Acoustic Echo Cancellation on the Adaptive Multi-Rate Speech Codec Parameters

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

[1]Aalborg University, Aalborg, Denmark [2]Nokia Siemens Networks, Milano, Italy [3]Politecnico di Milano, Milano, Italy

1 Motivation

- Speech enhancement techniques in mobile networks are done in the transcoding unit operating on the uncoded signal: the signal coming from the mobile terminals has to be decoded, enhanced and encoded again. These operations introduce delays other than being computationally intensive and prone to adding further quantization noise.
- The purpose of this work is to reduce the computational cost and delay of this inherently suboptimal scheme by transfering the acoustic echo cancellation operations to the Adaptive Multi-Rate (AMR) coded domain operating directly on the codec parameters.

2 AMR Codec

Analysis made on 20ms frame (160 samples @8kHz) divided further on in 5ms subframes, for each subframe are extracted:

- 10 short-term coefficients $\{a_i\}$ then converted in LSF $\{L_i\}$;
- pitch delay T_p and gain g_p (long-term predictor);
- algebraic codeword and gain g_{fc}.

For the n^{th} subframe, the AR model in which the excitation is passed through:

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

each subframe can then be represented by a 13 elements vector:

$$\underline{x}(n) = [g_{fc}(n), g_p(n), T_p(n), L_1(n), \ldots, L_{10}(n)]$$

3 Preliminary Activities to AEC

3.1 Echo Detector

To detect echo presence, a correlative measure is done on two segments of coded speech, one coming from the near-end $\underline{x}$ and one coming from the far-end $\underline{y}$, supposedly belonging to the same speaker. The VAD flag informs us about where the speech is present.

- First estimate of network delay $\hat{\tau}$:

$$\arg\max_{\tau} \sum_{i=1}^{13} \frac{E\left[(x_i(n+\tau) - \mu_{x_i})(y_i(n) - \mu_{y_i})\right]}{\sqrt{E\left[(x_i(n+\tau) - \mu_{x_i})^2\right] E\left[(y_i(n) - \mu_{y_i})^2\right]}}$$

  calculated for each possible delay $\tau = 6, \ldots, 50$ on each i^{th} feature and averaged.

- Iterative updating of the network delay estimate $\hat{\tau}$

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

  with $\tau = -20, \ldots, 20$. Done if at the n^{th} subframe, $VAD_y(n) = 1$ and $VAD_x(n+\hat{\tau}_0) = 1$. The two important values obtained are:

$$cc(n) = \max cc_{\underline{x},\underline{y}}(n,\tau)$$

$$\delta\hat{\tau}_0 = \arg\max_{\tau} cc_{\underline{x},\underline{y}}(n,\tau)$$

  $\delta\hat{\tau}_0$ is used to update the delay, $cc(n)$ is considered as the *echo-likelihood* parameter. The update of the delay by $\delta\hat{\tau}_0$ will be only done if $cc(n) > 0.85$.

3.2 Double Talk Detection

- Two Gaussian pdfs for $cc(n)$ in the presence and absence of double talk are defined and then weighted by $P(DTD) = 0.05$ and $P(\overline{DTD}) = 0.95$. An optimal fixed threshold is then found $cc_{DTD} = 0.42$.
- Cancellation algorithms will only work for $cc(n) > cc_{DTD}$.

4 AEC Algorithms

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 $cc(n) > cc_{DTD}$.

4.1 g_{fc} and g_p modifications

- Use of Normalized Least Mean Square algorithm with step-size $1.5 \cdot cc(n)$ (for convergence)
- General assumption: $g_y(n) \simeq g_e(n) + g_v(n) + g_{bn}(n)$
- Considering

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

  $\mathbf{h}$ is being adapted at time $n+1$ with the following NLMS procedure:

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

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

$$g_u(n) = g_y(n) - \hat{g}_y(n) = g_y(n) - \hat{\mathbf{h}}^T(n)\mathbf{g}_x(n)$$

4.2 T_p and $\{L_i\}$ modifications

- Eliminate the long-term information by randomizing the value of T_p:

$$T_{p,u} = T_r$$

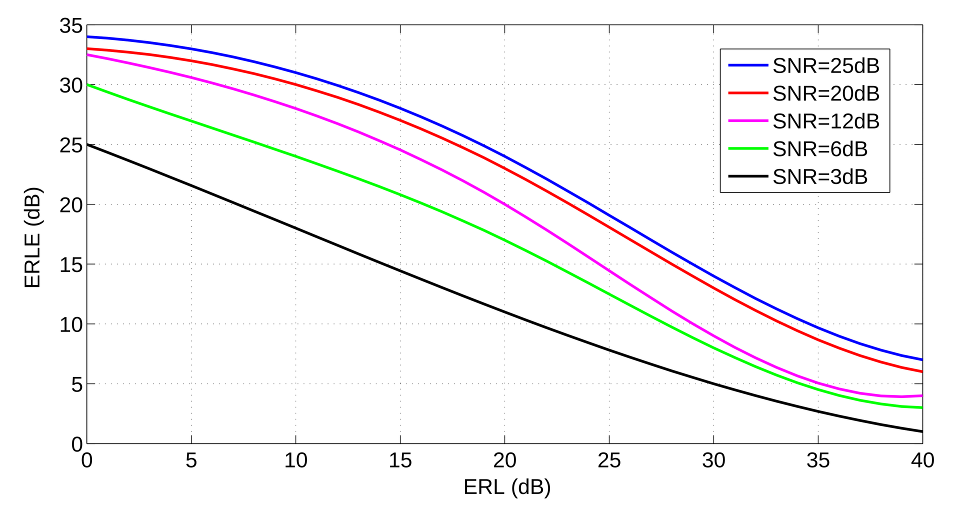
$$\Omega_{T_r} = \{17, 18, \ldots, 142, 143\}$$

- Whitening of the signal by morphing the LPC spectrum (done in the LSF domain):

$$\hat{L}_{i,u}(n) = cc(n)\frac{i\pi}{11} + (1 - cc(n))L_i(n)$$

5 Results

- The main problems of the AEC algorithm implemented happen as the ERL becomes too high or the SNR becomes too low, however in these cases the echo does not really affect the intelligibility of the conversation.
- $ERLE$ is mainly due to the modifications operating on the two gains. The other modifications act on a psycho-acoustic level.
- 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: all the informations we need for a segment of signal are contained in the AMR parameter vector.
- The performances are shown in the figure below.



Performance of the AEC algorithm in terms of $ERLE$, calculated for different values of ERL and SNR. Results are found by averaging different kinds of noise (car, street, wgn, babble, rain)

6 Conclusion

- It is possible to transpose AEC operations from time domain to parameters domain
- Suitable for implementation in speech enhancement equipments in voice networks and using AMR coded speech
- A good alternative to the existing AEC procedures

References

[1] H. Taddei *et al.*, "Noise Reduction on Speech Codec Parameters", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004.

[2] 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.