Encoder and Decoder

Description

Embodiments of the present invention refer to an encoder and decoder and to corresponding methods. In general, embodiments of the present invention are in the field of neural coding and bandwidth extension (BWE). Preferred embodiments refer to classical bandwidth extension for any speech coder (SC).

Bandwidth extension is a technique used in speech coding to enhance the quality of speech transmission in situations where the available bandwidth or the possible bit-rate is limited. In essence, it is a method of expanding the frequency range of a speech core-coder, like the classical speech coder CELP (Code-Excited Linear Prediction), beyond the Nyquist frequency of its internal sampling rate, which can improve the perceived quality of the reconstructed speech signal at the decoder side.

The same principle can be used for neural speech and audio coders. It has then the additional merit to limit the complexity required by the neural coder, using complex models based mainly on Deep Neural Networks (DNNs). However, efficient Bandwidth extension schemes for speech rely on a source-filter modeling of the speech production and additional exploit the fact that the same paradigm is used by the speech coder used for the baseband.

Bandwidth extension is very well studied and established technique, already deployed in different existing standards like, HeAAC defined in MPEG-4 Audio and 3GPP EVS. It is usually built over a baseband coder, aka core-coder, like a speech coder of type CELP or a generic transform-based audio coding, like AAC and TCX, used in HeAAC and EVS respectively. In consequence, bandwidth extension can be performed either in time domain, or in frequency domain or in both domains. However, the great majority of the techniques dissociate the modelling of the frequency fine structure, called excitation in Time Domain, and coarse spectral structure, also called spectral envelope.

For great bit saving, the principle is based on generating the fine structured high frequency (HF) content from the transmitted low frequency (LF) content from the baseband coder. The high frequencies are then spectrally shaped and/or post processed before being mixed at

the decoder side to the decoded baseband. The whole process can be steered by transmitted parameters.

Such techniques work particularly fine for extending WB speech coder, like CELP to SWB. It can be found in standards like in 3GPP AMR-WB+ [2] and 3GPP EVS [1]. In both cases the LF fine structure can be easily recovered and exploited by the BWE, since it be directly derived front the Linear Prediction residual or the associated coded excitation. Moreover, some well interpretable parameters representing salient acoustic features, like voicing factor, fundamental frequency or spectral envelope related parameters, can be directly inherited from the baseband coding scheme.

On the other hand, a neural speech coder, as in [3], is generally a monolithic block composed of several layers of neurons parameterized by learning and leading to a so-called latent representation in which quantization takes place.

The latent representation is generally learned and difficult to interpret which might have drawbacks, including their lack of direct usage for classical bandwidth extension. Therefore, there is need for an improved approach.

It is an objective of the present invention to improve the bandwidth extension.

This objective is solved by the subject-matter of the independent claims.

An embodiment of the present invention provides an encoder for coding an audio signal which comprises a band-limited signal portion and an extended-band signal portion. The encoder comprises the two central entities baseband encoder and bandwidth extension encoder. The baseband encoder is configured to encode the band-limited signal portion, wherein the baseband encoder comprises at least one learnable layer. The encoder comprises a linear prediction entity for performing linear prediction on the extended-band signal portion.

Embodiments of this encoder aspect are based on the finding that a baseband signal, like low frequency portion, or in general a band-limited signal portion can be preferably encoded using a neural encoder, i.e., an encoder comprising a learnable layer, while an extended-band signal portion, like a high frequency signal portion can be preferably encoded using a so-called bandwidth encoder being based on linear prediction. Due to the combination of

the two encoders, an efficient combination of neural speech coders and classical bandwidth extension techniques can be provided. According to embodiments, relevant information can be extracted that can steer or guide the (classical speech) BWE.

According to embodiments, the encoder comprises a preprocessor, like a filter so as to obtain the band-limited signal portion and the extended-band signal portion based on the (input) audio signal. For example, the preprocessor may comprise a transform entity like a Digital Fourier Transform (DFT) entity. Additionally or alternatively, the preprocessor may comprise a frequency domain truncation and normalization entity serving as lowpass filter in the transform domain and in an inverse time-frequency transform like an inverse DFT in a baseband path coupled with a baseband encoder. According to another variant, the preprocessor can additionally or alternatively comprise a frequency domain demodulator and a truncation entity in the transform domain and an inverse transform like a DFT in a bandwidth extension path coupled with a bandwidth extension encoder. All variants enable preferably to separate the band-limited signal portion and the extended-band signal portion starting from the audio signal, so that the baseband encoder and the bandwidth encoder can further process the separated signal portions. Preferably, the bandwidth extension models a high-band signal , wherein the baseband encoder codes a band-limited signal corresponding to a low-band signal. According to embodiments, the transform and the inverse transform are respectively a time-frequency transform and an inverse time-frequency transform. Alternatively, any sub-band decomposition or filterbank can be used to extract the two different signals conveyed to the baseband coder and the bandwidth extension.

Below, the baseband encoder will be defined with respect to optional details. The baseband encoder may, according to embodiments, comprise a neural coder and/or an auto-encoder architecture and/or a Vector Quantized Variational Auto-Encoder (VQ-VAE). For example, the baseband encoder may perform encoding based on a quantization of a latent representation of the band-limited signal portion. According to embodiments, the baseband encoder is trained or trainable using adversarial losses. According to embodiments, the baseband encoder is configured to obtain a latent, latent being quantized encoded by the baseband encoder. Note, preferably, i.e., according to preferred embodiments, the quantization of a latent is a vector quantization or a scalar quantization. For example, the quantization of the latent is structured into multi-stages or in a way of refining a residual of a previous quantized stage (also known as residual quantization). Due to the structure of

the latent, these kinds of quantization techniques are advantageous, wherein different quantization techniques are possible as well.

Below, optional features according to embodiments for the bandwidth extension encoder will be discussed. As discussed above, classical bandwidth extension may be used. For example, the bandwidth extension encoder may comprise a linear prediction coding analysis stage (LPC analysis stage) configured to perform a LPC analysis. Here, the audio signal may be partitioned into frames, wherein the LPC analysis stage is configured to perform LPC analysis at frame basis. Additionally or alternatively the bandwidth extension encoder is configured to perform interpolation and/or linear interpolation of the LPC coefficients between frames. Due to the usage of a bandwidth extension encoder being based on an LPC analysis, all advantages especially present for a bandwidth extension signal, like a higher frequency signal can be obtained for the encoder. For example, the LPC analysis entity comprises a short-term linear analysis to obtain a set of LPC coefficients, the LPC coefficients are then quantized and encoded by the bandwidth extension encoder. This means that the bandwidth extension encoder may comprise a quantization entity at the output. According to advantageous embodiments, the LPC coefficients are quantized in a transformed domain, like linear spectral frequencies (LSFs). In another embodiment, the LPC coefficients of the extended-band are not transmitted, but estimated at the decoder side. This estimator may comprise at least one learnable layer.

According to embodiments, the LPC analysis entity is configured to obtain a filter AHB(z). For example, the filter AHB(z) may be used to whiten the extended-band signal portion and to obtain a residual $e_{HB}(n)$ based on the formula

$$e_{HB}(n) = s_{HB}(n) - \sum_{i=1}^{M_{HB}} a_i \, s_{HB}(n-i) \quad .$$

Here, M_{HB} is the LPC order, n is a time-domain sample index of the audio signal, and are the LPC coefficients, which could be obtained after quantization and interpolation in the LSF domain. For this embodiments, the bandwidth extension encoder (or its quantization entity) is configured to code and quantize the energy after the residual signal $e_{HB}(n)$ had been obtained and/or after the prediction.

In general, the bandwidth extension entity is configured to code and quantize an energy of the residual exploiting information derived from the band-limited residual, so that energy of the residual in the extended-band is predicted by the energy of the linear prediction residual

signal computed on the band-limited signal for deriving a residual of the energy prediction, which is then quantized and/or an information output to the decoder. As discussed in context of the prior art, the latent representation is generally learned and difficult to interpret making it difficult to extract relevant information that can steer and guide a classical speech bandwidth encoder. By use of this, this drawback is solved, so that the described solution advantageously enables to steer and guide the bandwidth extension by relevant information extracted from the band-limited signal.

It should be noted that the quantization time resolution of the residual of the energy prediction or the energy of the residual signal is higher or equal to the quantization time resolution of the LPC coefficients or LSFs. Below, a preferred embodiment for a baseband encoder will be given. The baseband encoder comprises a format definer configured to define a first multi-dimensional band-limited signal representation of the band-limited signal, the first multi-dimensional band-limited representation of the band-limited signal including at least

- a first dimension so that a plurality of mutually subsequent frames is ordered according to the first dimension; and
- a second dimension so that a plurality of samples of at least one frame are ordered according to the second dimension, and
- the at least one learnable layer configured to process the first multidimensional band-limited signal representation of the band-limited signal, or processed version of the first multi-dimensional band-limited audio signal representation.

Another aspect of embodiments provide a decoder for decoding an audio signal comprising a band-limited decoded signal portion and an extended-band signal portion. The decoder comprises the two central entities baseband decoder and bandwidth extension decoder.

The baseband decoder is configured to decode an encoded audio signal comprising the band-limited signal portion, wherein the baseband decoder comprises at least one learnable layer to obtain the band-limited decoded signal portion. The band-limited extension decoder is configured to generate the extended-band signal portion based on a linear predictive synthesis filter and a generated extended-band excitation.

According to embodiments, the decoder comprises a LPC estimation entity configured to perform LPC estimation based on the band-limited decoded signal portion to obtain a LPC coefficients and a band-limited excitation. The bandwidth extension decoder is then

configured to generate the extended-band excitation based on the band-limited excitation. Embodiments of the aspect are based on the finding that an audio signal which is decoded, e.g. using the above-defined encoder, so that a band-limited decoded signal and an extended-band signal portion is present by use of a decoder comprising the baseband decoder and the bandwidth extension decoder. Again, the baseband decoder preferably is implemented as neural decoder, i.e., comprises at least one learnable layer. As discussed in context of the decoder, this embodiment enables as well to combine neural speech coding and classical bandwidth extension techniques so as to increase the efficiency.

According to embodiments, the excitation generation may be based on a technique called Waveform Envelope Synchronized Pulse Excitation (WESPE). According to further embodiments, the extended-band LPC are estimated based on an entity comprising at least one learnable layer. In this latest embodiment, the bandwidth extension decoder is not solely based on a classical approach, but also a neural decoder, i.e., a decoder comprising a learnable layer. The band-limited LPC estimation entity may, according to further embodiments, be configured to generate the excitation comprising an analysis of the band-limited decoded signal portion for getting an estimate of harmonicity and/or voicing factor. Due to this, the relevant information may be advantageously extracted from the band-limited band-limiteddecoded signal portion from the baseband decoder and exploited for the decoding of the extended-band excitation.

According to embodiments a WESPE coder may comprise:

- an envelope determiner for determining a temporal envelope of at least a portion of a linear prediction residual of the band-limited audio signal or an excitation modelling the linear prediction residual of the band-limited audio signal;
- an analyzer for analyzing the temporal envelope to determine certain values of the temporal envelope;
- an excitation generator (16) for generating the extended band excitation or a part of it, by placing pulses in relation to the determined certain values, wherein the pulses are weighted using weights derived from the temporal envelope.

According to embodiments, the band-limited LPC estimation entity comprises an input for receiving a decoded band-limited signal portion of the encoded audio signal.

According to embodiments, the baseband decoder comprises a learnable convolution layer and/or a learnable affine transform and/or a learnable recurrent layer and/or a weighting layer in a residual block of neural network and/or learnable element-wise modulation. According to embodiments, the decoder may comprise a combination entity for combining the band-limited decoded signal portion and the extended-band signal portion to obtain a reconstructed signal. For example, this combination entity may be connected to the two decoders or the output of the two decoders. For example, the combination entity may comprise a filterbank and/or block transform and/or a time domain upscale entity and/or a complex valued low delay filterbank (CLDFB) being configured to perform additional post-processing in a filterbank domain before combining and/or before transforming the constructed signal to a time domain and/or at a desired sampling rate.

The above-discussed encoder and decoder may be implemented as a method. Thus, an embodiment provides a method for coding an audio signal comprising a band-limited signal portion and an extended-band signal portion. The method comprises the two central steps

- encoding the band-limited signal portion using a baseband encoder, wherein the baseband encoder comprises at least one learnable layer; and
- performing linear prediction of the extended-band signal portion.

Another embodiment, provides a method for decoding an audio signal comprising a band-limited decoded signal portion and an extended-band signal portion. This method may comprise the two central steps:

- decoding an encoded audio signal comprising a band-limited signal portion, using at least one learnable layer to obtain the band-limited decoded signal portion;
- generating the extended-band signal portion based on a linear predictive synthesis filter and a generated extended-band excitation.

Of course, embodiments of the present invention may be computed and implemented. Thus, another embodiment provides a computer program for performing, when running on a processor the steps of the two method steps as above. Another embodiment provides a method for training the neural encoder and/or decoder. For example, the training may be performed on the decoder side, and the encoder side or by use of both sides.

Below, embodiments of the present invention will subsequently be discussed referring to the enclosed figures, wherein:

Fig 1 shows schematically a level zero of the split band encoder, involving the base band encoder and the BWE encoder according to embodiments;

Fig. 2 illustrates schematically a two-band system realized with block transform, e.g. DFT according to embodiments;

Fig. 3 shows a high-level architecture of an example of neural baseband encoder and decoder to discuss embodiments:

Fig. 4 shows a basic implementation of the baseband encoder and the baseband decoder according to embodiments;

Fig. 5 shows a schematic block diagram of a BWE encoder according to embodiments;

Fig. 6 shows a schematic representation of a level zero of the split band decoder, involving the baseband decoder and the BWE decoder according to embodiments; and

Fig. 7 shows schematically an overall block diagram of the encoding and decoding system combining a time domain classical speech BWE at neural coder.

Below, embodiments of the present invention will subsequently be discussed referring to the enclosed figures, wherein identical reference numerals are provided to objects having an identical or similar function, so that the description thereof is mutually applicable and interchangeable.

Fig. 1 shows a split band encoder 100 comprising the entity's baseband encoder 110, BWE encoder 120 and the optional entity for pre-processing (cf. reference numeral 130) and multiplexor 140.

The pre-processing entity 130 receives the audio signal s(n) and splits this audio signal s(n) into a limited-band portion and band-extended portion, e.g. the two portions $s_{lb}(n)$ and $s_{hb}(n)$. The band-limited signal portion $s_{lb}(n)$, e.g. a low band portion is provided to the baseband encoder 110, while the extended-band signal portion $s_{hb}(n)$, e.g. a high band portion is provided to the BWE encoder 120. Both encoders 110 and 120 perform an encoding as will

be discussed below, so as to output the two encoded signals for the baseband and the extended bandwidth portion to the multiplexing 140. The multiplexer uses the two signal so as to generate a bitstream b.

The input signal is first conveyed to a pre-processing block, which is in charge of performing several analyses like a pitch estimation, a voice activity detection but also to convey signals at a proper sampling rate to the subsequent coding modules, consisting in our case of the baseband coder 110 and bandwidth extension (BWE) encoder 120. For this a filter-bank, like a Quadrature Mirror Filters (QMF), pseudo QMF, modulated lapped or block transforms, or simply downsampling filters in time domain can be used.

The two signals conveyed to the baseband encoder 110 and the bandwidth extension (BWE) encoder 120 are usually at sampling rates lower than the sampling rate of the input signal **s(n).** The low band signal $\mathbf{s_{lb}(n)}$ is composed of frequencies below a cross-over frequency which is usually the corresponding Nyquist frequency of its sampling-rate. On the other hand, the high band signal $\mathbf{s_{hb}(n)}$ is composed of frequencies above a cross-over frequency which is usually the corresponding Nyquist frequency of its sampling-rate. The HB and LB cross-over frequencies are usually the same. Therefore and in the usual case the two signals are complementary in frequency representation of the input signal and at the same time the whole multi-rate system is critically sampled. As an example, $\mathbf{s_{lb}(n)}$ and $\mathbf{s_{hb}(n)}$ are both sampled at 16kHz, $\mathbf{s_{lb}(n)}$ retaining frequencies from 0 to 8 kHz, and $\mathbf{s_{hb}(n)}$ retaining frequencies from 8 to 16kHz. Another alternative is to have $\mathbf{s_{lb}(n)}$ sampled at 12.8 kHz, composed of frequencies from 0 to 6.4 kHz and $\mathbf{s_{hb}(n)}$ sampled at 16kHz composed of frequencies from 6.4 to 14.4 kHz. As in the filter-bank convention and in the subsequent description, the high-band signal (odd indexed band), is frequency reversed.

The low-band signal is conveyed to the baseband coder 110, which in our preferred case is a neural coder, similar to the Neural End-to-End Speech Coder (NESC). The $\mathbf{s_{lb}(n)}$ signal preferably contains a wideband or broadband signal sampled at 16 kHz.

Fig. 2 shows a possible implementation for the pre-processing 130.

The entity 130 comprises two paths, each comprising a truncation and a normalization entity 133a and 133b respectively and an inverse DFT 135a and 135b, respectively. At the input for receiving the audio signal s(n) a forward DFT 137 may be arranged. By use of this

structure, a two-band system 130 realized with block transforms, namely DFTs 137, 135a and 135b is realized.

The truncation and normalization 133a and 133b of DFT spectrum serves as lowpass and highpass filtering respectively and the Inverse DFT 135a is operating at a size corresponding to the target sampling rate for the low-band signal. For the high band, only the high frequencies are retained and copied and flipped to the baseband (aka known as demodulation) by the demodulation and truncation module 133b before being decimated by the Inverse DFT 135a with a size corresponding to the sampling-rate of high-band signal. Alternatively, the sub-band decomposition can be achieved by time-domain decimation, like with a polyphase filterbank, or a pseudo-QMF.

With respect to Fig. 3, the neural baseband coder to be used on the encoder side (cf. 110) and on the decoder side (cf. 210) will be discussed.

The encoder 110 may comprise a dual path convolutional recurrent neural network 111, convolutional residual blocks 112 and a residual quantization 130. The dual path convolutional recurrent neural network uses as shown on the right hand side, rolling windows 111r to be processed in a certain order. For example, the dual path convolutional evolutional neural network 111 may comprise a 1 x 1 convolutional entity at the input and the output with a gated-recurrent unit 111G (GRU) in between. Starting from the plurality of frames, the entity 111 outputs the latent.

The baseband coder 110 is preferably a neural coder like NESC. NESC is an end-to-end approach for coding speech, getting the waveform of the signal as input and outputting the decoded waveform. It operates at very low bit rates, usually from 3.2 kbps down to 0.8kbps. It is built on a kind of auto-encoder architecture, and uses a residual quantization 113 of the latent representation L in the bottleneck of the auto-encoder. Such an architecture benefits from being trained using adversarial losses, following the Generative Adversarial Networks (GAN) paradigm. It generates very natural speech quality even at very low bit-rates but at the expense of the a large number of parameter and significant overhead of complexity compared to classical speech coding schemes. Therefore, it is usually beneficial to limit the coded audio bandwidth to wide-band, i.e. operating the neural coder at 16kHz.

As can be seen with respect to Fig. 3, the encoder architecture comprising the encoder 110 and the decoder 220 is configured to perform the following processing: the encoder 110

receives a speech signal and codes same so as to output a bitstream B. The decoder 210 uses the bitstream B so as to decode same and outputting the decoded speech signal. The entities 111, 112, 113, 211, 212 and 213 belong to the baseband encoder 110 and the baseband decoder 210, respectively. Consequently, no bandwidth extension is used in the current example.

Starting from this, the neural codec having the entities 110 and 210 may be enhanced by the respective bandwidths and coders 120 (bandwidth extension encoder) and 220 (bandwidth extension decoder) as illustrated by Fig. 4.

Fig. 4 shows an encoder side 100 a preprocessor 130, a baseband encoder 110 and a BWE encoder 120. The preprocessor 130 is configured to split the signal S(N) into the two signals as **slb(n)** and **shb(n)**. The two entities 110 and 120 process these signals **slb(n)** and **shb(n)** respectively so as to output the bitstream B.

The decoder 200 uses the baseband decoder 210 and the bandwidths extension decoder 220. Both decode the bitstream B and output the respective signals $\mathbf{y_{lb}(n)}$ and $\mathbf{y_{hb}(n)}$, respectively to the post-processor 230. The post-processor is configured to obtain based on these two signals $\mathbf{y_{lb}(n)}$ and $\mathbf{y_{hb}(n)}$, the decoded audio signal **y(n)**. The exact decoding will be discussed in context of Fig. 6 with focus on the bandwidth extension decoding 220. Before discussing the decoder-side, the bandwidth extension encoding 120 will be discussed with respect to Fig. 5.

Fig. 5 shows a bandwidth extension encoder 120. It comprises LPC analysis 142, and LPC to LSF transformation 144 and LSF quantization 146 enabling to output the LSF parameters. In parallel to the calculation of the LSF parameters, energy parameters are determined using the entities 150, 152 (subframe windowing), 154 (energy computation) and 156 (energy quantization). The energy quantization 156 is based on the energy computation 154 and the energy prediction 160 which gets the signal from the entity 150 and from a baseband preprocessor 110. The entity 150 is connected with the input for the signal and the LSF quantization 146 via the entity 147. The BWE encoder 120 receives the high-band signal **shb(n)** in order to extract the main salient parameters from it, namely its spectral shape and its energy. To do this, it follows a source-filter model like in CELP coding scheme and exploits the Linear Predictive Coding (LPC). LPC is an adaptive filter that models the short-term linear prediction and, through duality between time and frequency domains, the spectral envelope of the signal. Quasi-optimality of LPC holds for near

stationary segments, which for audio and speech signal can be considered for a duration of about 20ms. Therefore, the signal is partitioned into 20ms frames, and the LPC analysis and parameter computation is performed at frame level. For smoothing the transition, the LPC coefficients are further interpolated between adjacent frames, at a subframe level of duration 4 or 5ms. The interpolation is performed by linear interpolation of LSFs.

An LPC analysis aka short-term linear analysis is performed on **shb(n)** to obtain a set of LPC coefficients. Since speech and in general audio shows less structure or formant structure in the high frequencies, fewer parameters are required than for the low-band signal. In our preferred mode, an order of 8 or 10 is used for a 16kHz sampled **shb(n)** signal.

The LPC analysis is performed as it can be done in baseband encoder, that means, by windowing the signal, computing the autocorrelation function up to a maximum lag corresponding to the order, before finding the optimal prediction coefficients with a recursive algorithm like Levinson-Durbin. It is worth noting that the LPC analysis windows of both low and high band can be the same and preferably time aligned, which will be an advantage in the subsequent processing steps, but also for exploiting the same lookahead. The so-obtained LPC coefficients are then quantized and coded. Once again, since the spectral envelope of the high-band is usually less structured and also perceptually less relevant, the quantization resolution can be lowered for the BWE coding compared to the baseband coding. For the quantization and the coding, a Vector quantization or a multi-stage vector quantization is preferably applied after conversion of LPC coefficients to LSFs. Precomputed LSF means, obtained during an offline analysis on a dataset, are removed before quantization as well as a 1st order prediction obtained from the previously transmitted set of LSFs. The LSF residuals are then vector quantized using from 8 to 16 bits per frame in our preferred embodiment. The quantized LSFs are converted to quantized LPC coefficients to form the LPC analysis filter $\hat{\mathbf{A}}_{\mathbf{hb}}(\mathbf{z})$ used to whiten the high-band signal and obtain the residual signal $\mathbf{e}_{\mathbf{hb}}(\mathbf{n})$:

$$e_{HB}(n) = s_{HB}(n) - \sum_{i=1}^{M_{HB}} a_i \, s_{HB}(n-i), n = 0, \ldots, L \quad -1$$

where $\mathbf{M}_{\mathbf{HB}}$ is the LPC order, and Lsub, the size of the subframe for which the LPC coefficients are constant ($\mathbf{L}_{\mathbf{sub}}$=80 samples for 5ms subframe at 16kHz).

The energy of $\mathbf{e_{hb}(n)}$ is then computed and coded per sub-frame of 4 to 5ms (5ms in our preferred mode) using rectangular and non-overlapping windows. This way, an energy parameter can be transmitted at every 4/5 ms.

In order to save transmitted bits, the energy is not coded and quantized directly, but after a prediction exploiting energy information derived from the low band. Only the residue of the energy prediction is then quantized. This information must be shared with the decoder, since the inverse prediction must be performed on the decoder side. For this purpose, a LPC analysis of the decoded baseband signal is added to the decoder.

Analysis of these two components, especially in the high frequencies of the low band, around the Nyquist frequency, gives a robust estimate of the high-band energy and the residual of the high-band LPC analysis. For a 20ms framing, a set of 4 energy parameters are then obtained, and can be coded for example with a vector quantization using 7 bits. For even lower bit demand, the energy can be averaged (geometrically in our preferred mode) over the frame size for the 4 subframes, to obtain 1 single value per frame to transmit. A 4bit quantization is then enough. In the extreme case, only the estimate can be used at the decoder without additional guidance from the encoder, corresponding then to a 0bit quantization.

Possible BWE parameters and bit allocations are illustrated by below table

	Resolution	Bits	Bit-rate (kbps)
LSF parameters	20ms	0/8/8/8/16	0/0.4/0.4/0.4/0.8
Energy parameters	5/20ms	0/0/4/7/7	0/0/0.2/0.35/0.35
Total		0/8/12/15/23	0/0.4/0.6/0.75/1.15

With respect to Fig. 6, a BWE decoder 220 will be discussed. The BWE decoder 220 is arranged parallel to the base band decoder 210. Both outputting their respective signal $\mathbf{y_{LB}}$ and $\mathbf{y_{HB}}$ to the post processor 230. The baseband decoder 210 and the BWE decoder 212 receive the bitstream B via a multiplexer 235.

From the transmitted parameters, i.e. the coded LPC coefficients and coded energies, an artificially generated excitation is energy normalized and scaled, and then spectrally shaped by the synthesis LPC filter **1/Âhb(z)**.

The generated **yhb(n)** signal is then combined with the decoded low-band signal **ylb(n)** to form the reconstructed signal **y(n**), as it is shown in Fig. 6. It can be achieved using a filter-bank, block transforms or time-domain up-sampling. In the preferred embodiment, a complex-valued low-delay filter bank (CLDFB) as in described EVS, is used, which allows to perform additional post-processing steps in the filter- bank domain before combining the two components and transforming the signal back to the time- domain and at the desired sampling rate.

Regarding the BWE decoder it should be mentioned that the excitation generation may be, according to embodiments, as follows. The HB excitation is usually generated artificially, in the sense that little or no parameters are transmitted for it. To generate a suitable excitation, the decoded low-band signal is used intensively. For this the decoded LB signal from the neural decoder is further analyzed in order to derive the Linear Prediction Coefficients and thus a Linear Predictive residual, which can be used as a LB excitation for generating the HB excitation. If both the extracted LB and target HB excitations are at the same sampling rate, one can simply copy the LB excitation to the HB excitation. This then corresponds to a mirroring replication in the frequency domain, since the high-band signal is frequency inverted in our case. According to embodiments, the LPC can be used further to reconstruct the gain for the HB excitation from the potential transmitted residual gain. Here it should be noted that there are multiple other ways to get a gain for the HB excitation.

This leads to decent quality, but also to some obvious problems: harmonicity is often overestimated, and generated harmonics in the high-band do not necessarily correspond to the natural subharmonics of the fundamental frequency. It is also possible to apply a non-linear operation by increasing the excitation and applying a non-linear operation, then sub-sampling the component at high frequency. This approach is the one adopted in EVS in the Time-Domain BWE. In our preferred version, the method known as WESPE [4] is adopted, giving greater control over the final result and the amount of harmonicity injected. WESPE is adopted in the invention to work in the above-described framework, i.e. the LPC residual domain and also applied over the code excitation of CELP. The method for generating the HB excitation uses another analysis of LB decoded signal for getting an estimate of the harmonicity also known as voicing factor. Note, according to further embodiments, zero

crossings or other methods may be used for estimating the harmonicity. Such a zero crossing can, therefore, be interpreted as voicing factor.

Starting from the above discussion, a preferred embodiment may be as illustrated by Fig. 7.

Fig. 7 shows the combination of the encoder 100' and the decoder 200'. The baseband encoder 110 and the base band decoder 212 are based on DNNs. The latent of the encoder 110 is output to the decoder 210. On the encoder side, a filterbank 130 is arranged which pre-processes the input signal for the encoder 110 and the BWE encoder 120'. It performs a perceived parameter estimation 142 and LPC encoding 143. The parameter encoder 143 uses LPC coefficients out of the input signal for the baseband encoder 110.

On the bandwidth extension decoder side 220' the parameter decoder 243, the excitation generation 244, the LPC synthesis 245 and optional LPC/voicing analysis 246. The LPC/voicing analysis 246 uses the output signal of the decoder 210. The bandwidth extension 220' and the decoder 210 output the respective signal to the filterbank 230.

Below, possible embodiments will be discussed in detail

An embodiment provides an encoder for encoding an audio signal comprising:

- a band-limited signal and
- an extended-band signal;
- the encoder comprising:
    - a baseband encoder configured to encode the band-limited signal including at least one learnable layer;
    - a bandwidth extension encoder which comprises a linear prediction of the extended-band signal.

Another embodiment provides a decoder for decoding an audio signal comprising:

- a band-limited decoded signal and
- an extended-band decoded signal,
- the decoder comprising:
    - a baseband decoder configured to decode the band-limited signal including at least one learnable layer;
    - LPC estimation based on a band-limited signal from a decoder

- a bandwidth extension decoder which comprises the generation of an excitation, input of linear predictive synthesis filter.

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, some one or more of the most important method steps may be executed by such an apparatus.

The inventive encoded audio signal can be stored on a digital storage medium or can be transmitted on a transmission medium such as a wireless transmission medium or a wired transmission medium such as the Internet.

Depending on certain implementation requirements, embodiments of the invention can be implemented in hardware or in software. The implementation can be performed using a digital storage medium, for example a floppy disk, a DVD, a 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.

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

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

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

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

A further embodiment of the inventive methods is, therefore, a data carrier (or a digital storage medium, or a computer-readable medium) comprising, recorded thereon, the computer program for performing one of the methods described herein. The data carrier, the digital storage medium or the recorded medium are typically tangible and/or non–transitionary.

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

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

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

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 .

In some embodiments, a programmable logic device (for example a field programmable gate array) may be used to perform some or all of the functionalities of the methods described herein. In some embodiments, a field programmable gate array may cooperate with a microprocessor in order to perform one of the methods described herein. Generally, the methods are preferably performed by any hardware apparatus.

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

References

[1] Bruhn, Stefan / Pobloth, Harald / Schnell, Markus / Grill, Bernhard / Gibbs, Jon / Miao, Lei / Järvinen, Kari / Laaksonen, Lasse / Harada, Noboru / Naka, Nobuhiko / Ragot, Stéphane / Proust, Stéphane / Sanda, Takako / Varga, Imre / Greer, Craig / Jelinek, Milan / Xie, Minjie / Usai, Paolo, "Standardization of the new 3GPP EVS codec", 2015, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015.

[2] Mäkinen, Jari / Bessette, Bruno / Bruhn, Stefan / Ojala, Pasi / Salami, Redwan / Taleb, Anisse "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services", 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005.

[3] Pia, Nicola / Gupta, Kishan / Korse, Srikanth / Multrus, Markus / Fuchs, Guillaume, "NESC: Robust Neural End-2-End Speech Coding with GANs" 2022-07

[4] WESPE, patent application US 2021/0287687 A1

Claims

1. Encoder (100) for coding an audio signal comprising a band-limited signal portion and an extended-band signal portion, the encoder (100) comprising:

a baseband encoder (110) configured to encode the band-limited signal portion, wherein the baseband encoder (110) comprises at least one learnable layer; and

a bandwidth extension encoder (120) comprising a linear prediction entity for performing linear prediction on the extended-band signal portion.

2. Encoder (100) according to claim 1, further comprising a pre-processor (130) and/or at least one filter so as to obtain the band-limited signal portion and the extended-band signal portion based on the audio signal.

3. Encoder (100) according to claim 2, wherein the pre-processor (130) comprises a transform entity like a DFT (Digital Fourier Transform) entity.

4. Encoder (100) according to claim 2 or claim 3, wherein the pre-processor (130) comprises a truncation and normalization entity serving as low pass filter in the transformed domain and an inverse time-frequency transform like an inverse DFT in a baseband path coupled to the baseband encoder.

5. Encoder (100) according to claim 2 or claim 3, wherein the pre-processor (130) comprises a demodulation and truncation entity in the transformed domain and an inverse transform like a DFT in a bandwidth extension path coupled bandwidth extension encoder.

6. Encoder (100) according to claim 5, wherein the transform and inverse transform are respectively a forward time-frequency transform and an inverse time-frequency transform.

7. Encoder (100) according to one of the previous claims, wherein the baseband encoder (110) comprises a neural coder and/or an auto-encoder architecture and/or a Vector-Quantized Variational Auto-Encoder (VQ-VAE).

8. Encoder (100) according to claim 7, wherein the baseband encoder (110) is configured to perform encoding based on a quantization of a latent representation of the band-limited signal portion.

9. Encoder (100) according to claim 7, wherein the baseband encoder (110) is trained or trainable using adversarial losses.

10. Encoder (100) according to claim 7, wherein the baseband encoder (110) is configured to obtain a latent representation of the band limited signal, the latent representation being quantized and coded by the baseband encoder.

11. Encoder (100) according to claim 7 or 8, wherein the quantization of the latent representation is a vector quantization or a scalar quantization; or

wherein the quantization of the latent representation is structured in multi-stages or in a way of refining a residual of a previously quantized stage (residual quantization).

12. Encoder (100) according to one of the previous claims, wherein the bandwidth extension encoder (120) comprises an LPC (linear prediction coding) analysis stage configured to perform LPC analysis on the extended-band and/or the limited-band signal.

13. Encoder (100) according to claim 12, wherein the audio signal is partitioned into frames and wherein the LPC analysis stage is configured to perform LPC analysis at frame basis.

14. Encoder (100) according to claim 12 or 13, wherein the bandwidth extension encoder (120) is configured to perform interpolation and/or linear interpolation of the LPC between frames.

15. Encoder (100) according to claim 12 or 13 or 14, wherein the LPC analysis entity comprises a short-term linear analysis to obtain a set of LPC coefficients, the LPC coefficients are quantized and coded by the bandwidth extension encoder.

16. Encoder (100) according to claim 10, wherein the LPC coefficients are quantized in a transformed domain and/or Linear Spectral Frequencies (LSFs).

17. Encoder (100) according to claim 12, 13, 14, 15 or 16, wherein the LPC analysis entity is configured to obtain a filter AHB(z); or

wherein the LPC analysis entity is configured to obtain a filter AHB(z) and wherein the filter AHB(z) with coefficients $a_i$is used to whiten the extended-band signal portion $s_{HB}(n)$ and to obtain a residual signal $e_{HB}(n)$ based on the formula

$$e_{HB}(n) = s_{HB}(n) - \sum_{i=1}^{M_{HB}} a_i \, s_{HB}(n-i)$$

wherein M_{HB} is the LPC order, and n a sample index of the audio signal.

18. Encoder (100) according to one of claims 12 to 17, wherein the bandwidth extension encoder (120) is configured to code and quantize the energy after obtaining the residual signal $e_{HB}(n)$ or after the prediction.

19. Encoder (100) according to one of the previous claims, wherein the bandwidth extension encoder (120) is configured to code and quantize the energy of the residual exploiting information derived from the band-limited residual signal so that the energy of the residual in the extended-band is predicted by the energy of the linear prediction residual signal computed on the band-limited signal for deriving an residue of the energy prediction, which is then quantized and/or an information output to a decoder.

20. Encoder (100) according to one the claims from 16 to 19, wherein the quantization time resolution of the residue of the energy prediction or the energy of the residual signal is higher or equal to the quantization time resolution of the LPC coefficients or LSFs.

21. Encoder (100) according to one of the previous claims, wherein baseband encoder (110) comprises:

a format definer configured to define a first multi-dimensional band-limited signal representation of the band-limited signal, the first multi-dimensional band-limited signal representation of the band-limited signal including at least:
a first dimension so that a plurality of mutually subsequent frames is ordered according to the first dimension; and

the at least one learnable layer configured to process the first multidimensional band-limited signal representation of the band-limited signal, or processed version of the first multi-dimensional band-limited audio signal representation.

22. Decoder (200) for decoding an audio signal comprising a band-limited decoded signal portion and an extended-band signal portion, the decoder comprising:

a baseband decoder (210) configured to decode an encoded audio signal comprising the band-limited signal portion, wherein the baseband decoder comprises at least one learnable layer to obtain the band-limited decoded signal portion;

a bandwidth extension decoder (220) being configured to generate the extended-band signal portion based on a linear predictive synthesis filter and a generated extended-band excitation.

23. Decoder (200) according to claim 22, comprising:

a baseband decoder (210) configured to decode an encoded audio signal comprising a band-limited signal portion;

a LPC estimation entity being configured to perform LPC estimation based on the band-limited decoded signal portion to obtain LPC coefficients and a band-limited excitation; and

wherein the bandwidth extension decoder (220) is configured to generate the extended-band excitation based on the band-limited excitation.

24. Decoder (200) according to claim 22 or 23, wherein the excitation generation comprises waveform envelope synchronized pulse excitation, WESPE.

25. Decoder (200) according to claim 22 or 23 or 24, wherein the extended band excitation generation comprises:

an envelope determiner for determining a temporal envelope of at least a portion of a linear prediction residual of the band-limited audio signal or an excitation modelling the linear prediction residual of the band-limited audio signal;

an analyzer for analyzing the temporal envelope to determine certain values of the temporal envelope;
an excitation generator (16) for generating the extended band excitation or a part of it, by placing pulses in relation to the determined certain values, wherein the pulses are weighted using weights derived from the temporal envelope.

26. Decoder (200) according to claim 22, 23, 24 or 25, wherein an extended-band LPC are estimated by an entity comprising at least one learnable layer.

27 Decoder (200) according to claim 23,24, 25 or 26, wherein the LPC estimation entity is configured to generate the excitation comprising an analysis of the band-limited decoded signal portion for getting an estimate of a harmonicity and/or voicing factor.

28. Decoder (200) according to one of claims 23 to 27, wherein the LPC estimation entity comprises an input for receiving a decoded limited band signal portion of the encoded audio signal.

29. Decoder (200) according to claims from 22 to 28, wherein the baseband decoder (210) comprises the learnable layer out of the following learnable layers

- a learnable convolutional layer,
- a learnable affine transform,
- a learnable recurrent layer,
- a weighting layer in a residual block of a neural network, and/or
- a learnable element-wise modulation.

30. Decoder (200) according to one of the claims 22 to 29, further comprising a combination entity for combining the band-limited decoded signal portion and the extended-band signal portion to obtain a reconstructed signal.

31. Decoder (200) according to claim 30, wherein the combination entity comprises a filterbank.

32. Decoder (200) according to claim 30, wherein the combination entity comprises a block transform.

33. Decoder (200) according to claim 30, wherein the combination entity comprises a time domain upscaling entity.

34. Decoder (200) according to claim 30, wherein the combination entity comprises a complex valued low delay filterbank (CLDFB) being configured to perform additional post-processing (230) in a filterbank domain before combining and/or before transforming the reconstructed signal to a time domain and/or at a desired sample rate.

35. Method for coding an audio signal comprising a band-limited signal portion and an extended-band signal portion, the method comprising:

encoding the band-limited signal portion using a baseband encoder, wherein the baseband encoder (110) comprises at least one learnable layer; and

performing linear prediction of the extended-band signal portion.

36. Method for decoding an audio signal comprising a band-limited decoded signal portion and an extended-band signal portion, the method comprising:

decoding an encoded audio signal comprising a band-limited signal portion, using at least one learnable layer to obtain the band-limited decoded signal portion;

generating the extended-band signal portion based on the based a linear predictive synthesis filter and a generated extended-band excitation.

37. Computer program for performing, when running on a processor, the method according to claims 35 and 36.