LOW-COMPLEXITY AND ROBUST CODING MODE DECISION IN THE EVS CODER

Emmanuel Ravelli[1]*, Christian R. Helmrich*[2]*, Guillaume Fuchs*[1]*, and Markus Multrus*[1]

[1] Fraunhofer Institut für Integrierte Schaltungen (IIS) [2] International Audio Laboratories Erlangen
Am Wolfsmantel 33, 91058 Erlangen, Germany

ABSTRACT

Several state-of-the-art switched audio codecs employ the closed-loop mode decision to select the best coding mode at every frame. The closed-loop mode selection is known to have good performance but also high complexity. The new approach we propose in this paper is a low-complexity version of the closed-loop approach, based on similar decisions which compute the coding distortion of each mode and select the one with the lowest distortion. Our approach differs mainly in the way the coding distortions are calculated. We are able to notably reduce the complexity by only estimating the distortions without encoding and decoding the input for each mode. The new approach was implemented in the EVS codec standard and evaluated both objectively and subjectively. Compared to the closed-loop approach, it yields similar performance and lower complexity.

Index Terms— Speech and audio coding, switched coding, mode decision, mode selection, closed-loop, open-loop.

1. INTRODUCTION

Most state-of-the-art speech-and-audio codecs (such as 3GPP AMR-WB+ [1, 2], MPEG-D USAC [3, 4], 3GPP EVS [5]) are based on a switched-coding design. This design allows the codec to switch, on a frame-by-frame basis, between different coding modes that are optimized for different content types. Generally, two main coding paradigms are employed. One is transform-based and provides best performance for music-like and noise-like input signals (like e.g. AAC [3, 4], TCX [1, 2], MDCT-based TCX [3, 4]). The other is CELP-based and provides best performance for speech- or transient-like input signals (mostly ACELP and its derivatives, see e.g. [5]). Using such kind of encoding modes, a switched codec would ideally be able to provide best output quality with any input content type.

However, the performance of a switched codec heavily depends on the mode decision taken at the encoder-side. Wrong decisions can significantly degrade the output quality, e.g. selecting a CELP-based coding mode on a multi-instrumental music frame would most probably make the output sound much noisier. Therefore, the selection of the encoding mode is critical for a switched codec, and it has to be carefully designed and tuned. Several solutions have already been proposed in the past, and some of them implemented in existing state-of-the-art switched audio codecs.

One well-known approach is the closed-loop mode decision [1, 2], introduced in AMR-WB+ and also adopted in the LPD-mode of USAC. This approach consists of encoding and decoding the current frame with all coding modes and selecting the mode which produces the lowest coding distortion. The closed-loop approach engenders a sequence of mode decisions which leads to near-optimal output quality, especially when the coding modes share the same perceptual model. In AMR-WB+, both ACELP and TCX optimize their coding in the perceptually weighted LP domain, a domain in which the comparison of the respective distortions give a very good indication of the best suited coding. As a posteriori decision, the closed-loop approach is expected to be robust for any other combinations of coding schemes even if they don't adopt the exact same objective function. However, this approach is also known to significantly increase the computational complexity of the encoder, due to the necessity of running at every frame all coding modes, including encoder and decoder. Such a high complexity can be problematic for telecommunication coders like EVS.

In this paper, we introduce a new approach that can produce similar performance as the closed-loop approach with lower complexity. This new approach has similarities with the closed-loop approach in the way that both use a measure of the distortion introduced by each coding mode to select the best one. Both approaches, however, differ in the way they measure the coding distortion of each mode. In the closed-loop approach, a complete encoding and decoding of all modes is needed in order to measure the coding distortions (see e.g. [2]). In the proposed approach, simple models of the coding modes are used to estimate the coding distortions. This way, the complexity introduced by encoding-decoding all modes can be avoided. The only additional complexity comes from the coding models, which is low in comparison. Additionally, the coding distortions estimated with the coding models are close enough to the "non-estimated" ones (obtained with the closed-loop approach) such that similar performance can be obtained.

Note that other low-complexity approaches have been proposed in the past [1, 2, 6, 7, 8], but they are all based on the same classic method where a set of features is used in combination with a statistical classifier, an approach clearly different from the new approach that we propose in this paper. An example of prior art is the open-loop mode decision implemented in AMR-WB+ [1, 2, 7]. In [7], the closed-loop and open-loop approaches implemented in AMR-WB+ are compared. The open-loop approach is shown to have indeed significantly lower computational complexity but listening test results indicate that the open-loop approach produces significantly lower output quality.

Evaluations given in section 3 show that a complexity reduction up to more than 30% is achievable for a quality evaluated as good as the conventional closed-loop decision. Moreover the low complexity decision introduces no additional delay over the coding schemes and is well suited for low-delay applications. For all these reasons, the decision technique was adopted in the recently standardized EVS for selecting on a 20 ms frame basis between the ACELP-based speech coding and the MDCT-based transform coding.

2. PROPOSED APPROACH

In this section, the proposed approach is described as it is implemented in the EVS standard. In that context, the proposed mode decision algorithm is used to select, at every frame, one of the two

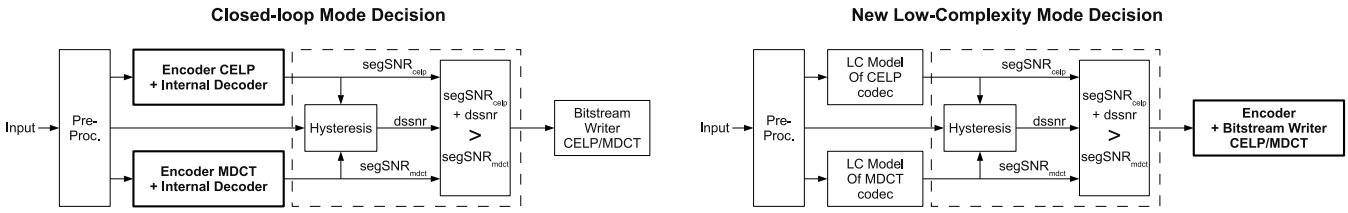

Fig. 1. Simplified block diagrams of the conventional closed-loop mode decision and the new low-complexity mode decision.

EVS coding modes, the MDCT-based encoding mode or the CELP-based mode, as shown in Figure 1.

Like in the closed-loop approach, a coding distortion measure is used to compute the distortion introduced by each encoding mode. We use in the EVS implementation the well-known segmental SNR in the weighted signal domain (used e.g. in AMR-WB+ or in the LPD-mode of USAC), given by

$$\text{segSNR} = \frac{\sum_{k=1}^{K} [10 \log_{10} \text{SNR}_k]}{K} \tag{1}$$

where K is the number of subframes and SNR_k is the SNR in the weighted signal domain of the subframe k. The SNR in a subframe of length N is given by

$$\text{SNR} = \frac{S}{D} = \frac{\sum_{n=1}^{N} (s_w(n))^2}{\sum_{n=1}^{N} (s_w(n) - \hat{s}_w(n))^2} \tag{2}$$

where s_w is the weighted input signal and $\hat{s}_w$ is the weighted output signal, obtained by filtering with a weighted version of the Linear Predictive (LP) analysis filter $A(z/\gamma)$ the input and output signal respectively. The weight γ depends of the internal sampling rate of ACELP is typically equal to 0.92 at 12.8 kHz.

The mode decision is then based on two estimates of the segmental SNR, one estimate corresponding to the MDCT-based coding mode, another estimate corresponding to the CELP-based coding mode. Based on these two estimates and on a hysteresis mechanism, a decision is taken. Formally, the decision can be simply described as follows

if $(\text{segSNR}_{\text{celp}} + dssnr > \text{segSNR}_{\text{mdct}})$ **then**
    The CELP-based coding mode is selected.
else
    The MDCT-based coding mode is selected.
end if

where $\text{segSNR}_{\text{mdct}}$ is the estimated segmental SNR of the MDCT-based coding mode (described in Sec. 2.1.1), $\text{segSNR}_{\text{celp}}$ is the estimated segmental SNR of the CELP-based coding mode (described in Sec. 2.1.2), and $dssnr$ is a value used to introduce a hysteresis in the decision (described in Sec. 2.2).

2.1. Estimation of the segmental SNR

In the closed-loop approach, the weighted output signal $\hat{s}_w(n)$ is obtained after running the corresponding encoder and decoder. The weighted output signal is then used to compute the segmental SNR using (2) and (1).

In the proposed approach, the coding distortion D in (2) is estimated employing only low-complexity operations and using signals and parameters already available from a pre-processing stage of EVS (as shown in Figure 1). The estimated coding distortion is then considered in (2) and (1) for getting an estimated segmental SNR.

2.1.1. MDCT-based coding mode

In the case of the MDCT-based coding mode, the coding distortion D_{mdct} is estimated directly in the MDCT domain and to keep the complexity low neither any decoding nor inverse MDCT is performed. The weighted input signal is computed in frequency domain by multiplying the MDCT spectrum by the frequency response of the weighted LP analysis filter, as it is done in the MDCT-based TCX of EVS [5].

The distortion in the weighted MDCT domain is then given by

$$W = \sum_{l=1}^{L_{\text{mdct}}} (c_w(l) - \hat{c}_w(l))^2 \tag{3}$$

where L_{mdct} is the MDCT length, c_w are the weighted MDCT coefficients and $\hat{c}_w$ are the quantized weighted MDCT coefficients. Assuming the weighted MDCT coefficients are quantized with a scalar uniform quantizer at high-rate, the distortion can be approximated as

$$\widetilde{W} = \frac{g^2}{12} L_{\text{mdct}} \tag{4}$$

where g is a global gain which is usually adjusted in order to reach a target bitrate (this target bitrate is actually here a hand-tuned constant independant of the codec bitrate) and that can be roughly estimated with low-complexity using an iterative algorithm described in [5]. The distortion in the weighted signal domain is then

$$D_{\text{mdct}} = \widetilde{W} \frac{\sqrt{2}}{L_{\text{mdct}}^2} N = \frac{g^2 \sqrt{2} N}{12 L_{\text{mdct}}} \tag{5}$$

Finally, the estimated SNR of the MDCT-based coding mode in a subframe is given by

$$\text{SNR}_{\text{mdct}} = \frac{S}{D_{\text{mdct}}} = \frac{\sum_{n=1}^{N} (s_w(n))^2}{D_{\text{mdct}}} \tag{6}$$

and the corresponding $\text{segSNR}_{\text{mdct}}$ is obtained via (1).

This simple model of the MDCT-based coding mode is valid most of the time. However it usually underestimates the segSNR on stationary and periodic music signals. To improve the model on such signals, a simple Long-Term-Prediction (LTP) is applied on the input time-domain signal just before the MDCT. This filter reduces the amplitudes of the harmonics in the MDCT domain, and consequently reduce D_{mdct} and increase the segSNR.

2.1.2. CELP-based coding mode

In the case of the CELP-based coding mode, we propose a simplistic model of the adaptive codebook and innovative codebook contributions (note that this model was empirically derived). The weighted output signal can be expressed as

$$\hat{s}_w(n) = \hat{s}_w^a(n) + \hat{s}_w^i(n) \tag{7}$$

where $\hat{s}_w^a$ is the adaptive codebook contribution and $\hat{s}_w^i$ the innovative contribution both computed in the weighted domain. The coding distortion D_{celp} can then be written as

$$D_{\text{celp}} = \frac{D_{\text{ada}}}{\text{SNR}_{\text{ino}}} \tag{8}$$

with

$$D_{\text{ada}} = \sum_{n=1}^{N} \left(s_w(n) - \hat{s}_w^a(n)\right)^2 \tag{9}$$

being the coding distortion introduced by the adaptive codebook and

$$\text{SNR}_{\text{ino}} = \frac{\sum_{n=1}^{N} \left(s_w(n) - \hat{s}_w^a(n)\right)^2}{\sum_{n=1}^{N} \left(\left(s_w(n) - \hat{s}_w^a(n)\right) - \hat{s}_w^i(n)\right)^2} \tag{10}$$

specifying the coding gain of the innovative codebook that can be approximated by a constant assuming a high number of pulses.

$$\text{SNR}_{\text{ino}} = \frac{1}{C_{\text{ino}}} \tag{11}$$

D_{ada} is approximated by

$$D_{\text{ada}} = \sum_{n=1}^{N} \left(s_w(n) - g s_w(n-T)\right)^2 \tag{12}$$

where T is an integer pitch-lag computed in the pre-processing stage of EVS and consequently available at no additional cost, and g is the gain which minimizes D_{ada}, obtained using

$$g = \frac{\sum_{n=1}^{N} s_w(n) s_w(n-T)}{\sum_{n=1}^{N} s_w(n-T) s_w(n-T)} \tag{13}$$

Finally, the estimated SNR of the CELP-based coding mode in a subframe is given by

$$\text{SNR}_{\text{celp}} = \frac{S}{D_{\text{celp}}} = \frac{\sum_{n=1}^{N} \left(s_w(n)\right)^2}{C_{\text{ino}} D_{\text{ada}}} \tag{14}$$

and the corresponding $\text{segSNR}_{\text{celp}}$ is obtained via (1).

2.2. Decision hysteresis

Segmental SNR is only an estimate of quality perception. It was observed that an ordinary comparison of $\text{segSNR}_{\text{celp}}$ and $\text{segSNR}_{\text{mdct}}$ and the selection of the coding mode exhibiting the larger of the two values (i.e. $dssnr$ = 0) results in frequent undesirable toggling between the CELP and MDCT modes. In particular, the temporary usage of the MDCT mode during clean speech, as well as a momentary switching to the CELP mode during background noise or musical passages, can lead to audible quality degradation even though the chosen mode yielded a greater segSNR. For this reason a simple hysteresis was introduced which minimizes frequent switching between the modes, especially in the abovementioned two cases.

The hysteresis is based on the observation that complex music and background noise exhibit high short-term temporal flatness, whereas the opposite is true for voiced and plosive speech. This is convenient, as frame-wise temporal flatness is already provided by the EVS time-domain transient detector (TDTD) [5, 9], which is a part of the pre-processing module depicted in Fig. 1. Moreover, $dssnr$ is modified based on the number P of consecutive CELP frames preceding the current one. Table 1 summarizes the computation of $ddsnr$, with P being a hand-tuned threshold (6 in EVS). As can be seen, when coding at least the last P_t frames with CELP, switching to MDCT is reduced if the frame is not temporally flat; and when at least one of the $P + 1$ previous frames was encoded with MDCT, MDCT is prefered if the frame is temporally flat.

Num. of prev. CELP frames	Frame is temporally flat	Frame is not temporally flat
$\geq P$	$ddsnr = 0$	$ddsnr = 2$
$\leq P$	$ddsnr = -2$	$ddsnr = 0$

Table 1. Value of $ddsnr$ as a function of the frame-wise temporal flatness and the number of previous CELP frames.

3. RESULTS

The algorithm described in the previous section was evaluated and compared with the closed-loop mode decision. Both schemes were implemented in the EVS codec [5] and both use the same hysteresis method discussed in 2.2. The results are presented in the following.

3.1. Mode decision statistics

Statistics on the decision were computed over 40 minutes audio files of clean speech, noisy speech, mixed content and music. The percentage of selected CELP obtained with both approaches is reported in Table 2. As it can be observed, the closed-loop decision is bit-rate dependent while the new version gives a constant decision. This consistency can be seen as an advantage in a system as the decision is not more function of an over or under-tuned module for certain bit-rates. Nevertheless, the behaviors of the two decisions are very similar. CELP is almost exclusively selected on clean and noisy speech on active voiced or active unvoiced segments. Inactive sections (silence or background noise) are usually handled by the transform coder. For mixed content, i.e. speech on music or speech between music, the decision is less systematic however speech-like sections are majority conveyed to CELP. For music, the transform coder is the preferred coding mode although CELP is usually selected on transients.

3.2. Objective quality evaluation

The tool for Perceptual Objective Listening Quality Assessment (POLQA [10]) was used for objectively evaluating the EVS codec quality on clean speech. 108 pair sentences of different languages were encoded using either the conventional closed-loop approach or the low-complexity version. Except on silence, CELP is expected to be selected most of the time on speech. A systematic difference in scores will indicate a misclassification problem. Figure 2 (a) shows that the average scores of the two decisions is very close to each other. The histogram of differential POLQA scores is given Figure 2 (b) and does not show any outlier. It can be concluded that, using the POLQA measurement tool, the proposed low-complexity approach performs as good as the closed-loop mode decision on clean speech.

3.3. Subjective quality evaluation

Listening tests were also conducted to subjectively evaluate the quality of the EVS codec using either the proposed mode decision or the closed-loop approach. Figure 3 shows the results of P.800 DCR listening tests [11] at SWB 16.4kbps and for three different contents, clean speech, noisy speech and mixed/music. The listening tests were conducted in germany using 16 naive listeners. Statistical analysis of the results show that the EVS codec with the new mode decision is statistically not worse than the EVS codec with the conventional closed-loop mode decision.

Condition	Decision	Clean Speech			Noisy Speech			Mixed Content				Music	
		V	U	I	V	U	I	V	U	I	M	I	M
SWB	LC-CL	98.8	95.4	13.2	98.3	91.1	16.2	77.9	68.2	5.5	18.2	2.6	16.6
SWB 13.2kbps	CL	97.9	78.1	6.6	87.2	51.1	4.7	77.3	50.5	7.6	12.7	5.5	18.6
SWB 16.4kbps	CL	98.3	87.2	9.0	89.1	59.4	7.1	77.2	56.6	7.9	13.3	7.6	19.7
SWB 24.4kbps	CL	98.8	95.5	25.4	89.9	73.9	21.0	76.9	66.9	13.1	17.1	19.6	22.3
SWB 32kbps	CL	95.7	88.8	8.9	79.2	57.6	7.2	69.2	58.9	5.7	12.9	12.1	20.6

Table 2. Percentage of selected CELP for different types of signals and portions: Voiced (V), Unvoiced (U), Inactive (I), Music (M).

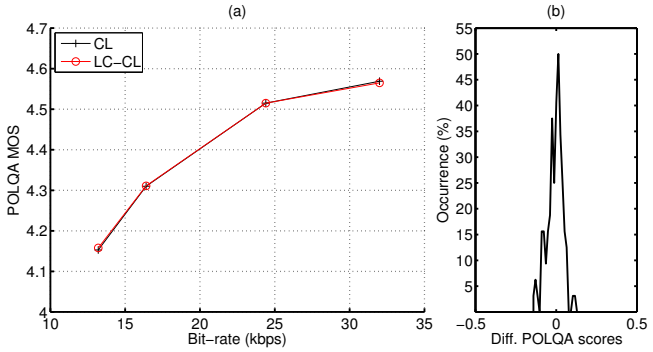

Fig. 2. POLQA mean scores and differential scores on clean speech SWB between the conventional closed-loop decision (CL) and the proposed low-complexity version (LC-CL).

3.4. Computational Complexity

The computational complexity was measured in WMOPS using a version of the EVS encoder implemented with the ITU-T fixed-point basic operators [12]. The input audio data was the same as the one used for computing the decision statistics (40 minutes audio files of clean speech, noisy speech, mixed content and music). The numbers given in Table 3 show that the proposed-approach is able to reduce the computational complexity of the EVS encoder by 24%-31% depending on the bitrate.

Condition	CL	LC-CL	Diff.
SWB 13.2kbps	74.9	56.9	-24.0%
SWB 16.4kbps	80.4	57.4	-28.6%
SWB 24.4kbps	88.7	62.0	-30.1%
SWB 32.0kbps	91.1	62.5	-31.4%

Table 3. Computational complexity of the EVS encoder (in WMOPS) using either the conventional closed-loop decision (CL) or the proposed low-complexity version (LC-CL).

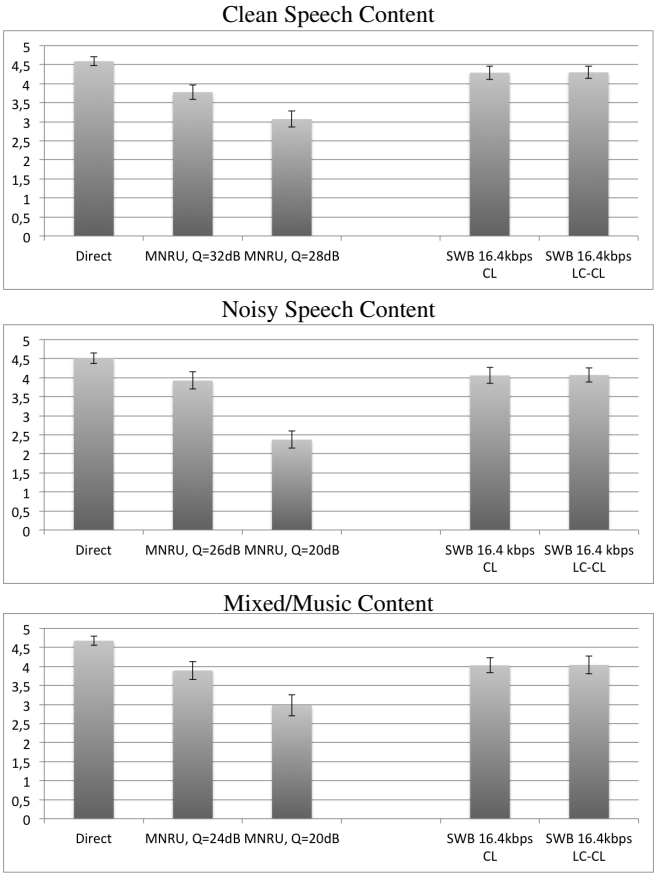

Fig. 3. P.800 DCR listening test results (means and confidence intervals) comparing the conventional closed-loop decision (CL) and the proposed low-complexity version (LC-CL) at SWB 16.4kbps.

4. CONCLUSION

In this paper, a low-complexity alternative to the traditional closed-loop decision is proposed. This new approach is part of the recent 3GPP EVS standard, where it makes the decision at every frame to use either the MDCT-based coding mode or the CELP-based coding mode. When compared to the conventional closed-loop decision, the results show that the new approach does not introduce any degradation in the performance and at the same time allows large savings in the computational complexity of the encoder.

5. REFERENCES

[1] *Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions*, 3GPP Technical Specification 26.290. [Online]. Available: http://www.3gpp.org/DynaReport/26290.htm

[2] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Proceedings of IEEE International Conference on Acoustic, Speech Signal Processing (ICASSP '05)*, vol. II, Mar. 2005, pp. 1109–1112.

[3] *Information technology – MPEG audio technologies – Part 3: Unified speech and audio coding*, ISO/IEC 23 003-3:2012. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:23003:-3:ed-1:v1:en

[4] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. R. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, K. Chong, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "The ISO/MPEG Unified Speech and Audio Coding Standard—Consistent High Quality for All Content Types and at All Bit Rates," *J. Audio Eng. Soc*, vol. 61, no. 12, pp. 956–977, Dec. 2013.

[5] *EVS Codec Detailed Algorithmic Description*, 3GPP Technical Specification 26.445. [Online]. Available: http://www.3gpp.org/DynaReport/26445.htm

[6] L. Tancerel, S. Ragot, V. Ruoppila, and R. Lefebvre, "Combined speech and audio coding by discrimination," in *Proceedings of IEEE Workshop on speech coding*, Sep. 2000, pp. 154–156.

[7] J. Makinen, A. Lakaniemi, and P. Ojala, "Low complex audio encoding for mobile, multimedia," in *Proceedings of IEEE 63rd Vehicular Technology Conference (VTC Spring)*, May 2006, pp. 461–465.

[8] J. K. Kim and N. S. Kim, "Improved frame mode selection for AMR-WB+ based on decision tree," *IEICE Transactions*, vol. 91-D, no. 6, pp. 1830–1833, 2008.

[9] C. R. Helmrich, G. Marković, and B. Edler, "Improved low-delay MDCT-based coding of both stationary and transient audio signals," in *Proceedings of IEEE International Conference on Acoustic, Speech Signal Processing (ICASSP '14)*, May 2014, pp. 6954–6958.

[10] *Perceptual objective listening quality assessment*, ITU-T Recommendation P.863, Sep. 2014. [Online]. Available: http://www.itu.int/rec/T-REC-P.863-201409-P/en

[11] *Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800, Aug. 1982. [Online]. Available: http://www.itu.int/rec/T-REC-P.800-199608-I/en

[12] *Software tools for speech and audio coding standardization*, ITU-T Recommendation G.191, Mar. 2010. [Online]. Available: http://www.itu.int/rec/T-REC-G.191-201003-I/en