CODED DOMAIN LEVEL CONTROL FOR THE AMR SPEECH CODEC

Antti Pasanen

Nokia Research Center,
P.O. Box 100, FIN-33721 Tampere, Finland
antti.ju.pasanen@nokia.com

ABSTRACT

This paper presents a coded domain level control technique for the Adaptive Multi-Rate (AMR) speech codec. Level control in the coded domain is done by directly modifying quantized speech parameters. The advantage of the method is an accurate speech level control with minimized system complexity and end-to-end delay when tandem coding can be avoided completely. Requantization optimization functions are derived for scalar and vector quantization of relevant parameters. The experimental results show that the presented technique makes possible to apply a desired gain in the coded domain maintaining high speech quality.

1. INTRODUCTION

In future speech transmission networks, speech enhancements, such as level control, noise suppression [1] and echo cancellation [2], are increasingly conducted in parameter level in the coded domain. Such methods are especially efficient when the originating and terminating connections were using the same speech codec without transcoding operations within the network. Tandem free operation (TFO) and transcoder free operation (TrFO) with parameter level speech processing lead to improvement in speech quality, savings in the processing power, transmission bandwidth and reductions in the end-to-end delay.

Speech level is one of the important factors affecting the perceived quality. On the network side level control algorithms are used to adjust the speech towards a desired level. Traditionally speech enhancement algorithms are carried out with PCM samples in the linear domain requiring tandem operation, i.e. an additional decoding and encoding process.

In this paper, we propose a technique that enables the level control of Adaptive Multi-Rate (AMR) coded speech directly by modifying quantized parameters. New optimization criteria are derived for the requantization of the level related parameters. Section 2 gives general overview of the AMR speech synthesis and describes the quantization of the fixed codebook gain. In Section 3 the coded domain level control technique is described and the requantization optimization criteria are derived. In Section 4 the experimental setup is explained and the results are presented. Finally, the work is briefly concluded in Section 5.

2. AMR SPEECH SYNTHESIS

In mobile communications, a low bit rate is desired while the speech quality should be preserved also in adverse conditions. The AMR codec fulfills these requirements and therefore 3GPP (3rd Generation Partnership Project) chose AMR as the mandatory speech codec for UMTS. The AMR codec has a frame length of 20 ms corresponding to 160 samples and each frame is divided into 4 subframes of equal length [3]. The codec has eight modes of operation with bit rates of 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 and 4.75 kbit/s and a low bit rate noise encoding mode for discontinuous transmission (DTX).

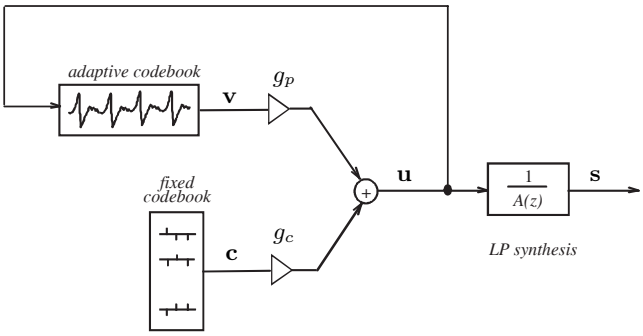

Fig. 1. Simplified block diagram of the CELP synthesis model.

The AMR codec is based on the Code-Excited Linear Predictive (CELP) coding model [4]. Figure 1 shows the general CELP synthesis model. It consists of the fixed codebook excitation $\mathbf{c}$ and the adaptive codebook excitation $\mathbf{v}$ and corresponding codebook gains g_c and g_p. The speech is reconstructed by filtering the total excitation signal $\mathbf{u}$ through the LP synthesis filter $1/A(z)$.

The transmitted data describes the following parameters: fixed codebook vector, fixed codebook gain, adaptive codebook delay, adaptive codebook gain and synthesis filter parameters. These parameters are quantized and encoded into

a speech frame and then transmitted to the decoder. At the decoder, the received parameters are decoded and speech is synthesized according to the CELP model.

Level information of the speech is transmitted using the codebook gains g_c and g_p. g_c is a multiplicative factor applied to the excitation signal while g_p controls the pitch contribution. The codebook gains are relatively independent of the other parameters and have a good range of quantization values. The proposed level control is performed by scaling g_c alone.

The fixed codebook gain g_c is quantized using the fixed codebook gain correction factor γ_{g_c} [3]. In the decoder the received gain correction factor $\hat{\gamma}_{g_c}$ adjusts the predicted fixed codebook gain g_c' to reconstruct the fixed codebook gain $\hat{g}_c$, i.e.

$$\hat{g}_c = \hat{\gamma}_{g_c} \cdot g_c'. \tag{1}$$

The fixed codebook gain is predicted for the k:th subframe as

$$g_c'(k) = 10^{0.05\left(\tilde{E}(k)+\bar{E}-E_I\right)},$$

where $\tilde{E}(k)$, $\bar{E}$ and E_I are predicted energy, mode dependent energy value and fixed codebook excitation energy, respectively. The predicted energy is found using MA-prediction of past correction factor values as

$$\tilde{E}(k) = \sum_{i=1}^{4} b_i \cdot 20 \log_{10} \hat{\gamma}_{g_c}(k-i),$$

where $[b_1, b_2, b_3, b_4] = [0.68, 0.58, 0.34, 0.19]$ are the MA-prediction coefficients.

The 12.2 kbit/s and 7.95 kbit/s modes employ Scalar Quantization (SQ) of the correction factor $\hat{g}_c$ while in the other modes the correction factor and the adaptive codebook gain are Vector Quantized (VQ). In the 4.75 kbit/s mode the gains from the 2 subframes are jointly vector quantized.

For the 12.2 kbit/s mode an open loop quantization of the correction factor is done and the quantization table search is performed by minimizing the quantization error ε_{sq} over $\hat{\gamma}_{g_c}^j$ for each subframe

$$\varepsilon_{sq} = \left(g_c - \hat{\gamma}_{g_c}^j \cdot g_c'\right)^2. \tag{2}$$

In the VQ modes a closed loop quantization is performed by minimizing the square of the weighted error between the original and reconstructed speech over $\hat{g}_p^j$ and $\hat{\gamma}_{g_c}^j$ for each subframe i.e.

$$\varepsilon_{vq} = \|\mathbf{x} - (\hat{g}_p^j \mathbf{y} + (\hat{\gamma}_{g_c}^j \cdot g_c')\mathbf{z})\|^2, \tag{3}$$

where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are the original weighted speech, synthesis filtered adaptive codebook vector and synthesis filtered fixed codebook vector respectively. For the 7.95 kbit/s mode an adaptive modified closed-loop criterion is used [3].

3. LEVEL CONTROL FOR AMR

In linear domain, level control is done by multiplying the signal by the desired gain α. The underlying idea of the coded domain level control is to scale the fixed codebook gain g_c by the gain α and keep the other parameters intact. This can be accomplished by modifying the quantized fixed codebook gain correction factor $\hat{\gamma}_{g_c}$. Let $\hat{\beta}(k-i)$ be the correction factor gain of the past subframes i.e.

$$\hat{\gamma}_{g_c}^{\text{new}}(k-i) = \hat{\beta}(k-i) \cdot \hat{\gamma}_{g_c}(k-i).$$

The new energy prediction for gain quantization is found in the k:th subframe as

$$\begin{aligned}
\tilde{E}^{\text{new}}(k) &= \sum_{i=1}^{4} b_i \cdot 20 \log_{10} \left(\hat{\gamma}_{g_c}^{\text{new}}(k-i)\right) \\
&= \sum_{i=1}^{4} b_i \cdot 20 \log_{10} \left(\hat{\beta}(k-i)\hat{\gamma}_{g_c}(k-i)\right) \\
&= \sum_{i=1}^{4} b_i \cdot 20 \log_{10} \hat{\beta}(k-i) + \\
&\qquad \sum_{i=1}^{4} b_i \cdot 20 \log_{10} \hat{\gamma}_{g_c}(k-i) \\
&= \sum_{i=1}^{4} b_i \cdot 20 \log_{10} \hat{\beta}(k-i) + \tilde{E}(k),
\end{aligned}$$

and the new fixed codebook gain prediction can be given as

$$\begin{aligned}
g_c'^{\,\text{new}}(k) &= 10^{0.05\left(\tilde{E}^{\text{new}}(k)+\bar{E}-E_I\right)} \\
&= 10^{0.05\left(\sum_{i=1}^{4} b_i 20 \log_{10} \hat{\beta}(k-i)+\tilde{E}(k)+\bar{E}-E_I\right)} \\
&= 10^{0.05\left(\sum_{i=1}^{4} b_i 20 \log_{10} \hat{\beta}(k-i)\right)} 10^{0.05\left(\tilde{E}(k)+\bar{E}-E_I\right)} \\
&= 10^{\left(\sum_{i=1}^{4} b_i \log_{10} \hat{\beta}(k-i)\right)} \cdot g_c'(k).
\end{aligned}$$

I.e. the adjustments of the past correction factors contribute to the $g_c'^{\,\text{new}}(k)$ which correspondingly scales the reconstructed fixed codebook gain $\hat{g}_c$ as shown in Eq. (1). Finally, the codebook index in the coded domain representing the gain is replaced with a new index corresponding to the new gain value.

3.1. Requantization for the Scalar Quantization

The 12.2 kbit/s and 7.95 kbit/s modes perform scalar quantization for the fixed codebook gain correction factor. The error function given in Eq. (2) can be used in the requantization. However, the fixed codebook gain g_c is only available in the encoder. In the requantization the g_c is replaced by the scaled quantized fixed codebook gain and the prediction is replaced by the new value $g_c'^{\,\text{new}}$. The error is minimized for

k:th subframe over $\hat{\gamma}_{g_c}^j$ as

$$\begin{aligned}\varepsilon_{sq_\alpha} &= \left(\alpha\hat{g}_c - \hat{\gamma}_{g_c}^j \cdot g_c'^{\,\text{new}}\right)^2 \\ &= \left(\alpha\hat{\gamma}_{g_c} g_c' - \hat{\gamma}_{g_c}^j \cdot 10^{\left(\sum_{i=1}^4 b_i \cdot \log_{10}\hat{\beta}(k-i)\right)} \cdot g_c'\right)^2 \\ &= g_c'^2 \left(\alpha\hat{\gamma}_{g_c} - \hat{\gamma}_{g_c}^j \cdot 10^{\left(\sum_{i=1}^4 b_i \cdot \log_{10}\hat{\beta}(k-i)\right)}\right)^2,\end{aligned}$$

where α is the target gain. Minimization of the ε_{sq_α} equals to minimization of

$$\varepsilon_{sq_\alpha^*} = \left(\alpha\hat{\gamma}_{g_c} - \hat{\gamma}_{g_c}^j \cdot 10^{\left(\sum_{i=1}^4 b_i \cdot \log_{10}\hat{\beta}(k-i)\right)}\right)^2.$$

The received correction factor $\hat{\gamma}_{g_c}$ and corresponding codebook index are replaced by the new correction factor value $\hat{\gamma}_{g_c}^{\text{new}}$ and codebook index minimizing the function $\varepsilon_{sq_\alpha^*}$.

3.2. Requantization for the Vector Quantization

Most of the AMR modes use vector quantization of the fixed codebook correction factor γ_{g_c} and adaptive codebook gain g_p. Requantization error criterion is designed according to the Eq. (3). However the original weighted speech $\mathbf{x}$ used in the encoder is not available afterwards. For the level control purposes $\mathbf{x}$ is approximated using the reconstructed speech i.e.

$$\mathbf{x} \approx \hat{g}_p\mathbf{y} + \hat{g}_c\mathbf{z},$$

where $\mathbf{y}$ and $\mathbf{z}$ are synthesis filtered adaptive codebook vector $\mathbf{v}$ and fixed codebook vector $\mathbf{c}$ respectively. Eventually the scaling of the fixed codebook gain will scale the adaptive codebook vector $\mathbf{v}$ accordingly. It is reasonable to approximate the new synthesis filtered adaptive codebook vector with $\mathbf{y}^{\text{new}} \approx \alpha\mathbf{y}$. Thus the requantization error function can be given as

$$\begin{aligned}\varepsilon_{vq_\alpha} &= \|\alpha\mathbf{x} - \left(\hat{g}_p^j\mathbf{y}^{\text{new}} + (\hat{\gamma}_{g_c}^j \cdot g_c'^{\,\text{new}})\mathbf{z}\right)\|^2 \\ &\approx \|\alpha\left(\hat{g}_p\mathbf{y} + \hat{g}_c\mathbf{z}\right) - \left(\hat{g}_p^j\alpha\mathbf{y} + (\hat{\gamma}_{g_c}^j \cdot g_c'^{\,\text{new}})\mathbf{z}\right)\|^2 \\ &= \|\left(\hat{g}_p - \hat{g}_p^j\right)\alpha\mathbf{y} + \left(\alpha\hat{g}_c - \hat{\gamma}_{g_c}^j \cdot g_c'^{\,\text{new}}\right)\mathbf{z}\|^2.\end{aligned}$$

Using the result from the previous section the minimization is equal to

$$\begin{aligned}\varepsilon_{vq_\alpha^*} = \|&\left(\hat{g}_p - \hat{g}_p^j\right)\alpha\mathbf{y} + \\ &\left(\alpha\hat{\gamma}_{g_c} - \hat{\gamma}_{g_c}^j \cdot 10^{\sum b_i \cdot \log_{10}\hat{\beta}(k-i)}\right)\mathbf{z}\|^2.\end{aligned}$$

The received correction factor $\hat{\gamma}_{g_c}$ and the adaptive codebook gain $\hat{g}_p$ are replaced by the new correction factor value $\hat{\gamma}_{g_c}^{\text{new}}$ and the adaptive codebook gain $\hat{g}_p^{\text{new}}$ which minimize $\varepsilon_{vq_\alpha^*}$. The requantization function determines how the quantization error is divided between the two parameters. In the 4.75 kbit/s mode, requantization is done by minimizing a weighted sum of subframe errors $\varepsilon_{vq_\alpha^*}$ [3].

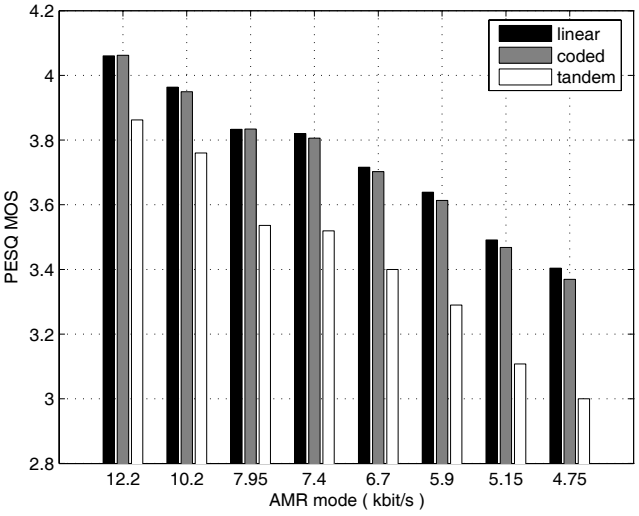

Fig. 2. Average PESQ MOS for the speech samples.

3.3. Silence Description Frames

The CELP coding model is utilized only during active speech. The average background noise level and spectral shape is transmitted in silence description (SID) frames between speech bursts [5]. When adjusting the speech level in coded domain, the information on the background noise level is important. If the signal level was adjusted only during active speech frames, the background noise level would change abruptly at the beginning and in the end of background noise only periods causing subjectively very annoying effects. Therefore, if the level of speech is adjusted, also the silence description frames should be adjusted accordingly. The averaged logarithmic frame energy parameter transmitted in the SID can be adjusted to obtain a suitable comfort noise level.

4. EXPERIMENTAL SETUP AND RESULTS

The performance of the coded domain speech level control technique was evaluated with subjective and objective criteria. The test set consisted of five male and five female speech samples of length 15 seconds. The test samples were first high pass filtered to simulate the sending frequency characteristics of a telephone handset. The levels were normalized to -16, -20, -32 and -36 dBov using a P.56 voltmeter [6]. For the evaluation the speech level was modified by -12, -8, $+8$ and $+12$ dB respectively. Three sets of samples were generated: 1) samples processed in the coded domain (coded), 2) AMR coded samples that were scaled in linear domain after decoding (linear), and 3) AMR coded samples that were scaled in the linear domain after the decoding and coded again (tandem). In the experiments, DTX was disabled.

The test samples were listened informally. The speech quality after transcoding (tandem) was found to be worse

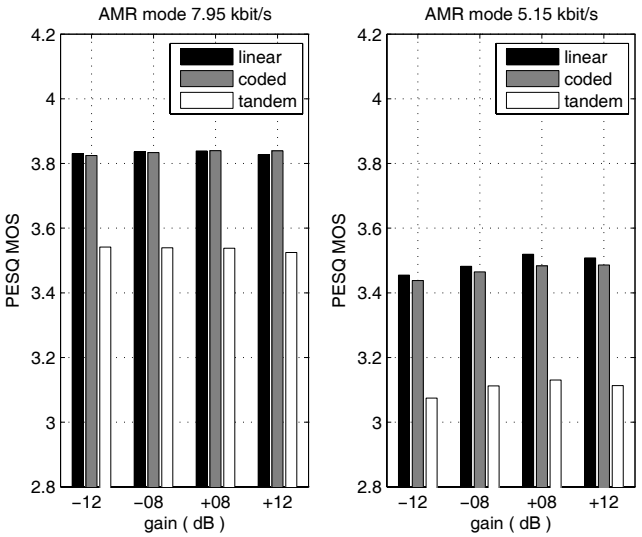

Fig. 3. Average PESQ MOS for two AMR modes.

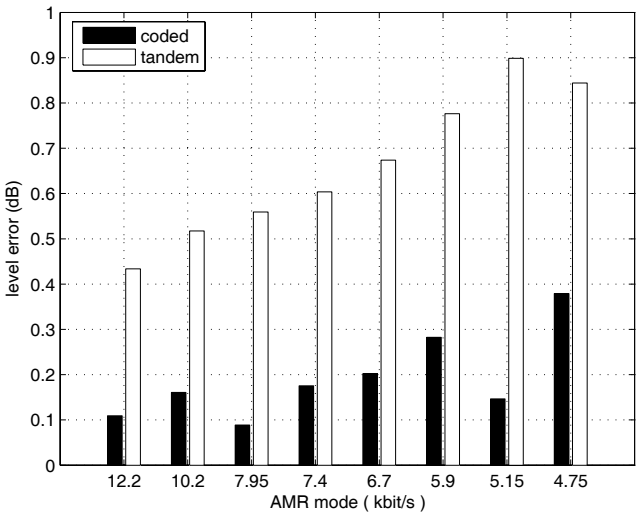

Fig. 4. Average absolute level differences compared to linear domain scaled reference samples.

than after coded domain processing (coded), especially at the lower bit rates. Actually no degradations could be heard while the samples processed in the coded domain (coded) were compared to the samples processed in the linear domain (linear).

In addition to the subjective analysis, objective Perceptual Evaluation of Speech Quality (PESQ)-method was utilized to predict the Mean Opinion Score (MOS) for the test samples [7]. In Figure 2 the average PESQ scores are shown for different AMR modes and processing types. Processing in the linear and coded domain (linear and coded) resulted about the same average PESQ scores. Transcoding (tandem) resulted in significantly lower scores than the others. In Figure 3 the average PESQ scores are shown for two AMR modes according to the used gain. Irrespective of the applied gain, the PESQ scores were found to be quite similar.

PESQ compensates for non-optimum signal levels in the input samples. Therefore the levels of the processed samples were analyzed using the voltmeter to verify the resulted gain. As a reference we used samples processed in the linear domain. The level error was computed as an absolute level difference (in dB) between the reference (linear) and processed samples (coded, tandem). As seen in Figure 4 the average level differences are clearly lower in all AMR modes when the coded domain processing (coded) is utilized.

5. CONCLUSIONS

Technique for controlling the level of AMR coded signal in the coded domain was discussed. The method consists of replacing the relevant parameters in the bit stream according to the level control. New criteria are derived for the requantization of the related parameters. The performance of the coded domain processing was evaluated using PESQ MOS, voltmeter and informal listening of the processed speech samples. It was shown that the performance of the coded domain processing does not differ from level control of linear domain signal. The advantage is that, the method preserves the quality and introduces significant savings in complexity and end-to-end delay since transcoding can be avoided in the system.

6. REFERENCES

[1] Hervé Taddei, Christophe Beaugeant, and Mickael de Meuleneire, "Noise reduction on speech codec parameters," in *IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, 2004, vol. 1, pp. 497–500.

[2] Ravi Chandran and Daniel J. Marchok, "Compressed domain noise reduction and echo suppression for network speech enhancement," in *43rd IEEE Midwest Symbosium on Circuits and Systems*, Aug. 2000, vol. 1, pp. 10–13.

[3] *AMR Speech Codec; Transcoding Functions*, 3GPP TS 26.090. June 2002.

[4] A. M. Kondoz, *Digital Speech Coding for Low Bit Rate Communications Systems*, John Wiley & Sons, Chichester, 1994.

[5] *AMR Speech Codec; Comfort Noise Aspects*, 3GPP TS 26.092. June 2002.

[6] *Objective Measurement of Active Speech Level*, ITU-T P.56. Mar. 1996.

[7] *Perceptual Evaluation of Speech Quality (PESQ): an Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs*, ITU-T P.862. Feb. 2001.