverted for the output signal [70]. A similar approach is applied in the Opus codec [71], though only with Givens-rotations without permutations. A simple quantization such as 1 bit quantization is applied to obtain high performance at low complexity and very low bitrates [8]. The randomized quantization methods of the second coding concepts are unique in the way they allow quantization and coding of signals without a lower limit on bitrate, while simultaneously providing the best SNR per bit ratio. Concurrently, the second coding concepts provide the benefits of vector coding by joint processing of multiple samples, without significant penalty on complexity.

[0100] Digital compression of speech signals for transmission and storage applications, known as speech coding, is a classic topic within speech processing and modern speech coding standards achieve high efficiency in their respective application scenarios [1], [2], [3], [4], [5]. Though these standards are high-fidelity products, they are constrained to configurations with a single encoder. Designs which would allow using the microphones of multiple independent devices could improve signal quality, and moreover, it would allow a more natural interaction with the user-interface as the speaker would no more be constrained to a single device. If the codec can flexibly use all available hardware, then the user does not need to know which devices are recording, releasing mental capacity from attention to devices to the communication at hand.

[0101] Such an ideal user interface is possible only if devices cpoperate in the speech coding task. The aim is that, through cooperation, the acoustic signal should be flexibly captured and transmitted to one or several decoders or fusion centers. Clearly, one thus requires a distributed speech and audio codec. A distributed system however requires substantial modifications to existing codec designs; most notably, 1) the increase in algorithmic complexity due to added nodes becomes an issue and 2) concepts to ensure that each transmitted bit conveys unique information are needed. Specifically, conventional codec designs are based on an intelligent encoder and a simple decoder, whereby a majority of the computational complexity resides at the encoder. In a distributed system, the overall computational complexity increases linearly with both the encoder complexity as well as the number of nodes, whereby it is important to keep encoder complexity low to be able to use a large number of nodes. If one can move the main intelligence of the codec from the encoder to the decoder, then the overall complexity of the system would thus be much lower.

[0102] A majority of speech coding standards are based on the code-excited linear prediction (CELP) paradigm [1]. It is based on an analysis-by-synthesis loop, where the perceptual quality of a large number of different quantizations are evaluated to optimize output quality. While this approach provides the best quality for bitrate trade-off, its usefulness in distributed coding is limited by its computational complexity, rigid design and error propagation issues. Frequency domain methods, on the other hand, have not yet reached quite the same efficiency as CELP, but it is clear that coding in the frequency domain is computationally much simpler. Moreover, since most noise attenuation and spatial filtering methods are defined in the frequency domain [6], it will be straightforward to implement such methods if frequency-domain coding is used. Another issue is the amount of interaction between encoder nodes. Clearly communication between nodes requires some administration, whereby it would be beneficial to minimize or even avoid interaction between nodes if possible. If nodes only transmit data and interaction between nodes is avoided, then the overall system structure is simpler and the computational complexity required for said interaction is avoided. The question is thus whether interaction between nodes is required for coding efficiency. At high bit-rates (say 100 kbits s), very small differences in the signal, such as variations in delay, background noise or sensor noise, would be sufficient to make quantization noise between nodes uncorrelated [7], whereby each node will provide unique information. However, experience with lower bit-rates (such as 10 kbits s) has shown that low-energy areas of the signal are often quantized to zero, whereby quantization errors are perfectly correlated with the input signal [1]. Multiple nodes transmitting zeros would then convey no new information about the signal, whereby there is little advantage of using multiple devices.

[0103] An objective is to develop a distributed codec for speech and audio, where coding efficiency is optimized, but which can also be applied on any device, including simple mobile or even wearable devices with limited CPU and battery resources. For an overall design for such a method, see [8], [9]. The approach is based on randomizing the signal before quantization, such that quantization error expectations between devices are uncorrelated. It is assumed that the randomizer uses a random-number generator whose seed is communicated from the encoder to the decoder either offline or sufficiently seldom that it has a negligible effect on the bitrate. Overall, the randomizer in this context is similar to dithering and was inspired by the 1 bit quantization used in compressive sensing [10], [11].

[0104] Randomization has several distinct benefits in the codec of the second coding concepts:

1) In low-bitrate coding (below 10 kbits s), only a limited number of quantization levels are available which can be encoded with the available bits. With decreasing bitrate, the quantized signal distribution thus becomes increasingly sparse, granular and biased. By applying a randomization and its inverse before and after quantization, respectively, one can hide the undesirably sparse structure. Similarly as dithering, one can thus retain the signal distribution, without any penalty on the signal to noise ratio.

2) In perceptual audio coding a too low number of quantization levels for speech and audio signals leads to artifacts known as musical noise, where components which sporadically appear and disappear become coherent sound objects in their own right. A standard approach for avoiding musical noise in audio codecs is noise filling, a method similar to dithering, where noise is added to spectral areas quantized to zero [12]. In the described approach, by quantization in randomized domain, errors become incoherent and the reduction in SNR caused by noise filling can be avoided.

3) Randomization of the signal can also work as a component of encryption [13]. It provides diffusion in a similar way as the Hill cipher [14], that is, it distributes the contribution of the input vector evenly onto the bitstream.

4) In distributed coding,two alternative approaches can be applied. If nodes encode separate subspaces (or cosets in the vocabulary of distributed coding), then increasing the bitrate by 1 bit/sample yields a 6 dB improvement in quality. The downside in an ad-hoc network is that then the nodes have to negotiate which subspaces/cosets to transmit, which requires extract bandwidth, administration and increases the risk of eavesdropping. Moreover, spatio-temporal filtering such as beamforming is impossible if cosets do not overlap. On the other hand, if nodes are independent, then one can get a 3 dB improvement from a doubling of the number of nodes, as long as the quantization errors are uncorrelated [6]. Randomized quantization yields quantization errors which are uncorrelated, whereby in difference to conventional quantization, the 3 dB improvement when doubling the number of microphones is achieved.

5)When transmission errors corrupt some transmitted bits, conventional entropy coders (e.g. arithmetic coding) will loose all data after the first corrupted bit. With entropy coding according to the second coding concepts which is enabled by the randomizing scheme, there is no serial dependency of bits, whereby the signal can be reconstruct also when some bits are corrupted. The transmission errors will then be visible as noise in the reconstructed signal, which can be attenuated by conventional noise attenuation methods such as Wiener filtering [15], [16]. The details of these benefits are discussed in the following.

[0105] In comparison, conventional single-device quantization and coding methods all suffer from some constraints. Entropy coders such as Huffman or arithmetic coders with uniform quantization do not scale to very low bitrates (less than 2 bits/sample), since the output signal distribution becomes unnaturally sparse [1], [17]. Lattice quantization does reduce quantization error, but does not solve the issue of granularity at low bitrates and moreover, it does not easily lend itself to arbitrary probability distributions [1], [18]. Vector coding is optimal in accuracy and does not suffer much from sparsity, but computational complexity is high and it is challenging to encompass variable bitrates [19].

[0106] Moreover, achieving robustness to transmission errors is difficult with all of the above methods. Distributed source coding methods, on the other hand, do provide methods for optimal joint encoding [20], [21]. These methods make use of the correlation between the microphone signals by binning in order to reduce the rates using the results from distributed source coding. The most common way of implementing this is by using error-correcting codes, but in a practical setup due to complexity and other considerations, such implementations will be highly suboptimal, leading to higher complexity without significant gains. For these reasons, no focus is put on binning. Specifically, distributed source coding methods for ad-hoc networks are not available, which would include signal-adaptive perceptual modeling, which would solve the above mentioned problems with sparsity and which would simultaneously apply signal-adaptive source models. All of these properties are required features of a speech and audio codec to retain a competitive performance for the single-channel case.

[0107] The fields of speech and audio coding [1], [22], and distributed source coding e.g. [20], [21], are well-understood topics. Work on wireless acoustic sensor networks has however not yet made it to consumer products of the type discussed here [23]-[25]. Some works have addressed higher bitrates [26], or with the assumption that nodes are fully connected and co-operating [7], [27], though both approaches lack a perceptual model. Some early work do apply a perceptual model [28], [29], but do not include other elements of mainstream speech and audio codecs, such as source modeling. A somewhat similar problem is design of hands-free devices and the associated beamforming tasks [6], which can be applied in a distributed system [30], though methods usually require accurate synchronization of nodes to give competitive performance [31]. Similar methods have also been tried for hearing-aids, though distributed source coding gave there only a negligible improvement [32]. More generally, distributed processing of speech and audio can be also applied for classification and recognition tasks [33]-[35]. In comparison, the distributed scheme of the second coding concepts is now a complete system, with the exception of speech source modeling, may, e.g., be incorporated into the system, though this has not been studied intensely ([36], [37]).

[0108] Specifically, while a rudimentary entropy codec is not applied, in the current experiments, a model of the spectral magnitude envelope of harmonic structure nor spatio-temporal correlations is not included. While many of the prior works have admirably fine theoretical analyses, here, the design has no barriers against creating a practical system whose single-channel quality is near state-of-the-art while simultaneously providing a benefit when increasing the number of nodes. That is, by including the above mentioned source models, the single-channel performance should be similar to the performance of the TCX mode of the EVS standard [2]. It should be emphasized that not all details of best practices are included in lossy distributed source coding, since it has been opted to take incremental steps in order to retain the single-channel quality near the state-of-the-art.

[0109] The current contribution addresses the complexity bottleneck of the system, namely, the randomization and its inverse. The algorithmic complexity of generic quantization together with randomization is $\mathcal{O}(N^2)$. Moreover, at the decoder, the original approach required the inversion of an $N \times N$ matrix, which gives a complexity of $\mathcal{O}(N^3)$ using Gaussian elimination [38].

[0110] Algorithmic complexity shall be improved, and the randomization properties and coding efficiency shall be retained or improved as much as feasible. In addition to distributed coding, randomization and decorrelation are used also in many other fields of speech, audio and signal processing in general. For example, in upmixing of audio signals from a low to a higher number of channels, concepts for generating uncorrelated source signals are needed [39].

[0111] Randomization methods of the second coding concepts may find application in any such applications which require low-complexity methods for generation of uncorrelated signals. Notably, moreover, randomization can be used in single-channel codecs to diffuse unnaturally sparse quantization levels which appear at low bitrates.

[0112] In the following, randomized entropy coding of the second coding concepts is described.

[0113] The main objective of coding is to quantize and encode an input signal $x \in \mathbb{R}^{N\times 1}$, with a given number of bits B, such that it can be decoded with highest accuracy possible. The objective of randomization, on the other hand, is to make sure that the resynthesized signal retains the continuous distribution of the original signal and that the quantization error is uncorrelated Gaussian noise. In other words, whereas quantization by construction yields a signal with a discrete distribution, a signal which follows a similar distribution as the original signal shall be obtained. Moreover, the aim is that if the signal is quantized and coded at multiple nodes, then the quantization errors of the outputs would be uncorrelated.

[0114] Fig. 11 illustrates a flow diagram of the randomization process, where P is a random (orthogonal) matrix and $Q[\cdot]$ is a quantizer.

[0115] It is needed to introduce randomness in the signal without reducing accuracy of the reconstruction. To achieve such randomness, three aspects of linear randomizing transforms shall be discussed (see Fig. 11). First, it is shown that orthonormal projections are optimal for the error criterion. Secondly, random permutations for diffusing information across the input vector are discussed. Finally, it is demonstrated that low-order random rotations can be used in block-matrices to diffuse quantization levels.

[0116] As objective design criteria for the randomization, the following concepts are used:

1) The accuracy of reconstruction is measured by the minimum mean square error min $E[\|e\|^2]$, where $e = x - \hat{x}$ is the quantization error and $\hat{x}$ is the quantized signal.

2) To measure the correlation between randomized vectors, the normalized covariance between the original x and its randomized counterpart Px is measured. If the randomization is effective, then the normalized covariance should behave like the normalized covariance $\frac{x^T y}{\|x\|\,\|y\|}$ between two uncorrelated signals *x and y.* Specifically, the mean of the normalized covariance should be zero and its variance $\frac{1}{N}$ for Gaussian signals (see below for details).

3) The accuracy with which the distribution of the output signal follows the distribution of the input signal can be measured with the Kullback-Leibler (KL) divergence. However, since analytic analysis of divergences is difficult, the KL-divergence is applied only experimentally below.

4) Finally, algorithmic complexity is characterized with the Big-O notation.

[0117] Now, orthonormal randomization of the second coding concepts is considered.

[0118] To introduce randomness in the signal without compromising accuracy, the signal is multiplied with a random orthogonal matrix P before quantization. At the decoder, multiplication with the inverse P^T is conducted (see Fig. 11). It is important that the transform is orthonormal, since it preserves signal energy, such that the transform provides perfect reconstruction and a minimal white noise gain. Specifically, let the output signal be $\hat{x} = P^T Q[Px]$ where $Q[\cdot]$ signifies quantization.

[0119] If the quantization is perfect, $u = Q[u]$, then $\hat{x} = P^T Q[Px] = P^T Px = x$. In other words, the randomizing transform P does not corrupt the signal in any way; all information is retained and the signal can be perfectly reconstructed. Moreover, if the quantization error is $v = u - \hat{u}$, then the output error energy will be

$$\begin{aligned}\|e\|^2 &:= \|x - \hat{x}\|^2 = \left\|x - P^T Q[Px]\right\|^2 \\ &= \|Px - Q[Px]\|^2 = \|v\|^2.\end{aligned} \tag{2.1}$$

[0120] In other words, since for an orthogonal P: $\|e\|^2 = \|Pe\|^2$, it follows that quantization error energy in the transform domain is exactly equal to the error in the output domain. If the constraint would be relaxed from orthogonal matrices, and consider matrices P whose samples are uncorrelated and have unit variance, then $E[P^T P] = I$ would result, and the matrices would be orthogonal with respect to the expectation.

[0121] However, as is known from random matrix theory [40], the eigenvalue distribution of such matrices is significantly off unity. It follows that P^T would not be an accurate inverse of the transform but one would have to use the actual inverse P^{-1} or a pseudo-inverse. Consequently, the inverse transform P^{-1} would emphasize the quantization errors corresponding to small eigenvalues, whereby the inverse transform would, on average, increase the error energy. Orthogonal random matrices are thus preferable to random matrices since they provide perfect reconstruction and unit white noise gain. Orthogonal matrices have also computational benefits with respect to noise attenuation at the decoder side, see below.

[0122] Now, random permutations of the second coding concepts is considered.

[0123] Permutations are computationally fast operations which correspond to orthogonal matrices, whereby their use in randomization is interesting. As a matter of definition, a permutation perfectly diffuses input information over the output vector if the output location of all input samples are uniformly distributed over the whole vector. Specifically, if an input sample ξ_h has an input location $h \in [0, N - 1]$, then the probability that it will appear at location k after permutation is $p(k) = \frac{1}{N}$, whereby the input and output locations are not correlated. The covariance c_P of the original signal x and the permuted signal

$$c_P = x^T y = x^T P x = \sum_{k=1}^{N} \xi_k \xi_{q(k)}. \qquad (2.2)$$

where $q(k)$ is the permutation function. If defining the set of fixed points of the permutation as $S = \{k|k = q(k)\}$, that is, this is the set of samples which do not move due to the permutation, whereby

$$c_P = \sum_{k\in S} |\xi_k|^2 + \sum_{k\notin S} \xi_k \xi_{q(k)}. \qquad (2.3)$$

[0124] While the expectation of the latter sum is zero $E[\xi_k\xi_h] = 0$, for $k \neq h$, the former sum has a non-negative expectation. It follows that the expectation of the covariance has a non-negative bias $E[cp] \geq 0$, which is in contradiction with the objective. The set S is, however, small when N is large, whereby the deviation is not large. Nevertheless, for the sake of completeness, the problem can be remedied.

[0125] Let $\Lambda_\pm$ be a diagonal matrix whose diagonal elements are randomly chosen as ± 1. Clearly this matrix is also orthogonal. One can thus apply a randomization $y = \Lambda_\pm Px$, whereby the correlation is

$$c_P = x^T y = x^T \Lambda_\pm P x = \sum_{k=1}^{N} \pm \xi_k \xi_{q(k)}. \qquad (2.4)$$

[0126] If both signs have equal probability, then clearly $E[\pm\xi_k\xi_{q(k)}] = 0$ and $E[c_P] = 0$ as required. The combination of random signs and permutations thus decorrelates the signal in the sense that the covariance has zero mean and one simultaneously achieves perfect diffusion. Moreover, multiplication by $\Lambda_\pm$ can be readily generalized to orthogonal block-matrices, which are discussed below.

[0127] Random permutations can be easily generated at algorithmic complexity $\mathcal{O}(N \log N)$ [41]. A heuristic approach is for example to apply a sort algorithm, such as merge sort, on a vector of random uncorrelated samples. The sort operation then corresponds to a permutation, which is uniformly distributed over the length of the vector.

[0128] Now, blockwise random rotations are considered.

[0129] Multiplication with matrices has in general an algorithmic complexity of $\mathcal{O}(N^2)$. To reduce complexity, consider $N \times N$ random orthogonal block-diagonal matrix rotations B of the

$$B = \begin{bmatrix} Q_1 & & & 0 \\ & Q_2 & & \\ & & \ddots & \\ 0 & & & Q_K \end{bmatrix}. \quad (2.5)$$

where K is the number of blocks and Q_k are random orthogonal matrices of size $N_k \times N_k$ such that $\sum_{k=1}^{K} N_k = N$. The complexity of the transform is $\mathcal{O}(\sum_{k=1}^{K} N_k^2)$. Clearly the random sign operation $\Lambda_\pm$ is a block-matrix of this form with $N_k = 1$ and $K = N$.

[0130] Specifically, consider size 2 × 2 orthogonal matrices Q of the form

$$Q = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}. \quad (2.6)$$

[0131] The related covariance is

$$\begin{aligned} c_Q := x^T Q x &= \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}^T \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \\ &= (\xi_1^2 + \xi_2^2)\cos\alpha. \end{aligned} \quad (2.7)$$

[0132] If α is uniformly distributed on $\alpha \in [0, 2\pi]$, then clearly c_Q has zero mean, $E[c_Q] = \|x\|^2 E[\cos\alpha] = 0$ as desired. Moreover, if $y = Qx = [\eta_1, \eta_2]^T$, then the parameters η_k follow the arcsine-distribution on the range $|\eta_k|^2 \le 2\|x\|^2$.

[0133] In other words, by applying 2 × 2 blocks of random rotations on the quantized signal ^u, one can diffuse the quantization levels to make the distribution of the output less sparse. Note that since the outputs of the 2 × 2 transforms have symmetric distributions, signrandomization by $\Lambda_\pm$ becomes redundant and can be omitted. The 2 × 2 matrices were chosen since they are simple to implement and any larger orthonormal rotation can be implemented as a combination of 2 × 2 rotations.

[0134] Diffusion of the quantization levels is however not yet complete; the arcsine-distribution is less spiky than the distribution of the quantized signal, but it is still far from the normal distribution.

[0135] To obtain an output distribution which better resembles the normal distribution, a sequence of permutations P_k and block rotations B_k is therefore applied as

$$P = \prod_{k=1}^{M} B_k P_k. \quad (2.8)$$

where each block rotation B_k is of the form in equation (2.5) with 2 × 2 rotations Q_k of the form in equation (2.6). Each consecutive randomization $B_k P_k$ will then further diffuse the output distribution. A number of rotations M are experimentally determined such that the output distribution is sufficiently diffused. However, the algorithmic complexity of applying the above rotation will in any case be $\mathcal{O}(MN)$, while generation of such permutations has complexity $\mathcal{O}(N \log N)$.

[0136] Now, conventional randomization methods with respect to the second coding concepts are considered.

[0137] Generating random orthogonal matrices, with uniformly distributed rotations is not as easy as it would seem. With 2 × 2 matrices such as those in equation (2.6) one can get uniform distribution if α is uniformly distributed on [0, 2π], however, with $N > 2$ such heuristic approaches are not as simple anymore. The reason is easy to understand in the 3 × 3 case, where uniform rotations along each axis would yield a higher concentration of points around the poles of the sphere.

[0138] In the general case, however, the problem is equivalent with choosing N random points on the unit *N*-sphere, such that the corresponding unit vectors are orthonormal. One can thus choose *N* random vectors of size $N \times 1$ from a distribution with spherical symmetry, and orthogonalize the set of vectors. A numerically stable and efficient algorithm which generates such orthonormal vectors is the QR-algorithm [38]. Specifically, at first, an $N \times N$ matrix X with uncorrelated and normally distributed samples with zero mean and equal variance is generated.

[0139] Secondly, the QR-algorithm is applied to obtain an orthogonal matrix P_{QR}. The overall algorithmic complexity of this approach is $\mathcal{O}(N^2)$ [38], [41].

[0140] A simplification of the QR-algorithm is to apply Householder transformations with each of the columns of X [42]. While this approach is efficient for small matrices, in informal experiments it is found that for large matrices, the simplification does not provide a uniform distribution. The subgroup algorithm presented in [43], [44] has also been tried. Though this algorithm is faster than the Householder-based algorithm by a factor of 2, unfortunately however, it suffers from the same issues as the Householder based algorithm. The QR-algorithm applied on random matrices thus remains the high-quality and -complexity method for reference.

[0141] Now, algorithmic complexity of the second coding concepts is considered.

[0142] In application of each of the orthogonal matrices of the second coding concepts, there are two sources of algorithmic complexity; one which emerges from generation of the matrix and a second which is related to application of the matrix and its inverse. Furthermore, one should evaluate storage requirements for the generated matrices. In each case, it shall be assumed that one has access to a pseudo-random number generator, which produces pseudo-random sequence of scalars ξ_k with independent and identically distributed values from the uniform distribution on $\xi_k \in$ [0, 1].

[0143] The simplest case is generation of random signs for the matrix $\Lambda_\pm$. For each of the *N* diagonal samples, a random sign with equal probability is needed, which can be obtained by thresholding ξ_k at 0.5. Given the sequence ξ_k the algorithmic complexity for generation and application is thus $\mathcal{O}(N)$.

[0144] For the block rotations, the number of 2 × 2 blocks is $N/2$, whereby $N/2$ random scalars are needed to generate the matrix B. Application of B involves $N/2$ multiplications by a 2 × 2 matrix at complexity $\mathcal{O}(2N)$, as well as $N/2$ evaluations of cos α and sin α at $\mathcal{O}(N)$, though evaluation of trigonometric functions can have a high constant multiplier for the complexity.

[0145] Application of permutations is straightforward; it is essentially a mapping of sample indices, whereby it does not involve computations other than moving operations, $\mathcal{O}(N)$. Here, the permutation indices and the permuted vector have to be stored, whereby the storage requirements are $\mathcal{O}(2N)$. Generation of permutations can be applied by sorting a vector of random scalars ξ_k with, for example, the merge sort algorithm [41]. It requires also a storage of $\mathcal{O}(2N)$, but not at the same time as the application of the permutation, whereby it does not add to the overall storage requirements.

[0146] The algorithmic complexity for generating the permutation is $\mathcal{O}(N \log N)$ [41]. To generate a random orthogonal matrix, the QR-algorithm can be applied with arbitrary accuracy with an algorithmic complexity of $\mathcal{O}(N^2)$ and storage $\mathcal{O}(N^2)$ [38]. Application of the randomization and its inverse are then simply multiplications by a matrix and its transpose, both at complexity $\mathcal{O}(N^2)$. It however requires N^2 random scalars as input, whereby also the complexity of generating pseudo-random numbers becomes an issue. Moreover, the random values at the input should have rotational symmetry, whereby the uniformly distributed scalars ξ_k variables are not sufficient. It is thus necessary to apply a transform such as the inverse cumulative normal distribution on ξ_k to obtain normally distributed variables, which comes at a considerable extra computational cost. Each of the above algorithms assume that one has access to a sequence of pseudo-random numbers ξ_k. If one chooses to generate the randomization on-line, then the complexity of generating pseudo-random numbers shall be considered. The algorithmic complexity of generating pseudo-random numbers is in general linear $\mathcal{O}(N)$ with the number N of scalars to be generated [45]. A commonly used generator is the Mersenne-twister, though there are also lower-complexity versions available [46], [47]. The trade-off is that if the random sequence is not generated on-line, then it needs to be stored. In any case, it is assumed that the seed of the random sequence is communicated either off-line, or sufficiently seldom such that it does not induce a significant penalty on the bit-rate.

[0147] In summary, algorithmic complexity of generating random matrices is $\mathcal{O}(MN\log N)$, while their application has $\mathcal{O}(MN)$, where M is the number of iterations (typically M = 4) and N is the vector length. If the random coefficients are not generated on-line but stored, then a storage of 32 M N is necessary.

[0148] A typical speech and audio codec, such as the TCX mode of the EVS, would use a step of 20 ms between windows [1], [2], whereby the spectra would be of length N = 256 at a sampling rate of 12.8 kHz, N = 320 at 16 kHz or N = 882 at 44.1 kHz.

[0149] A typical frequency domain codec will have no components which require a complexity more than $\mathcal{O}(N\log N)$. Since the randomization of the second coding concepts is also $\mathcal{O}(N\log N)$, in terms of algorithmic complexity, this corresponds to conventional TCX codecs. The complexity bottleneck thus returns to the rate-loop of the entropy codec [1], [48].

[0150] In the following, applications in distributed coding of speech and audio of the second coding concepts are described.

[0151] As a demonstration of the randomizer of the second coding concepts, the randomized quantizer was applied for coding of the fine-spectral structure in a distributed speech and audio codec. The overall structure is similar to that of the TCX mode in 3GPP EVS [2] and the implemented codec structure is illustrated in Fig. 12.

[0152] Fig. 12 illustrates a structure of the (a) encoder and (b) decoder of one node of the distributed speech and audio codec. Dashed boxes indicate modules which were not included in current experiments to facilitate isolated evaluation of randomization coefficients, while working memory must be always at least 2N coefficients.

[0153] First, the MDCT time-frequency transform and half-sine windowing on the input signal [22] was applied to obtain spectral representations x_k of each time-frame at a node k. Here the window length was 20 ms with 10 ms overlap, and the sampling rate was 12.8 kHz.

[0154] The sampling rate was chosen to match the core-rate of EVS in wide-band mode [2]. In EVS, the remaining bandwidth is coded with bandwidth-extension methods, to obtain a total sampling rate of 16 kHz.

[0155] Then, the signal envelope and perceptual model is analyzed, using the LPC-based approach as in EVS [1], [2]. The signal is then perceptually weighted and multiplied with random rotation matrices to obtain randomized vectors. As a last step of the encoder, the signal is quantized and coded.

[0156] A fixed-bitrate entropy coder with 2 bits/sample was used as follows: the distribution was split into four quantization cells such that each cell had equal probability and each input sample was quantized to the nearest quantization cell. The quantization levels are thus fixed and the system does not require a rate-loop.

[0157] This corresponds to entropy coding the spectral fine structure alone with a bitrate of 25.6 kbits/s. Perceptual envelopes and signal gain are usually transmitted with a rate in the range 2-3 kbits/s, whereby the overall bitrate is approximately 28 kbits s. This does not strictly qualify as a low-bitrate codec, but on the other hand, an explicit source model or a rate-loop is not implemented, here. The inventors' experience with the EVS codec suggests that inclusion of a source model (such as a fundamental frequency model) and a proper rate-loop, would reduce bitrate to below 10 kbits s without reduction in perceptual quality, whereby this experiment is representative for low-bitrate coding. Conversely, the bitrate was chosen such that it roughly corresponds to the accuracy achieved in TCX in the EVS standard well below 10 kbit/s.

[0158] At the decoder, the operations are reversed, with the exception of the perceptual model estimation, which is assumed to be transmitted. One should focus on the performance of the randomizer, whereby in the current application, an explicit source model, rate-loop or quantization the perceptual model was not included. The quality of the output signal could be further enhanced with noise attenuation techniques such as Wiener filtering [6], [49]. Here, however, noise attenuation was omitted such that tuning parameters can be avoided and the comparison of different methods can be kept fair.

[0159] Fig. 13 illustrates an encoder/decoder structure of the distributed speech and audio codec with N encoder nodes and a single decoder node.

[0160] To demonstrate performance in a multi-node scenario, a distributed codec was implemented as illustrated in Fig. 13, where each individual encoder node follows the configuration of Fig. 12.

[0161] As described above, the decoder can then contain the inverse randomizations and a "merge & enhance" -block can implement Wiener filtering independently from the randomization.

[0162] As long as the quantization errors are orthonormal and quantization accuracy is uniform, one would not gain anything from joint processing of the randomization and enhancement, whereby independent blocks give optimal performance. It was decided to not implement Wiener filtering here, since it would present additional perceptual tuning factors, whereby the design of a fair comparison would be difficult. Instead merely the mean of the two channels was used and only the objective quality of the two-channel case was considered.

[0163] In the following, conducted experiments with respect to the second coding concepts are described.

[0164] To evaluate the randomization methods of the second coding concepts, objective experiments corresponding to the performance measures described above were performed as well as subjective perceptual experiments with the distributed speech and audio codec described above.

[0165] Now, statistical properties of the second coding concepts are described.

[0166] To quantify the influence that randomization has on the coding error magnitude, $N \times N$ matrices with N = 100 were created, such that 1) the matrices P_o were orthogonal $P_o^T P_o = I$ and 2) the matrices P_r were orthogonal with respect to the expectation $E[P_r^T P_r] = I$. Specifically, a random number generator was used to produce uncorrelated, normally distributed samples with zero mean and variance $\frac{1}{N}$, which form the entries of P_r. It follows that $E[P_r^T P_r] = I$. By applying the QR-algorithm on P_r, a random orthogonal matrix is then obtained, which is defined as P_o.

[0167] K = 1000 vectors x of length N are generated, whose samples are uncorrelated and follow the normal distribution.

[0168] Each vector was randomized with the two matrices $u_o = P_o x$ and $u_r = P_r x$, quantized with the sign quantization $\hat{u}o$ =sign(u_o) and $\hat{u}_r$ = sign(u_r) and the randomization was reverted by $\hat{x}_o = P_o^T \hat{u}_o$ and $\hat{x}_r = P_r^T \hat{u}_r$. As reference, quantization without randomization is used as x_{ref} = sign(x). Finally, each of the vectors $x_o.x_r$ and x_{ref} were individually scaled for minimum error, and the quantization error energy for each vector was calculated.

[0169] The results of this experiment are listed in Table I.

Table I:

Method	Orth	Rand	None
SNR (dB)	4.52	2.25	4.52

[0170] Table I illustrates a signal-to-noise ratio (SNR) of sign quantization with orthonormal randomization (Orth), randomization with a random matrix (Rand), and without randomization (None).

[0171] Clearly randomization with the random matrix P_r yields a much lower SNR upon reconstruction, whereas randomization with the orthogonal matrix P_o has no influence on accuracy. The results thus exactly follow the results above and randomization by orthogonal matrices should always be used.

[0172] To determine the efficiency of randomization in terms of decorrelation, normalized covariances $\lambda = \frac{x^T A x}{\|x\|^2}$ are then calculated for the different orthogonal matrices A.

[0173] Fig. 14 depicts an illustration of histograms of the normalized covariance $\lambda = \frac{x^T A x}{\|x\|^2}$ for different $N \times N$ orthogonal matrices: The random permutation matrix P, the random sign matrix $\Lambda_\pm$, the combination $\Lambda_\pm P$ as well as the block-matrix (with random 2 × 2 rotations) in combination with permutation BP, evaluated over K = 10 000 matrices. The dashed line indicates the theoretical distribution of normalized covariance between random Gaussian vectors, scaled to fit to each histogram.

[0174] In particular, Fig. 14 illustrates the histogram of K = 1000 iterations of the normalized covariances for matrices of different sizes N = {4, 8, 256} as well as for random permutations P, random signs $\Lambda_\pm$ and random permutations with random signs $P\Lambda_\pm$ As a reference, the theoretical distribution illustrated with a dashed line is used (see below for details).

[0175] It was observed that the distribution of the random permutation P follows the theoretical distribution (dashed line) at higher N. However, it is biased to positive values especially at lower N.

[0176] The random sign $\Lambda_\pm$ yields a covariances whose variance (width of histogram) is higher than the theoretical distribution. Clearly neither method is sufficient alone in decorrelating the signal.

[0177] The combination of random signs and a permutation $\Lambda_\pm P$ performs much better in that the bias to positive values is removed and the variance is similar to the theoretical distribution. At N = 4, however, it was observed that the input signal x cannot be treated as a random variable anymore, but since samples of x frequently get multiplied with itself (though with random sign), peaks were obtained in the histogram corresponding to ±1 and 0. The situation is further improved by replacing the random signs with block-matrices B, where the 2 × 2 blocks are calculated with uniformly distributed angles a. The peaks at ±1 and 0 for N = 4 have almost completely disappeared and the histogram nicely follows the theoretical distribution. Overall, it was found that the decorrelation performance of randomization improves with increasing vector length, as well as when using a combination of at least two orthogonal matrices.

[0178] The third objective performance measure is the ability of randomization to diffuse the quantization levels in the output signal. Specifically, the aim is that the distribution of the output is transformed from a sparse distribution to something which resembles the input distribution. It has already been found that application of random permutations and block-matrix rotations is an effective combination, whereby it is an aim is to evaluate how many such pairs have to be applied to get proper diffusion. To that end it is defined

$$P_M = \prod_{k=1}^{M} B_k P_k, \qquad (2.9)$$

where Bk and Pk are random block-matrix rotations and permutations, respectively. Fig. 15 illustrates the output histogram when applying sign-quantization and the inverse rotation PTM for different values of M. The use of sign quantization has been chosen here, since it is the worst-case in terms of sparsity.

[0179] In particular, Fig. 15 depicts normalized histograms of the output samples after M consecutive randomizations, as well as the theoretical distributions of normalized Gaussian (NNorm, dashed line) and Laplacian (NLap, gray line), for a vector of length N = 16.

[0180] It can be observed in Fig. 15 that the original quantized signal has a sparse distribution, where all samples are ±1, but each consecutive randomization makes it resemble more the normalized Gaussian distribution (dashed line). Note that the normalized Gaussian is here the reference distribution, since the normalized covariance has a limited range (see below for details).

[0181] At M = 4 iterations the histogram has already converged to a unimodal distribution and at M = 16 the histogram is very close to the normalized Gaussian distribution. The rate of convergence depends, however, on the vector length N.

[0182] As a final test of statistical properties, the rate of convergence to the normalized Gaussian distribution was tested with increasing number of iterations and different vector lengths N. As a measure of convergence, the Kullback-Leibler divergence between the normalized Gaussian distribution and the histogram of the output was used. As reference, randomization based on the QR-algorithm was used. The results are illustrated in Fig. 16(a).

[0183] Fig. 16 illustrates a convergence of distribution with increasing number of rotations M to (a) the normalized Gaussian and (b) the normalized Laplacian, as measured by the Kullback-Leibler divergence, for different vector lengths N. As a reference, randomization with the QR-algorithm is depicted with crosses "×," representing a high-complexity and high-performance target level.

[0184] It can be seen that convergence is faster for the shorter vectors, as is to be expected, since in a large space one needs more rotations to reach all possible dimensions. The performance of the QR algorithm is illustrated with crosses '×' and it can be seen that after 16 iterations, for all vector lengths N, the randomizers of the second coding concepts have more or less reached the diffusion of the QR algorithm. In fact, for N = 4 and N = 8, the performance saturates already after 5 and 7 iterations, respectively.

[0185] It is however clear that speech signals are not normally distributed, but it can often be assumed that spectral components follow the Laplace distribution [50], illustrated by the gray line in Fig. 15. Adding further rotations will not reduce the distance to a normalized Laplacian distribution, but it will saturate at some point, as is illustrated in Fig. 16(b). The divergence between the obtained histograms and the target distribution levels off after a few iterations. Moreover, when applying noise attenuation at the decoder, such as Wiener filtering, the distribution will be further modified. It can therefore be concluded that as few as *M* = 4 iterations should be sufficient to diffuse the quantization levels of a speech signal.

[0186] Now, applications in a speech and audio codec of the second coding concepts are described.

[0187] To evaluate the performance of randomization in a practical application, the speech and audio codec was calculated according to the second coding concepts, whose generic structure was described above as follows. As mentioned before, the LPC-based perceptual model was copied as-is from EVS [2]. The perceptually weighted frequency representation was then quantized with randomization using the QR-algorithm, the low-complexity randomization of the second coding concepts (equation (2.9)) with M = 4 iterations, as well as without randomization.

[0188] For the randomized signals, an assumption of Gaussian distribution and for the signal without randomization was used, and the Laplacian distribution was used. Previous experiments have shown that the Laplacian works best for speech signals [48], [50]. Above it has, however, been shown that randomized signals are closer to Gaussian, whereby the choice of distributions is well-warranted. Informal experiments confirmed these choices as the best in terms of perceptual SNR. Here, the perceptual SNR refers to the signal to noise ratio between the perceptually weighted original and quantized signals [22].

[0189] As test-material, 6 samples (3 male and 3 female) from the TIMIT corpus [51] were randomly chosen. For each coded sample, the window-wise perceptual SNR in decibel was calculated, and the mean perceptual SNRs of respective methods was calculated, which are listed in Table II.

Table II:

Method	Perceptual SNR (dB)	
	Single node	Two nodes
None	4.08	4.08
QR	8.79	11.67
Proposed botrule	6.42	8.02

[0190] Table II illustrates a perceptual SNRs of each evaluated method.

[0191] Even though all methods use entropy coding with the same bitrate, rather large differences in perceptual SNR are obtained. The QR method is over 4.8 dB better than no randomization, and the low-complexity of the second coding concepts falls in between the two. Informal experiments show that an increase in the number of iterations used for creating the randomization of the second coding concepts, will improve the SNR. The number of iterations is therefore directly proportional to complexity and SNR. It should be noted, however, that source modeling has not been applied explicitly here (such as that in [48]), which would most likely increase the performance of all methods, but especially the version without randomization. The obtained results should therefore be treated as provisional results until a source model has been implemented.

[0192] To evaluate the perceptual influence of the randomization on the quantization noise, a MUSHRA listening test was conducted [52] (Regarding MUSHRA, see also [56]). Thirteen subjects, aged between 22 and 53, were asked to evaluate the quality of the different approaches. Seven of the thirteen test persons referred to themselves as expert listeners.

[0193] As test items, the same six sentences of the TIMIT database were used as above. For each item, five conditions were used: no randomization, the fast randomization approach of the second coding concepts and as an upper bound, randomization using the QR approach, as well as a 3.5 kHz low pass signal as a lower anchor, and the hidden reference, in accordance with the MUSHRA standard.

[0194] The results of the listening test are presented in Fig. 17.

[0195] Fig. 17 illustrates the results of the MUSHRA test, given for the different items, and an average over all items. The reference was omitted as it was always rated to 100.

[0196] The results show that there is a clear trend that randomization improves the perceived quality, as both the QR and the fast randomization approach are rated higher than the approach without randomization.

[0197] Moreover, with the exception of item 3, the QR approach has high scores.. The coding quality of all methods is in the same range as the anchor, which is arguably low, even for a low bitrate codec. However, since the experiments did not include proper source modeling, the perceptual results overall should be treated as preliminary. In any case, for all items combined, the 95% confidence intervals do not overlap, and there is a clear difference between all three conditions under test, where the second coding concepts perform on average about 20 MUSHRA points better than the conventional, and the QR approach can improve the quality by approximately 15 points.

[0198] To determine whether there is a statistically significant difference between the ratings of the second coding concepts (Median = 42.5) and the lower anchor (Median = 40), hypotheses testing was applied. Since a Shapiro-Wilk test of the score differences ($W = 0.917, p < 0.01$) as well a visual inspection of Q-Q-plots indicated non-normally distributed data, a Wilcoxon signed rank test was performed which indicated no significant difference (V = 1469, p = 0.87) between the anchor and the second coding concepts.

[0199] However, it is unclear whether a comparison between the second coding concepts and lower anchor is relevant anyway, since the characteristics of the distortions in the two cases are very different, rendering a comparison difficult, and as a source model has not yet been included, the absolute quality level was rather arbitrarily chosen. The anchor thus serves only as way to roughly characterize the absolute quality level used in the experiment.

[0200] Fig. 18 illustrates the difference scores of the performed MUSHRA test, where the second coding concepts were used as a reference point. The lower anchor and the hidden reference were omitted.

[0201] The difference scores in Fig. 18 support the findings of the above analysis. Taking the second coding concepts as the reference point, the second coding concepts performed always significantly better than the conventional. Moreover, with the exception of item 3, the QR approach performed good. It is unclear why QR does not have an advantage for item 3, but it is suspected this is merely a statistical outlier.

[0202] In any case, the low-complexity second coding concepts are always better than no randomization, which was a target. This argument also validates the choice of not using source modeling; by source modeling, quantization accuracy can be improved, but the experiments show that the perceptual quality of a codec can be improved by randomization even with a fixed quantization accuracy.

[0203] Finally, to determine how well quantization errors are decorrelated, the randomization concepts were applied on two independent encoders (without difference in delay and without background noises) and took the mean of the outputs. In theory, taking the mean of two signals with uncorrelated noises should increase SNR by 3 dB. From Table II it can be seen that randomization with the QR-algorithm almost reaches this target level, with an improvement of 2.88 dB. The low-complexity randomizer of the second coding concepts achieves an improvement of 1.6 dB. It is thus again a compromise between complexity and quality, as the higher-complexity QR-method gives better SNR than the low-complexity randomizer of the second coding concepts. In a real-life scenario, one can expect to see higher numbers, since any differences in acoustic delay and background/sensor noises would further contribute to decorrelate the quantization errors.

[0204] In the following, conclusions with respect to the second coding concepts are presented.

[0205] Quality of speech and audio coding can be improved in terms of both signal quality and ease of interaction with the user-interface by including, in the coding process, all connected hardware which feature a microphone. For this purpose, a distributed speech and audio codec design may, e.g., be used which is based on randomization of the signal before quantization [8]. The complexity bottle-neck of the codec of the second coding concepts, that is, the randomizer, is considered.

[0206] The low-complexity randomizer of the second coding concepts may, e.g., be based on a sequence of random permutations and 2 × 2 block-rotations. Experiments show that by successive randomizations, high-accuracy decorrelation is obtained, such that the covariance of the original and the randomized signal behaves like uncorrelated signals, and such that the quantization levels of the output signal are diffused.

[0207] The randomization of the second coding concepts has multiple benefits for low-bitrate coding, distributed coding, perceptual performance, robustness and encryption. Experiments confirm these benefits by showing that randomization improves perceptual SNR and subjective quality. Though inclusion of a randomizer shows here an SNR improvement of 2.4 dB, this benefit is expected to be reduced when a proper source model is included. However, it is shown that if quantization errors are randomized, taking the mean of signals improves SNR as much as 2.8 dB, whereby one can always improve quality by adding more microphones.

[0208] The algorithmic complexity of the randomizer of the second coding concepts is $\mathcal{O}(N \log N)$, where the main complexity is due to generation of random permutations. If the permutations are generated off-line the overall complexity is $\mathcal{O}(MN$, where M is the number of iterations. Typically M = 4 is sufficient, whereby complexity is $\mathcal{O}(N)$. Storage requirements are in all cases $\mathcal{O}(N)$. It is believed that the complexity of the encoder therefore becomes viable even on low-performance nodes, such as wearable devices. Generation of the randomizer requires a sequence of pseudo-random numbers. It is assumed that the seed of the pseudo-random number generator is either known at both the encoder and decoder or seldom communicated as side-info.

[0209] Overall, here, the distributed speech and audio codec takes a large step forward to become a full system. Only source modeling was here omitted from a full system, due to space constraints. The randomizer is, however, a necessary and important part of the overall design, whereby finding a low-complexity solution was crucial.

[0210] In the following, a distribution of normalized generalized Gaussians of the second coding concepts are described.

[0211] When a signal which follows a Gaussian or Laplacian distribution is normalized by its norm, its range becomes limited.

[0212] Consequently, normalization of a signal changes its distribution and in the following, the form of such distributions shall be determined. In interest of generality, the generalized normal distribution is considered, which includes both the Gaussian and Laplacian distributions as special cases.

[0213] Suppose *x* is an $N \times 1$ vector whose entries x_k are uncorrelated and follow the generalized normal distribution with zero mean and equal variance σ^2 :

$$f(x_k) = \frac{p}{2b\Gamma(1/p)} \exp\left(-\left|\frac{x_k}{b}\right|^p\right). \qquad (2.10)$$

where the scaling factor is $b = \sigma^2 \sqrt{\frac{\Gamma(1/p)}{\Gamma(3/p)}}$ and where $\Gamma(\cdot)$ is the Gamma function [53].

[0214] By normalizing *x* with its $\mathcal{L}_p$-norm $\|x\|_p$, a new random variable $y = \frac{x}{\|x\|_p}$ is obtained, which is closely related to *x* but does not follow the generalized normal distribution. In particular, in difference to x, the entries y_k of *y* have a limited range

$$\sum_{k=1}^{N} |y_k|^p = 1. \qquad \text{whereby} \qquad y_k \in [-1. +1]. \qquad (2.11)$$

[0215] To derive the marginal distributions of y_k, one can start by considering the entries x_k of x. Let $\gamma_k = 2\left|\frac{x_k}{b}\right|^p$, whereby one can find the distribution of y_k by substitution of variables

$$f(\gamma_k) = 2f(x_k)\frac{dx_k}{d\gamma_k} = \frac{\gamma^{\frac{1}{p}-1} e^{-\frac{\gamma_k}{2}}}{b\Gamma(1/p)} \sim \chi^2\left(\frac{2}{p}\right). \qquad (2.12)$$

[0216] In other words, y_k follows the Chi-squared distribution with $\frac{2}{p}$ degrees of freedom. Then, it is defined

$$\lambda_k := |y_k|^p = \frac{|x_k|^p}{\|x\|_p^p} = \frac{|x_k|^p}{|x_k|^p + \sum_{h\neq k} |x_h|^p}. \qquad (2.13)$$

[0217] Since the x_k's follow the generalized normal distribution, then $|x_k|^p$ and $\Sigma_{h\neq k}|x_h|^p$ will follow the Chi-squared distribution with $\frac{2}{p}$ and $(N-1)\frac{2}{p}$ degrees of freedom, respectively. Ratios such as λ_k of Chi-squared distributed variables will follow the Beta-distribution with parameters $\alpha = \frac{1}{p}$ and $\beta = \frac{N-1}{p}$, or specifically [54], Section 4.2.

$$f(\lambda_k) = \frac{\Gamma\left(\frac{N}{p}\right)}{\Gamma\left(\frac{1}{p}\right)\Gamma\left(\frac{N-1}{p}\right)}\lambda_k^{\frac{1}{p}-1}(1-\lambda_k)^{\frac{N-1}{p}-1}. \tag{2.14}$$

[0218] Moreover, from equation (2.11) it follows that

$$\sum_{k=1}^{N}\lambda_k = 1. \qquad \text{and} \qquad 0 \le \lambda_k \le 1. \tag{2.15}$$

[0219] The joint distribution of the λ_k's therefore follows the Dirichlet distribution with $\alpha_k = \frac{1}{p}$, [55].

[0220] One can then substitute $\lambda_k = |y_k|^p$ to get the distribution of y_k as

$$f(y_k) = \frac{1}{2}f(\lambda_k)\frac{d\lambda_k}{dy_k} = \frac{\Gamma\left(\frac{N}{p}\right)(1-|y_k|^p)^{\frac{N-1}{p}-1}}{2\,\Gamma\left(1+\frac{1}{p}\right)\Gamma\left(\frac{N-1}{p}\right)}. \tag{2.16}$$

[0221] This is the marginal distribution of the normalized Gaussian y_k for any k. It can readily be seen that it is a scaled and translated version of a symmetric Beta distribution. It should be noted, however, that the entries y_k are correlated with each other due to equation (2.15).

[0222] The distribution is symmetric around zero, whereby the mean is zero *E*[*yk*] = 0 and the variance is (omitting subscripts for brevity)

$$E\left[|y|^2\right] = \int_{-1}^{1} f(y)\,|y|^2 dy_k = \frac{\Gamma\left(\frac{N}{p}\right)\Gamma\left(\frac{3}{p}\right)}{\Gamma\left(\frac{1}{p}\right)\Gamma\left(\frac{N+2}{p}\right)}. \tag{2.17}$$

where the substitution λ = |*y*|^p was used. The integrand is similar to the Beta-distribution, whereby solution is simple. In particular, the variances

$$E[|y|^2] = \begin{cases} \frac{1}{N} & \text{for } p = 2 \\ \frac{2}{(N+1)N} & \text{for } p = 1. \end{cases} \tag{2.18}$$

are obtained.

[0223] In the following, embodiments are described.

[0224] Fig. 1a illustrates an audio encoder for encoding an audio signal, wherein the audio signal is represented in a spectral domain, according to an embodiment.

[0225] The audio encoder comprises a spectral envelope encoder 110 configured for determining a spectral envelope of the audio signal and for encoding the spectral envelope.

[0226] Moreover, the audio encoder comprises a spectral sample encoder 120 configured for encoding a plurality of spectral samples of the audio signal.

[0227] The spectral sample encoder 120 is configured to estimate an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the plurality of spectral samples depending on the spectral envelope.

[0228] Moreover, the spectral sample encoder 120 is configured to encode each spectral sample of the plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

[0229] In some of the embodiments, it may not be necessary to estimate an estimated bitrate for each of the plurality of spectral samples. For example, if it is estimated that for one of the spectral samples, the estimated bitrate is not greater than a threshold value, than it may be concluded that following spectral samples are also not greater than a threshold value.

[0230] In some alternative embodiments, however, the estimated bitrate may, e.g., be determined for each of the plurality of spectral samples.

[0231] In an embodiment, if the estimated bitrate needed for encoding a spectral sample of the one or more spectral samples is greater than a threshold value, the spectral sample encoder 120 may, e.g., be configured to encode said spectral sample according to the first coding rule, and if said estimated bitrate needed for encoding said spectral samples is smaller than or equal to a threshold value, the spectral sample encoder 120 may, e.g., be configured to encode said spectral sample according to the second coding rule.

[0232] According to an embodiment, the threshold value may, e.g., be 1 bit/sample (1 bit/spectral sample).

[0233] In an embodiment, the spectral sample encoder 120 may, e.g., be configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on an estimated variance of said spectral sample which depends on the spectral envelope.

[0234] According to an embodiment, the spectral sample encoder 120 may, e.g., be configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on the equation

$$b_k = \frac{1}{2}\log_2(4.1159\sigma_k^2)$$

wherein b_k is a k th spectral samples of the one or more spectral samples, wherein σ_k^2 is an estimated variance of said spectral sample.

[0235] In an embodiment, the spectral sample encoder 120 may, e.g., be configured to encode the spectral samples that are encoded according to the second coding rule by quantizing said spectral samples that are encoded according to the second coding rule employing an orthogonal matrix. Moreover, the spectral sample encoder 120 may, e.g., be configured to encode the spectral samples that are encoded according to the first coding rule by quantizing said spectral samples that are encoded according to the first coding rule without employing the orthogonal matrix.

[0236] According to an embodiment, the spectral sample encoder 120 may, e.g., be configured to encode the spectral samples that are encoded according to the second coding rule using:

$$Q_B[Ax],$$

wherein x is a vector comprising the spectral samples, wherein $Q_B[\cdot]$ is a quantizer defined as

$$Q_B[y] := \gamma \begin{bmatrix} \text{sign}(y_0) \\ \text{sign}(y_1) \\ \vdots \\ \text{sign}(y_{B-1}) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

wherein y_0 y_1, y_{B-1} are quantized values resulting from quantization, wherein B indicates a total bitrate, wherein γ is a scaling coefficient, wherein A is the orthogonal matrix.

[0237] Fig. 1b illustrates an audio decoder for decoding an encoded audio signal according to an embodiment.

[0238] The audio decoder comprises an interface 130 configured for receiving an encoded spectral envelope of the audio signal and configured for receiving an encoded plurality of spectral samples of the audio signal.

[0239] Moreover, the audio decoder comprises a decoding unit 140 configured for decoding the encoded audio signal by decoding the encoded spectral envelope and by decoding the encoded plurality of spectral samples.

[0240] The decoding unit 140 is configured to receive or to estimate an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the encoded plurality of spectral samples.

[0241] Moreover, the decoding unit 140 is configured to decode each spectral sample of the encoded plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples of the encoded plurality of spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

[0242] In some embodiments, the decoding unit 140 may, e.g., receive the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples from the audio encoder that encoded the plurality of spectral samples.

[0243] In some alternative embodiments, however, the decoding unit 140 may, e.g., estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on the spectral envelope in a same way as the audio encoded that encoded the spectral samples has estimated the estimated bitrate.

[0244] In an embodiment, if the estimated bitrate needed for encoding a spectral sample of the one or more spectral samples is greater than a threshold value, the decoding unit 140 may, e.g., be configured to decode said spectral sample according to the first coding rule, and if said estimated bitrate needed for encoding said spectral samples is smaller than or equal to a threshold value, the decoding unit 140 may, e.g., be configured to decode said spectral sample according to the second coding rule.

[0245] According to an embodiment, the threshold value may, e.g., be 1 bit/sample (1 bit/spectral sample).

[0246] In an embodiment, the decoding unit 140 may, e.g., be configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on an estimated variance of said spectral sample which depends on the spectral envelope.

[0247] According to an embodiment, the decoding unit 140 may, e.g., be configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on the equation

$$b_k = \frac{1}{2}\log_2(4.1159\sigma_k^2)$$

wherein b_k is a k th spectral samples of the one or more spectral samples, wherein σ_k^2 is an estimated variance of said spectral sample.

[0248] In an embodiment, the decoding unit 140 may, e.g., be configured to decode the spectral samples that are decoded according to the second coding rule by employing an orthogonal matrix. Moreover, the decoding unit 140 is configured to decode the spectral samples that are decoded according to the first coding rule without employing the orthogonal matrix.

[0249] According to an embodiment, if the estimated bitrate needed for encoding a spectral sample of the one or more spectral samples is smaller than another threshold value, the decoding unit 140 may, e.g., be configured to employ spectral noise shaping for decoding the encoded plurality of spectral samples.

[0250] Fig. 1c illustrates a system according to an embodiment. The system comprises an audio encoder 105 according to Fig. 1a and an audio decoder 125 according to Fig.1b.

[0251] The audio encoder 105 is configured to feed a encoded spectral envelope of an encoded audio signal and an encoded plurality of spectral samples of the encoded audio signal into the audio decoder.

[0252] The audio decoder 125 is configured to decode the encoded audio signal by decoding the encoded spectral envelope and by decoding the encoded plurality of spectral samples.

[0253] In the following, particular embodiments of the present invention are described.

[0254] Dithering methods however provide an alternative approach, where both accuracy and energy are retained. A hybrid coding approach is provided where low-energy samples are quantized using dithering, instead of the conventional uniform quantizer. For dithering, 1 bit quantization is applied in a randomized sub-space. It is moreover demonstrated that the output energy can be adjusted to the desired level using a scaling parameter. Objective measurements and listening tests demonstrate the advantages of the provided concepts.

[0255] In the following, an application of the provided randomization for dithered quantization in frequency-domain coding of speech and audio is provided, to allow coding at very low bitrates without excessive sparseness or low energy in the output. Perceptual listening tests demonstrate that the provided dithered quantizer gives the best performance.

[0256] In the following, quantization concepts of embodiments are described.

[0257] The performance of dithered quantization methods in comparison to conventional uniform quantization and in combination with entropy coding are considered. In the TCX mode of EVS [2], [48], entropy coding and uniform quantization is implemented assuming that the sample distribution is Laplacian, and the sample variance is estimated using the linear predictive envelope. The quantization accuracy is determined in a rate-loop such that the bit-budget is used as effectively as possible. In a vector of samples, trailing zeros are truncated. The scaling of the signal is determined after quantization, such that the output signal-to-noise ratio is optimized. This implementation of uniform quantization may, e.g., be used as a baseline system.

[0258] The above-described second coding concept comprises an approach for dithering and encoding data at low bit-rates (less than 1 bit/sample) based on random rotations and which is defined as follows [70]. Supposing there is a vector $x \in \mathbb{R}^{N \times 1}$, which shall be encoded with *B N* bits. Using a random orthogonal matrix *A*, one can then quantize

$$\hat{x} = A^T Q_B[Ax]. \qquad (3.1)$$

where $Q_B[\cdot]$ is a quantizer defined as

$$Q_B[y] := \gamma \begin{bmatrix} \text{sign}(y_0) \\ \text{sign}(y_1) \\ \vdots \\ \text{sign}(y_{B-1}) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (3.2)$$

and γ is a scaling coefficient. In other words, it uses a 1 bit-quantizer, where the *B* first samples are quantized with the sign of the input sample, for a total bitrate of *B*. The quantized $\hat{x}$ then has an approximately normal distribution and the variance is $E[\|\hat{x}\|^2] = \gamma^2 \frac{B}{N}$. It has furthermore been shown that the orthogonal matrix *A* can be approximated by a low-order rotation such that the algorithmic complexity remains linear $O(N)$.

[0259] It can then readily been show that γ can be chosen according to a number of criteria, for example:

1. $\gamma_{\text{MMSE}} = \sigma\sqrt{\frac{2}{\pi}}$ is the minimum mean square error (MMSE) scaling for normal input of variance σ^2. This thus corresponds to Wiener filtering the quantized signal.
2. $\gamma_{\sigma^2} = \sigma\sqrt{\frac{N}{B}}$ retains the variance σ^2 of the original signal. This corresponds to quantization on the surface of an *N*-dimensional hyper-sphere, which is normalized to original energy.

[0260] Clearly, γ can thus be tuned according to perceptual criteria, for a balance between accuracy and how well the quantizer retains the signal distribution and variance.

[0261] This approach to quantization provides error-diffusion similar to Floyd-Steinberg-type methods. However, in difference, error is not diffused forward to following components, but instead, it is diffused among the samples of a vector. In signal processing terms, Floyd-Steinberg-type methods are thus similar to infinite impulse responses (IIR) filters, whereas the proposed method is more like a finite impulse response (FIR) operation.

[0262] In the following speech and audio coding frameworks of embodiments are provided.

[0263] To evaluate the error characters of different quantizers, they should be implemented in a speech and audio codec which allows a fair comparison. This task is less straightforward than one might expect. The main issue is that codecs regularly use ad hoc tricks to overcome saturation effects of the arithmetic coder [1], [2], [48]. Namely, for example, high-frequency samples quantized to zero are typically truncated from the spectrum above the last non-zero sample. By omitting the transmission of zero samples one can save a considerable amount of bits, which can instead be used for coding low-frequency components. The performance of the arithmetic coder, in isolation, does therefore not accurately reflect the performance of the overall codec.

[0264] To obtain a fair comparison, a state-of-the-art baseline system is implemented following the simplified structure of the 3GPP Enhanced Voice Services (EVS) [2], [1], [48] (see Fig. 3).

[0265] Fig. 3 illustrates a diagram of the speech and audio encoder. The gray box is modified in by embodiments.

[0266] For frequency-domain coding, here the MDCT-transform with a window length of 30 ms, 50% overlap, a half-sine window and pre-emphasis with a filter $P(z) = 1 - 0.68z^{-1}$ is used. At a sampling rate of 16 kHz, the magnitude envelope is modeled with a linear predictive model of order M = 20, which is used as an estimate of the variance of each frequency component, and which is further fed into a conventional arithmetic coder with an assumption of a Laplacian distribution. Quantization in the perceptually weighted domain is applied as in [1]. It should be noted that a deadzone-quantizer was not implemented, even if it is known to improve signal-to-noise ratio [64], because it also amplifies the saturation effect at high frequencies. A deadzone-quantizer would therefore have unfairly penalized the baseline codec in terms low-rate performance.

[0267] Conventional codecs include noise-fill and bandwidth-extension methods to reduce the bitrate and to compensate for the energy-loss at high frequencies. To allow a straightforward and fair comparison between methods, bandwidth-extension was not included in the codec. The noise-fill algorithm is applied at frequencies above 1.6 kHz, on all spectral components which are quantized to zero, where noise with a random sign was added, and the magnitude was adjusted to match that obtained with the proposed dithering method with gain MMSE. The noise-fill used in EVS uses advanced signal analysis to fine-tune noise-filling, but this simplified method was chosen to make the test easy to reproduce. All parameters should anyway be tuned to the particular configuration of the final codec, whereby further perceptual tuning of parameters is not worthwhile for these experiments.

[0268] For the bitrate of the codec, it was assumed that spectral envelope, gain and other parameters are encoded with 2.6 kbits/s, whereby the remaining bits can be used for encoding the spectrum. Further, for simplicity and reproducability, any other parameters of the signal were not quantized. It should be noted, however, that bitrate calculations here are provided only to assist the reader in getting a realistic impression of performance, as the bitrate of side-information can vary in particular implementations of codecs.

[0269] In the following, a hybrid coder according to embodiments is described.

[0270] The combination of uniform quantization and arithmetic coder saturates at low bitrates, whereby it is proposed to replace the conventional approach by dithered coding for spectral samples whose bitrate is below 1 bit/sample. It is thus a hybrid entropy coder, which uses uniform quantization and arithmetic coding following [48] in high-energy areas of the spectrum and dithered coding at the low-energy areas.

[0271] The baseline entropy coder uses the linear predictive envelope to estimate the variance σ_k^2 of frequency components [48]. Note that this envelope has to be scaled such that the expected bitrate of a signal which follows that envelope, matches the target bitrate. Based on the variance σ_k^2 of the k th component, one can then estimate the expected bitrate of a sample as $b_k = \frac{1}{2}\log_2(4.1159\sigma_k^2)$ but limited to $b_k \geq 0$. For spectral components with $b_k > 1$ quantization and arithmetic coding is used, for $b_k \leq 1$ dithered coding is applied (see Fig. 4). The bit-allocation between uniform and dithered quantization is derived directly from the expected bitrate b_k.

[0272] Fig. 4 illustrates hybrid coding of spectral components.

[0273] All low-energy samples are thus collated into a vector x and are quantized with equation (3.1). Implicitly, it is thus assumed that vector *x* follows the normal distribution with uniform variance. To improve accuracy, embodiments further subdivide x according to their variance.

[0274] In [70], it was demonstrated that the randomization matrix *A* of sufficient quality can be readily generated with 4 iteration of *N*/2 random 2 × 2 rotations and length *N* permutations, when the bitrate is *B* = *N*. However, with *B* << *N*, a majority of samples are zeros, whereby the number of iterations were increased to 8 such that the output distribution remains normal.

[0275] Random rotations between the non-zero and zeroed values could be readily used to reduce the computational complexity without effect on the statistics of the output signal.

[0276] In the following, conducted experiments with respect to embodiments are described.

[0277] To evaluate the performance the provided hybrid codec, in comparison to uniform quantization, three types of experiments were performed. Firstly, the performance of dithering in isolation with synthetic input is considered. Secondly, speech from the TIMIT corpus was encoded and performance was evaluated by objective criteria. Finally, using samples from the TIMIT corpus, a MUSHRA subjective listening test was performed to determine perceptual preference among methods [52].

[0278] Output distribution of the provided dithered quantization (equation (3.1)) in comparison to uniform quantization is illustrated in Fig. 5.

[0279] Fig. 5 illustrates collated histograms of K = 10000 vectors of unit variance Gaussian input, quantized by uniform quantization and entropy coding as well as the provided dithered coder (N = 32, $\gamma\sigma^2$), with 1 bit/sample.

[0280] Here, normally distributed K = 10000 vectors of length N = 32 were encoded with B = 32 bits, and the gain factor $\gamma\sigma^2$ was used. It can readily be observed that uniform quantization is unable to retain the shape of the original distribution, whereas the distribution of the output of the provided dithered codec exactly matches that of the input.

[0281] Performance of the provided coder for a single frame of speech is illustrated in Fig. 6.

[0282] Fig. 6 depicts an illustration of performance for a typical speech spectrum at 13.2 kbits/s, wherein (a) depicts an input signal spectrum and its envelope, wherein (b) depicts the bit-rate estimated from the envelope and the threshold where quantizers are switched, and wherein (c) the quantized spectra with conventional uniform quantization and entropy coding in comparison to the provided, dithered coder.

[0283] The spectral magnitude envelope is estimated using linear predictive modelling in Fig. 6(a), the expected bit-rate for each frequency is estimated from the envelope using the method developed in [48] in Fig. 6(b) and a threshold is applied to determine the choice of quantizer. Finally, in Fig. 6(c), the quantized output of the conventional method is compared with the provided concepts, where the gain factor was _2 . It can be clearly seen that whereas for the conventional approach, all frequencies above 2 kHz are quantized to zero, the provided concepts retain the spectral shape also at the higher frequencies.

[0284] For objective evaluation of performance on real speech, the entire TIMIT database (training and evaluation) was encoded with different combinations of quantization and entropy coding [51]. Namely, it was applied: 1. uniform quantization with arithmetic coding following [48] (Conventional), 2. a dithering simulation by adding white noise to obtain same signal to noise ratio as the conventional approach (Dithering), 3. the provided hybrid codec using γMMSE (1 bit MMSE) and 4. Using $\gamma\sigma^2$ (1 bit EM). The mean output energy across frequencies for each method is illustrated in Fig. 2.

[0285] It can be seen that all modifications of conventional arithmetic coding bring the amount of energy closer to the original energy envelope. The dithering simulation saturates at the perceptual noise floor near -20 dB, which is higher than the original energy envelope. Informal listening confirms that such dithering has a noisy character, where the conventional is muffled. The two provided 1 bit dithering concepts are closer to the original energy envelope, such that the MMSE approach $\gamma MMSE$ is clearly below the original while the energy matching method $\gamma\sigma^2$ approximates nicely the desired energy envelope.

[0286] The average signal to noise ratios (SNR) in the perceptual domain for the conventional and provided concepts are listed in Table III.

Table III:

	1 bit MMSE	1 bit EM	Conventional
SNR (dB)	10.75	10.46	8.93

[0287] Table III illustrates a mean signal to noise ratio in the perceptually weighted domain for the conventional and the two provided concepts.

[0288] Clearly the 1 bit MMSE approach reaches the highest SNR, since it was designed to optimize SNR. However, as the conventional method is also designed to optimize SNR, it is surprising that such a large improvement of 1.81 dB is obtained. The energy-matching approach $\gamma\sigma^2$ looses slightly in SNR to the MMSE approach, but not the difference is not large 0.29 dB. A subjective listening test is needed to determine it is more important to preserve envelope shape or optimize SNR.

[0289] Finally, to determine subjective preference among methods, a MUSHRA listening test was performed [52]. In the test, 6 samples (3 male and 3 female) randomly chosen from the TIMIT corpus [51] were included. In addition to the above methods, Conventional, Dithered, 1 bit MMSE and 1 bit EM, also cases where the conventional uniform coder is enhanced by noise filling in post-processing were included. It was not included in the previous tests because it is a blind post-processing method in the sense that it adds noise without any transmitted information from the input signal, whereby it reduces accuracy even if it is designed to improve perceptual quality. In the listening test, 14 normal hearing subjects in the age-range 26 to 43 years attended. Fig. 7 illustrates the results.

[0290] Fig. 7 depicts results of a subjective MUSHRA listening test, comparing the provided 1 bit dithered quantizers with conventional arithmetic coding, as well as a synthetic dithering serving as an anchor.

[0291] It was observed for all items, that the provided dithered 1 bit quantizers have a higher mean than the other methods. Moreover, in the mean over all items (the "All" column), the provided dithered 1 bit quantizers have a statistically significant difference to the antecedent methods. Conventional arithmetic coding without noisefill also shows a statistically significant reduction in quality in comparison to all other methdods. To further determine whether listeners have a preference among the two provided dithered quantizers, the differential MUSHRA scores with noisefill as a reference (see Fig. 8) was calculated.

[0292] Fig. 8 illustrates differential scores of a subjective MUSHRA listening test, comparing the provided 1 bit dithered quantizers with conventional arithmetic coding, as well as a synthetic dithering serving as an anchor. Differences are calculated with the noisefill as reference.

[0293] However, the differential scores revealed no additional details.

[0294] Now, a discussion the provided concepts and conclusions follow.

[0295] The applicability of dithering in signal processing has been demonstrated., Conventional methods in coding, such as noisefill and bandwidth-extension do not attempt to optimize SNR. In contrast, embodiments may, e.g., use a recently developed method for dithering and coding which is applicable to very low bitrates [70]. The approach is based on a random rotation, sign-quantization in the randomized domain, and an inverse transform. It is proposed to apply it in combination with conventional uniform quantization and entropy coding, such that only frequency components where one can afford to use very little accuracy, are coded with the dithered quantizer.

[0296] By using dithering one can avoid the characteristic problem of conventional frequency-domain codecs, where higher frequencies are often quantized to zero such the output sounds muffled. In other words, the output is not unnaturally sparse.

[0297] Objective and subjective experiments demonstrate that the method gives a significant improvement in perceptual quality.

[0298] 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. Some or all of the method steps may be executed by (or using) a hardware apparatus, like for example, a microprocessor, a programmable computer or an electronic circuit. In some embodiments, one or more of the most important method steps may be executed by such an apparatus.

[0299] Depending on certain implementation requirements, embodiments of the invention can be implemented in hardware or in software or at least partially in hardware or at least partially in software. The implementation can be performed using a digital storage medium, for example a floppy disk, a DVD, a Blu-Ray, 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. Therefore, the digital storage medium may be computer readable.

[0300] 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.

[0301] 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.

[0302] Other embodiments comprise the computer program for performing one of the methods described herein, stored on a machine readable carrier.

[0303] 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.

[0304] 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. The data carrier, the digital storage medium or the recorded medium are typically tangible and/or non-transitory.

[0305] 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.

[0306] 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.

[0307] A further embodiment comprises a computer having installed thereon the computer program for performing one of the methods described herein.

[0308] A further embodiment according to the invention comprises an apparatus or a system configured to transfer (for example, electronically or optically) a computer program for performing one of the methods described herein to a receiver. The receiver may, for example, be a computer, a mobile device, a memory device or the like. The apparatus or system may, for example, comprise a file server for transferring the computer program to the receiver.

[0309] 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.

[0310] The apparatus described herein may be implemented using a hardware apparatus, or using a computer, or using a combination of a hardware apparatus and a computer.

[0311] The methods described herein may be performed using a hardware apparatus, or using a computer, or using a combination of a hardware apparatus and a computer.

[0312] 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.

References

[0313]

[1] T. Bäckström, Speech Coding with Code-Excited Linear Prediction. New York, NY, USA: Springer, 2017.

[2] TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12), 3GPP, 2014.

[3] TS 26.190, Adaptive Multi-Rate (AMR-WB) Speech Codec, 3GPP, 2007.

[4] MPEG-D (MPEG Audio Technologies), Part 3: Unified Speech and Audio Coding, ISO/IEC 23003-3:2012, 2012.

[5] M. Bosi et al., "ISO/IECMPEG-2 advanced audio coding," J. Audio Eng. Soc., vol. 45, no. 10, pp. 789-814, 1997.

[6] J. Benesty, M. Sondhi, and Y. Huang, Springer Handbook of Speech Processing. New York, NY, USA: Springer, 2008.

[7] A. Zahedi, J. Østergaard, S. H. Jensen, S. Bech, and P. Naylor, "Audio coding in wireless acoustic sensor networks," Signal Process., vol. 107, pp. 141-152, 2015.

[8] T. Bäckström, F. Ghido, and J. Fischer, "Blind recovery of perceptual models in distributed speech and audio coding," in Proc. Interspeech, 2016, pp. 2483-2487.

[9] T. Bäckström and J. Fischer, "Coding of parametric models with randomized quantization in a distributed speech and audio codec," in Proc. ITG Fachtagung Sprachkommunikation, 2016, pp. 1-5.

[10] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in Proc. IEEE Inf. Sci. Syst. 42nd Ann. Conf., 2008, pp. 16-21.

[11] A. Magnani, A. Ghosh, and R. M. Gray, "Optimal one-bit quantization," in Proc. IEEE Data Compression Conf., 2005, pp. 270-278.

[12] M. M. Truman, G. A. Davidson, M. C. Fellers, M. S. Vinton, M. A. Watson, and C. Q. Robinson, "Audio coding system using spectral hole filling," U.S. Patent 7 447 631, Nov. 4, 2008.

[13] S. Vaudenay, "Decorrelation: a theory for block cipher security," J. Cryptol., vol. 16, no. 4, pp. 249-286, 2003.

[14] S. Saeednia, "How to make the Hill cipher secure," Cryptologia, vol. 24, no. 4, pp. 353-360, 2000.

[15] C.-C. Kuo and W. Thong, "Reduction of quantization error with adaptive Wiener filter in low bit rate coding," in Proc. 11th Eur. IEEE Signal Process. Conf., 2002, pp. 1-4.

[16] G. E. Øien and T. A. Ramstad, "On the role of Wiener filtering in quantization and DPCM," in Proc. IEEE Norwegian Signal Process. Symp. Workshop, 2001.

[17] J. Rissanen and G. G. Langdon, "Arithmetic coding," IBM J. Res.Develop., vol. 23, no. 2, pp. 149-162, 1979.

[18] J. D. Gibson and K. Sayood, "Lattice quantization," Adv. Electron. Electron Phys., vol. 72, pp. 259-330, 1988.

[19] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression. New York, NY, USA: Springer, 1992.

[20] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," IEEE Signal Process. Mag., vol. 21, no. 5, pp. 80-94, Sep. 2004.

[21] Z. Xiong, A. D. Liveris, and Y. Yang, "Distributed source coding," in Handbook on Array Processing and Sensor Networks. New York, NY, USA: Wiley-IEEE Press, 2009, pp. 609-643.

[22] M. Bosi and R. E. Goldberg, Introduction to Digital Audio Coding and Standards. Norwell, MA, USA: Kluwer, 2003.

[23] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in Proc. 18th IEEE Symp. Commun. Veh. Technol. Benelux, 2011, pp. 1-6.

[24] I. F. Akyildiz, T. Melodia, and K. R. Chowdury, "Wireless multimedia sensor networks: A survey," IEEE Wireless Commun., vol. 14, no. 6, pp. 32-39, Dec. 2007.

[25] B. Girod, A. M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," Proc. IEEE, vol. 93, no. 1, pp. 71-83, Jan. 2005.

[26] F. de la Hucha Arce, M. Moonen, M. Verhelst, and A. Bertrand, "Adaptive quantization for multi-channel Wiener filter-based speech enhancement in wireless acoustic sensor networks," Wireless Commun. Mobile Comput., vol. 2017, 2017, Art. no. 3173196.

[27] A. Zahedi, J. Østergaard, S. H. Jensen, P. Naylor, and S. Bech, "Coding and enhancement in wireless acoustic sensor networks," in Proc. IEEE Data Compression Conf., 2015, pp. 293-302.

[28] A. Majumdar, K. Ramchandran, and L. Kozintsev, "Distributed coding for wireless audio sensors," in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., 2003, pp. 209-212.

[29] H. Dong, J. Lu, and Y. Sun, "Distributed audio coding in wireless sensor networks," in Proc. IEEE Int. Conf. Comput. Intell. Secur., 2006, vol. 2, pp. 1695-1699.

[30] G. Barriac, R. Mudumbai, and U. Madhow, "Distributed beamforming for information transfer in sensor networks," in Proc. 3rd Int. Symp. Inf. Process. Sens. Netw., 2004, pp. 81-88.

[31] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2003, vol. 4, p. IV-840-3.

[32] O. Roy and M. Vetterli, "Rate-constrained collaborative noise reduction for wireless hearing aids," IEEE Trans. Signal Process., vol. 57, no. 2, pp. 645-657, Feb. 2009.

[33] S. Bray and G. Tzanetakis, "Distributed audio feature extraction for music," in Proc. Int. Conf. Music Inf. Retrieval, 2005, pp. 434-437.

[34] N. Rajput and A. A. Nanavati, "Distributed speech recognition," in Speech in Mobile and Pervasive Environments. New York, NY, USA: Wiley, 2012, pp. 99-114.

[35] D. Pearce, "Distributed speech recognition standards," in Automatic Speech Recognition on Mobile Devices and Over Communication Networks. London, U.K.: Springer, 2008, pp. 87-106.

[36] S. Korse, T. Jahnel, and T. Bäckström, "Entropy coding of spectral envelopes for speech and audio coding

using distribution quantization," in Proc. Interspeech, 2016, pp. 2543-2547.

[37] S. Das, A. Craciun, T. Jahnel, and T. Bäckström, "Spectral envelope statistics for source modelling in speech enhancement," in Proc. ITG Fachtagung Sprachkommunikation, 2016, pp. 1-5.

[38] G. H. Golub and C. F. van Loan, Matrix Computations, 4th ed. Baltimore, MD, USA: John Hopkins Univ. Press, 2004.

[39] V. Pulkki, "Spatial sound reproduction with directional audio coding," J. Audio Eng. Soc., vol. 55, no. 6, pp. 503-516, 2007.

[40] A. Edelman and N. R. Rao, "Random matrix theory," Acta Numerica, vol. 14, pp. 233-297, 2005.

[41] D. Knuth, The Art of Computer Programming. Reading, MA, USA: Addison-Wesley, 1998.

[42] D. E. Knuth, "Volume 2: Seminumerical algorithms," in The Art of Computer Programming, 3rd ed. Reading, MA, USA: Addison-Wesley, 2007.

[43] P. Diaconis and M. Shahshahani, "The subgroup algorithm for generating uniform random variables," Probab. Eng. Inf. Sci., vol. 1, no. 1, pp. 15-32, 1987.

[44] G. W. Stewart, "The efficient generation of random orthogonal matrices with an application to condition estimators," SIAM J. Numer. Anal., vol. 17, no. 3, pp. 403-409, 1980.

[45] P. L'Ecuyer, "Pseudorandom number generators," in, Encyclopedia of Quantitative Finance. New York, NY, USA: Wiley, 2010.

[46] M. Matsumoto andT. Nishimura, "Mersenne twister:A623-dimensionally equidistributed uniform pseudo-random number generator," ACM Trans. Model. Comput. Simul., vol. 8, no. 1, pp. 3-30, 1998.

[47] A. Jagannatam, "Mersenne twister-A pseudo random number generator and its variants," Dept. Elect. Comput. Eng., George Mason Univ., Fairfax, VA, USA, 2008.

[48] T. Bäckström and C. R. Helmrich, "Arithmetic coding of speech and audio spectra using TCX based on linear predictive spectral envelopes," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Apr. 2015, pp. 5127-5131.

[49] J. Fischer and T. Bäckström, "Wiener filtering in distributed speech and audio coding," IEEE Signal Process. Lett., 2017, submitted for publication.

[50] T. Bäckström, "Estimation of the probability distribution of spectral fine structure in the speech source," in Proc. Interspeech, 2017, pp. 344-348.

[51] J. S. Garofolo, et al., TIMIT: Acoustic-Phonetic Continuous Speech Corpus. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[52] Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems, ITU-R Recommendation BS.1534, 2003.

[53] S. Nadarajah, "A generalized normal distribution," J. Appl. Statist., vol. 32, no. 7, pp. 685-694, 2005.

[54] C. Walck, Handbook on Statistical Distributions for Experimentalists. Stockholm, Sweden: Univ. Stockholm, 2007.

[55] A. Bela, A. Frigyik, and M. Gupta, "Introduction to the Dirichlet distribution and related processes," , Dept. Elect. Eng., Univ. Washington, Seattle, WA, USA, Tech. Rep. UWEETR-2010-0006, 2010.

[56] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)," in Proc. 1st

Web Audio Conf., Paris, France, 2015.

[57] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kj¨orling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, K. S. Chong, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "The ISO/MPEG unified speech and audio coding standard - consistent high quality for all content types and at all bit rates," Journal of the AES, vol. 61, no. 12, pp. 956-977, 2013.

[58] ITU-T G.718, "Frame error robust narrow-band and wideband embedded variable bitrate coding of speech and audio from 8-32 kbit/s," 2008.

[59] J. Mäkinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in Proc. ICASSP, 2005, vol. 2, pp. 1109-1112.

[60] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Advanced audio coding," in 101 AES Convention, 2012.

[61] G. Fuchs, M. Multrus, M. Neuendorf, and R. Geiger, "Mdct-based coder for highly adaptive speech and audio coding," in European Signal Processing Conference (EUSIPCO 2009), 2009, pp. 24-28.

[62] I. H. Witten, R. M Neal, and J. G. Cleary, "Arithmetic coding for data compression," Communications of the ACM, vol. 30, no. 6, pp. 520-540, 1987.

[63] ITU-T Recommendation P.800, Methods for subjective determination of transmission quality, 1996.

[64] G. Fuchs, C. R. Helmrich, G. Markovic´, M. Neusinger, E. Ravelli, and T. Moriya, "Low delay LPC and MDCT-based audio coding in the EVS codec," in Proc. ICASSP. IEEE, 2015, pp. 5723-5727.

[65] S Disch, A. Niedermeier, C. R. Helmrich, C. Neukam, K. Schmidt, R. Geiger, J. Lecomte, F. Ghido, F. Nagel, and B. Edler, "Intelligent gap filling in perceptual transform coding of audio," in Audio Engineering Society Convention 141. Audio Engineering Society, 2016.

[66] J. Vanderkooy and S. P. Lipshitz, "Dither in digital audio," Journal of the Audio Engineering Society, vol. 35, no. 12, pp. 966-975, 1987.

[67] R. W. Floyd and L. Steinberg, "An adaptive algorithm for spatial gray-scale," in Proc. Soc. Inf. Disp., 1976, vol. 17, pp. 75-77.

[68] M Li, J. Klejsa, and W. B. Kleijn, "Distribution preserving quantization with dithering and transformation," IEEE Signal Processing Letters, vol. 17, no. 12, pp. 1014-1017, 2010.

[69] T. Bäckström, "Enumerative algebraic coding for ACELP," in Proc. Interspeech, 2012.

[70] T. Bäckström and J. Fischer, "Fast randomization for distributed low-bitrate coding of speech and audio," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 26, no. 1, January 2018.

[71] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the OPUS codec," in Audio Engineering Society Convention 135. Audio Engineering Society, 2013.

Claims

1. An audio encoder for encoding an audio signal, wherein the audio signal is represented in a spectral domain, wherein the audio encoder comprises:

   a spectral envelope encoder (110) configured for determining a spectral envelope of the audio signal and for encoding the spectral envelope, and
   a spectral sample encoder (120) configured for encoding a plurality of spectral samples of the audio signal,

wherein the spectral sample encoder (120) is configured to estimate an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the plurality of spectral samples depending on the spectral envelope, and
wherein the spectral sample encoder (120) is configured to encode each spectral sample of the plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

2. An audio encoder according to claim 1,
wherein, if the estimated bitrate needed for encoding a spectral sample of the one or more spectral samples is greater than a threshold value, the spectral sample encoder (120) is configured to encode said spectral sample according to the first coding rule, and if said estimated bitrate needed for encoding said spectral samples is smaller than or equal to a threshold value, the spectral sample encoder (120) is configured to encode said spectral sample according to the second coding rule.

3. An audio encoder according to claim 2,
wherein the threshold value is 1 bit/sample.

4. An audio encoder according to one of the preceding claims,
wherein the spectral sample encoder (120) is configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on an estimated variance of said spectral sample which depends on the spectral envelope.

5. An audio encoder according to one of claims 1 to 3,
wherein the spectral sample encoder (120) is configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on the equation

$$b_k = \frac{1}{2}\log_2(4.1159\sigma_k^2)$$

wherein b_k is a *k* th spectral samples of the one or more spectral samples,

wherein σ_k^2 is an estimated variance of said spectral sample.

6. An audio encoder according to one of the preceding claims,
wherein the spectral sample encoder (120) is configured to encode the spectral samples that are encoded according to the second coding rule by quantizing said spectral samples that are encoded according to the second coding rule employing an orthogonal matrix, and
wherein the spectral sample encoder (120) is configured to encode the spectral samples that are encoded according to the first coding rule by quantizing said spectral samples that are encoded according to the first coding rule without employing the orthogonal matrix.

7. An audio encoder according to 6,
wherein the spectral sample encoder (120) is configured to encode the spectral samples that are encoded according to the second coding rule using:

$$Q_B[Ax],$$

wherein *x* is a vector comprising the spectral samples,
wherein $Q_B[\cdot]$ is a quantizer defined as

$$Q_B[y] := \gamma \begin{bmatrix} \mathrm{sign}(y_0) \\ \mathrm{sign}(y_1) \\ \vdots \\ \mathrm{sign}(y_{B-1}) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

wherein y_0 y_1, y_{B-1} are quantized values resulting from quantization,
wherein B indicates a total bitrate,
wherein γ is a scaling coefficient,
wherein A is the orthogonal matrix.

8. An audio decoder for decoding an encoded audio signal, wherein the audio decoder comprises:

an interface (130) configured for receiving an encoded spectral envelope of the audio signal and configured for receiving an encoded plurality of spectral samples of the audio signal, and
a decoding unit (140) configured for decoding the encoded audio signal by decoding the encoded spectral envelope and by decoding the encoded plurality of spectral samples,
wherein the decoding unit (140) is configured to receive or to estimate an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the encoded plurality of spectral samples, and
wherein the decoding unit (140) is configured to decode each spectral sample of the encoded plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples of the encoded plurality of spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

9. An audio decoder according to claim 8,
wherein, if the estimated bitrate needed for encoding a spectral sample of the one or more spectral samples is greater than a threshold value, the decoding unit (140) is configured to decode said spectral sample according to the first coding rule, and if said estimated bitrate needed for encoding said spectral samples is smaller than or equal to a threshold value, the decoding unit (140) is configured to decode said spectral sample according to the second coding rule.

10. An audio decoder according to claim 9,
wherein the threshold value is 1 bit/sample.

11. An audio decoder according to one of claims 8 to 10,
wherein the decoding unit (140) is configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on an estimated variance of said spectral sample which depends on the spectral envelope.

12. An audio decoder according to one of claims 8 to 10,
wherein the decoding unit (140) is configured to estimate the estimated bitrate needed for encoding for each spectral sample of the one or more spectral samples depending on the equation

$$b_k = \frac{1}{2}\log_2(4.1159\sigma_k^2)$$

wherein b_k is a k th spectral samples of the one or more spectral samples,

wherein σ_k^2 is an estimated variance of said spectral sample.

13. An audio decoder according to one of claims 8 to 12,
wherein the decoding unit (140) is configured to decode the spectral samples that are decoded according to the second coding rule by employing an orthogonal matrix, and
wherein the decoding unit (140) is configured to decode the spectral samples that are decoded according to the first coding rule without employing the orthogonal matrix.

14. An audio decoder according to one of claims 8 to 13,
wherein, if the estimated bitrate needed for encoding a spectral sample of the one or more spectral samples is smaller than another threshold value, the decoding unit (140) is configured to employ spectral noise shaping for decoding the encoded plurality of spectral samples.

15. A system comprising:

an audio encoder (105) according to one of claims 1 to 7, and
an audio decoder (125) according to one of claims 8 to 14,
wherein the audio encoder (105) is configured to feed a encoded spectral envelope of an encoded audio signal and an encoded plurality of spectral samples of the encoded audio signal into the audio decoder, and
wherein the audio decoder (125) is configured to decode the encoded audio signal by decoding the encoded spectral envelope and by decoding the encoded plurality of spectral samples.

16. A method for encoding an audio signal, wherein the audio signal is represented in a spectral domain, wherein the method comprises:

determining a spectral envelope of the audio signal and for encoding the spectral envelope, and
encoding a plurality of spectral samples of the audio signal,
wherein encoding the plurality of spectral samples of the audio signal is conducted by estimating an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the plurality of spectral samples depending on the spectral envelope, and
wherein encoding the plurality of spectral samples of the audio signal is conducted by encoding each spectral sample of the plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

17. A method for decoding an encoded audio signal, wherein the method comprises:

receiving an encoded spectral envelope of the audio signal and for receiving an encoded plurality of spectral samples of the audio signal, and
decoding the encoded audio signal by decoding the encoded spectral envelope and by decoding the encoded plurality of spectral samples,
wherein decoding the encoded audio signal is conducted by receiving or by estimating an estimated bitrate needed for encoding for each spectral sample of one or more spectral samples of the encoded plurality of spectral samples, and
wherein decoding the encoded audio signal is conducted by decoding each spectral sample of the encoded plurality of spectral samples, depending on the estimated bitrate needed for encoding for the one or more spectral samples of the encoded plurality of spectral samples, according to a first coding rule or according to a second coding rule being different from the first coding rule.

18. A computer program for implementing the method of claim 16 or 17 when being executed on a computer or signal processor.

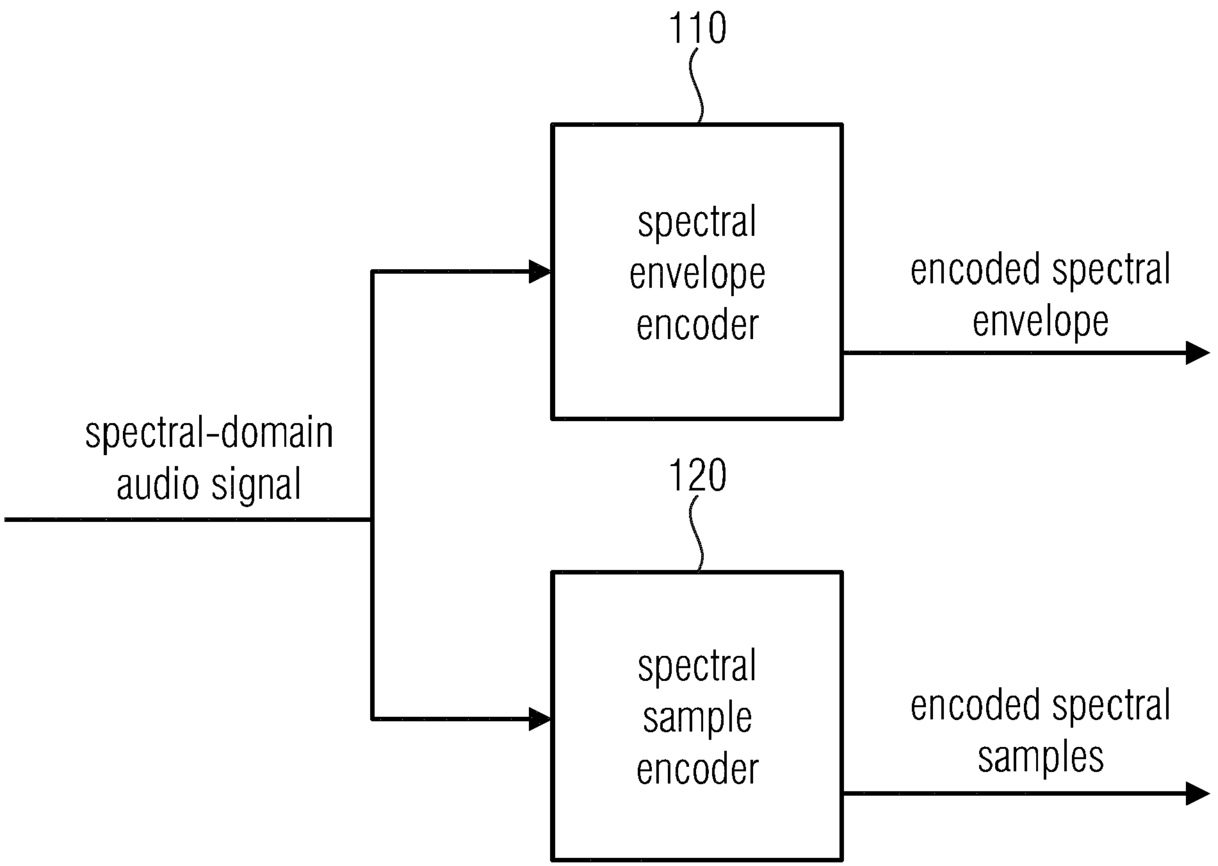

110
spectral envelope encoder
120
spectral sample encoder
spectral-domain audio signal
encoded spectral envelope
encoded spectral samples

Fig. 1a

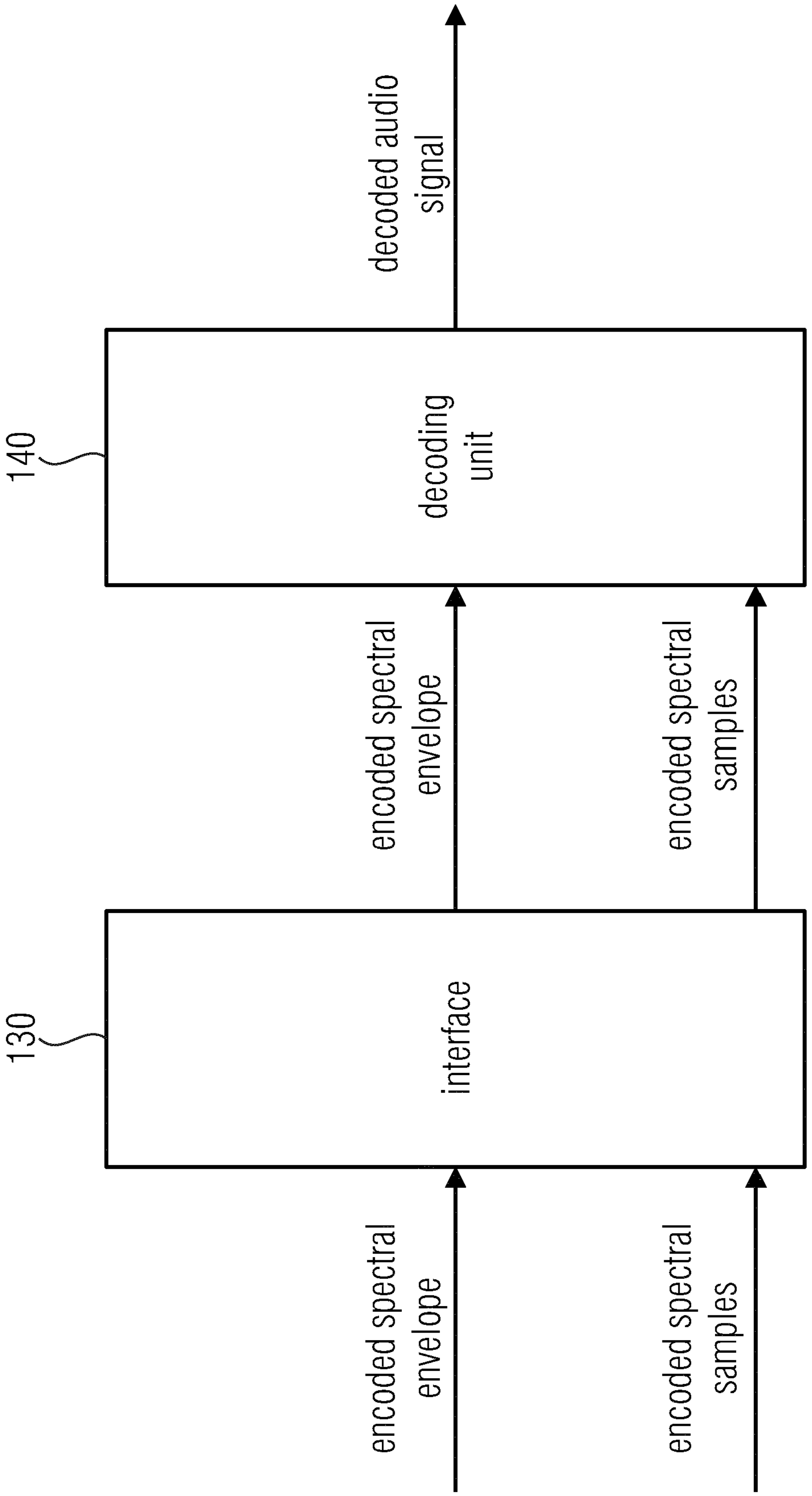

encoded spectral envelope
encoded spectral samples
130
interface
encoded spectral envelope
encoded spectral samples
140
decoding unit
decoded audio signal

Fig. 1b

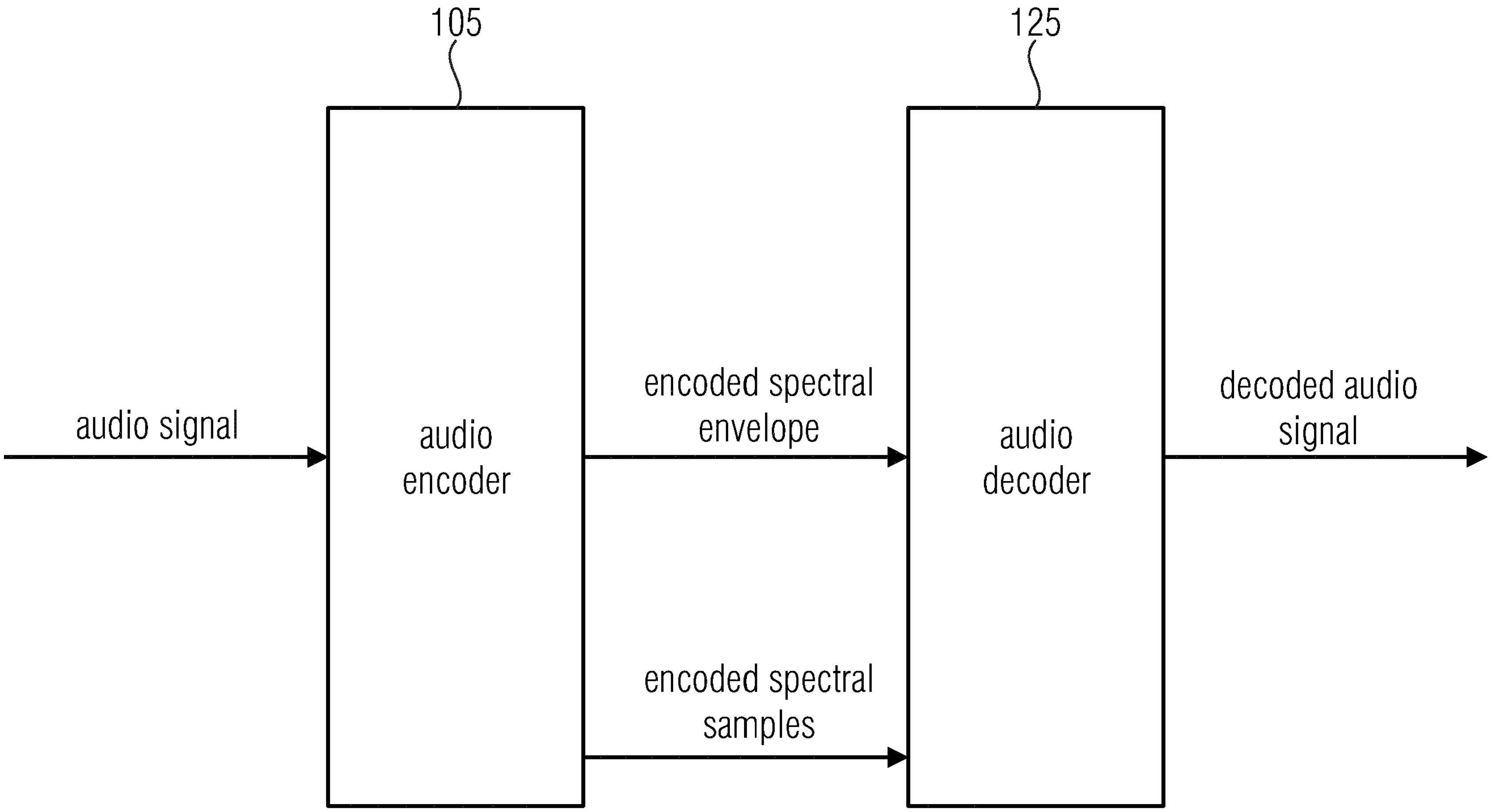

105
125
audio signal
audio
encoder
encoded spectral
envelope
encoded spectral
samples
audio
decoder
decoded audio
signal

Fig. 1c

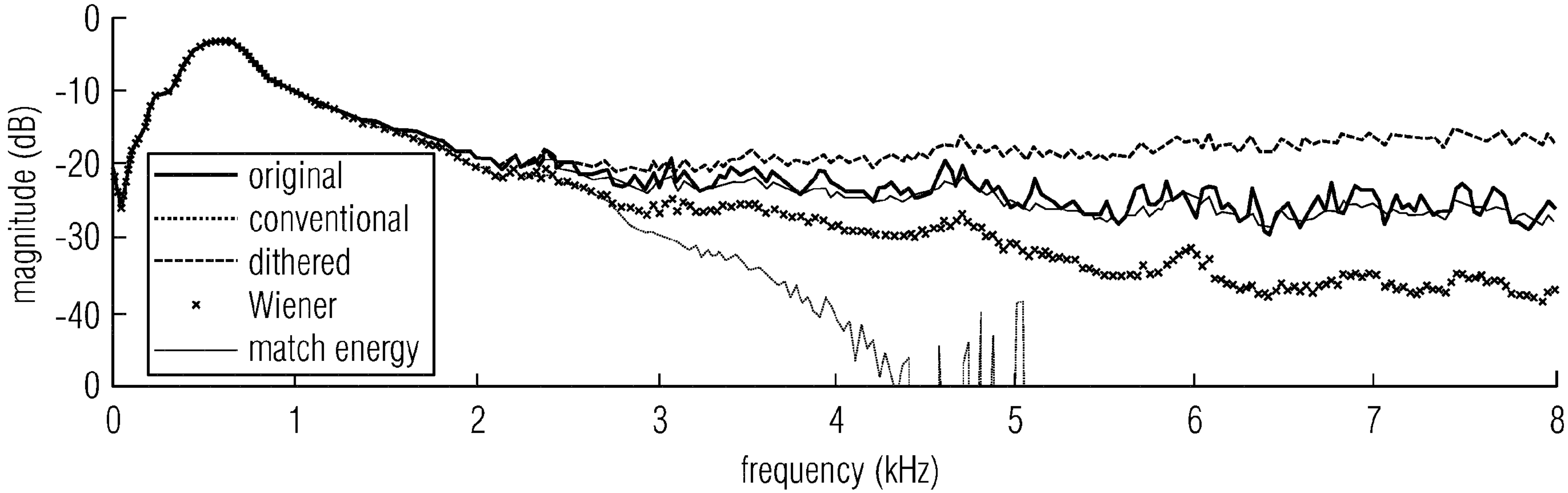

magnitude (dB)
0
-10
-20
-30
-40
0
0
1
2
3
4
5
6
7
8
frequency (kHz)
original
conventional
dithered
Wiener
match energy

Fig. 2

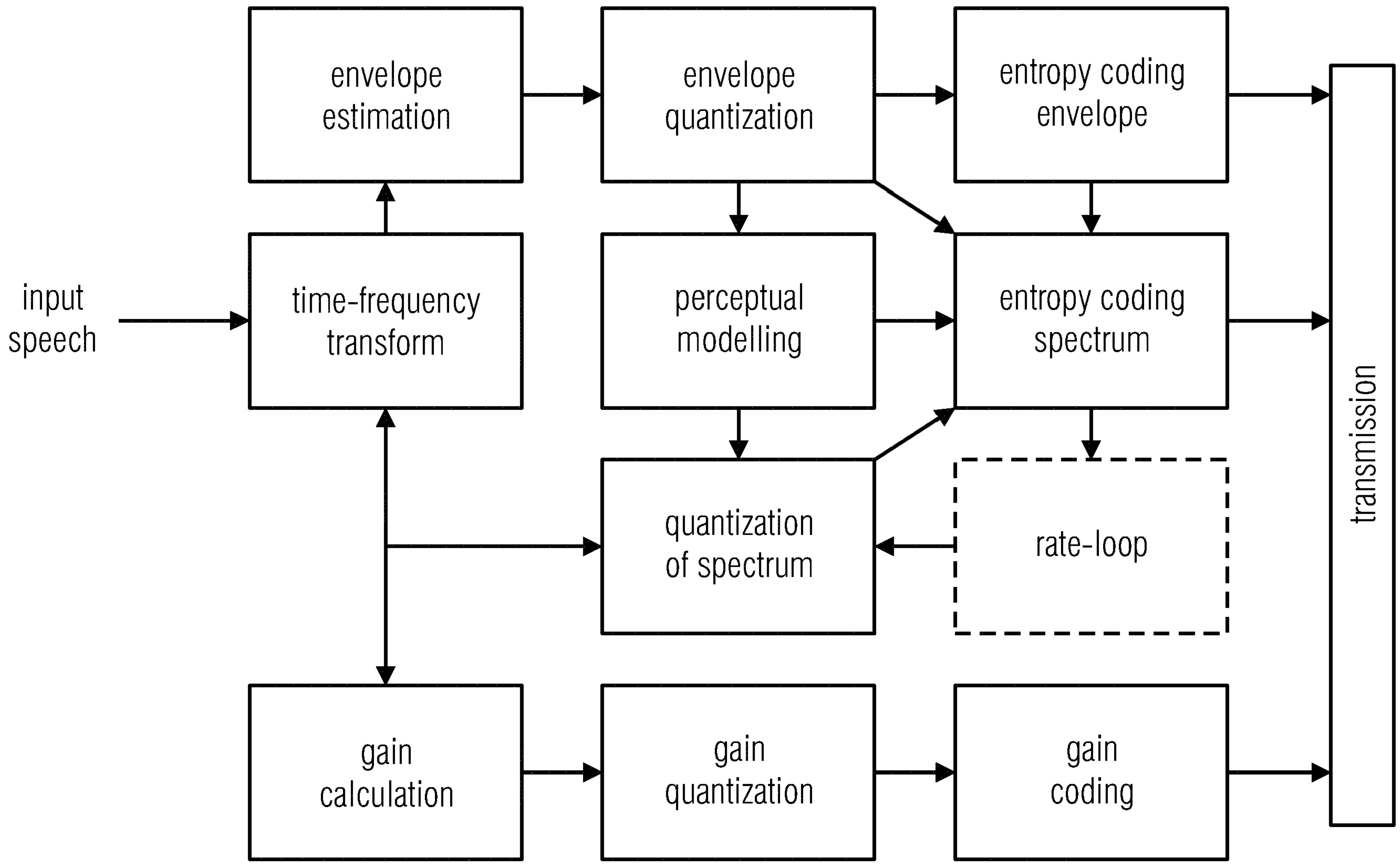

input
speech
envelope
estimation
time-frequency
transform
gain
calculation
envelope
quantization
perceptual
modelling
quantization
of spectrum
gain
quantization
entropy coding
envelope
entropy coding
spectrum
rate-loop
gain
coding
transmission

Fig. 3

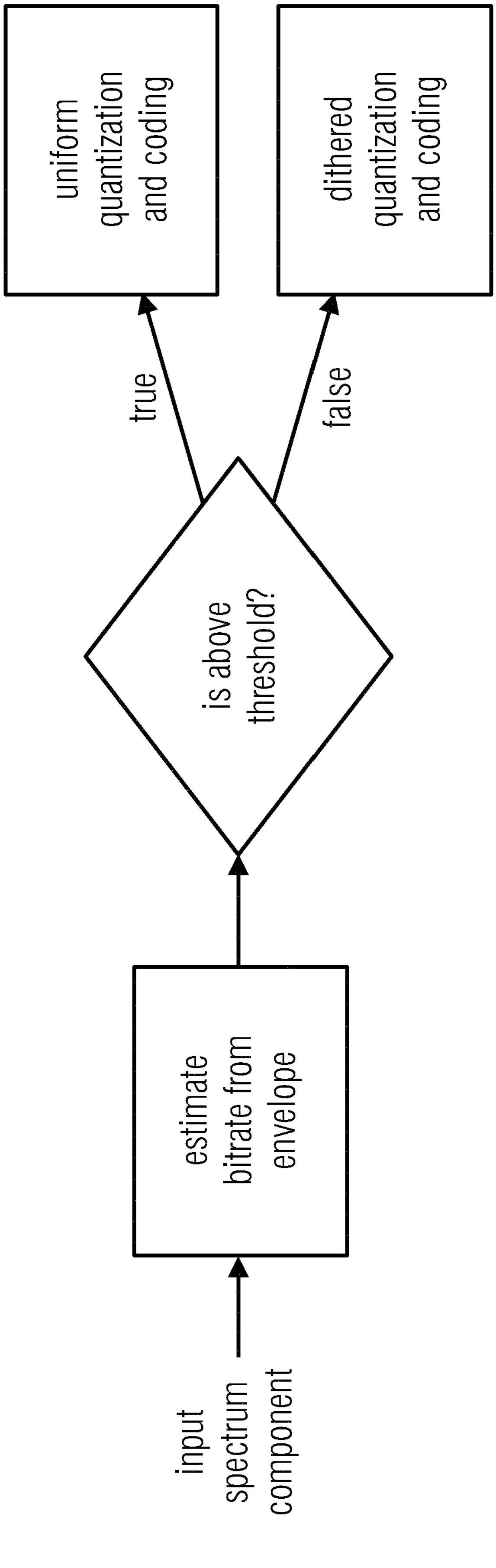

input spectrum component
estimate bitrate from envelope
is above threshold?
true
false
uniform quantization and coding
dithered quantization and coding

Fig. 4

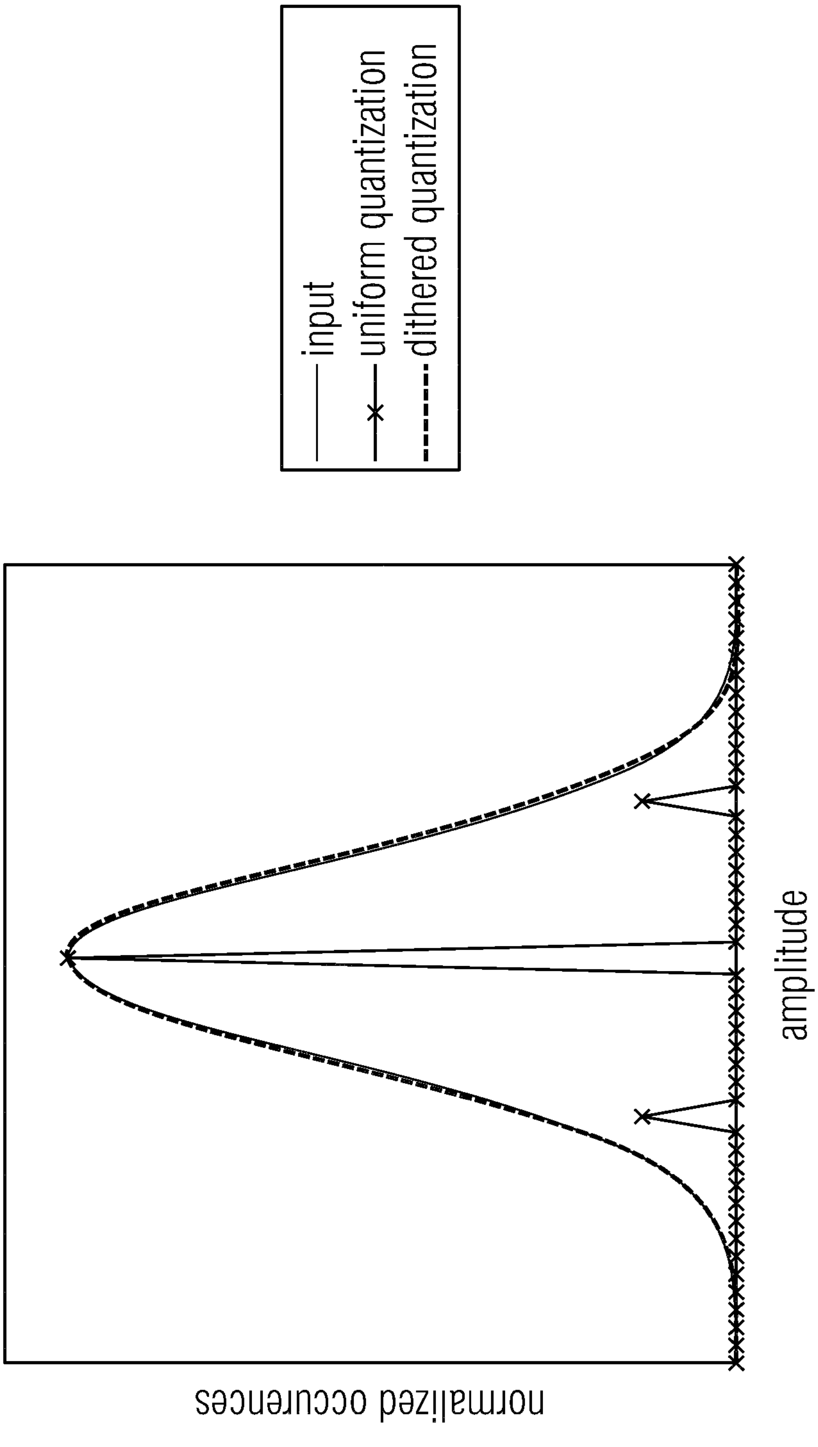

input
uniform quantization
dithered quantization
normalized occurences
amplitude

Fig. 5

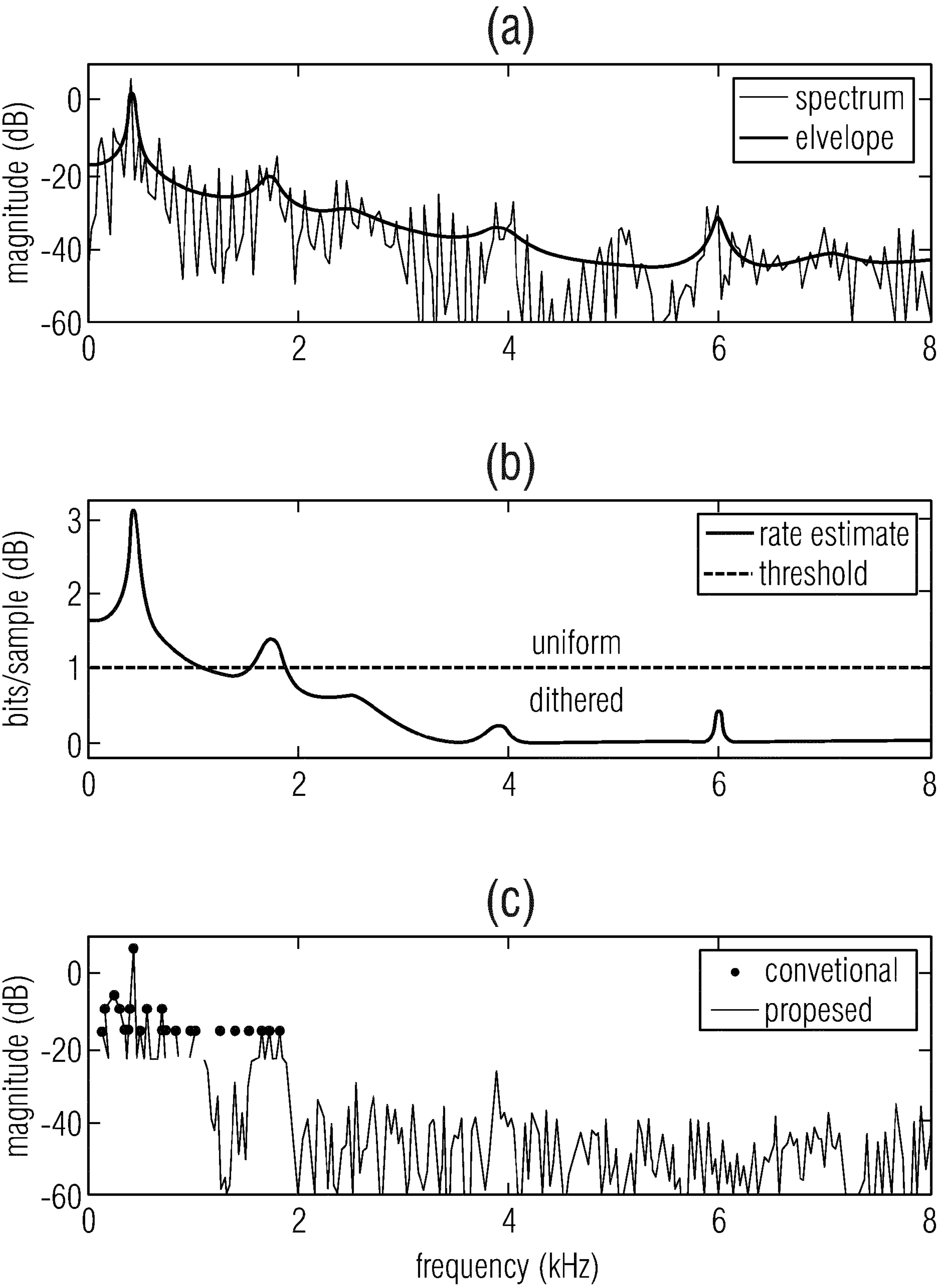

(a)
magnitude (dB)
0
-20
-40
-60
0
2
4
6
8
spectrum
elvelope
(b)
bits/sample (dB)
3
2
1
0
0
2
4
6
8
rate estimate
threshold
uniform
dithered
(c)
magnitude (dB)
0
-20
-40
-60
0
2
4
6
8
frequency (kHz)
● convetional
propesed

Fig. 6

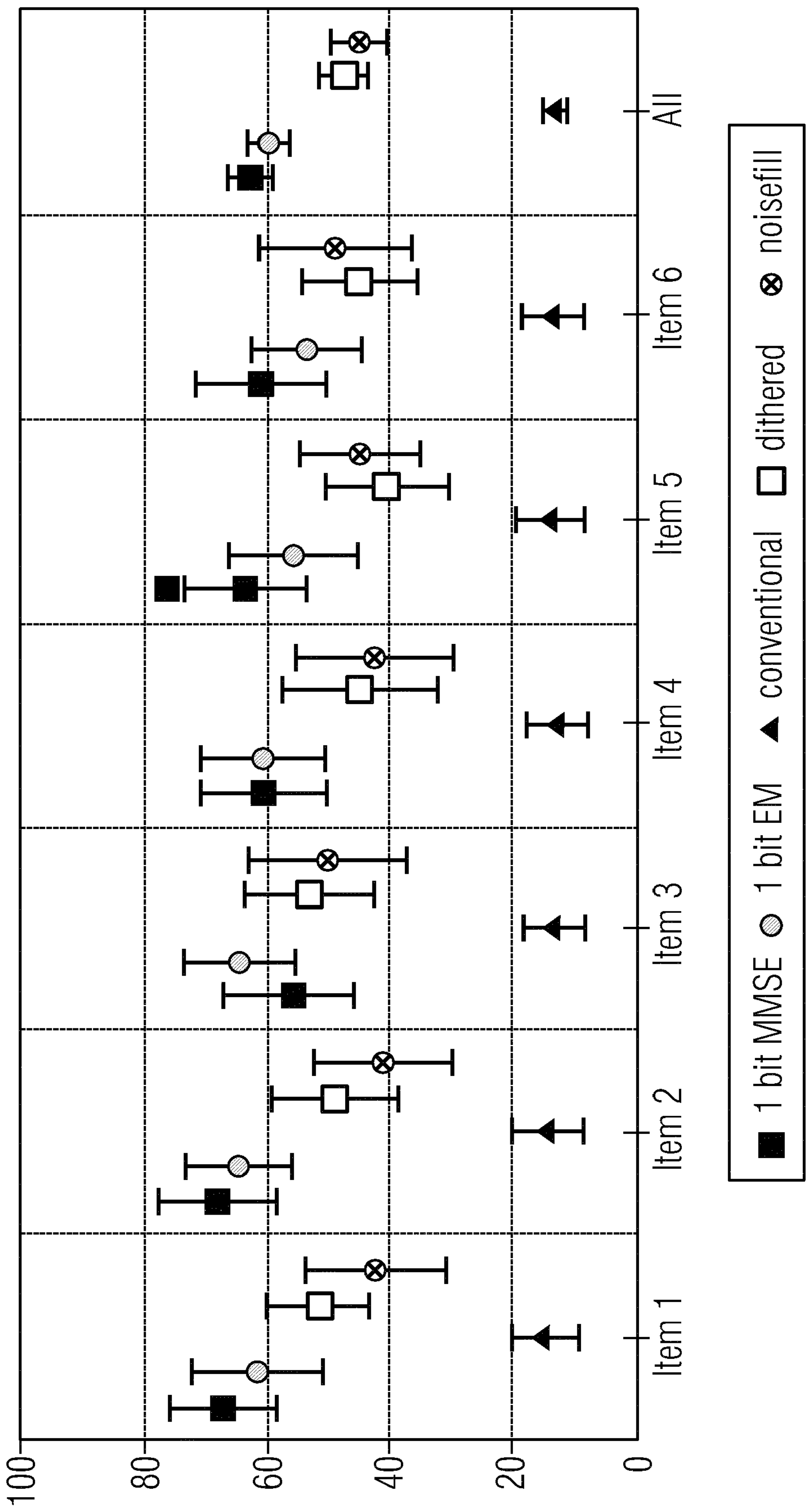

100
80
60
40
20
0
Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
All
■ 1 bit MMSE
● 1 bit EM
▲ conventional
□ dithered
⊗ noisefill

Fig. 7

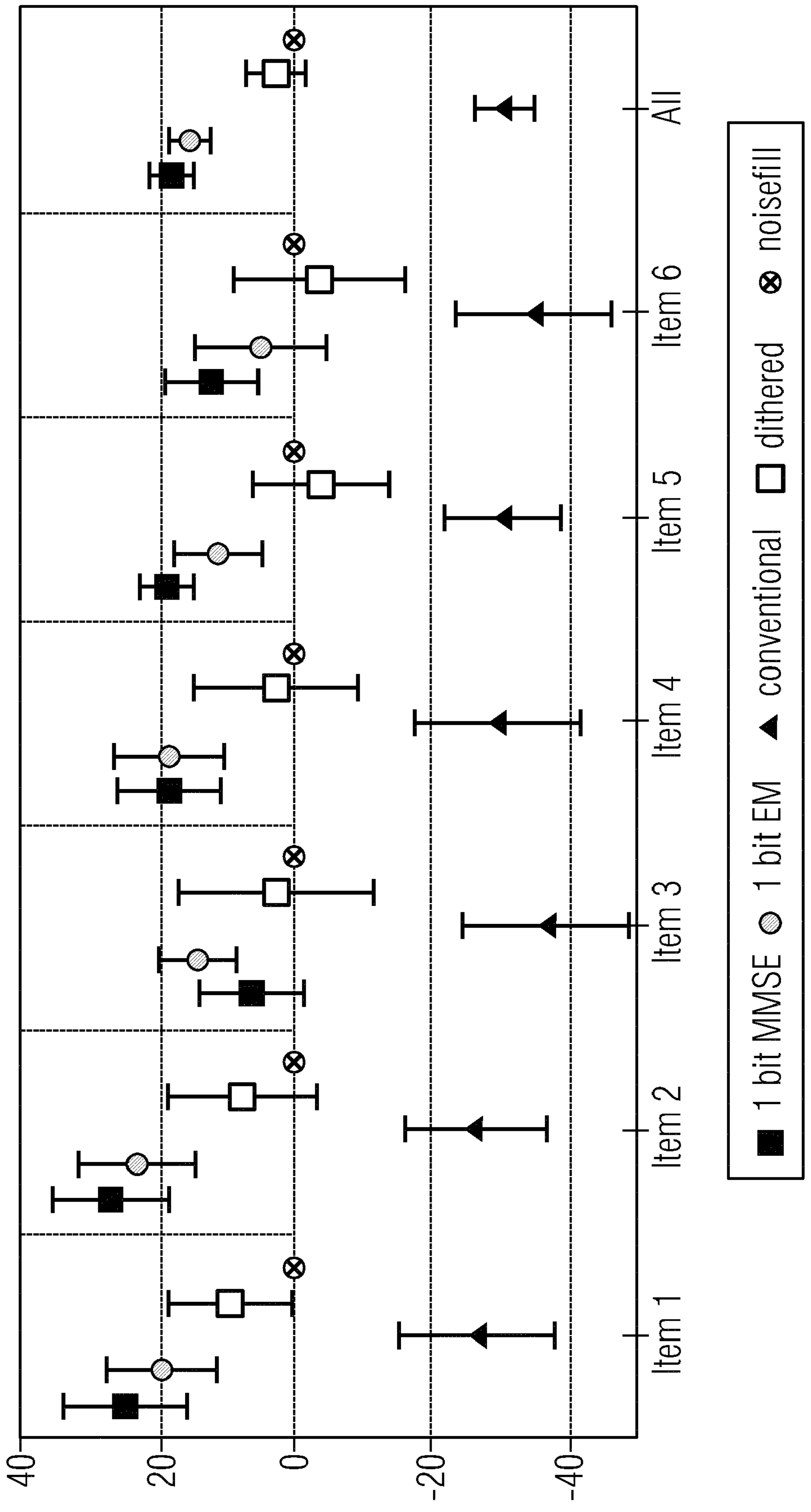

40
20
0
-20
-40
Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
All
■ 1 bit MMSE
● 1 bit EM
▲ conventional
□ dithered
⊗ noisefill

Fig. 8

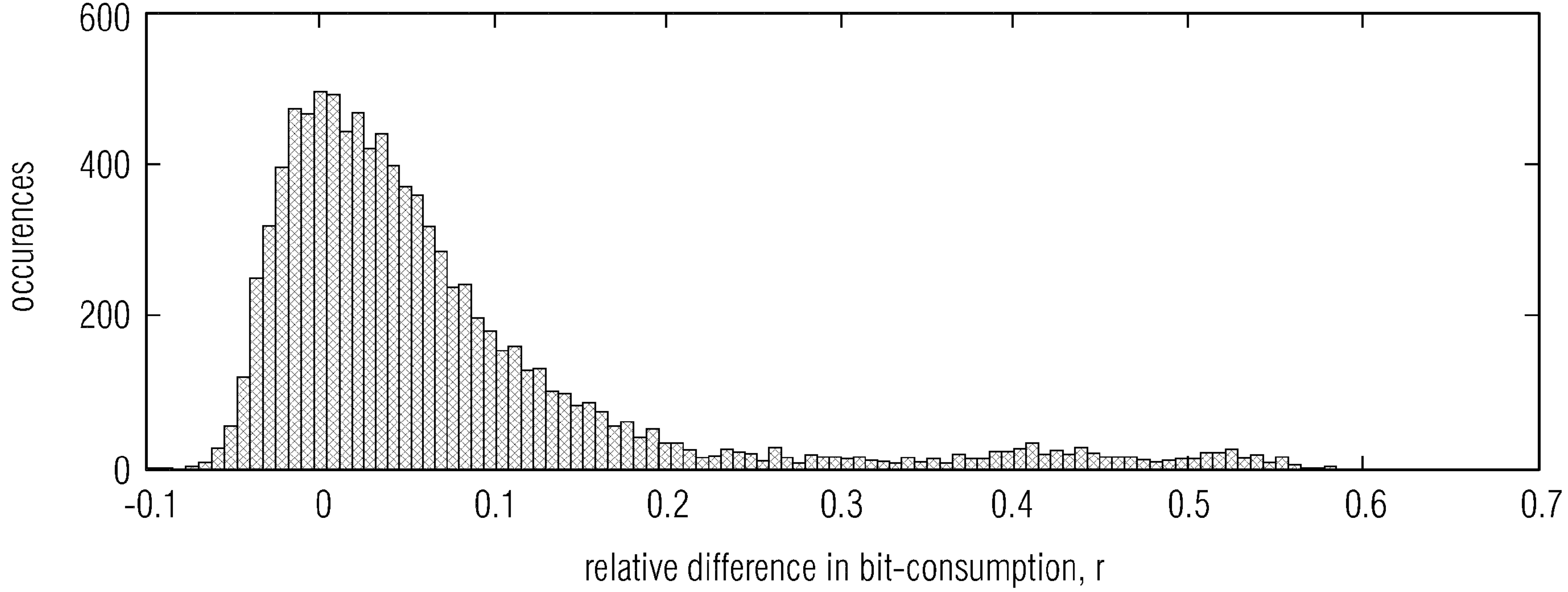

600
400
200
0
occurences
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
relative difference in bit-consumption, r

Fig. 9

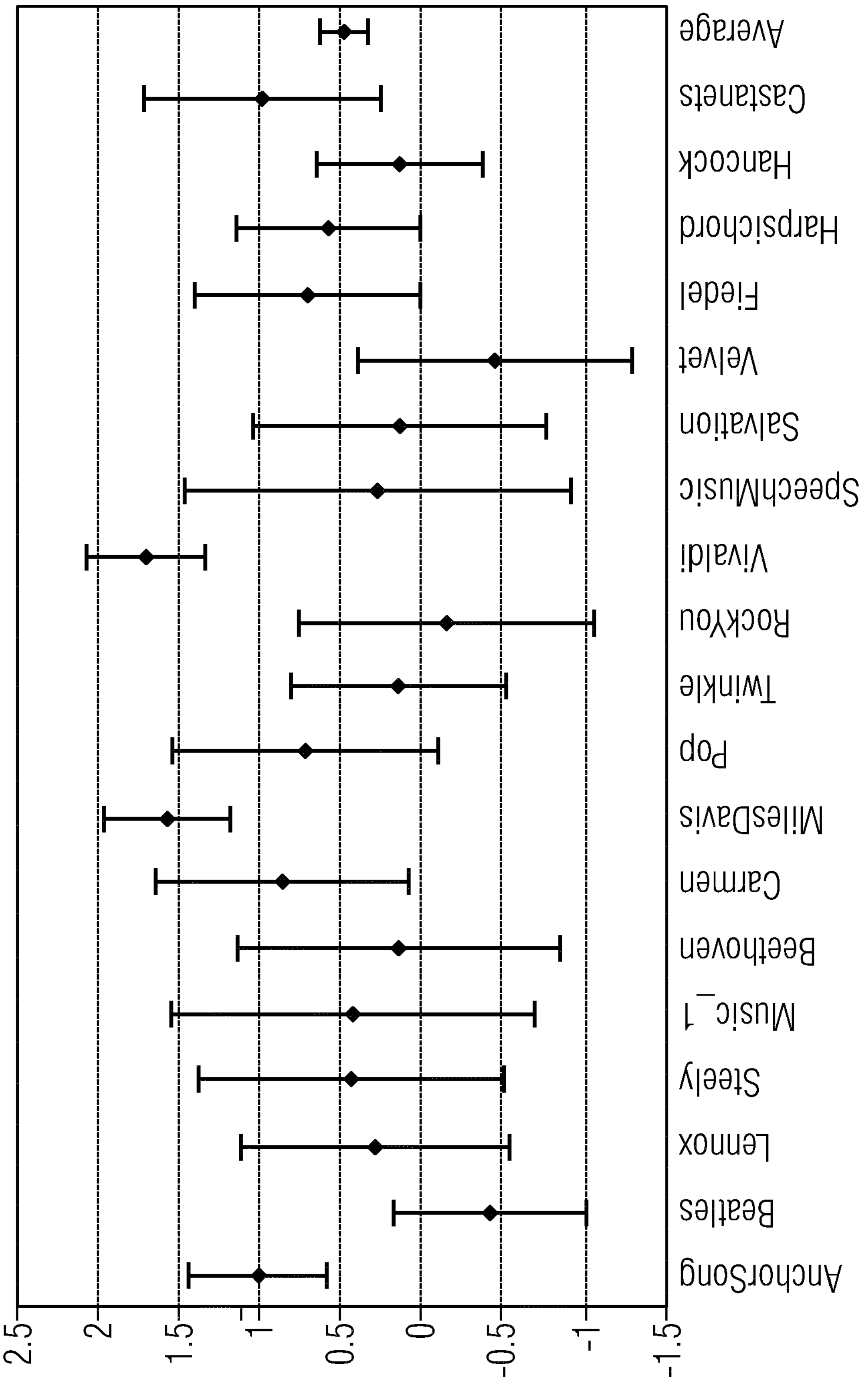

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
AnchorSong
Beatles
Lennox
Steely
Music_1
Beethoven
Carmen
MilesDavis
Pop
Twinkle
RockYou
Vivaldi
SpeechMusic
Salvation
Velvet
Fiedel
Harpsichord
Hancock
Castanets
Average

Fig. 10

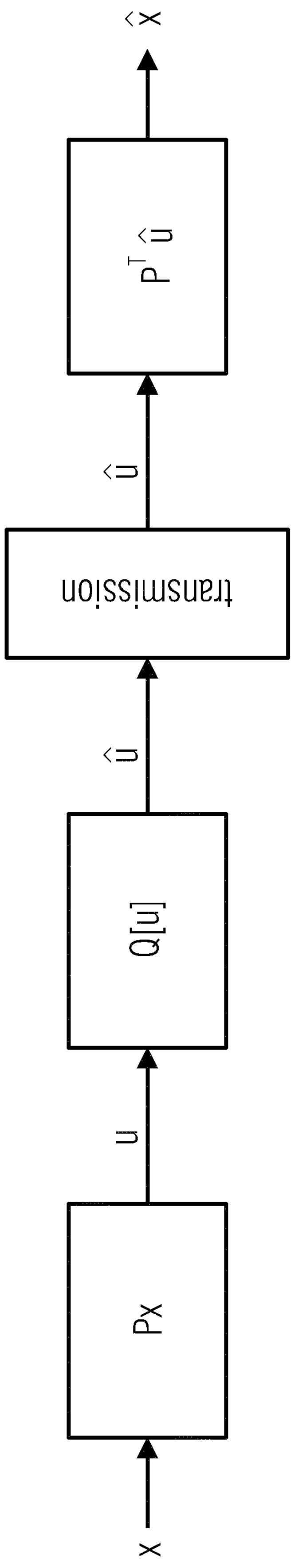
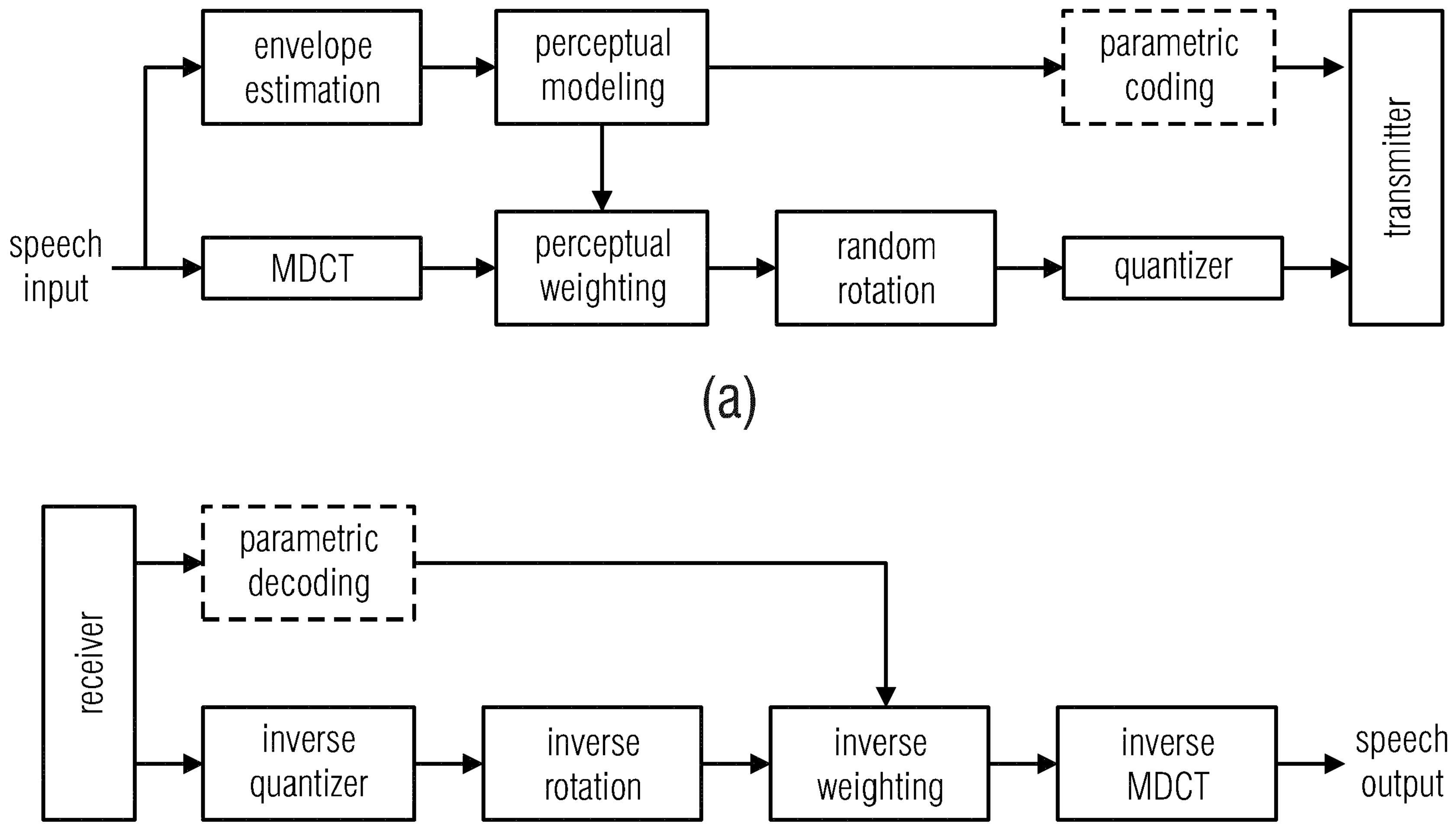


Fig. 11


speech input
envelope estimation
MDCT
perceptual modeling
perceptual weighting
random rotation
parametric coding
quantizer
transmitter
receiver
parametric decoding
inverse quantizer
inverse rotation
inverse weighting
inverse MDCT
speech output

(a)

(b)

Fig. 12

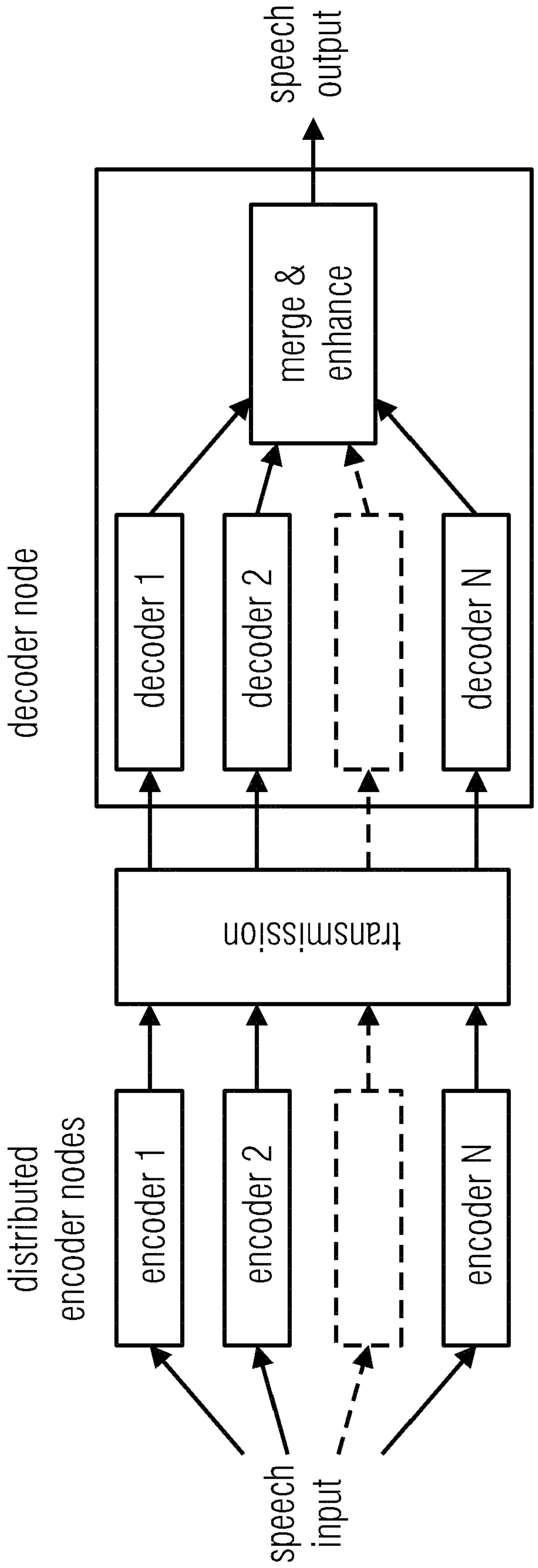

speech
input
distributed
encoder nodes
encoder 1
encoder 2
encoder N
transmission
decoder node
decoder 1
decoder 2
decoder N
merge &
enhance
speech
output

Fig. 13

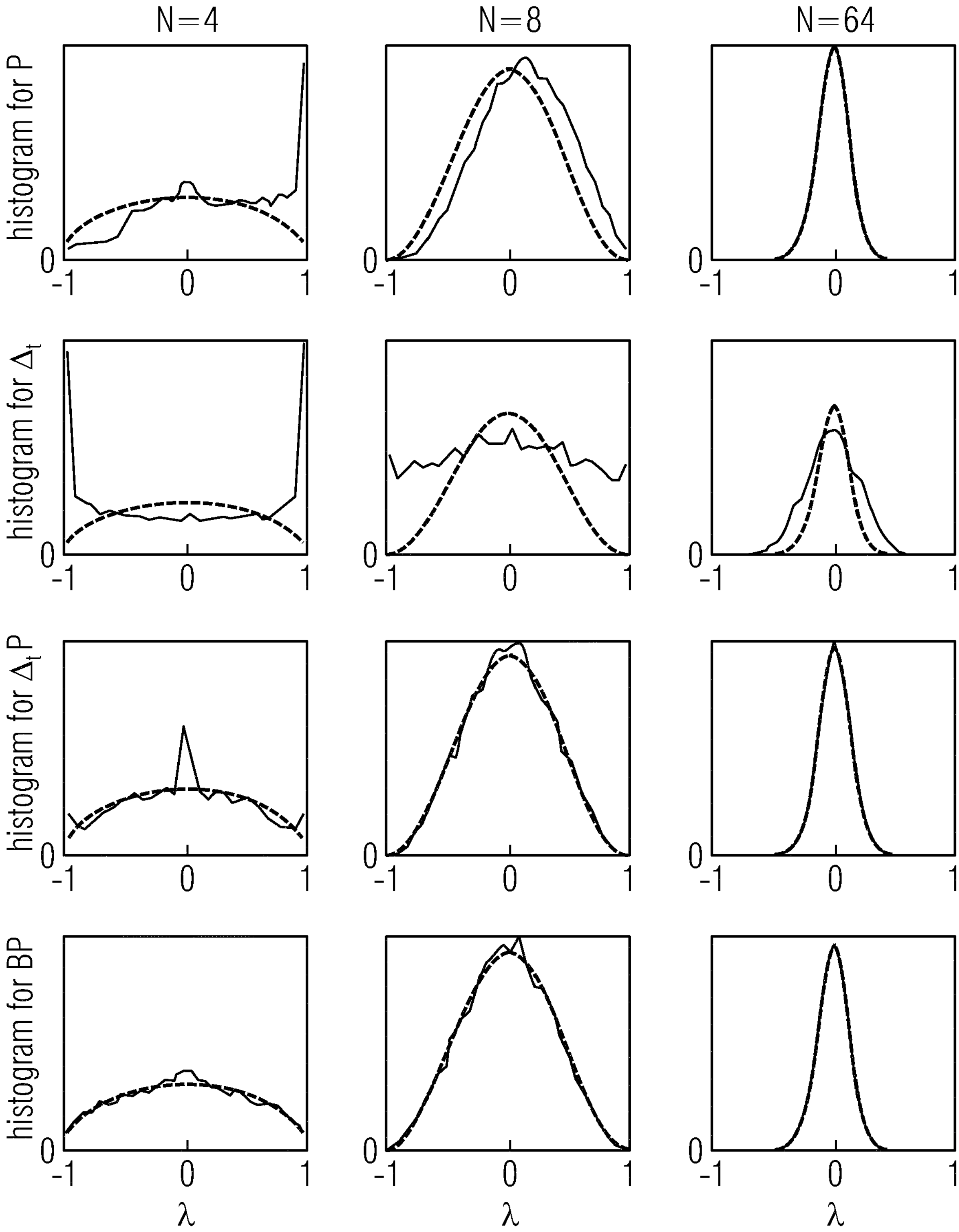

N=4
N=8
N=64
histogram for P
histogram for Δ_t
histogram for Δ_tP
histogram for BP
0
-1
0
1
λ

Fig. 14

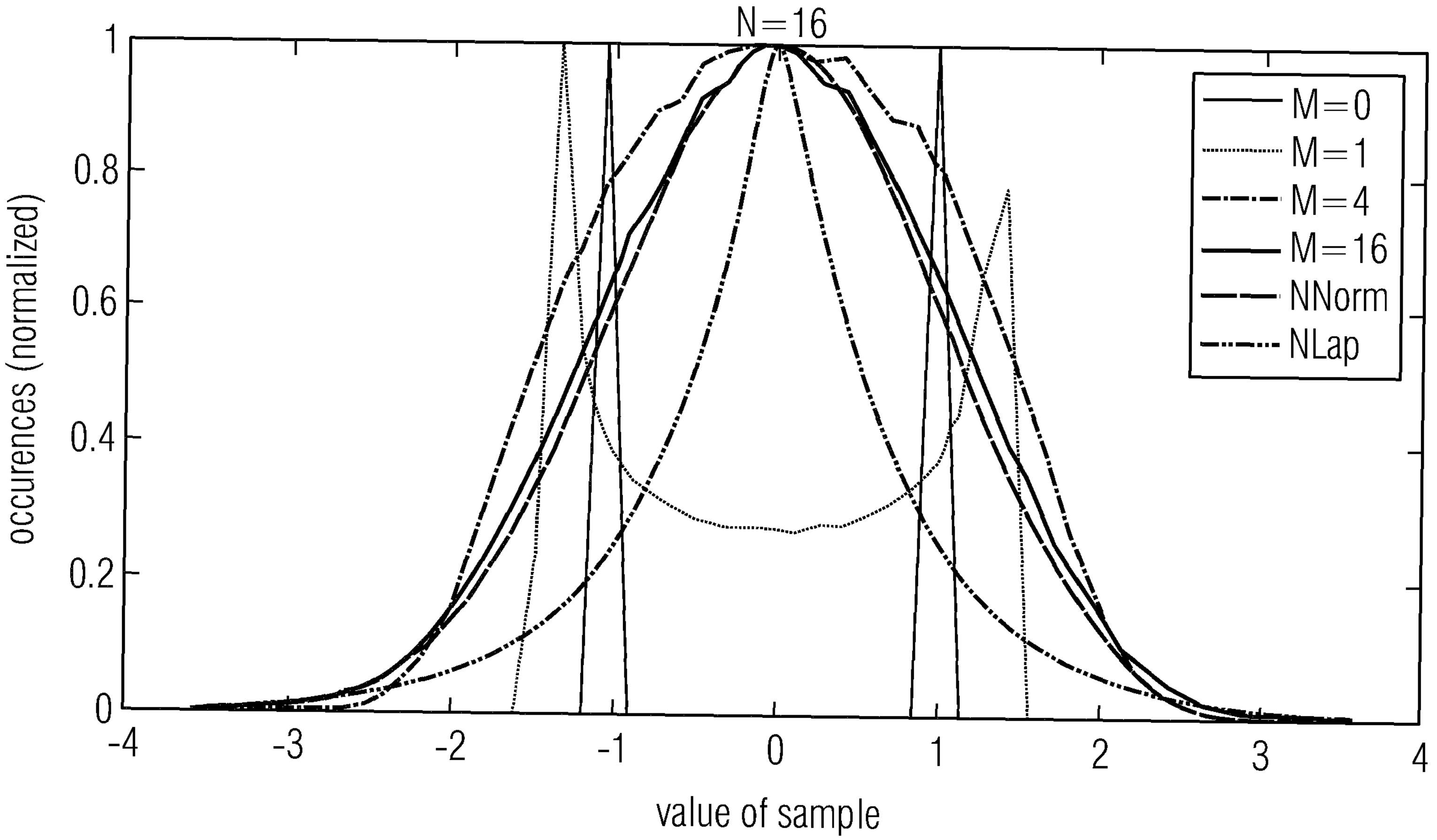

N=16
M=0
M=1
M=4
M=16
NNorm
NLap
occurences (normalized)
1
0.8
0.6
0.4
0.2
0
-4
-3
-2
-1
0
1
2
3
4
value of sample

Fig. 15

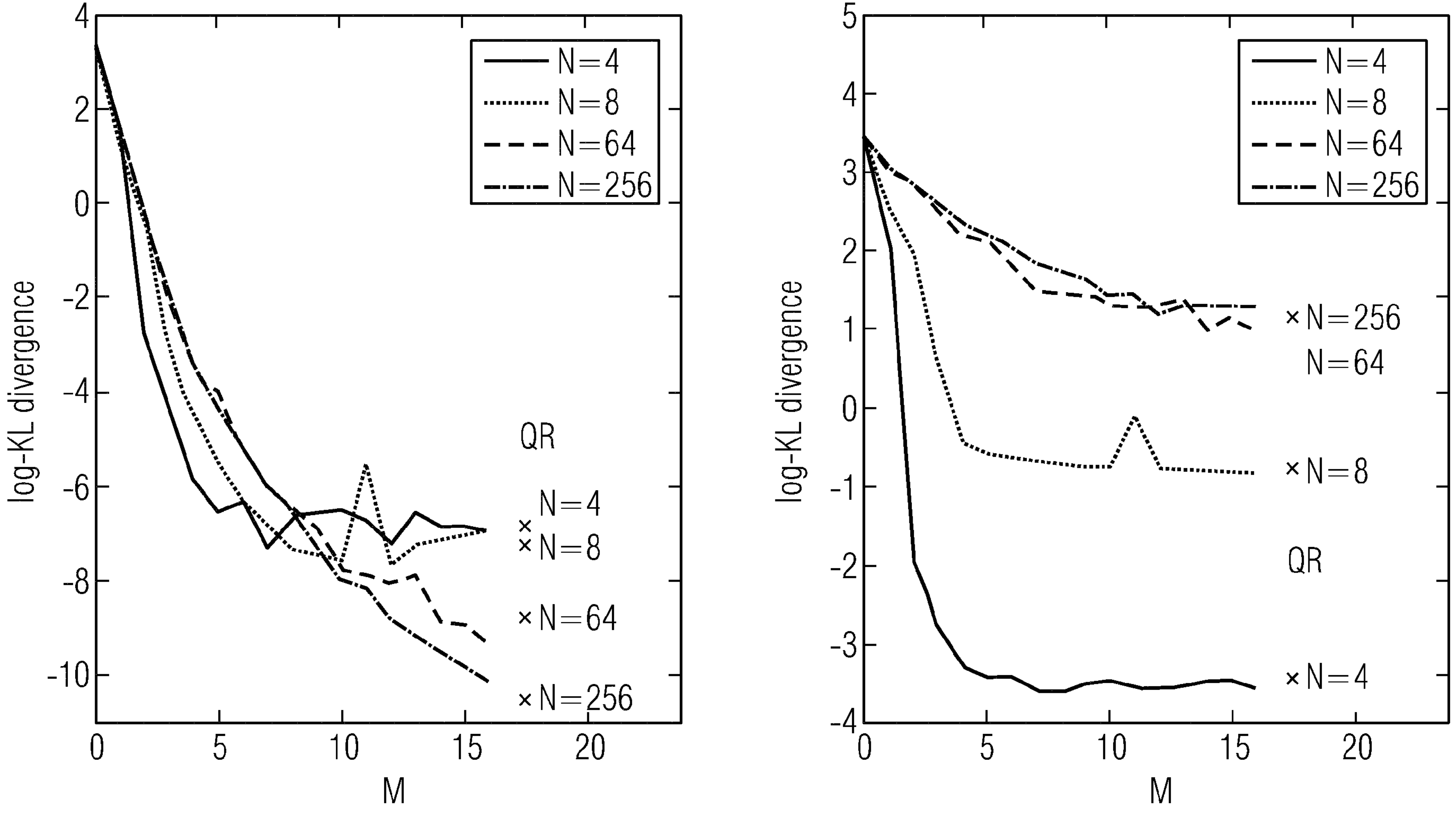

log-KL divergence
4
2
0
-2
-4
-6
-8
-10
0
5
10
15
20
M
N=4
N=8
N=64
N=256
QR
× N=4
× N=8
× N=64
× N=256
log-KL divergence
5
4
3
2
1
0
-1
-2
-3
-4
0
5
10
15
20
M
N=4
N=8
N=64
N=256
× N=256
N=64
× N=8
QR
× N=4

Fig. 16

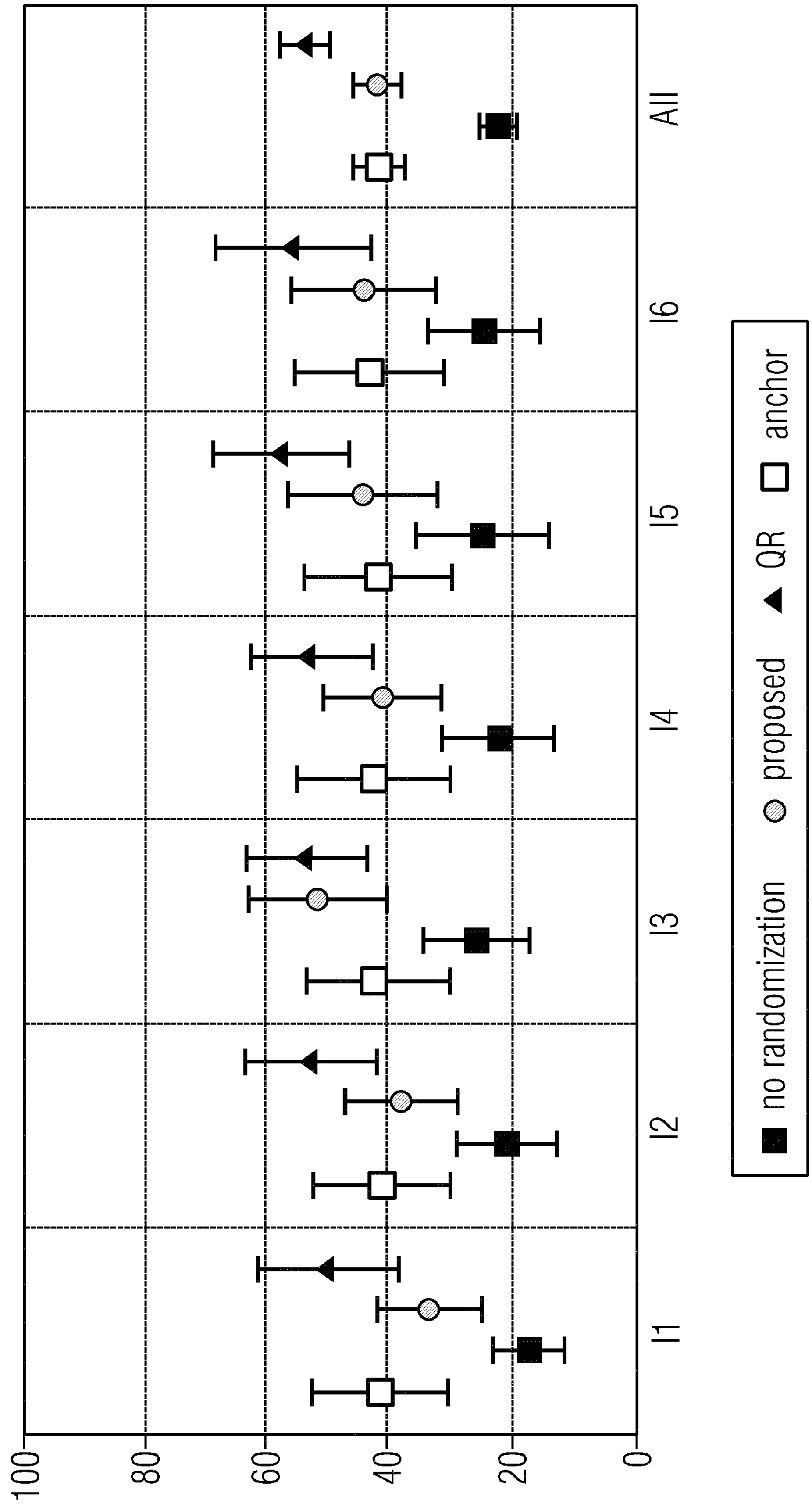

100
80
60
40
20
0
I1
I2
I3
I4
I5
I6
All
■ no randomization
● proposed
▲ QR
□ anchor

Fig. 17

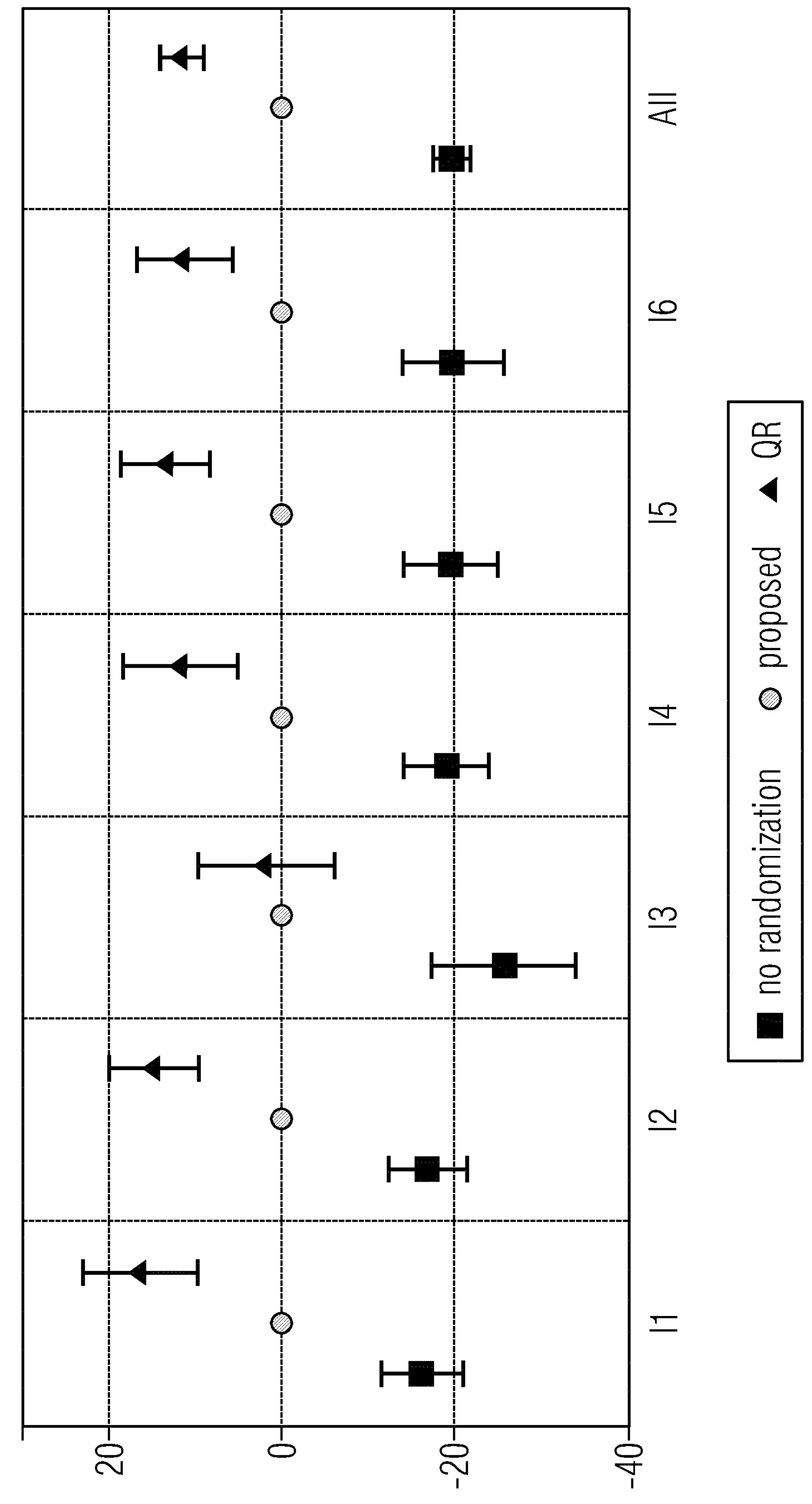


Fig. 18

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 18 18 7597

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	WO 2014/161994 A2 (DOLBY INT AB [NL]) 9 October 2014 (2014-10-09) * figures 1a,1b * * page 25, line 32 - page 26, line 26 * * page 27, line 15 - page 30, line 5 * * page 32, line 4 - page 33, line 3 * * page 33, lines 23-30 * * page 42, lines 18-25 * * page 46, line 32 - page 47, line 6 * * page 53, line 26 - page 54, line 15 * -----	1-18	INV. G10L19/22 ADD. G10L19/032 G10L19/00
X	BACKSTROM TOM ET AL: "Arithmetic coding of speech and audio spectra using tcx based on linear predictive spectral envelopes", 2015 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), IEEE, 19 April 2015 (2015-04-19), pages 5127-5131, XP033064629, DOI: 10.1109/ICASSP.2015.7178948 [retrieved on 2015-08-04] * page 5128 - page 5129 * ----- -/--	1,8, 15-18	TECHNICAL FIELDS SEARCHED (IPC) G10L

The present search report has been drawn up for all claims

1

Place of search	Date of completion of the search	Examiner
Munich	12 April 2019	Ramos Sánchez, U

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

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 18 18 7597

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	TOM BACKSTROM ET AL: "Fast Randomization for Distributed Low-Bitrate Coding of Speech and Audio", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, USA, vol. 26, no. 1, 1 January 2018 (2018-01-01), pages 19-30, XP058381855, ISSN: 2329-9290, DOI: 10.1109/TASLP.2017.2757601 * page 21, left-hand column, last paragraph - page 23, left-hand column, paragraph 4 * * page 24, left-hand column, paragraph 3 - right-hand column, paragraph 2 * * page 20, left-hand column, lines 11-13, paragraph 2 * * see references to "sign-quantization"; page 25 * -----	1-18	
A	BOUFOUNOS P T ET AL: "1-Bit compressive sensing", INFORMATION SCIENCES AND SYSTEMS, 2008. CISS 2008. 42ND ANNUAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 19 March 2008 (2008-03-19), pages 16-21, XP031282831, ISBN: 978-1-4244-2246-3 * page 17, left-hand column * * see 'A. Measurement Model'; page 18, left-hand column * ----- -/--	7	TECHNICAL FIELDS SEARCHED (IPC)

The present search report has been drawn up for all claims

1

Place of search	Date of completion of the search	Examiner
Munich	12 April 2019	Ramos Sánchez, U

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

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 18 18 7597

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
T	TOM BÄCKSTRÖM ET AL: "Dithered Quantization for Frequency-Domain Speech and Audio Coding", INTERSPEECH 2018, 1 January 2018 (2018-01-01), pages 3533-3537, XP055579878, ISCA DOI: 10.21437/Interspeech.2018-46 * page 3533 - page 3535, left-hand column * -----		
			TECHNICAL FIELDS SEARCHED (IPC)

1

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	12 April 2019	Ramos Sánchez, U

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

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO.

EP 18 18 7597

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.

12-04-2019

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2014161994 A2	09-10-2014	BR 112015025009 A2	18-07-2017
		CN 105144288 A	09-12-2015
		EP 2981961 A2	10-02-2016
		EP 3217398 A1	13-09-2017
		ES 2628127 T3	01-08-2017
		HK 1215751 A1	09-09-2016
		JP 6158421 B2	05-07-2017
		JP 6452759 B2	16-01-2019
		JP 2016519787 A	07-07-2016
		JP 2017182087 A	05-10-2017
		KR 20150139518 A	11-12-2015
		KR 20170078869 A	07-07-2017
		RU 2015141996 A	13-04-2017
		RU 2017143614 A	14-02-2019
		US 2016042744 A1	11-02-2016
		US 2018211677 A1	26-07-2018
		WO 2014161994 A2	09-10-2014

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.

Patent documents cited in the description

- US 7447631 B **[0313]**

Non-patent literature cited in the description

- **T. BÄCKSTRÖM.** Speech Coding with Code-Excited Linear Prediction. Springer, 2017 **[0313]**
- TS 26.445, EVS Codec Detailed Algorithmic Description. 3GPP Technical Specification (Release 12). 3GPP, 2014 **[0313]**
- Adaptive Multi-Rate (AMR-WB) Speech Codec. TS 26.190. 3GPP, 2007 **[0313]**
- Part 3: Unified Speech and Audio Coding. *MPEG-D (MPEG Audio Technologies),* 2012 **[0313]**
- **M. BOSI et al.** ISO/IECMPEG-2 advanced audio coding. *J. Audio Eng. Soc.,* 1997, vol. 45 (10), 789-814 **[0313]**
- **J. BENESTY ; M. SONDHI ; Y. HUANG.** Springer Handbook of Speech Processing. Springer, 2008 **[0313]**
- **A. ZAHEDI ; J. ØSTERGAARD ; S. H. JENSEN ; S. BECH ; P. NAYLOR.** Audio coding in wireless acoustic sensor networks. *Signal Process.,* 2015, vol. 107, 141-152 **[0313]**
- **T. BÄCKSTRÖM ; F. GHIDO ; J. FISCHER.** Blind recovery of perceptual models in distributed speech and audio coding. *Proc. Interspeech,* 2016, 2483-2487 **[0313]**
- **T. BÄCKSTRÖM ; J. FISCHER.** Coding of parametric models with randomized quantization in a distributed speech and audio codec. *Proc. ITG Fachtagung Sprachkommunikation,* 2016, 1-5 **[0313]**
- **P. T. BOUFOUNOS ; R. G. BARANIUK.** 1-bit compressive sensing. *Proc. IEEE Inf. Sci. Syst. 42nd Ann. Conf.,* 2008, 16-21 **[0313]**
- **A. MAGNANI ; A. GHOSH ; R. M. GRAY.** Optimal one-bit quantization. *Proc. IEEE Data Compression Conf.,* 2005, 270-278 **[0313]**
- **S. VAUDENAY.** Decorrelation: a theory for block cipher security. *J. Cryptol.,* 2003, vol. 16 (4), 249-286 **[0313]**
- **S. SAEEDNIA.** How to make the Hill cipher secure. *Cryptologia,* 2000, vol. 24 (4), 353-360 **[0313]**
- **C.-C. KUO ; W. THONG.** Reduction of quantization error with adaptive Wiener filter in low bit rate coding. *Proc. 11th Eur. IEEE Signal Process. Conf.,* 2002, 1-4 **[0313]**
- **G. E. ØIEN ; T. A. RAMSTAD.** On the role of Wiener filtering in quantization and DPCM. *Proc. IEEE Norwegian Signal Process. Symp. Workshop,* 2001 **[0313]**
- **J. RISSANEN ; G. G. LANGDON.** Arithmetic coding. *IBM J. Res.Develop.,* 1979, vol. 23 (2), 149-162 **[0313]**
- **J. D. GIBSON ; K. SAYOOD.** Lattice quantization. *Adv. Electron. Electron Phys.,* 1988, vol. 72, 259-330 **[0313]**
- **A. GERSHO ; R. M. GRAY.** Vector Quantization and Signal Compression. Springer, 1992 **[0313]**
- **Z. XIONG ; A. D. LIVERIS ; S. CHENG.** Distributed source coding for sensor networks. *IEEE Signal Process. Mag.,* September 2004, vol. 21 (5), 80-94 **[0313]**
- Distributed source coding. **Z. XIONG ; A. D. LIVERIS ; Y. YANG.** Handbook on Array Processing and Sensor Networks. Wiley-IEEE Press, 2009, 609-643 **[0313]**
- **M. BOSI ; R. E. GOLDBERG.** Introduction to Digital Audio Coding and Standards. Kluwer, 2003 **[0313]**
- **A. BERTRAND.** Applications and trends in wireless acoustic sensor networks: A signal processing perspective. *Proc. 18th IEEE Symp. Commun. Veh. Technol,* 2011, 1-6 **[0313]**
- **I. F. AKYILDIZ ; T. MELODIA ; K. R. CHOWDURY.** Wireless multimedia sensor networks: A survey. *IEEE Wireless Commun.,* December 2007, vol. 14 (6), 32-39 **[0313]**
- **B. GIROD ; A. M. AARON ; S. RANE ; D. REBOLLO-MONEDERO.** Distributed video coding. *Proc. IEEE,* January 2005, vol. 93 (1), 71-83 **[0313]**
- **F. DE LA HUCHA ARCE ; M. MOONEN ; M. VERHELST ; A. BERTRAND.** Adaptive quantization for multi-channel Wiener filter-based speech enhancement in wireless acoustic sensor networks. *Wireless Commun. Mobile Comput.,* 2017, vol. 2017 **[0313]**
- **A. ZAHEDI ; J. ØSTERGAARD ; S. H. JENSEN ; P. NAYLOR ; S. BECH.** Coding and enhancement in wireless acoustic sensor networks. *Proc. IEEE Data Compression Conf.,* 2015, 293-302 **[0313]**

- **A. MAJUMDAR ; K. RAMCHANDRAN ; L. KOZINTSEV.** Distributed coding for wireless audio sensors. *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.,* 2003, 209-212 **[0313]**
- **H. DONG ; J. LU ; Y. SUN.** Distributed audio coding in wireless sensor networks. *Proc. IEEE Int. Conf. Comput. Intell. Secur.,* 2006, vol. 2, 1695-1699 **[0313]**
- **G. BARRIAC ; R. MUDUMBAI ; U. MADHOW.** Distributed beamforming for information transfer in sensor networks. *Proc. 3rd Int. Symp. Inf. Process. Sens. Netw.,* 2004, 81-88 **[0313]**
- **R. LIENHART ; I. KOZINTSEV ; S. WEHR ; M. YEUNG.** On the importance of exact synchronization for distributed audio signal processing. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 2003, vol. 4, IV-840-3 **[0313]**
- **O. ROY ; M. VETTERLI.** Rate-constrained collaborative noise reduction for wireless hearing aids. *IEEE Trans. Signal Process.,* February 2009, vol. 57 (2), 645-657 **[0313]**
- **S. BRAY ; G. TZANETAKIS.** Distributed audio feature extraction for music. *Proc. Int. Conf. Music Inf. Retrieval,* 2005, 434-437 **[0313]**
- Distributed speech recognition. **N. RAJPUT ; A. A. NANAVATI.** Speech in Mobile and Pervasive Environments. Wiley, 2012, 99-114 **[0313]**
- Distributed speech recognition standards. **D. PEARCE.** Automatic Speech Recognition on Mobile Devices and Over Communication Networks. Springer, 2008, 87-106 **[0313]**
- **S. KORSE ; T. JAHNEL ; T. BÄCKSTRÖM.** Entropy coding of spectral envelopes for speech and audio coding using distribution quantization. *Proc. Interspeech,* 2016, 2543-2547 **[0313]**
- **S. DAS ; A. CRACIUN ; T. JAHNEL ; T. BÄCKSTRÖM.** Spectral envelope statistics for source modelling in speech enhancement. *Proc. ITG Fachtagung Sprachkommunikation,* 2016, 1-5 **[0313]**
- **G. H. GOLUB ; C. F. VAN LOAN.** Matrix Computations. John Hopkins Univ. Press, 2004 **[0313]**
- **V. PULKKI.** Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc.,* 2007, vol. 55 (6), 503-516 **[0313]**
- **A. EDELMAN ; N. R. RAO.** Random matrix theory. *Acta Numerica,* 2005, vol. 14, 233-297 **[0313]**
- **D. KNUTH.** The Art of Computer Programming. Addison-Wesley, 1998 **[0313]**
- Seminumerical algorithms. **D. E. KNUTH.** The Art of Computer Programming. Addison-Wesley, 2007, vol. 2 **[0313]**
- **P. DIACONIS ; M. SHAHSHAHANI.** The subgroup algorithm for generating uniform random variables. *Probab. Eng. Inf. Sci.,* 1987, vol. 1 (1), 15-32 **[0313]**
- **G. W. STEWART.** The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM J. Numer. Anal.,* 1980, vol. 17 (3), 403-409 **[0313]**
- Pseudorandom number generators. **P. L'ECUYER.** Encyclopedia of Quantitative Finance. Wiley, 2010 **[0313]**
- **M. MATSUMOTO ; T. NISHIMURA.** Mersenne twister:A623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.,* 1998, vol. 8 (1), 3-30 **[0313]**
- Mersenne twister-A pseudo random number generator and its variants. **A. JAGANNATAM.** Dept. Elect. Comput. Eng. George Mason Univ, 2008 **[0313]**
- **T. BÄCKSTRÖM ; C. R. HELMRICH.** Arithmetic coding of speech and audio spectra using TCX based on linear predictive spectral envelopes. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* April 2015, 5127-5131 **[0313]**
- **J. FISCHER ; T. BÄCKSTRÖM.** Wiener filtering in distributed speech and audio coding. *IEEE Signal Process. Lett.,* 2017 **[0313]**
- **T. BÄCKSTRÖM.** Estimation of the probability distribution of spectral fine structure in the speech source. *Proc. Interspeech,* 2017, 344-348 **[0313]**
- **J. S. GAROFOLO et al.** TIMIT: Acoustic-Phonetic Continuous Speech Corpus. *Linguistic Data Consortium,* 1993 **[0313]**
- Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems. *ITU-R Recommendation BS.1534,* 2003 **[0313]**
- **S. NADARAJAH.** A generalized normal distribution. *J. Appl. Statist.,* 2005, vol. 32 (7), 685-694 **[0313]**
- **C. WALCK.** Handbook on Statistical Distributions for Experimentalists. Univ. Stockholm, 2007 **[0313]**
- Introduction to the Dirichlet distribution and related processes. **A. BELA ; A. FRIGYIK ; M. GUPTA.** Tech. Rep. UWEETR-2010-0006. Dept. Elect. Eng., Univ. Washington, 2010 **[0313]**
- **M. SCHOEFFLER ; F.-R. STÖTER ; B. EDLER ; J. HERRE.** Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA). *Proc. 1st Web Audio Conf.,* 2015 **[0313]**
- **M. NEUENDORF ; M. MULTRUS ; N. RETTELBACH ; G. FUCHS ; J. ROBILLIARD ; J. LECOMTE ; S. WILDE ; S. BAYER ; S. DISCH ; C. HELMRICH.** The ISO/MPEG unified speech and audio coding standard - consistent high quality for all content types and at all bit rates. *Journal of the AES,* 2013, vol. 61 (12), 956-977 **[0313]**
- Frame error robust narrow-band and wideband embedded variable bitrate coding of speech and audio from 8-32 kbit/s. *ITU-T G.718,* 2008 **[0313]**
- **J. MÄKINEN ; B. BESSETTE ; S. BRUHN ; P. OJALA ; R. SALAMI ; A. TALEB.** AMR-WB+: a new audio coding standard for 3rd generation mobile audio services. *Proc. ICASSP,* 2005, vol. 2, 1109-1112 **[0313]**

- **M. BOSI ; K. BRANDENBURG ; S. QUACKENBUSH ; L. FIELDER ; K. AKAGIRI ; H. FUCHS ; M. DIETZ ; J. HERRE ; G. DAVIDSON ; Y. OIKAWA.** ISO/IEC MPEG-2 Advanced audio coding. *101 AES Convention,* 2012 **[0313]**
- **G. FUCHS ; M. MULTRUS ; M. NEUENDORF ; R. GEIGER.** Mdct-based coder for highly adaptive speech and audio coding. *European Signal Processing Conference (EUSIPCO 2009),* 2009, 24-28 **[0313]**
- **I. H. WITTEN ; R. M NEAL ; J. G. CLEARY.** Arithmetic coding for data compression. *Communications of the ACM,* 1987, vol. 30 (6), 520-540 **[0313]**
- Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800,* 1996 **[0313]**
- Low delay LPC and MDCT-based audio coding in the EVS codec. **G. FUCHS ; C. R. HELMRICH ; G. MARKOVIĆ ; M. NEUSINGER ; E. RAVELLI ; T. MORIYA.** Proc. ICASSP. IEEE, 2015, 5723-5727 **[0313]**
- **S DISCH ; A. NIEDERMEIER ; C. R. HELMRICH ; C. NEUKAM ; K. SCHMIDT ; R. GEIGER ; J. LECOMTE ; F. GHIDO ; F. NAGEL ; B. EDLER.** Intelligent gap filling in perceptual transform coding of audio. *Audio Engineering Society Convention 141. Audio Engineering Society,* 2016 **[0313]**
- **J. VANDERKOOY ; S. P. LIPSHITZ.** Dither in digital audio. *Journal of the Audio Engineering Society,* 1987, vol. 35 (12), 966-975 **[0313]**
- **R. W. FLOYD ; L. STEINBERG.** An adaptive algorithm for spatial gray-scale. *Proc. Soc. Inf. Disp.,* 1976, vol. 17, 75-77 **[0313]**
- **M LI ; J. KLEJSA ; W. B. KLEIJN.** Distribution preserving quantization with dithering and transformation. *IEEE Signal Processing Letters,* 2010, vol. 17 (12), 1014-1017 **[0313]**
- **T. BÄCKSTRÖM.** Enumerative algebraic coding for ACELP. *Proc. Interspeech,* 2012 **[0313]**
- **T. BÄCKSTRÖM ; J. FISCHER.** Fast randomization for distributed low-bitrate coding of speech and audio. *IEEE/ACM Trans. Audio, Speech, Lang. Process.,* January 2018, vol. 26 (1 **[0313]**
- **J.-M. VALIN ; G. MAXWELL ; T. B. TERRIBERRY ; K. VOS.** High-quality, low-delay music coding in the OPUS codec. *Audio Engineering Society Convention 135. Audio Engineering Society,* 2013 **[0313]**