LINEAR PREDICTION BASED COMFORT NOISE GENERATION IN THE EVS CODEC

Zhe Wang[*], *Lei Miao*[*], *Jon Gibbs*[*], *Tomas Toftgård*[†], *Martin Sehlstedt*[†], *Stefan Bruhn*[†], *Venkatraman Atti*[§], *Vivek, Rajendran*[§], *Duminda Dewasurendra*[§]

[*]Huawei Technologies, [†]Ericsson AB, [§]Qualcomm Technologies, Inc.

ABSTRACT

A Discontinuous transmission (DTX) system, which is widely adopted in speech codecs, is an important function for speech communication systems that can reduce the transmission bandwidth by at least a half. Within a DTX system, the comfort noise generation (CNG) plays a key role in the overall quality. Critical performance parameters with respect to the CNG including the transition quality from active to comfort noise (CN) frame, the quality of CN spectrum estimation, wider bandwidth rendering and the DTX efficiency have all been found to be very important. This paper describes a series of new technologies developed for the EVS codec aiming to address the performance of the CNG: A new hangover based CN analysis technique provides improved CNG transition quality. A new entropy based CN spectrum estimation technique and a new hybrid CNG scheme improve the CN spectrum estimation. Finally, a novel bandwidth extension technique for efficient rendering of high-frequency CN and a novel technique improving the DTX efficiency by controlling the DTX hangover length are described.

Index Terms— *CNG, DTX, comfort noise, linear prediction, EVS*

1. INTRODUCTION

Comfort noise generation (CNG) is a technology commonly used in speech communication systems. The CNG, together with a voice activity detector (VAD) and the DTX scheme, is used to reduce the transmission rate by simulating the background noise during inactive signal periods. This is achieved by encoding the characteristics of the background noise using features known as CN parameters and forming a special speech frame known as a silence insertion descriptor (SID) frame with a much lower bit rate than active speech frames. The CN parameters are utilized at the decoder to regenerate the background noise with as much fidelity as possible, by respecting the spectral and temporal content of the background noise presented to the encoder. Some typical CNG technologies can be found in the ITU-T Recommendations G.729B [1], G.729.1C [2], G.718 [3], or in the 3GPP Specifications for AMR [4] and AMR-WB [5]. All these technologies generate CN by using the analysis/synthesis approach making use of linear prediction (LP). The 3GPP EVS codec [6] also makes use of CNG to provide increased capacity. In the EVS codec, the CNG algorithm reproduces high quality CN by choosing between a linear prediction-domain based coding mode (LP-CNG) and a frequency-domain based coding mode (FD-CNG), depending upon the input signal characteristics. The LP-CNG, similar to the conventional approach, utilizes residual signal and LP parameters to represent the background noise whereas the FD-CNG uses spectral energies of critical bands along with a global gain.

In this paper, we focus on the description of the novel technologies used by the LP-CNG mode. In section 2, an overview of the LP-CNG is described; in section 3, an entropy based LSF (Line Spectral Frequencies) estimation approach used by the low-band CNG is presented; in section 4, a description of an extended CN analysis for smoother CNG is provided; in section 5, a hybrid scheme of LP-CNG used to compensate the CN spectrum is described; in section 6, a bandwidth extension technology used by the high-band CNG is presented; in section 7, a smart DTX hangover control mechanism is described and in section 8, subjective test results showing the performance of the described CNG technologies are presented. Finally, conclusions are provided in section 9.

2. OVERVIEW OF THE LP-CNG IN EVS

Mirroring the speech coding modes of the EVS codec for active signals, the LP-CNG operates on a split-band basis with the coding consisting of both a low-band and a high-band analysis/synthesis encoding stage. The low-band analysis is performed on an input signal sampled at 12.8 or 16 kHz depending on the operational bit rate and the bandwidth of the input signal. As a result of this analysis, CN parameters including the low-band excitation energy, the low-band LSF spectrum and the low-band excitation frequency envelope are encoded and transmitted to the decoder. The high-band analysis is performed on one of two frequency regions; 6.4 - 14.4 kHz or 8 - 16 kHz, depending on the low-band sampling rate, which is spectrally reversed and decimated to 16 kHz sampling rate. In contrast to the low-band encoding, no parameter modeling of the high-band noise spectrum is performed for the high-band signal. Only the energy of high-band signal is encoded and transmitted to the decoder and the high-band noise spectrum is generated purely at the decoder side. Both the low-band and the high-band CN is synthesized by filtering an excitation through a synthesis filter. The low-band excitation is derived from the received low-band excitation energy and the low-band excitation frequency envelope. The low-band synthesis filter is derived from the received LP parameters in the form of line spectral frequency (LSF) coefficients. The high-band excitation is obtained using energy which is extrapolated from the low-band energy and the high-band synthesis filter is derived from a decoder side LSF interpolation. The high-band synthesis is spectrally flipped and added to the low-band synthesis to form the final CN signal.

3. ENTROPY BASED LSF ESTIMATION

The LP spectrum (the LSF vector in case of EVS) to be quantized and transmitted in the SID frame in a LP based CNG scheme is usually an averaged LP spectrum over the CN averaging period. In EVS, the number of consecutive frames without transmission preceding the current SID frame determines the CN averaging

period, which additionally is upper-bounded to 8. The averaged LP spectrum is more robust to short-time spectral variation and is therefore considered to be a better representation of the noise spectrum than any single frame LP spectrum. An improved method can be found in [5] where only a majority of the LP spectra (the LSP vectors) in the CN averaging period are used for averaging. Two outliers which result in the maximum overall spectral distances to all the other LSP vectors in the CN averaging period are omitted from the averaging. However, it is not always the case that the majority spectra represent the best spectrum to be reproduced by the CN. For example, a typical street noise case may consist of vehicle noise with dense horns, where the horns may be accidently averaged into the averaged LP spectrum with a majority based approach. An assumption, based upon observation, is that most practical background noises $n(t)$ may be decomposed into stationary and non-stationary (or transient) noise components, i.e.

$$n(t) = c(t) + t(t) . \quad (1)$$

The stationary noise component, $c(t)$, which represents the noise that typically has quasi-stationary spectrum and energy over time, is more suitable for the SID frame low rate transmission. Additionally, the transient noise component, $t(t)$, is changing rapidly in energy and/or spectrum and cannot be captured by the SID frames. Thus, our goal should always be to have the spectrum representing the stationary noise component captured in the SID frame. To exclude the transient noise component from the LP spectrum estimation, an entropy based approach has been developed for the EVS codec, where it has been assumed that the noise frame containing both the stationary and transient components is more structural in the spectrum than the noise frame containing only the stationary component, thus resulting in lower spectral entropy. To save complexity, a method to estimate the noise frame spectral entropy, by using the LSF vector parameters calculated in the codec preprocessing, has been developed. The spectral entropy of the *i-th* noise frame in the CN averaging period is estimated by a parameter C_i as

$$C_i = \sum_{k=0}^{M} (\Delta_i(k) - \Delta)^2 , \quad (2)$$

where M=16 is the order of the LP filter, Δ denotes the bandwidth of the partition if the signal bandwidth is divided by M equally spaced LSF coefficients i.e. $\Delta = fs/(M+1)$, where fs is the Nyquist bandwidth of the signal. For the EVS codec, fs is 6.4 kHz for the 12.8 kHz core and 8 kHz for the 16 kHz core. $\Delta_i(k)$ denotes the bandwidth of the *k-th* partition divided by LSF coefficients of the *i-th* LSF vector over the signal bandwidth, that is

$$\Delta_i(k) = \begin{cases} lsf_i(0) & for\, k = 0 \\ fs - lsf_i(M-1) & for\, k = M \\ lsf_i(k) - lsf_i(k-1) & for\, k = 1,2,..., M-1 \end{cases} , \quad (3)$$

where $lsf_i(k)$ is the *k-th* LSF coefficient of the *i-th* LSF vector. A more structured spectrum will result in *Ci* having a higher value, and thus *Ci* will be negatively correlated to the spectral entropy. This is because the more the actual LSF spectrum deviates from a white noise spectrum, the more structural the actual spectrum is, thus representing the lower entropy. The two outliers are found in the CN averaging period by seeking the two maximum *Ci*, and the averaged LSP vector to be transmitted in the SID frame is calculated by

$$\overline{lsp}(k) = \frac{1}{N-2} \cdot \sum_{i=0, i\neq o1, i\neq o2}^{N-1} lsp_i(k) , \quad (4)$$

where N is the length of CN averaging period and $o1$ and $o2$ denote the indices of the two outliers.

4. EXTENDED COMFORT NOISE ANALYSIS

To enable smooth DTX transitions, codecs such as EFR [7] delay the CNG a certain number of hangover frames to obtain actively encoded noise frames that can be used for CN analysis. In the EVS codec, a similar approach is used with an extended CN analysis allowing more efficient DTX. Hangover frames are added based on encoder specific features [6], and to indicate what frames are relevant for CN analysis a 3-bit counter value $burstho_{ix}$ of the number of consecutive hangover frames is transmitted to the decoder in the very first SID frame of an inactive period.

During actively encoded periods, two buffers, ho_{lsp} and enr_{dec}, of fixed size are kept updated with the latest encoded frame's line spectral pairs (LSP) vector, which is another representation of the LSF vector, and excitation (LP-residual) energy. In the mentioned first SID frame, the $burstho_{ix}$ most recent vectors of ho_{lsp} and enr_{dec} are copied into the buffers $ho_{lsp-hist}$ and enr_{hist} which are used for the CN analysis. Parameters from previous hangover periods and SID frames may still remain in these buffers, especially for short bursts of active speech coding when no or only a few hangover frames are added.

To reduce the influence of eventual speech frames included in the hangover period, a selection of the m hangover frames having a LP residual energy not being more than 0.13 dB above and not more than 1.5 dB below the LP residual energy of the most recent buffered frame is made. An age weighted average energy of the m selected enr_{hist} entries is computed as $enr_{hist-ave-weighted}$ with the relative weights

$$\mathbf{w_{ho-enr}} = \begin{bmatrix} 0.2, & 0.16, & 0.128, & 0.1024, & 0.08192,... \\ 0.065536, & 0.0524288, & 0.01048576 \end{bmatrix} , \quad (5)$$

applied such that more recent energies contribute more to the average than less recent energies. In addition, the m $ho_{lsp-hist}$ vectors that correspond in time to the past residual energies in enr_{hist} used for the calculation of $enr_{hist-ave-weighted}$ are saved in the buffer $ho_{lsp-hist-sel}$.

The CN excitation energy E_{CN} is subsequently derived from the obtained average excitation energy and the current SID frame excitation energy $\hat{\bar{E}}$ according to

$$E_{CN} = \alpha \cdot enr_{hist-ave-weighted} + (1-\alpha)\hat{\bar{E}} , \quad (6)$$

where α determines the weights of the components. When there is a sufficient number of buffer entries used for the computation of $enr_{hist-ave-weighted}$ (more than 3), the average is considered stable enough and a weight $\alpha = 0.95$ is used, but if this is not the case a slightly larger weight is given to the SID frame excitation energy by setting $\alpha = 0.8$.

In the same way as LSP outliers are removed in the SID parameter analysis, depending on the number m of selected

vectors, zero, one or two LSP vectors in $ho_{lsp-hist-sel}$ with the lowest spectral entropy are excluded before an average hangover LSP vector $ho_{lsp-ave-weighted}$ is calculated as the arithmetic mean. The final LSP vector $\mathbf{lsp_{CNG}}$ used for CNG is then computed as

$$\mathbf{lsp_{CNG}}(k) = \beta \cdot ho_{lsp-ave-weighted}(k) + (1-\beta) \cdot \overline{\mathbf{lsp}}(k), \quad (7)$$

where $\overline{\mathbf{lsp}}$ is the current SID frame LSP vector and β determines the weights of the components. If the SID frame LSP vector is not differing too much from the average LSP vector an update in the direction of the SID LSP vector is made by setting $\beta = 0.8$, otherwise $\beta = 1.0$. More details can be found in [6].

For each SID frame, the SID parameters are stored in the buffers enr_{hist} and $ho_{lsp-hist}$ to be used in the extended CN analysis for future transitions between active coding and CNG. Additionally, as the background characteristics might change over time, old buffer elements are disregarded after a certain period of active coding such that only more recent characteristics are considered in the analysis. In the EVS codec, the oldest elements are excluded from the buffers after each half second of actively coded frames.

5. A HYBRID SCHEME OF LP-CNG

While conventional LP based CNG schemes have been successfully deployed in many previous generation codecs, it has been observed that the LP spectrum is sometimes deficient when representing the spectra of some noise types as part of codecs operating with wider bandwidths and at the very high codec qualities obtainable now. Car noise is such a noise type, with very high energy concentrated at low frequencies. With a conventional LP-CNG approach, the sharp spectrum of the car noise at very low frequencies cannot be faithfully reproduced in the CN with sufficient frequency resolution. To overcome this deficiency, a novel hybrid scheme based on the pure LP based CNG scheme has been developed. Besides the LP spectrum and the excitation energy computed in the time-domain, the frequency envelope of the excitation signal is also computed in the frequency-domain and transmitted as part of the SID frame. At the decoder side, the excitation signal frequency envelope is utilized to generate an excitation signal representing the spectral details of the CN. This excitation is then combined with the white noise excitation obtained in the usual manner to obtain the final excitation signal used to excite the CNG synthesis filter. In the encoder, the excitation frequency envelope is obtained by first Fourier transforming (FT) the LP residual signal, preferably using a Fast Fourier Transform (FFT) of length 256, and then obtaining the energies of the first 20 frequency bins (excluding the 0 Hz (DC) component) as the frequency envelope E. The frequency envelope vector transmitted as part of the SID frame is an averaged envelope over the CN averaging period calculated in a similar way to the LSF estimation with two outliers, identified as described in section 3, also removed from the averaging process. By receiving the quantized frequency envelope at the decoder side, a smoothed frequency envelope is calculated at each CN frame through a low-pass auto-regressive (AR) filtering procedure

$$\widetilde{E}(k) = 0.9\widetilde{E}^{[-1]}(k) + 0.1\hat{E}(k), \quad k = 0,...,19, \quad (8)$$

where $\hat{E}$ is the de-quantized frequency envelope decoded from the latest SID frame, $\widetilde{E}$ is the smoothed frequency envelope and superscript [-1] denotes the value from the previous frame. A white noise sequence $e_r(n)$ is a random excitation generated in a conventional manner [5]. The frequency envelope of the random excitation corresponding to the one transmitted in the SID frame is calculated and the difference envelope between it and the smoothed envelope $\widetilde{E}$ is computed as

$$D_E(k) = MAX\left[\widetilde{E}(k) - E_r(k), 0\right], \qquad k = 0,...,19, \quad (9)$$

where $E_r(k)$ is the frequency envelope calculated from the random excitation $e_r(n)$. It is then possible to generate another random sequence of 256 points and to consider this as a set of transform coefficients of a 256-point FT with random energy and phase. By altering the random sequence so that it has the same energies in its first 20 FT bins (excluding the DC component) as the difference envelope $D_E(k)$, and clearing the other frequency bins all to 0, a time-domain sequence $e_d(n)$ is then obtained by inverse transforming the altered sequence of FT coefficients. The time sequence thus can be regarded as a low-passed excitation representing the low frequency spectral details of the CN. The final excitation signal used to excite the CNG synthesis filter is then obtained by

$$e(n) = e_r(n) + e_d(n), \qquad n = 0,1,...,255. \quad (10)$$

6. NOVEL BANDWIDTH EXTENSION TECHNOLOGY FOR HIGH-BAND CNG

During super-wideband (SWB) operation of the EVS codec, high perceptual quality in the inactive portions of speech at the decoder side must be maintained to produce a natural sounding output. For this purpose, a high-band CN synthesis is added to the low-band LP-CNG synthesis output. This does additionally ensure smooth transitions between active and inactive speech.

However, the generation of this high-band CN synthesis at the decoder is performed without transmitting additional parameters from the encoder to the decoder in order to model the high-band spectral characteristics of noise frames. Instead, the high-band LSF parameters of the active speech frames preceding the current inactive frames at the decoder are used to model the high band after interpolation. The hangover setting in the VAD algorithm ensures that the active speech segments used for the spectral characteristic estimation of inactive frames sufficiently capture the background noise characteristics without significant impact from the talk spurt.

De-quantized high-band LSF vectors of order 10, corresponding to the last two active speech frames with super-wideband content are buffered at the decoder. Let the high-band LSF vectors corresponding to past active frames (N-1) and (N) be denoted by $\rho_{N-1,k}, k = 1 \cdots 10$ and $\rho_{N,k}, k = 1 \cdots 10$ respectively. Then the high-band LSF vector of the (N+M)th inactive frame is interpolated as

$$\rho_{N+M,k} = T \cdot \rho_{N,k} + (1-T) \cdot \rho_{N-1,k} \quad k = 1,...,10, \quad (11)$$

where the interpolation factor,

$$T = \min(\frac{M}{32}, 1), \quad (12)$$

is computed using the number of inactive frames M leading up to the current inactive frame (N+M) since the last active frame N. This interpolated LSF vector $\rho_{N+M,k}$ is converted to LP coefficients as the synthesis filter to be used in the high-band CNG synthesis. This allows for reliable high-band CN synthesis at the

decoder to accurately represent the background noise, without transmitting extra bits from the encoder side.

The energy of the high-band CN is estimated at the decoder side with guidance from the high-band energies encoded and transmitted in the SID frames. However, in contrast to the encoding of the low-band energy parameter, the high-band energy is encoded and transmitted within the SID frame with a much lower rate, that is, only a few SID frames will contain the quantized high-band energies while most of the SID frames do not deliver this information even for SWB operation. With such a scheme, for SID frames without high-band energy, the bits reserved for encoding the high-band energy can be used for improving other CN parameters, for example the low-band frequency envelope. The transmission rate reduction of the high-band energy encoding is achieved by taking advantage of the split-band processing. Since the low-band CNG is operating independently from the high-band CNG, the low-band noise energy is always available at the decoder. Thus the high-band noise energy E_h can always be derived from the low-band energy E_l given the current high-band to low-band energy ratio (HLR) R by

$$E_h = R \cdot E_l . \quad (13)$$

Also, it has been found that the HLR of a noise signal is usually quasi-stationary. Therefore, the HLR does not need to be transmitted frequently but only need be transmitted once the current HLR deviates from the last transmitted HLR by a significant step. In this way, the transmission of HLR is equivalent to the transmission of the high-band energy since the HLR can always be derived at the decoder side when the quantized high-band energy is received.

So, the energy of the high-band CN is obtained at the decoder by

$$E_h = \gamma \widetilde{E}_h^{[-1]} + (1-\gamma)\left(\hat{E}_h - \hat{E}_l + \widetilde{E}_l\right), \quad (14)$$

where $\widetilde{E}_h$, $\widetilde{E}_l$ are the smoothed logarithmic energy for the high-band and the low-band CN respectively, $\hat{E}_h$ denotes the quantized high-band energy in the last received SID frame, $\hat{E}_l$ denotes the smoothed logarithmic energy for the low-band CN at the frame of the last received SID where $\hat{E}_h$ is decoded, γ is a smoothing factor.

7. LP BASED DTX HANGOVER CONTROL

In this section, a technology which is not directly related to the CNG but still affects the quality/capacity of the overall DTX/CNG system is presented. The DTX hangover, which is an extension of the active period at the end of a speech segment is usually applied to the first several background noise frames after a speech segment to facilitate the estimation of the CN parameters at the decoder side. While the DTX hangover helps to improve the quality of transition from actively coded frame to CN frame, it is also a source which makes the system capacity reduced. The length of the DTX hangover is usually made a fixed length as in [4] and [5]. A more elaborated approach may be to adapt the DTX hangover length to some high level noise characteristics such as the long-term SNR. However, none of the previous approaches is really efficient and intended to relate the problem to the actual quality of the CN. Below an efficient approach to adaptively controlling the DTX hangover length based on CN prediction/evaluation at the encoder side is presented. The basic principle is that if the encoder is aware that the current DTX hangover is sufficient for reproducing a high

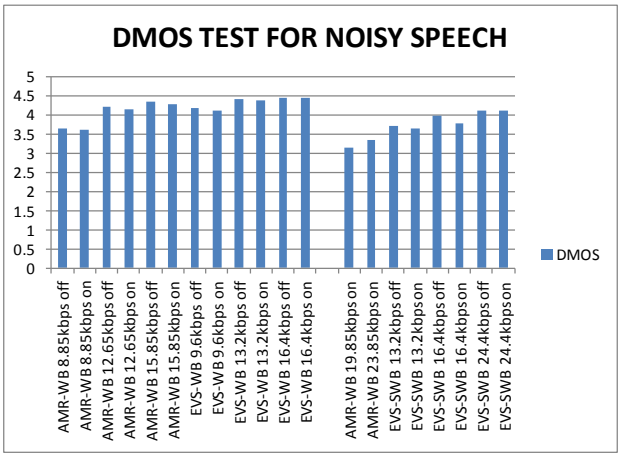

Fig. 1: Subjective test results (WB - 20dB SNR street noise, SWB - 20dB SNR office noise) (on = DTX enabled, off = DTX disabled)

quality CN transition at the decoder side, then there is no need to use a longer DTX hangover. In the present algorithm, a local CN generator is contained in the encoder. The local CN generator works in the same or a similar manner to that of the decoder. In the encoder, from the first DTX hangover frame, the encoder makes the assumption that the current frame will be encoded as a SID frame and based on this assumption the local CN generator creates the local CN which is a prediction of the CN which will be obtained in the decoder. If the local CN is close to the latest local CN generated before the current active burst in both energy and spectrum, then the encoder is confident that the quality of the CN is good and the DTX hangover may be terminated here. Otherwise, the DTX hangover increases by one frame and the prediction/measurement is repeated until a good CN is found or the maximum hangover duration is reached.

8. PERFORMANCE EVALUATION

Subjective tests compliant to Recommendation ITU-T P.800 [8] have been conducted to evaluate the performance of the EVS CNG. Since the CNG works in a switched manner, Figure 1 shows the DMOS results for noisy inputs where the LP-CNG is used. Both results for nominal wide-band (WB) and SWB inputs are presented. The WB test was conducted in English with 20dB street noise and the SWB test was conducted in Finnish with 20dB office noise. The results show no statistically significant degradation for the EVS codec with DTX enabled comparing to the same condition without DTX activated for different bitrates and bandwidths. The results also show clear superiority of the EVS codec utilizing DTX over the AMR-WB codec at similar bitrates.

9. CONCLUSION

In this paper, we have presented a series of new technologies developed for the EVS codec for improving the performance of CNG. The presented technologies successfully improve many important technical aspects related to the CNG including the CNG transition quality, the noise spectrum estimation, the CNG bandwidth extension and the DTX system capacity. Subjective test results show the success of these improvements in that the DTX system for the EVS codec performs as well with DTX on as it does with DTX off. Furthermore, the results also demonstrate the clear superiority of the EVS codec with DTX enabled over similar operating point for AMR-WB.

10. REFERENCES

[1] ITU-T G.729 Annex B, A silence compression scheme for G.729 optimized for terminals conforming to ITU-T Recommendation V.70. International Telecommunication Union (ITU), Series G., 2007

[2] ITU-T G.729.1 Annex C, DTX/CNG scheme. International Telecommunication Union (ITU), Series G., 2008

[3] ITU-T G.718, Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s. International Telecommunication Union (ITU), Series G., 2008

[4] 3GPP Technical Specification, TS 26.090, "Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions", v.12.0.0, 2014.

[5] 3GPP Technical Specification, TS 26.190, "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions", v.12.0.0, 2014.

[6] 3GPP Spec., Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description, TS 26.445, v.12.0.0, Sep 2014.

[7] ETSI, EN 300 728, "Digital cellular tele-communications system (Phase 2+); Comfort noise aspects for Enhanced Full Rate (EFR) speech traffic channels, (GSM 06.62 version 8.0.1 Release 1999)", V8.0.1, Nov 2000.

[8] ITU-T P.800, Methods for Subjective Determination of Transmission Quality. International Telecommunication Union (ITU), Series P., August 1996.