ECHO REDUCTION BASED ON SPEECH CODEC PARAMETERS

Emmanuel Thepie[1], *Christophe Beaugeant*[2], *Hervé Taddei*[2], *Nicolas Dütsch*[3], *Dominique Pastor*[4]

[1]BenQ Mobile, Munich, Germany (Fapi.Thepie@BenQ.com)
[2]Siemens AG, Munich, Germany
[3]Technische Universität München, Munich, Germany
[4]ENST Bretagne, Brest, France

ABSTRACT

To improve speech quality, disturbing background noise and acoustic echo are attenuated by using signal processing techniques prior to speech encoding. A different approach consists of embedding and performing noise reduction and/or echo cancellation into the speech codec. This embedding decreases the complexity and allows integration of noise reduction and echo cancellation in the network without adding any delay or creating the so-called tandem effect. In this paper we propose a solution based on this principle. We introduce an innovative estimation of the signal to echo ratio based on a linear model of the fixed gain parameters of the speech codec. Listening test results validate the good quality for such a low complexity system.

1. INTRODUCTION

In mobile phone environment, external disturbing signals (environmental noise, acoustic echo) corrupt the useful speech signal. In presence of these types of disturbing signals, low bitrate speech codecs do not perform so well. Indeed, such codecs are well designed for encoding single clean speech signal, but are not suitable to encode other kinds of signal, as for example noisy speech. To improve speech quality, noise reduction (NR) and echo compensation (EC) are strongly recommended.

Recently, speech enhancement has been implemented in the network in order to allow operators to deliver to their customers signal with constant quality whatever terminal is used. Usually, this requires to decode the bitstream, to perform NR and EC in the time and/or frequency domain and to re-encode the signal. This creates the so-called tandem effect that impacts the quality. Another approach consists of embedding NR and EC into the speech codec. It highly reduces the complexity compared to the former technique. Another substantial advantage is that the tandeming effect is also very reduced, as the NR/EC integration in the network does not require the decoding and re-encoding of the signal, but just the modification of a few bitstream parameters.

In [1] [2], the fixed gain of the speech codec was shown to be a relevant parameter to decrease background noise. Further study in [3] extended the analysis to the EC problem, introducing a gain loss control in the parameter domain. This paper proposes an extension of [3] by proposing a smoothed modification rule of the fixed gain.

2. ACOUSTIC ECHO

Echo is due to the acoustic coupling between the phone transducers. It creates feedback of the far-end speech through the whole communication path. Due to delay introduced by the network, the far-end user experiences the annoying effect of hearing his own voice with a delay of around 200 to 500 ms. Fig. 1 depicts the coupling model. The loudspeaker signal $x(t)$ is coupled to the microphone through the acoustic path $h(t)$ and the resulting echo $e(t)$ is considered to be the result of the convolution of $x(t)$ and $h(t)$. In this paper we only focus on the echo, accordingly we neglect noise ($n(t) = 0$). The microphone signal $y(t)$ is the addition of the useful speech signal $s(t)$ and of the echo signal $e(t)$. Taking into account discrete signals, we can write that $y(n) = s(n) + e(n)$.

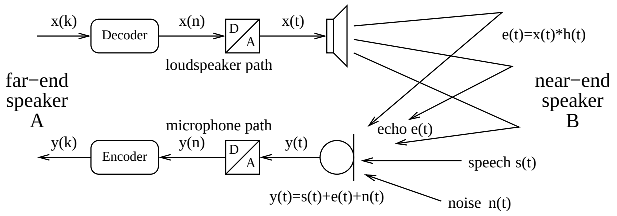

Figure 1: Model of acoustic and mechanic echo

3. GAIN MODIFICATION FILTER

3.1. General assumption

To study echo reduction embedded in speech codecs, we use the Adaptive Multi Rate (AMR) codec [4] at 12.2 kbit/s. In the z-domain, the AMR synthesis filter transfer function applied to the microphone coded signal $y(k)$

can be written as in [2]:

$$S(z) = \frac{g_y(m)}{\left(1 - g_a(m) \cdot z^{-T(m)}\right)\left(1 + \sum_{i=1}^{M} c_i(m) \cdot z^{-i}\right)} \quad (1)$$

with M being the order of the linear prediction filter, m the subframe index, c_i the Linear Prediction Coefficients, $g_a(m)$ the adaptive gain, T the current pitch delay and $g_y(m)$ the fixed gain value. Referring to this formula, g_y can be seen as a multiplicative factor directly applied to the signal. As a result, reducing $g_y(m)$ decreases the signal amplitude. By applying a scalar factor $G(m)$ to the fixed gain, $G(m)$ depending on the amplitudes of the echo and of the useful signal, we can expect to reduce the processed signal amplitude according to the echo amplitude. The process is as follows, we apply a gain to $g_y(m)$:

$$g_s(m) = G(m) g_y(m) \quad (2)$$

and we replace the corrupted fixed gain $g_y(m)$ with $g_s(m)$ in the coded parameter domain. We consider that $G(m)$ leads to a ”perfect” echo reduction so that $g_s(m)$ would be the fixed speech gain obtained if there were no echo (pure clean speech condition). To do so, we propose a new method to compute $G(m)$ based on a joint function as introduced in the next section.

3.2. Approximation of the joint function

Our basic idea consists in computing the fixed gain $g_y(m)$ of the input signal $y(n)$ as a joint function $f()$ depending on the speech gain, $g_s(m)$, and on the echo gain $g_e(m)$ (i.e. the gain obtained if there were no useful signal): $g_y(m) = f(g_s(m), g_e(m))$. As a result:

$$G(m) = \frac{g_s(m)}{f(g_s(m), g_e(m))} \quad (3)$$

We assumed that $f()$ can be written as a linear function based on three parameters a, b and c. Despite the simplicity of the retained model, our simulations (see section 5) justified this strong hypothesis of linear model:

$$f(g_s(m), g_e(m)) = a \cdot g_s(m) + b \cdot g_e(m) + c \quad (4)$$

When the input signal is zero the output should also be zero, therefore $c = 0$. Introducing the Speech to Echo Ratio (SER) in the codec parameter domain as the ratio between the fixed gains of the speech and of the echo:

$$SER(m) = \frac{g_s(m)}{g_e(m)} \quad (5)$$

The weighting gain of Eq. (3) can be written as follow:

$$G(m) = \frac{SER(m)}{b + a \cdot SER(m)} \quad (6)$$

This last expression can be interpreted as a weighted Wiener filtering on the gain, showing the similarity of our method to the filter developped in ’classical’ frequency domain noise reduction.

3.3. Linear Coefficients approximations

During echo only period, $e(n) \neq 0$ and $s(n) = 0$, $g_y(m)$ is the echo gain. According to Eq. 6, during this echo only period, b is equal to one. In the same way, when $e(n) = 0$ and $s(n) \neq 0$ we can find out that $a = 1$. Accordingly, by defining single-talk periods as periods when ($e(n) \neq 0$, $s(n) = 0$) or ($e(n) = 0$, $s(n) \neq 0$), we can choose:

$$a\,|_{\text{singletalk}} = 1 \qquad , \qquad b\,|_{\text{singletalk}} = 1 \quad (7)$$

During double talk periods, $s(n) \neq 0$ and $e(n) \neq 0$, a and b are estimated by computing a set of gains $\mathbf{G}_\nu = [g_\nu(0) \ \cdots \ g_\nu(l-1)]^T$, obtained from several experimental scenarios, with $\nu \in \{s, e, y\}$ and l the number of subframes. This set of gains verifies the equation:

$$[\mathbf{G_s} \quad \mathbf{G_e}] \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{G_y} \quad (8)$$

We assume that $\mathbf{z} = [a \quad b]^T$ stays constant during double talk periods. The over determined system of equations (8) is solved in the least-squares sense by the pseudo-inverse X^+ (also called Moore-Penrose generalized inverse [5]), of the matrix $\mathbf{X} = [\mathbf{G_s} \quad \mathbf{G_e}]$:

$$\mathbf{X}^+ = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \qquad \mathbf{z} = \mathbf{X}^+ \, \mathbf{G_y} \quad (9)$$

where X^T is the transposed matrix of X. In order to verify the solution that was obtained in Eq. 8 for a and b, we compute a normalized error by comparing the exact value of the microphone fixed gain with the estimated value. This verification was done using six different typical carkit systems:

$$error = \frac{\sum_{\kappa=0}^{l-1} \left(g_y(\kappa) - (a \cdot g_s(\kappa) + b \cdot g_e(\kappa))\right)^2}{\sum_{\kappa=0}^{l-1} g_y^2(\kappa)} \quad (10)$$

Results are shown in Tab. 1. As our database scenario was based on similar environments (acoustic in car), the optimal vector $\mathbf{z}$ is almost the same for the six echo paths:

$$a\,|_{\text{doubletalk}} \approx 1 \qquad \text{and} \qquad b\,|_{\text{doubletalk}} \approx \frac{4}{3} \quad (11)$$

Finally, the filter expression in Eq. (6) is simplified as:

$$G(k) = \frac{SER(k)}{\zeta + SER(k)} \quad (12)$$

with ζ equal to 1 in single talk and $4/3$ in double talk.

3.4. SER estimation

The SER is computed recursively, similarly to [2]:

$$SER(m) = \beta \, \frac{G(m-1) g_y(m-1)}{g_e(m)} + (1 - \beta) \, \frac{g_y(m)}{g_e(m)} \quad (13)$$

Such a formula allows us to compute the SER without the need to estimate $g_s(m)$, only the computation of $g_e(m)$ is required. Different experiments showed that the use of $\beta \approx 0.9$ leads to an accurate estimation of the SER.

carkit	c_1	c_2	c_3	c_4	c_5	c_6
”optimal” a	1.03	0.99	0.97	0.98	0.96	0.97
”optimal” b	1.33	1.39	1.36	1.34	1.38	1.27
$error$ (%)	9.30	6.85	3.35	8.27	5.75	3.25

Table 1: Linear coefficients in double-talk mode

4. FIXED GAIN ESTIMATION OF THE ECHO

To estimate the fixed gain $g_e(m)$ of the echo signal as in [6], we multiply the loudspeaker gain from the present or past (shifted by $m_{opt}(m)$ sub-frames) by $g_{opt}(m)$:

$$\hat{g}_e(m) = g_{opt}(m) g_x(m - m_{opt}(m)) \quad (14)$$

with $g_x(m)$ being the fixed gain of the loudspeaker signal. The values $m_{opt}(m)$ and $g_{opt}(m)$ are computed in two steps: echo mode detection according to a correlation analysis and determination of the filter parameters $g_{opt}(m)$ and $m_{opt}(m)$.

4.1. Echo mode detection

The echo mode detection is done through a correlation analysis between the gains of the loudspeaker $g_x(m)$ and of the microphone paths $g_y(m)$. To estimate the signal energy, we directly used the fixed gains information. First, encoder and decoder smoothed fixed gains $\hat{g}_i$ are computed as follow:

$$\hat{g}_i(m) = \gamma\ \hat{g}_i(m-1) + (1-\gamma)\ g_i(m) \quad i \in \{x, y\} \quad (15)$$

where γ is a smoothing factor, typically $\gamma = 0.9$.
Echo is detected if:

$$\hat{g}_x(m) > max(t, \hat{g}_y(m)\ ) \quad (16)$$

where $t = 10$. Eq. (16) verifies that the far-end speaker is talking by comparing its fixed gain to a fixed threshold t measured in the loudspeaker path and by verifying also that the loudspeaker fixed gain is bigger than the microphone fixed gain. This is typically the case when we assume that the coupling between the loudspeaker and the microphone reduces the signal energy. To detect double-talk mode, fixed gains $g_x(m)$ and $g_y(m)$ are analyzed using the normalized cross-correlation function:

$$\varphi_{g_x g_y}(i) = \frac{\sum_{j=0}^{N-i-1} g_x(j+i) g_y(j)}{\sqrt{\sum_{j=0}^{N-1} g_x(j)^2 \sum_{j=0}^{N-1} g_y(j)^2}} \quad (17)$$

N is the length of the cross-correlation analysis. The maximum of the correlation function as well as its corresponding lag are searched:

$$c_{\max}(m) = \max_i \varphi_{g_x g_y}(i) \quad (18)$$

$$\ell_{\max}(m) = arg \max_i \varphi_{g_x g_y}(i) \quad (19)$$

Finally, if $c_{max}(m)$ is bigger than a threshold $t_a = 0.75$ and Eq. (16) is fulfilled, echo only period is assumed, then $m_{opt}(m)$ and $g_{opt}(m)$ are adapted as described in 4.2.

4.2. Determination of the filter parameters

The parameters $m_{opt}(m)$ and $g_{opt}(m)$ are determined in a similar way as in [6] with a difference being that we use the fixed gains as a sufficiently good representation of the energy. In a first stage the optimal sub-frame shift $m_{opt}(m)$ is specified. Therefore a short-term lag $\ell_{st}(m)$, taking into account the rapid fixed gain variations is computed during echo periods:

$$\ell_{st}(m) = \hat{\alpha}(m)\ \ell_{st}(m-1) + (1 - \hat{\alpha}(m))\ \ell_{\max}(m) \quad (20)$$

where the smoothing factor, $\hat{\alpha}(m)$, is a function of the correlation coefficient $c_{max}(m)$:

$$\hat{\alpha}(m) = \begin{cases} -\frac{\alpha-\delta}{1-t_a} c_{\max}(m) + \frac{\alpha-\delta.t_a}{1-t_a} & \text{if} \quad c_{\max} > t_a \\ \alpha & \text{else} \end{cases} \quad (21)$$

where δ and α are smoothing factors, with $\alpha = 0.96$ and $\delta = 0.25$. As a result, the short-term lag is adapted slower or faster depending on the correlation between the gains of the loudspeaker and of the microphone.
During echo period, an average value of the short term lag is computed ($\mu = 0.995$):

$$\bar{\ell_{st}}(m) = \mu\ \bar{\ell_{st}}(m-1) + (1-\mu)\ \ell_{st}(m) \quad (22)$$

This averaged lag is used during non-echo period as a convergence point for the short term lag according to:

$$\ell_{st}(m) = \alpha\ \ell_{st}(m-1) + (1-\alpha)\ \bar{\ell_{st}}(m) \quad (23)$$

It means that if a non echo period is really short, $\ell_{st}(m)$ keeps a value close to the last computed value. In that case, for the next echo period we consider that the echo path did not change dramatically, and we use nearly the same value as the last short term lag. If the non echo period is long, $\ell_{st}(m)$ converges to the averaged short term lag $\bar{\ell_{st}}(m)$. The echo path may have changed considerably. As there is no way to estimate this change, we use a conservative value as the next short term lag when the next echo period starts. Finally, the optimal sub-frame shift is obtained by:

$$m_{opt}(m) = \text{round}(\ell_{st}(m)) \quad (24)$$

The ratio of the microphone gain and the shifted loudspeaker gain is calculated during echo periods:

$$g_{opt}(m) = \frac{g_y(m)}{g_x(m - m_{opt}(m))} \tag{25}$$

As the gain of the echo should be smaller than the gain of the loudspeaker, if this ratio is bigger than one, it is set to its previous value.
During non-echo periods, the short-term multiplication factor is updated with its long-term value in similar ways as for the computation of $m_{opt}(m)$. Finally, the obtained values of $m_{opt}(m)$ and $g_{opt}(m)$ are used to estimate the echo gain $\hat{g}_e(m)$ according to Eq. (14).

5. EXPERIMENTAL RESULTS

An ACR (Absolute Category Rate) listening test [7] was conducted. 10 naive and expert listeners participated. They scored each scenario defined as a conversation between a near-end speaker and a far-end speaker, using Mean Opinion Scores (MOS) ranging from 1 (unacceptable) to 5 (excellent). The listening test files contain both single talk and double talk periods. They are classified in 4 groups of 15 scenarios each. Group A is composed of clean speech files without echo, group B of files with unprocessed echo, group C of files processed with our echo reduction method based on codec parameters (CP) and group D of files enhanced using a "standard" Normalized-Least-Mean-Square (NLMS) method based on [8]. Echo is simulated using three different impulse responses of car named h_i. Mean scores and their standard deviation are displayed in Tab. 2.
When considering mean scores through all scenarios, the listening test results show that the quality of our echo reduction based on CP method is assessed slightly below the quality of the NLMS. It shows that our relatively simple solution brings results not far away to the intensively studied NLMS method. Moreover, the obtained averaged MOS (3.15) indicates an absolute quality of fair/good and is far better that the assesment of the unprocessed signal (1.69). In addition, the results are highly dependent on the kind of impulse response that was used and the resulting SER of the scenarios. We have measured mean SER during double talk periods and obtained the following values: 11 dB for h_1, 15 dB for h_2 and 17 dB when using h_3. We can see in Tab. 2 that the CP method was not as good as NLMS for lower SER whereas it was rated as good or even better for SER>15 dB.

$Group$	$GrpA$	$GrpB$	$GrpC$	$GrpD$
h_1	$4.66/0.63$	$1.6/0.93$	$2.68/0.87$	$4.1/0.78$
h_2	$4.58/0.62$	$1.64/0.72$	$3.1/0.7$	$2.82/0.80$
h_3	$4.56/0.57$	$1.84/0.88$	$3.68/0.76$	$3.72/0.82$
$Total$	$4.6/0.69$	$1.69/0.85$	$3.15/0.88$	$3.56/0.99$

Table 2: Mean and Standard Deviation Opinion Score

6. CONCLUSION

This paper presents an echo reduction method embedded in speech codec. This method is directly based on the modification of the speech codec parameters. We modify the fixed gain using weighting rules comparable to Wiener filtering depending on an innovative estimation of the SER. The SER estimation is based on a linear model of the joint function as described in Section 3.2. One main advantage is that its complexity is very low compare to 'classical' solution as NLMS. The listening test shows that the performance of our proposed echo reduction is highly influenced by the SER. We find out that the quality of our method is better in high SER condition than classical NLMS method. The proposed system still requires enhancement in low SER condition. One possibility would be to obtain a better estimation of the fixed gain of the echo, for instance by applying a non linear model when computing the joint function. Another possibility would be to combine the proposed solution with the one depicted in [2]. Behavior and influence of the other codec parameters like Linear Prediction Coefficients are also under investigation.

7. REFERENCES

[1] R. Chandran and D.J. Marchok, "Compressed domain noise reduction and echo suppression for network speech enhancement," in *Proc.of the 43rd IEEE Midwest Symposium on Circuits and Systems*, 2000, vol. 1, pp. 10–13.

[2] H. Taddei, C. Beaugeant, and M. de Meuleneire, "Noise reduction on speech codec parameters," *ICASSP*, May 2004.

[3] C. Beaugeant, N. Duetsch, and H. Taddei, "Gain loss control based on speech codec parameters," *EUSIPCO*, September 2004.

[4] 3GPPP TS 26.071, *Mandatory Speech Codec Speech Processing Functions; General Description*, June 2002.

[5] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1996.

[6] P. Heitkämper, "An adapatation control for acoustic echo cancellers," in *IEEE Signal Processing Letters*, April 1997, vol. 2, pp. 170–172.

[7] ITU-T P.800, *Methods for Subjective Determination of Transmission Quality*, 1996.

[8] S. Haykin, *Adaptive Filter Theory*, chapter 5, Least-Mean-Square Adaptive Filters, Prentice Hall, 4^{th} edition, 2002.