A NOVEL SINUSOIDAL APPROACH TO AUDIO SIGNAL FRAME LOSS CONCEALMENT AND ITS APPLICATION IN THE NEW EVS CODEC STANDARD

Stefan Bruhn, Erik Norvell, Jonas Svedberg, Sigurdur Sverrisson

SMN, Ericsson Research, Ericsson AB
164 80, Stockholm, Sweden

ABSTRACT

The new 3GPP codec for Enhanced Voice Services (EVS) comprises a collection of frame loss concealment techniques, each specifically designed for the different coding modes of that codec. One of them, called "Phase Error Concealment Unit (Phase ECU)", was developed for the High Quality (HQ) MDCT coding mode. Despite this target application, Phase ECU is a generic stand-alone tool operating on a buffer of the previously decoded and reconstructed time signal. Its framework is based on the sinusoidal analysis and synthesis paradigm. Besides a description of the basic technology we present optimizations and adaptations required for meeting the challenging 3GPP EVS codec performance requirements, and that make the method robust for a broad range of audio signals under various frame loss conditions from isolated frame erasures to severe burst loss. Test results are reported that show significant improvements over traditional techniques.

Index Terms— *EVS, speech/audio coding, error concealment*

1. INTRODUCTION

Frame loss occurs frequently in packet-switched communication and media distribution systems. Existing speech and audio coding standards rely on frame loss concealment (FLC) techniques to mitigate quality impairments resulting from that. Current state-of-the-art FLC techniques for conversational speech codecs typically apply the concept of freezing and extrapolating of codec parameters of a previously received frame. Prominent examples are the linear predictive mobile communication codecs AMR [1] and AMR-WB [2] for which FLC techniques are specified in [3] and [4], respectively.

Audio codecs typically apply frequency domain coding techniques. Even in that case similar FLC strategies to those of speech codecs are applied, at least when the audio codec operates under a strict delay constraint. The frequency domain codec parameters from a previously received frame are frozen or suitably extrapolated and then used when generating a substitution of the lost frame. Examples are the FLC schemes of the 3GPP audio codecs AMR-WB+ [5] and Enhanced aacPlus [6] and of the ITU-T G.719 codec [7]. Audio codecs not operating under strict delay constraints have a higher degree of freedom for FLC. Interpolative techniques can be applied that reconstruct the erased segment from both sides. Examples for such techniques are given in [8] and [9].

However, even though suggested by these examples as well as by conceptual and complexity considerations, it is not imperative for FLC methods to apply the same model as the codec. There may rather be a larger degree of freedom with better optimization possibilities if there is no such structural constraint. A codec-independent modelling technique that is of interest in the context of this paper is sinusoidal modelling with examples in [8][9][10][11].

This paper presents Phase ECU as a new FLC technique standardized together with the new 3GPP codec for enhanced Voice Services (EVS) [12][13][14]. First, the framework underlying our method is introduced, followed by a description of optimizations and adaptations that are essential to make the method practically applicable for a broad range of signals and single and burst loss conditions. Performance evaluation results presented in the end of the paper show significant improvements over traditional techniques.

2. FRAMEWORK

The FLC method of this paper is based on the sinusoidal analysis and synthesis paradigm [15]. It operates with a sinusoidal model under the assumption that the audio signal is composed of a limited number of K individual sinusoidal components, each component characterized by magnitude a_k, frequency f_k and phase φ_k. With the discrete time index n and the sampling frequency f_s, the audio signal $s(n)$ is expressed as the following multi-sine signal:

$$s(n) = \sum_{k=1}^{K} a_k \cos(2\pi \frac{f_k}{f_s} n + \varphi_k). \quad (1)$$

Applied to the task of FLC the sinusoidal model parameters a_k, f_k and φ_k can be derived locally in the decoder based on a correctly received segment of the signal prior to the frame loss. The lost frame can then be synthesized according to eq. (1) with the model parameters unchanged. While it is possible to apply this principle directly, this would require estimating all model parameters where a potential drawback is the estimation errors related to a limited measurement time period and the parameter variation within that period. The following implicit approaches avoid this problem to a large extent.

Parikh et al. [10] proposed an implicit method operated in DFT domain that alleviates the need to estimate sinusoidal magnitudes or phases. In the analysis step, after Hanning windowing, zero-padding and re-arranging (time shift), a previously decoded frame is transformed to DFT domain in which DFT coefficients corresponding to magnitude spectrum peaks are identified. The subsequent synthesis step involves IDFT of the DFT spectrum after all DFT coefficients not belonging to spectral peaks have been zeroed. The synthesized frame is then rescaled and synchronised with the analysis frame. A consequence of the coefficient zeroing is that the sinusoids in the synthesized frame are periodic with the DFT block length and that the effect of the Hanning window is removed. The periodic continuation of the frame can hence yield a substitution signal at the location of the lost frame without discontinuities at repeated frame boundaries. Problems with this method are however related to the limited DFT frequency

resolution that causes frequency errors of the synthesized sinusoids. Moreover, the zeroing of DFT coefficients leads to magnitude errors that cannot be fully compensated by the rescaling. Discontinuities at the boundaries between received and substituted frames are the consequence. Moreover, Parikh et al. report problems that the synthesized signal may become too tonal.

Hou et al. [11] presented a modification of this method. The problem of the limited DFT resolution was identified and solved by so-called inter-peak compensation. Like [10] the method retains the DFT coefficients related to the magnitude spectrum peaks and, in addition, the neighboring DFT coefficients. The remaining DFT coefficients are zeroed. This modification increases the accuracy of the sinusoidal frequencies but affects the property of the sinusoids to be periodic with the DFT block length. As a remedy a linear phase adjustment related to the time offset between the analysis frame and the substitution frame locations is introduced, which results in that the signal frame after IDFT matches to the location of the lost frame. Conceptual problems caused by the zeroing of DFT coefficients remain. Despite inter-peak compensation there is the need to make magnitude adjustments in order to prevent energy loss. In addition the problem with too strong tonality of the synthesized signal is not solved.

2.1. Phase ECU

The presented method aims to overcome the limitations of the discussed references. In particular problems like energy loss and tonal artefacts of the synthesized signal after FLC are avoided.

2.1.1. Sinusoidal Analysis

The first step of the method is sinusoidal analysis of a segment (analysis frame) $y_{-1}(n), n = 0 \dots L-1$, of the previously received signal $y(n)$, starting from time index n_{-1}, i.e. $y_{-1}(n) = y(n + n_{-1})$. The analysis identifies constituent sinusoidal components. Of main importance for the FLC method to work properly is to find the exact frequencies of the sinusoids. A good trade-off between measurement accuracy, complexity and the fact that practical audio signals are hardly stationary, is to use an analysis frame length L in the order of 20-40 ms. We use 32 ms where it is to be noted that the EVS codec frame size is 20 ms [13]. The analysis frame length corresponds to 512, 1024 and 1536 samples for f_s= 16, 32 and, respectively, 48 kHz. The analysis is done on the magnitude of the DFT spectrum $Y_{-1,w}(m)$, that is defined as:

$$Y_{-1,w}(m) = \mathrm{DFT}\{w(n) \cdot y_{-1}(n)\}. \quad (2)$$

Function $w(n)$ denotes the window function with which the analysis frame is weighted. Its choice is a trade-off between the width of the main lobe of the window spectrum and the side-lobe rejection. An excellent discussion about the choice of window functions in the context of sinusoidal analysis in DFT domain is found in [16]. In the context of this paper accurate estimation of the sinusoidal frequencies is a priority. In addition, for the task of the substitution frame generation it is also desirable to use a window that is flat over a long range of samples. Consequently a combination of Hamming (37.5%) and rectangular (62.5%) window is used with Hamming-type rising and falling edge shapes.

2.1.2. Substitution frame generation

The sinusoidal concealment methods of the cited literature [10][11] generate the substitution frame for the lost frame by IDFT of essentially a DFT line spectrum where DFT coefficients not belonging to spectral peaks are zeroed. This is unsuitable for reconstructing constituent sinusoids which are not periodic in the DFT block length L, i.e. which frequencies do not coincide with the DFT frequency grid.

Our approach understands substitution frame generation as a transformation of a previously reconstructed 'prototype frame' to the location of the lost frame, where the transformation is based on the sinusoidal model assumption. While in principle analysis and prototype frames could be chosen differently, for practical reasons (complexity) we choose them identically.

Let $y_0(n)$ be the unavailable segment for which a substitution frame has to be generated (i.e. $y(n + n_0) \coloneqq y_0(n)$), and let $y_{-1}(n)$ be the prototype frame. The DFT spectrum $Y_{-1,w}(m)$ of the windowed prototype frame can be expressed as the convolution of the discrete time Fourier transform $W(\Omega)$ of the window $w(n)$ with the discrete time Fourier transform $Y_{-1}(\Omega)$ of the prototype frame, subsequently sampled at the grid points of the DFT:

$$Y_{-1,w}(m) = \int_{2\pi} \delta(\Omega - 2\pi\tfrac{m}{L}) \cdot (W(\Omega) * Y_{-1}(\Omega)) d\Omega\,. \quad (3)$$

By using the sinusoidal model assumption for the prototype frame and by correspondingly substituting $Y_{-1}(\Omega)$ with the discrete time Fourier transform expression of eq. (1), this can be written as

$$Y_{-1,w}(m) = \frac{1}{2}\sum_{k=1}^{K} a_k \left(W\left(2\pi\left(\frac{m}{L}+\frac{f_k}{f_s}\right)\right) e^{-j\varphi_k} + W\left(2\pi\left(\frac{m}{L}-\frac{f_k}{f_s}\right)\right) e^{j\varphi_k} \right). \quad (4)$$

This can be understood as a superposition of frequency-shifted versions of the window function spectrum, where the shift frequencies are the sinusoidal frequencies. The superposition is then sampled at the DFT grid points.

Equation (4) is the basis for the substitution frame generation with the Phase ECU. As an approximation the window spectrum is truncated such that the shifted window spectra are strictly non-overlapping. Let us define integer intervals M_k:

$$M_k = \left[\mathrm{round}\left(\frac{f_k}{f_s} L\right) - m_{\mathrm{left},k}, \mathrm{round}\left(\frac{f_k}{f_s} L\right) + m_{\mathrm{right},k} \right], \quad (5)$$

where $m_{\mathrm{left},k}$ and $m_{\mathrm{right},k}$ fulfill the constraint that the intervals are not overlapping. Now, for each k and $m \in M_k$ the above expression (4) reduces to the following approximation:

$$\hat{Y}_{-1,w}(m) = \frac{a_k}{2} W\left(2\pi\left(\frac{m}{L}-\frac{f_k}{f_s}\right)\right) e^{j\varphi_k}. \quad (6)$$

Under the assumption that the sinusoidal model parameters remain constant, the correspondingly approximated DFT of the windowed lost frame is identical with the only difference that the phases of the sinusoids advance linearly with the respective sinusoidal frequencies and the time difference $n_0 - n_{-1}$ between the lost frame and the prototype frame:

$$\Delta\varphi_k = 2\pi \cdot \frac{f_k}{f_s}(n_0 - n_{-1}), k = 1 \dots K. \quad (7)$$

The approximated DFT coefficients of the lost frame for each k and $m \in M_k$ are hence given by

$$\hat{Y}_{0,w}(m) = \frac{a_k}{2} W\left(2\pi\left(\frac{m}{L}-\frac{f_k}{f_s}\right)\right) e^{j(\varphi_k + \Delta\varphi_k)}. \quad (8)$$

Thus, comparing equations (6) and (8) it is apparent that the DFT coefficients of the substitution frame belonging to any of the intervals M_k are readily obtained by phase shifting the corresponding DFT coefficients of the prototype frame by $\Delta\varphi_k$.

For the DFT coefficients outside the intervals M_k we assume that they are not related to any sinusoid. Still, they contribute to the perceptual impression of the signal through their power level.

Hence, unlike suggested by the cited literature they are not set to zero. Rather, their magnitudes are retained while their phases are randomized in order to avoid any undesirable periodicities.

These steps greatly facilitate the calculation of the DFT spectrum of the substitution frame. The spectrum is simply obtained by retaining the magnitude spectrum of the prototype frame and linearly evolving or randomizing the phase spectrum.

Subsequently the frame is transformed by means of IDFT. The frame could then be merged in time domain with the previously synthesized signal and the signal following the loss. However, in the framework of MDCT coding the conceptually more obvious and advantageous way is to follow the proposal in [10]. This means that the frame is windowed with the ALDO window of the HQ MDCT [13] and subsequently time domain aliased. After this operation the frame can readily be used instead of a regularly received and inversely transformed MDCT frame.

2.2. Optimizations

2.2.1. Peak search and refinement

A problem with sinusoidal analysis in DFT domain is the limited resolution which is $f_s/2L$. This means that the true sinusoidal frequency will potentially deviate by this amount from the corresponding peak frequency of the DFT magnitude spectrum. Such an estimation error may result in a significant phase error of the reconstructed sinusoid in the substitution frame, which may cause discontinuities at the boundaries between received and substituted frames. Zero-padding of the windowed analysis frame prior to DFT is one possibility to address this problem [10][11][16] as this leads to interpolation of the spectrum at more grid points. However, this increases the complexity and is not a very efficient technique. We rather apply parabolic interpolation in the magnitude spectrum domain, as proposed in [16].

2.2.2. Delta range around identified peak

The size of the intervals M_k around the identified spectral peaks is a tuning parameter. We found that best quality for a large range of audio signals is obtained when limiting $m_{\text{left},k}$ and $m_{\text{right},k}$ to a value of 6. Given the frame length and chosen window this corresponds to retaining a delta range of 187.5Hz around the spectral peaks, which includes the main lobe and the four largest side lobes. The discarded side lobes are at minimum ca. 50dB below the main lobe. Hence, the window truncation has in practice no significant impact for tonal signals which main frequencies are sufficiently separated. For signals with tonal components close to each other, like e.g. complex music, the size of the intervals is reduced in order to avoid overlap. Consequently a stronger effect of the window truncation may be expected. However, due to masking, complex music signals are less sensitive to possible reconstruction inaccuracy and hence the window truncation does not have a strong practical consequence in that case either.

2.2.3. Placement of analysis/prototype frame

For the application of our method to the HQ MDCT coding mode of the EVS codec it is important to note that the used ALDO MDCT windows are designed for low-delay operation. Their overlap is relatively short. The placement of the analysis/prototype frame must ensure a good trade-off between using an as recent reconstruction signal portion as possible and avoiding too strong time domain alias distortions if the frame extends into the overlap region. The Phase ECU of the EVS codec has an experimentally optimized placement of the analysis/prototype frame that partially extends into the time-aliased overlap region with the lost frame.

3. ADAPTATION CONTROL

In order to achieve best possible performance in an application in the EVS codec and to make the method suitable for a broad range of audio signals as well as for various frame loss conditions including burst loss the following adaptations were introduced.

3.1. Signal adaptive control

A consequence of our approach to retain a significant portion of the prototype frame spectrum surrounding the identified peaks is the desirable property that the generated substitution frame essentially still exhibits the window shape. Even more, the time envelope of the windowed prototype frame is retained. This is however generally not desirable if the frame comprises a substantial level change, e.g. in case of onset or offset. If a non-flat time envelope is reproduced, the consequence may be a saw-tooth like repetition of the level change with corresponding audible impact.

This raises the more general question of substitution frame generation for frames that follow transients. It can be argued that sinusoids of a transient prototype frame should not be retained in the substitution frame in order not to produce erroneous periodicities. A particular problem for music signals is however that transition is not a global frame characteristic. Rather, there may be frequency ranges with transient behavior and other ranges that may be fairly stationary.

This consideration suggests that the FLC should operate in a frequency selective manner. Frequency bands that show no transient behavior can be processed according to the basic method described above, but transient bands should not. Our approach is using a frequency-band selective transient detector and a frequency-band selective adaptation of the sinusoidal substitution frame generation.

We use a transient detector that operates in the DFT domain based on two partial frames of the analysis frame. These partial frames are taken from the beginning and the end of the analysis frame and then transformed into frequency domain after proper windowing. Based on perceptual considerations the resulting spectra are split into bands approximately following the size of the human auditory critical bands. Then the ratio of the respective band energies is calculated. A transient condition in a band is detected if the absolute value of the respective band energy ratio exceeds a threshold.

The frequency-band selective adaptation of the substitution frame generation in case of a detected transient in a frequency band involves attenuating the spectral magnitudes of that band in case of an offset. For detected onsets no magnitude modification takes place. Moreover, it has been experimented with randomization of the phases for transient bands. Phase randomization has the effect of discontinuing sinusoids of the prototype frame in the respective band while still preserving the spectral envelope. The experimental evaluation was however inconclusive if this kind of phase randomization can lead to quality advantages.

3.2. Burst loss handling

The described method including transient handling is found to perform satisfactorily even for burst frame losses. However, in case of longer loss bursts occasional tonal sounds due to sustained periodicity are observed. This problem is effectively solved by increasing attenuation and phase randomization of the substitution frame DFT coefficients, depending on the length of the loss burst. The adapted substitution frame spectrum is given by

$$\hat{Y}_{0,adap}(m) = \alpha \cdot \hat{Y}_{0,w}(m) \cdot e^{j\vartheta(m)}. \quad (9)$$

The magnitude attenuation is done with a successively (exponentially) decreasing factor α starting from the 3rd or 4th successive frame loss, depending on whether the signal is classified as speech or music. The phase adaptation is done through an increasing additive dithering component $\vartheta(m) = 2\pi\rho\ \mathrm{rand}(\cdot)$ where ρ is linearly scaled up from 0 to 1, also starting from the 3rd or 4th successive frame loss. The somewhat delayed adaptation for music is based on experimental findings that music generally suffers less from sustained periodicities than speech.

The above-described magnitude adaptation may be perceived as muting or signal drop outs in case of very long loss bursts. This affects the overall quality impression of e.g. music or the ambient noise of a speech signal. We address this problem by compensating the frame energy loss through the addition of a noise signal with spectral characteristics similar to those of the prototype frame. To that end eq. (9) is further extended to the following expression:

$$\hat{\hat{Y}}_{0,adap}(m) = \hat{Y}_{0,adap}(m) + \beta \cdot \overline{Y}(m) \cdot \mathrm{e}^{j\eta(m)}. \quad (10)$$

Here $\beta(m)$ is a magnitude scaling factor and $\eta(m)$ is a random phase, and $\overline{Y}(m)$ is a magnitude spectrum representation of the prototype frame. It can be assumed that the two additive terms in eq. (10) are uncorrelated. Hence, β is preferably chosen as

$$\beta = \sqrt{1-\alpha^2}. \quad (11)$$

The additive noise signal $\overline{Y}(m)$ should have spectral characteristics similar to those of the prototype frame. Though, in order to avoid tonal artifacts it has been found beneficial to use a low-resolution representation of the prototype frame magnitude spectrum, obtained by frequency-group-wise averaging. This representation is readily obtained by averaging the respective frequency band energies of the two partial frames that are available from the transient analysis step above. The perceptual quality could still be impacted from high-frequency noise in case higher frequency bands exhibit too strong energy. This is avoided by additionally attenuating such bands.

The described adaptations of the method gradually replace the signal generated by the sinusoidal FLC method by a suitably shaped noise signal of same power level. It can be argued that for very long loss bursts the signal should still be muted. Otherwise a receiver might produce annoying sustained noise signals e.g. in case of a lost connection. A suitable choice is to start gradual muting of the substitution frames with a degree of 6 dB per frame after 10 lost frames in a row (corresponding to 200 ms).

4. PERFORMANCE

During EVS codec development our method was subject to various quality assessments where it was evaluated against a multitude of competing development candidates. Based on these evaluations Phase ECU was chosen as a central component of the FLC for the HQ MDCT coding mode. For signal bandwidths from wideband (f_s=16 kHz) to fullband (f_s=48 kHz) it is used in case of single frame losses depending on signal characteristics; for burst losses starting from the second consecutive loss it is used exclusively. Comprehensive evaluation results of the EVS codec including our method under frame loss conditions are part of the EVS codec selection results [17] and the 3GPP EVS Codec Performance Characterization Technical Report [18].

A reduced P.800 DCR test was conducted during EVS development, where the Phase ECU was compared to the concealment method based on the techniques of the ITU-T G.719 codec [7]. Both our and the reference method was applied in conjunction with the same MDCT-based codec mode and compared for 10% frame erasure rate (FER). The test comprised 14 super-wideband (SWB) music items presented to 6 expert listeners using a diotic headset playback. The results shown in Figure 1 indicate significant quality enhancements with our method by more than 0.5 DCR MOS points over the baseline reference.

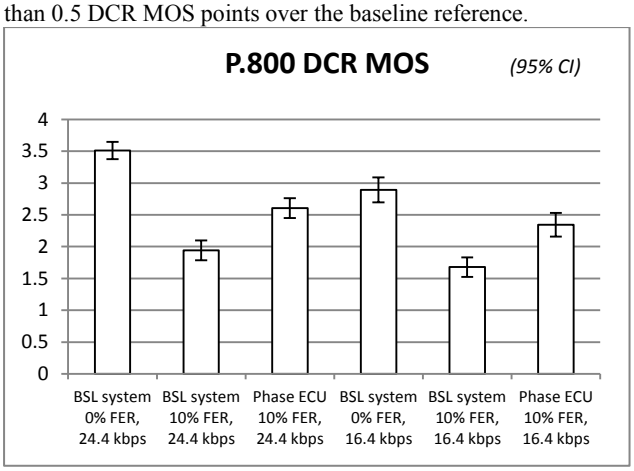

Figure 1: Subjective evaluation results comparing Phase ECU against G.719 FLC

An assessment of the quality in the presence of severe burst frame losses was also performed. The test contained SWB mixed and music material, encoded at 32 kbps and subjected to 25% frame loss. Burst frame losses ranging from 1 to 10 losses in a row occurred with likelihood inversely proportional to the burst length. The reference system was a basic Phase ECU scheme, though without additive noise mixing according to section 3.2. Figure 2 shows overall results from an AB test with 28 items presented to 6 expert listeners. A clear and statistically significant improvement is observed for the test system with additive noise mixing. Per-item analysis reveals strong listener preference especially for the music items while for the mixed items containing speech the preference is less pronounced, though still existing.

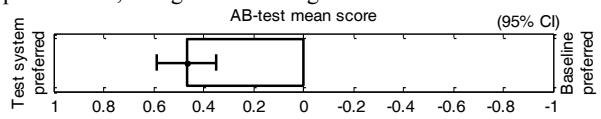

Figure 2: AB test results Phase ECU with noise mix (test system) and without (reference)

5. CONCLUSION

A novel FLC method based on sinusoidal analysis and synthesis has been presented. Due to its merits in terms of performance it has become integral part of the EVS codec error concealment standard [14]. The method is well-proven and robust for a large range of audio signals and very suitable even for severe burst errors. This makes it even a strong candidate as stand-alone FLC tool for other already existing or future audio codecs. As the method operates on a buffer of the previously decoded and reconstructed audio signal the integration into an audio decoder implementation is straightforward.

6. REFERENCES

[1] 3GPP TS 26.071, "Mandatory speech CODEC speech processing functions; AMR speech Codec; General description".

[2] 3GPP TS 26.171, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description".

[3] 3GPP TS 26.091, "Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Error concealment of lost frames".

[4] 3GPP TS 26.191, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Error concealment of erroneous or lost frames".

[5] 3GPP TS 26.290, "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions".

[6] 3GPP TS 26.402, "General audio codec audio processing functions; Enhanced aacPlus general audio codec; Additional decoder tools".

[7] ITU-R Recommendation G.719, "Low-complexity, full-band audio coding for high-quality, conversational applications".

[8] Sang-Uk Ryu, Kenneth Rose, "Advances in Sinusoidal Analysis/Synthesis-based Error Concealment in Audio Networking", AES Convention 116, May 2004.

[9] M. Bartkowiak, B. Latanowicz, "Mitigation of long gaps in music using hybrid sinusoidal+noise model with context adaptation", International Conference on Signals and Electronic Systems (ICSES), 2010.

[10] Vipul N. Parikh, Juin-Hwey Chen, Gerard Aguilar, "Frame Erasure Concealment Using Sinusoidal Analysis-Synthesis and Its Application to MDCT-Based Codecs", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000.

[11] Huan Hou, Weibei Dou, "Real-time Audio Error Concealment Method Based on Sinusoidal Model", International Conference on Audio, Language and Image Processing (ICALIP), 2008.

[12] 3GPP TS 26.441, "Codec for Enhanced Voice Services (EVS); General overview".

[13] 3GPP TS 26.445, "Codec for Enhanced Voice Services (EVS); Detailed algorithmic description".

[14] 3GPP TS 26.447, "Codec for Enhanced Voice Services (EVS); Error concealment of lost packets".

[15] Robert J. McAulay, Thomas F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Transactions on Acoustics Speech and Signal Processing, 09/1986.

[16] Julius O. Smith III, Xavier Serra "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation", Proceedings of the International Computer Music Conference (ICMC-87, Tokyo), Computer Music Association, 1987.

[17] Tdoc S4-141065, "Report of the Global Analysis Lab for the EVS Selection Phase", Dynastat, Inc.

[18] 3GPP TR 26.952, "Codec for Enhanced Voice Services (EVS); Performance characterization".