ACOUSTIC ECHO CANCELLATION ON THE ADAPTIVE MULTI-RATE SPEECH CODEC PARAMETERS

Daniele Giacobello[1], *Danilo Neri*[2], *Luca Prati*[2], *Sergio Brofferio*[3]

[1]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
[2]Nokia Siemens Networks, Cinisello Balsamo, Milano, Italy
[3]Dipartimento di Elettronica e Informazione, Politecnico Di Milano, Milano, Italy
dg@es.aau.dk, {danilo.neri,luca.prati}@nsn.com, sergio.brofferio@polimi.it

ABSTRACT

In this paper we present a novel solution to the problem of acoustic echo cancellation in the context of mobile communications. Traditionally, speech enhancement techniques in mobile networks are done in the transcoding unit operating on the uncoded signal. This means that the signal coming from the mobile terminals has to be decoded, enhanced and encoded again; these operations obviously introduce delays other than being computationally intensive and particularly prone to adding further quantization noise. The aim of this work is to reduce the computational costs and delays of this inherently suboptimal scheme; in order to do so, we transfer the acoustic echo cancellation operations to the coded domain by operating directly on the codec parameters.

Index Terms— Acoustic Echo Cancellation, Adaptive Multi-Rate Codec

1. INTRODUCTION

Voice Quality Enhancement (VQE) has assumed a great deal of importance in the context of mobile communications systems. The quality of the signal results to be greatly compromised due to numerous phenomena, among these the presence of acoustic echo due to the coupling of loudspeaker and microphone in the mobile terminals. A great deal of research has taken place in the past decades in order to mitigate this effect and a number of advanced techniques for Acoustic Echo Cancellation (AEC) are today a integral part of every speech transmission device (see, e.g., [1]). However, most of these algorithms are limited to the uncoded domain and result to be inappropriate when dealing with speech that has been previously encoded in the mobile terminal. The procedure in this case is to decode the signal, perform the AEC, and encode back the signal. However, by doing this, high computational costs and delays are introduces and this procedure, other than being costly, does not take advantage of the information already present in the coded speech parameters. A more efficient solution would be to work directly on the coded speech.

In this paper we will discuss a novel acoustic echo canceller that works directly on the Adaptive Multi-Rate (AMR) codec parameters, being this the standard speech codec adopted for GSM and UMTS networks [2]. After giving a brief overview on the AMR codec and the physical phenomenon of the echo in mobile communications, we will present the algorithms used to perform the AEC along with the experimental results. We will conclude our paper with the results, advantages and drawbacks of this technique.

2. OVERVIEW OF THE ADAPTIVE MULTI-RATE CODEC

The AMR was chosen by the 3GPP consortium as the mandatory codec for the UMTS mobile networks working with speech sampled at 8 kHz. Its main advantage is to be a multimodal coder, working on different rates from 12.2 kbit/s to 4.75 kbit/s, with the possibility of changing rate during the voice transmission by interacting with the channel coder. In our studies, mainly centered on the analysis of parameters, we worked on the 12.2 kbit/s mode (AMR 122) considering straightforward the extension to lower bit rates. Below, we will give a very brief overview on the main aspects of the encoder.

The AMR codec is based on the Algebraic Code Excited Linear Prediction (ACELP) paradigm [3] that refers to a particular approach for finding the most appropriate residual excitation after the linear prediction (LP) analysis. The speech waveform, after being sampled at 8 kHz and quantize with 16 bits, is divided into frames of 20 ms (160 samples) where each frame contains 4 subframes of equal length. The codec then uses a 10^{th} order linear predictive analysis on a subframe basis and then transform the coefficients obtained into Line Spectral Frequencies (LSF) [4] for more robust quantization.

After passing the signal through the LP filters, a residual signal is obtained. The codec then looks for a codeword that best fits the residual. There are two codebooks in the ACELP paradigm: an adaptive codebook and an algebraic codebook (also called fixed codebook). The parameters of the adaptive codebook are the pitch gain and pitch period; these are found through a closed-loop long-term analysis. The parameters of the fixed codebook are found analyzing the residual signal subtracted of its pitch excitation. The calculations make possible to find a codeword with only 10 non-zero coefficients out of 40 and a gain (on a subframe basis).

The decoder performs the synthesis of the speech using the transmitted parameters. The excitation that is passed through the LP filter is created by combining the fixed codeword, multiplied by its gain, and the adaptive codeword. It has been shown [5] that a good approximation for the transfer function of the n^{th} subframe is given by:

$$H_n(z) = \frac{g_{fc}(n)}{\left(1 - g_p(n) z^{-T_p(n)}\right)\left(1 - \sum_{i=1}^{10} a_i(n) z^{-i}\right)}, \quad (1)$$

where $g_{fc}(n)$ is the fixed codebook gain, $g_p(n)$ and $T_p(n)$ are the parameters of the pitch excitation and $\{a_i(n)\}$ are the linear prediction coefficients. Each n^{th} subframes belonging to a continuous speech signal $x(t)$ is therefore represented by a 13-elements vector:

$$\underline{x}_{PAR}(n) = [g_{fc}(n), g_p(n), T_p(n), lsf_1(n), \ldots, lsf_{10}(n)], \quad (2)$$

where $\{lsf_i(n)\}$ are the line spectral frequency, transformation of the linear prediction coefficents $\{a_i(n)\}$.

3. ACOUSTIC ECHO IN MOBILE NETWORKS

The physical phenomenon of the echo that we will analyze regards the mobile terminal. The coupling between microphone and loudspeaker usually takes place due to the reduced dimensions of the terminal, usually lower than 7-8 *cm*. The signal gets out the loudspeaker and gets in in the microphone in two ways: through air propagation or through the non-rigid behavior of the terminal chassis. In the frequency interval considered, we can approximate the total impulse response of the echo as:

$$h(t) \simeq \alpha \cdot \delta(t - \tau_e - \tau_0), \tag{3}$$

where τ_e and τ_0 represent respectively the delays introduced by the echo path and the network, while α is the attenuation and it is proportional to the Echo Return Loss parameter (ERL): $\alpha = 10^{-ERL/20}$. To simplify our notation, we have assumed that the echo signal is just an attenuated and delayed version of the signal itself, as in fact should be except for some small non-linearities introduced by the decoding-encoding procedure. Furthermore, we can assume that $\tau_e = 0$, considering the really short path of the echo signal, and assume that the network delay τ_0 is the only one present. The model of system is shown in figure 1. A highly reliable Voice Activity Detector (VAD) that works directly on the AMR parameters has also been implemented and included in our algorithm and it was subject of a deeper analysis [6].

It should be noted that the same echo cancellers are implemented on both the near-end and far-end.

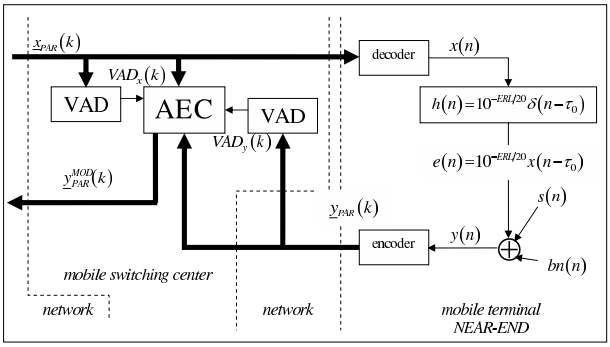

Fig. 1. *Model of the system in which the AEC is evaluated. The signal coming into the microphone at the near-end $y(n)$ is composed of echo $e(n)$, near-end signal $s(n)$ and background noise $bn(n)$. $\underline{x}_{PAR}(k)$, $\underline{y}_{PAR}(k)$ and $\underline{y}^{MOD}_{PAR}(k)$ represent respectively the far-end, near-end and modified near-end AMR parameters vectors for the k^{th} subframe.*

4. PRELIMINARY ACTIVITIES TO AEC

4.1. Echo Detector

In the wireless communication systems environment, voice packets can have delays that go approximately from 30 ms to 250 ms (6 to 50 AMR subframes), consequently, the echo detection assumes an important role. The implementation of an echo detector is here divided in two steps, a first step where an initial estimation of the delay is found and a second step where the the initial estimation is adaptively refined and updated considering that delays can change during the course of the conversation. In our framework the delays calculated have a temporal interval of $\delta\tau_0 = 5ms$. We have seen from the experimental analysis of the algorithm that the estimation will tend to stabilize itself on the closest value to the real delay, therefore not creating real problems to our estimation.

The first estimation is done by operating a correlative measure on two segments of coded speech, one coming from the near-end and one coming from the far-end, supposedly belonging to the same speaker. The VAD flag informs us about where the speech is present. For example, if the far-end signal segment is included in a subframe interval of $[m, m + L]$, the measurement will be done on the interval $[m + 6, m + L + 50]$ of the near-end signal, where we chose $L \geq 100$. If $L < 100$ we will wait for the next segment, as the correlative measure can be biased by having speech segment that is too small. Considering $\underline{x}(n)$ and $\underline{y}(n)$ the parameters vectors of the n^{th} subframe, respectively of the far-end and near-end, and $x_i(n)$ the i^{th} element of the parameter vector in (2), we calculate the find the mean values and the cross-covariance:

$$r_{\underline{x},\underline{y}}(\tau) = \frac{1}{13}\sum_{i=1}^{13}\frac{E\left[(x_i(n+\tau)-\mu_{x_i})(y_i(n)-\mu_{y_i})\right]}{\sqrt{E\left[(x_i(n+\tau)-\mu_{x_i})^2\right]E\left[(y_i(n)-\mu_{y_i})^2\right]}}, \tag{4}$$

calculated for each possible delay $\tau = 6, \ldots, 50$. The value of τ that maximizes the argument in (4) will be our first estimation of the delay.

In order to find a good estimation of the initial delay, we made an *ad-hoc* rule, only if:

$$\max r_{\underline{x},\underline{y}}(\tau) > 0.6 \quad \text{or} \quad \frac{\max r_{\underline{x},\underline{y}}(\tau)}{E[r_{\underline{x},\underline{y}}(\tau)]} > 2.5 \tag{5}$$

we define our estimation of the value $\hat{\tau}_0 = \max_\tau r_{\underline{x},\underline{y}}(\tau)$, otherwise we will move to the following segment distant 50 subframes. This rule basically asks the correlative measure in (4) to either have a maximum value that well defines the presence of correlation between the two segments or to have a well-defined peak. With this rule, we highly reduce the probability of having a bad estimation. Furthermore, the standard deviation of the estimator, that here represents the possible range in which the decision takes place, will start being unacceptable only for $ERL > 30dB$ and $SNR < 15dB$ values for which the presence of echo becomes perceptually unimportant and therefore having the AEC not working does not reduce the psycho-acoustic level of the conversation.

The second step of the algorithm for echo detection is an iterative method to update the estimation of the delay. In this part, the presence of a good VAD is also important; only if, at the n^{th} subframe, the VAD flag of the near-end ($VAD_y(n) = 1$) and the far-end aligned with the initial estimation ($VAD_x(n + \hat{\tau}_0) = 1$), we will perform the cross-correlation measure:

$$cc_{\underline{x},\underline{y}}(n,\tau) = \frac{1}{13}\sum_{i=1}^{13}\frac{E\left[x_i(m)y_i(m+\tau)\right]}{\sqrt{E[x_i^2(m)]E[y_i^2(n)]}}, \tag{6}$$

with $\tau = -20, \ldots, 20$. The segments taken to calculate (6) are $[n - \hat{\tau}_0 - 25, n + \hat{\tau}_0 + 25]$ for the far-end signal and $[n - 25, n + 25]$ for the near-end signal. We now take the two important values of (6):

$$\begin{aligned} cc(n) &= \max cc_{\underline{x},\underline{y}}(n,\tau), \\ \delta\hat{\tau}_0 &= \arg\max_\tau cc_{\underline{x},\underline{y}}(n,\tau). \end{aligned} \tag{7}$$

The update of the delay by $\delta\hat{\tau}_0$ will be only done if $cc(n) > 0.85$, considering this a good value for the cross-correlation between the two analyzed signals. The value $cc(n)$ can be seen as a echo likelihood value and we will use it to perform the echo cancellation.

4.2. Double Talk Detection

The importance of including a Double Talk Detector (DTD) to perform AEC has already been widely demonstrated in particular to prevent the Least Mean Square (LMS) algorithm from diverging and to avoid cancellation of speech information. The best choice is to *freeze* the operations and not perform any manipulation on the signal.

DTDs are usually based on correlative measures between the near-end and far-end signals [7]. In the previous section we have already found the echo-likelihood parameter $cc(n)$ as correlative measure (6): studying the statistical behavior of this feature, we will be able to perform a reliable double talk detection. The values of the echo-likelihood parameter $cc(n)$ have been shown to have a well defined Gaussian behavior in both presence or absence of double talk. Two estimated Gaussian probability density functions have been found analyzing different working condition ($ERL = 10 \div 30$ dB, $SNR_y = 10 \div 30$ dB and $SNR_y = 10 \div 30$ dB). Considering that the presence of double talk is somehow rare and usually estimated around 5%, weighting the two pdfs by $P(DTD) = 0.05$ and $P(\overline{DTD}) = 0.95$, we were able to define an optimal fixed threshold $cc_{DTD} = 0.42$. The total error probability, found averaging false-alarm and miss probabilities and calculated over the near-end to far-end ratio (NFR), was around $3 \div 7\%$ showing the algorithm to be reliable. In different conditions from the one analyzed, the echo-likelihood is usually low and it will prevent the cancellation algorithms from operating on the signal.

5. ECHO CANCELLATION ALGORITHMS

In this section we will shows how our algorithm processes the AMR parameters in order to perform the echo cancellation. The basic principle is to partially decode the voice packets in order to define each segment of speech by its parameter vector (2) and therefore by the simplified transfer function (1). The conditions for the cancellation algorithms to be operative are that the voice activity detectors on the aligned temporal axis are both high $VAD_x(n + \hat{\tau}_0) = 1$ and $VAD_y(n) = 1$ and only the echo is present $cc(n) > cc_{DTD}$.

5.1. Fixed Codebook Gain and Adaptive Codebook Gain Modifications

The $g_{fc}(n)$ can be considered as a multiplicative factor applied to the n^{th} subframe transfer function. This parameter controls the overall level of the synthesized signal at the speech decoder, thus its attenuation will result in the reduction of the echo. The degree of attenuation has been made proportional to the likelihood of the echo, represented by $cc(n)$ by setting the step-size of the Normalized Least Mean Square (NMLS) algorithm $\mu = 1.5 \cdot cc(n)$, that also prevents the algorithm from diverging. The idea behind this, starts from this assumption:

$$\begin{aligned} g_y(n) &= f(g_e(n), g_v(n), g_{bn}(n)) \\ &\simeq g_e(n) + g_v(n) + g_{bn}(n), \end{aligned} \tag{8}$$

where g_e represents the fixed codebook gain of the echo signal, g_{bn} the one of the background noise and g_v the gain of the signal at the near-end in the case of double talk. We'll also assume that:

$$g_e(n) = \sum_{l=0}^{L-1} g_x(n)h(l) = \mathbf{h}^T g_x(n), \tag{9}$$

where $\mathbf{h}$ is the filter that it's being adapted at time $n + 1$ with the following NLMS procedure:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) + 1.5 \cdot cc(n) \frac{g_y(n) - \hat{g}_y(n)}{\mathbf{g}_x^T(n)\mathbf{g}_x(n)} \mathbf{g}_x(n). \tag{10}$$

Thus, the signal g_u coming out of the canceller will be:

$$g_u(n) = g_y(n) - \hat{g}_y(n) = g_y(n) - \hat{\mathbf{h}}(n)\mathbf{g}_x(n). \tag{11}$$

In theory, having the echo detector performing the time alignment and the having a flat frequency response of the echo path, a one tap adaptive filter should be enough to perform the cancellation. However, due to the variance of the gain, having only one tap can bring to local errors in the estimation and also, in the way the signal has been coded, exists a dependency between adjacent samples of $g_{fc}(n)$. In our empirical studies, a five taps adaptive filter $\hat{\mathbf{h}}$ has been shown to offer the best trade-off between cancellation and ability to quickly adapt; in fact, in this case, the system distance has shown the lowest variance.

The adaptive codebook gain $g_p(n)$ does not influence the energy level of the signal as much as the fixed codebook gain $g_{fc}(n)$, but it maintains itself high in the conditions of speech and therefore in the presence of echo as well. In this case, we used the same NLMS algorithm applied on the codebook gain (10).

5.2. Pitch Period and Line Spectral Frequencies Modifications

In the context of our work, a perfect correspondence between the system and the adaptive filter that performs the echo cancellation is clearly not possible. The algorithms to attenuate the two codebooks gains greatly reduce the energy of the echo signal; nevertheless the echo is not completely reduced due to the spectral characterization that has basically remained untouched with the two gains modifications. Therefore, the solution is to modify the parameters responsible for this, the pitch period and the line spectral frequencies. The values of the pitch period parameter have shown to have a really small variance around the true value of the pitch period in the presence of voiced speech and also in the presence of unvoiced speech (due to the dependence of the calculations of one sample of T_p from the adjacent ones). On the other hand, when only noise is present then T_p behaves in a total random way exhibiting a uniform probability distribution over the possible values that it can assume; this is due to the fact that a pitch period does not exists but it is still being calculated. Our algorithm eliminates the long-term information by randomizing the value of T_p when echo is present. In other words, the output will be:

$$\begin{aligned} T_{p,u} &= T_r, \\ \Omega_{T_r} &= \{17, 18, \ldots, 142, 143\}, \end{aligned} \tag{12}$$

where T_r is taken from a uniform proability space Ω_{T_r}.

The modification of the line spectral frequencies is the most important part to eliminate the spectral characterization of the speech frame. As already shown [8], the line spectral frequencies tend to cluster around the formants if the speech is vocalized while for unvoiced speech and noise they are more uniformly distributed along their domain $[0, \pi]$. The extreme case in which the analyzed signal

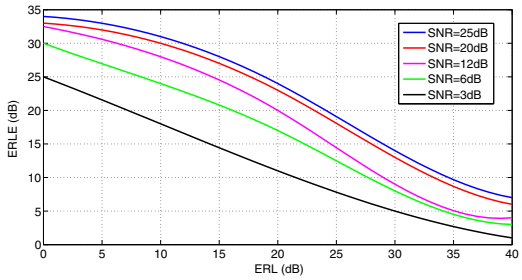

Fig. 2. *Performances of the AEC algorithm in terms of $ERLE$, calculated for different values of ERL and SNR*

has a flat spectrum, the i^{th} lsf will be equal to $i \cdot (\pi/(p+1))$, where $p = 10$ is the prediction order. The hypothesis, amply demonstrated in [8], can allow us to perform the whitening of the spectrum just by changing the value of the i^{th} lsf with its value in the flat spectrum case. In order to make the algorithm work more smoothly we decided to adapt the substitution of the values with the information coming from the echo-likelihood parameter $cc(n)$, the replacement of the i^{th} lsf will be:

$$\hat{lsf}_{i,u}(n) = cc(n)\frac{i\pi}{11} + (1 - cc(n))lsf_i(n) \qquad (13)$$

This is really similar to the bandwidth expansion process used in LPC coding with $\gamma_i = (1 - cc(n))^i$. In fact, if $cc(n) = 1$ the transfer function would become an all-pass filter while if $cc(n) = 0$ it would not be modified. In our algorithm though, the morphing will not take place at all if $cc(n) < cc_{DTD}$. An observation has to be made; in our algorithm we try to whiten the spectrum of the echo signal to transform it into white noise. However, in the context of mobile communication, the background noise is usually not white. Instead of averaging the i^{th} line spectral frequency with its white noise value, we will average it with a value, calculated adaptively when $VAD_y(n) = 0$.

It is important to notice that it is hard to calculate the objective improvement of this two latter techniques. Testing the algorithm in different conditions, the power of the echo was reduced by $2 \div 3dB$. The real improvement is subjective and mainly due to the modifications of the spectrum of the subframe.

6. RESULTS

The measures of the performances of the algorithm have been done in the uncoded domain confronting the value of the echo return loss enhancement $ERLE$ with the echo return loss ERL at the near-end. Various values of SNR have been analyzed averaging the performances obtained with different kinds of noise (car, street, wgn, babble, rain). The residual ERL will be equal to the sum of ERL and $ERLE$. The results are shown in figure 2. The results are therefore comparable to the mandatory specifications for AEC in the uncoded domain. The main problems of the AEC implemented happen as the ERL becomes too high or the SNR becomes too low, however in these cases the echo does not really act on the intelligibility of the conversation. It's important to notice that the $ERLE$ is mainly due to the modifications operated on the two gains; when the others operate modifications on a psycho-acoustic level, for which a deeper analysis has to be made (for example using the Mean Opinion Score). The main advantage of this technique is the simplicity in which is possible to modify the energy level (working on the gains) and the spectrum of the echo signal; in fact, all the information we need for a segment of signal is contained in the AMR parameter vector.

7. CONCLUSIONS

In this paper we have discussed several techniques to perform echo cancellation in the compressed domain. In particular we have shown that is possible to transpose these operations from time domain to parameters domain. These techniques are suitable for implementation in speech enhancement equipments in mobile networks and other kind of networks working with AMR-coded speech. Given the interesting results of all the algorithms tested on the UMTS network, we can see these as a good alternative to the existing AEC procedures.

8. REFERENCES

[1] J. Benesty, T. Gansler, D. R. Morgan, M. M. Sondhi, S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer-Verlag, Berlin, 2001.

[2] 3GPP, *TS 26.071; AMR Speech Codec: General Description*, Version 6.0.0, 2005.

[3] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, Wiley, 2003

[4] F. Itakura, Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals, *Journal of the Acoustic Society of America*, vol. 57, S35(A), 1975.

[5] H. Taddei,C. Beaugeant,M. de Meuleneire, Noise Reduction on Speech Codec Parameters, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004.

[6] D. Giacobello, *Study and Evaluation of Innovative Algorithms for Voice Quality Enhancement in Speech Signals Encoded Using ACELP (Algebraic Code Excited Linear Prediction)*, M.Sc. Thesis, Politecnico Di Milano, Milano, Italy, July 2006.

[7] J. Benesty, D. R. Morgan, J. H. Cho, A New Class of Doubletalk Detectors Based on Cross-Correlation, *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 8, pp. 168-172, March 2000.

[8] B. W. Kleijn, T. Backstrom, P. Alku, On Line Spectral Frequencies, *IEEE Signal Processing Letters*, vol. 10, no. 3, march 2003, pp. 75-77.