14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

MEASUREMENT OF THE EFFECTS OF NONLINEARITIES ON THE NETWORK-BASED LINEAR ACOUSTIC ECHO CANCELLATION

*Ted S. Wada**, *Biing-Hwang Juang, and Rafid A. Sukkar*†

Center for Image and Signal Processing, Georgia Institute of Technology, Atlanta, GA 30332
† Tellabs, Inc., 1415 W. Diehl Road, Naperville, IL 60563
e-mail: (twada, juang)@ece.gatech.edu, rafid.sukkar@tellabs.com

ABSTRACT

It is well known that an over-driven loudspeaker would produce a nonlinearity that limits the performance of an acoustic echo canceler (AEC). In contrast, only a handful of studies have been documented on the effect of speech coding nonlinearity on the AEC. This paper investigates the combined effect of both types of nonlinearities in the network-based AEC framework as opposed to when the AEC is performed at the source of echo such as a cellular handset. The simulation results show that while a mild saturation-type loudspeaker nonlinearity causes the echo return loss enhancement (ERLE) to go down significantly, it is the nonlinear speech coding distortion on the acoustic echo signal that ultimately reduces the achievable ERLE. The results also point to the fact that a low bit-rate speech codec is capable of synthesizing a perceptually acceptable speech signal but does it in a way that is untractable by traditional linear AEC algorithms.

1. INTRODUCTION

In order to meet the customer satisfaction, cellular handset manufactures today integrate many popular features, such as multimedia playback or global navigation system capabilities, into their products. As a consequence, they mostly overlook a minor yet important and computationally intensive task like the AEC in order to minimize the handset's power consumption and manufacturing cost. Furthermore, echo problems in telecommunication are historically considered a network issue, and it is often not easily justified for handset manufacturers to take upon themselves the issue for no obvious benefit to the user of their equipment.

Therefore, the responsibility of AEC implementation is often relegated to the network providers. The AEC must then be performed at a central location somewhere in the network, as illustrated in Figure 1. The encoding and the decoding at the base station are not required in a tandem-free operation (TFO) cellular network, but such network must perform the AEC in coded domain and is not considered in this paper; otherwise, they are required when transcoding takes place in the network, which occurs most of the time, or when the remote call is made through a landline.

There are two types of nonlinearities that potentially limit the AEC performance in the network like the one in Figure 1. One is the nonlinear acoustic coupling between the loudspeaker and the microphone of a handset. The other is the nonlinear speech coding distortion applied to both the far-end and the near-end signals. There are numerous papers published already on the topic of nonlinear AEC, but few of them, if any, address the two types of nonlinearities together in a single network-based AEC framework.

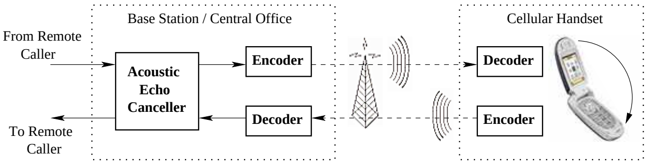

Figure 1: Implementation of AEC in the network.

For instance, a mean reduction of 12 dB in the ERLE due to the saturation-type loudspeaker nonlinearity on an Alcatel handset is reported in [1]. In such case, a cascading of the polynomial Volterra filter with a linear adaptive filter [1], a partial adaptive structure with the time-delay neural network (TDNN) [2], or the adaptive orthogonalized power filter [3] can be used to achieve roughly 5 dB increase in the ERLE, but none of these nonlinear adaptive filtering methods take into account the speech coding nonlinearity. On the other hand, over 50 dB reduction in the ERLE is attributed to the nonlinear speech coding distortion alone when the AEC is performed in a simulated cellular network [4], yet no solution to the problem has been published to date due to the difficulty in characterizing the speech coding nonlinearity. In addition, the characteristics of coded speech have been used to improve the network-based AEC performance with the use of a post-filter based on the statistical information from a speech encoder [5]. However, [5] does not consider either the acoustic coupling nonlinearity or the speech coding nonlinearity, both of which can inhibit a linear adaptive filter from reaching the optimum solution.

In this paper, we will examine the AEC in the network through simulations to see how the two types of nonlinearities together affect the linear AEC performance. We will also quantify the degree of nonlinear distortion caused by several speech codecs and by specific components within a codec in order to further characterize the effect of speech coding nonlinearity on linear AEC. The overall goal is to numerically assess the effects of nonlinearities to gain a better understanding of the problem so that much more effective network-based AEC scheme, whether it be linear or nonlinear itself, may be developed in the future.

The paper is organized as follows. First, we discuss in Section 2 the possible sources of nonlinearities in acoustic coupling and speech coding. Next, we present in Section 3 the simulation method for assessing the nonlinearities, followed by the results and analyses in Section 4. Finally, we end with the conclusions in Section 5.

* Partial work performed during Summer 2005 internship at Tellabs.

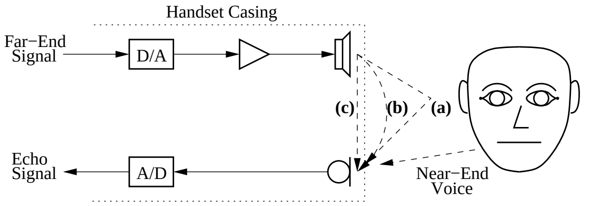

Figure 2: Three distinct acoustic echo paths.

2. SOURCES OF NONLINEARITIES

2.1 Acoustic Coupling

Figure 2 shows three distinct acoustic echo paths between the loudspeaker and the microphone of a cellular handset: (a) a reverberation in a room during the speaker-phone mode or a reflection off of near-end speaker's head during the handset mode, (b) a direct coupling through the air, or (c) a mechanical coupling through the handset itself. While (a) and (b) can be modeled together with a single impulse response, (c) most likely does not behave in a linear fashion and cannot be characterized simply by an impulse response. The loudspeaker with a saturation characteristic is also a major source of nonlinearity, which is usually modeled with a memoryless polynomial function. Another possible source of nonlinearity is the microphone that can suffer from both the over-driving and the saturation problems.

Both the mechanical coupling and the loudspeaker saturation will have a significant effect on the AEC when the far-end signal is played back at a high volume on a handset with an inexpensive loudspeaker or with a casing that is not properly designed to reduce the mechanical coupling. However, it is difficult to simulate a mechanical coupling without working with a set of differential equations, thus such analysis is out of the scope of this paper. We will also ignore the nonlinearity due to over-driving or saturation at the microphone since the far-end signal played out at the loudspeaker of a handset is most likely at a level much lower than the near-end speaker's voice.

2.2 Speech Coding

Most of the current speech codecs used in wireless communications are based on the linear prediction coding analysis-by-synthesis (LPC-ABS) approach. Code-excited LPC (CELP) is one type of LPC-ABS codec. The CELP encoding and decoding processes are represented by the schematics in Figure 3. Basically, the CELP encoder searches iteratively for the best encoding parameters (i.e. the codeword $c_i(n)$, the codeword gain g_c, the pitch gain g_p, and the pitch delay T) by perceputally weighting the error $e(n)$ between the original speech $s(n)$ and the decoded speech $\tilde{s}(n)$ and minimizing the energy of the weighted error $y(n)$. More detailed information on LPC-ABS and CELP can be found in [6].

As illustrated in Figure 3, the CELP speech coding process is altogether nonlinear and sub-optimal. One possible source of nonlinearity is the adaptive post-filter, which improves the perceptual quality of the synthesized speech $\hat{s}(n)$ but does not necessarily give a better match to the original speech. By the same token, the perceptual weighting filter may also affect the AEC performance since it is an integral part of the encoding parameter search process. Another source of nonlinearity is the quantization of encoding parameters, for which the quantization noise is not additive anymore and cannot be treated as the output of some random process. Other nonlinearity factors include how the channel coding bits are distributed among the quantized encoding parameters and how the individual components within the encoder, such as the fixed and the adaptive codebooks, are implemented in different LPC-ABS codecs.

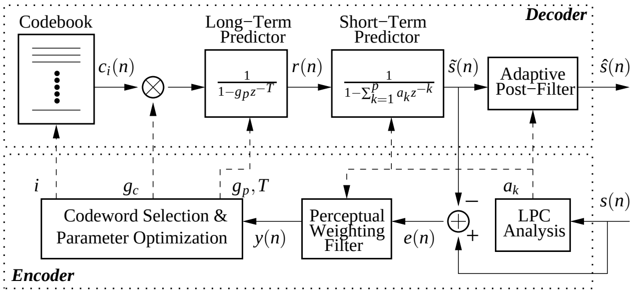

Figure 3: Encoding and decoding schemes of CELP.

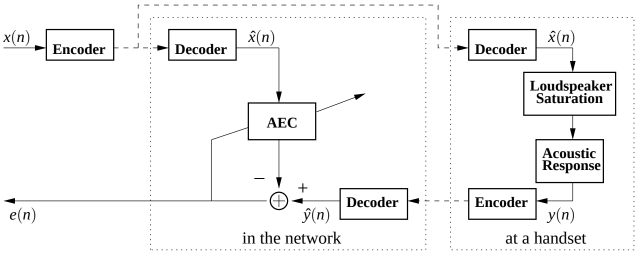

Figure 4: Network-based AEC configuration with loudspeaker saturation and speech coding.

3. SIMULATION METHOD

Figure 4 shows the network-based linear AEC configuration that is implemented during simulations. We assume here that $x(n)$ (the far-end signal) is also encoded, which would be the case for a cellular-to-cellular call, and that the decoder in a handset and the decoder in the network are identical. For simplification purposes, we also assume that the wireless channel (i.e. the path between encoder and decoder) is ideal and does not impose any communication delay or packet loss, that there is no delay due to codec processing time, and that there is no double-talk situation.

The goal is to measure the ERLE through simulations for various signal types and conditions, listed below:

- Male speech, female speech, or white noise.
- Voiced or unvoiced speech.
- With or without speech coding.
- With or without loudspeaker saturation.
- With or without post-filtering.

Only the post-filtering in the decoder in the network that operates on encoded $y(n)$ (the acoustic echo signal) is removable since the decoder in the network that operates on encoded $x(n)$ simulates the decoding in a handset, which is out of the network's control.

The ERLE is defined as

$$\mathrm{ERLE} = 10\log_{10}\frac{\sum_{n=1}^{N}\hat{y}^2(n)}{\sum_{n=1}^{N}e^2(n)} \quad (\mathrm{dB}), \tag{1}$$

Codec Name	Bit-Rate (kbps)	Codec Type	Frame Size (ms)
G.723.1	5.3 6.3	ACELP MP-MLQ	30
G.728	16	LD-CELP	2.5
G.729	8	CS-ACELP	10
GSM AMR	4.75, 5.15, 5.9, 6.7 7.4, 7.95, 10.2, 12.2	ACELP	20

Table 1: LPC-ABS speech codecs and their features.

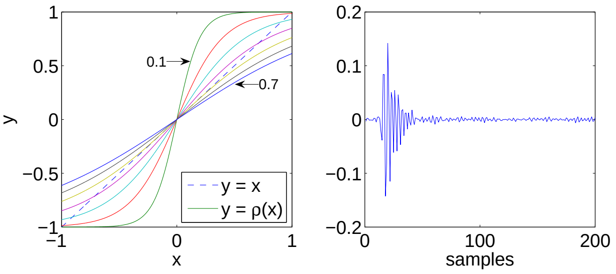

Figure 5: Loudspeaker saturation function $\rho(x)$ for α = 0.1, 0.2, …, 0.7 (left) and the acoustic impulse response (right) used during simulations.

where $\hat{y}(n)$ is the acoustic echo signal with or without speech coding distortion, $e(n)$ is the residual echo signal, and N is the sample size. For calculating the average ERLE, only the last one-half of the data are used in order to ensure sufficient convergence.

16-bit waveforms sampled at 8 kHz are used as the far-end signal. Male and female speeches are obtained from the TIMIT database [7], which comes with phoneme labels that can be used for voiced/unvoiced speech classification. A white noise is generated from the Gaussian distribution with zero mean and unit variance. Each signal is analyzed for 30 seconds and is scaled to 99% of the maximum volume range in order to observe the loudspeaker saturation effect.

The speech codecs tested during simulations are ITU G.723.1 [8], ITU G.728 [9], ITU G.729 [10], and GSM AMR [11]. They are all based on LPC-ABS and are widely used in satellite, wireless, or voice-over-IP (VoIP) communications. The corresponding bit-rates, codec types, and frame sizes are listed in Table 1.

The acoustic coupling is modeled by a loudspeaker saturation followed by a linear acoustic impulse response. The loudspeaker saturation is implemented with a soft-decision function

$$\rho(x) = \frac{1-\exp(-x/\alpha)}{1+\exp(-x/\alpha)}, \quad -1 \le x \le 1, \qquad (2)$$

where the parameter α determines the degree of saturation. The linear acoustic response is determined from an actual acoustic echo recorded from a cellular handset. The impulse response is truncated to 200 samples long so that the maximum peak occurs at the 20$^{\text{th}}$ sample, and it is scaled such that it produces roughly 10 dB echo return loss (ERL). The plot of $\rho(x)$ for several values of α and the plot of the impulse response used during simulations are shown in Figure 5.

The linear AEC is implemented by using the normalized least mean square (NLMS), the fast (or frequency) block LMS (FBLMS), and the resursive least square (RLS) algorithms [12]. The adaptation step sizes are set to 0.5 and 0.017 for NLMS and FBLMS, respectively, and the forgetting factor of 1 is used with RLS. The block size of 200 samples (same as the impulse response length) is used with FBLMS.

4. SIMULATION RESULTS

The ERLE plots in Figure 6 were obtained from implementing NLMS, FBLMS, and RLS on a female speech in the newtork-based AEC setting with only loudspeaker saturation (α = 0.4) and no speech coding. They show that NLMS can provide better result at low volume level than the other algorithms, which illustrates the robustness of NLMS against the saturation-type loudspeaker nonlinearity. On the other hand, the ERLE plots in Figure 7 were obtained when there is only speech coding (GSM AMR 12.2 kbps) and no loudspeaker saturation, and they show that the ERLE is consistently well below 20 dB for all three AEC algorithms tested. It means that nonlinear adaptive filters based on LMS-type structure, such as the Volterra filter and the power filter, would also be insufficient to handle the speech coding nonlinearity.

The combined effect of the loudspeaker and speech coding nonlinearities on the network-based AEC (FBLMS) can be seen in Figure 8, in which the average ERLE obtained from a female speech is plotted as a function of the parameter α and the GSM AMR bit-rate. The plot shows that unless α is very small (i.e. when there is a severe loudspeaker saturation), most of the AEC performance degradation encountered in a real-life situation can likely be attributed to the nonlinear speech coding distortion.

Table 2 provides the average ERLE obtained from male speech, female speech, and white noise (WN) signals when the AEC (FBLMS) was performed in a handset (i.e. when the encoder-decoder pair in the echo path was removed) with post-filtering (PF) and with or without loudspeaker saturation (LSS) (α = 0.4 for LSS). The table shows that a speech coding on the far-end signal does not practically affect the handset-based AEC performance, where the minimum average ERLE for the male speech with no loudspeaker saturation is 40.5 dB, which is only 11.5 dB down from the baseline of 52.0 dB. In fact, there is a slight increase in the average ERLE for most of the loudspeaker saturation cases, more so for a white noise than for the other signal types. This is likely due to the increased energy of the far-end signal in mid-volume range after going through the soft-decision function, as the distribution of a white noise is more heavily tailed than that of a speech signal.

Tables 3, 4, and 5 provide the average ERLE obtained from male speech, female speech, and white noise (WN) signals, respectively, when the AEC (FBLMS) was performed in the network with different combinations of PF and LSS (α = 0.4). The three tables show that the highest average ERLE when there is a speech coding with post-filtering is around 16 dB, with or without the loudspeaker nonlinearity.

Tables 3, 4, and 5 also show that although the post-filtering effect seems small in general, there is still an overall decrease in the average ERLE for male and female speeches when there is post-filtering, where the reduction can be as much as 3 dB for some cases (e.g. female speech coded with G.728 with no loudspeaker saturation). This suggests the need to take a closer look at the perceptual weighting filter as well to see how it affects the AEC performance.

In addition, Tables 3, 4, and 5 show that the average

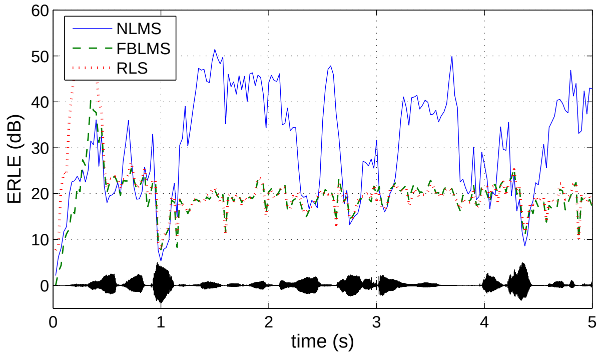

Figure 6: ERLE from the network-based AEC for a female speech when there is only loudspeaker saturation (α = 0.4) and no speech coding.

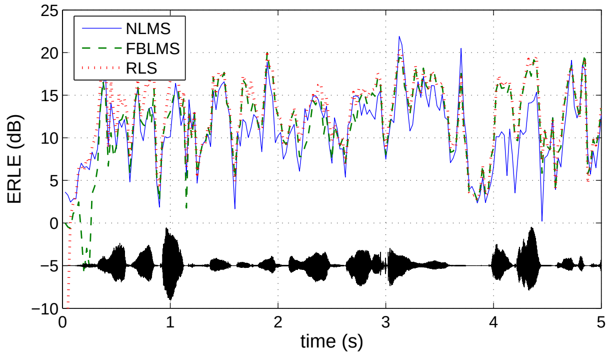

Figure 7: ERLE from the network-based AEC for a female speech when there is only speech coding (GSM AMR 12.2 kbps) and no loudspeaker saturation.

ERLE is reduced the most for a white noise than for the other types of signal in the network-based AEC, whereas the slightly higher average ERLE for a female speech than for a male speech is consistent with the general results obtained when a stochastic gradient-type algorithm is used. This is probably because the codebooks used by low bit-rate codecs are adequate enough to model the excitation signals that drive the voiced speech production but tend to be too sparse to describe many possible random noises.

Furthermore, Table 6 provides the average ERLE calculated separately for voiced and unvoiced portions of male and female speeches (62% and 60% of the male and female speeches, respectively, were voiced), and it can be observed from the table that the average ERLE for an unvoiced speech is just as low as what is obtained from a white noise. The problem can potentially be made worse by the artificial silence insertion algorithm used by some speech codecs (e.g. GSM AMR's discontinuous transmission (DTX) option, which was not implemented during simulations), and it may necessitate the background/foreground filtering in order to avoid the adaptive filter divergence during unvoiced or silent portions of a speech.

Finally, Tables 3 through 6 collectively show that there is a strong correlation between the ERLE and the bit-rate. Also, there are several sizeable jumps in the average ERLE for GSM AMR at serveral bit-rates (e.g. from 6.7 to 7.4 kbps and from 7.95 to 10.2 kbps). These results stress the need to investigate how the the encoding parameter quantization and the channel coding bit allocation are implemented in a specific speech codec so that the information may be used to improve the AEC algorithm.

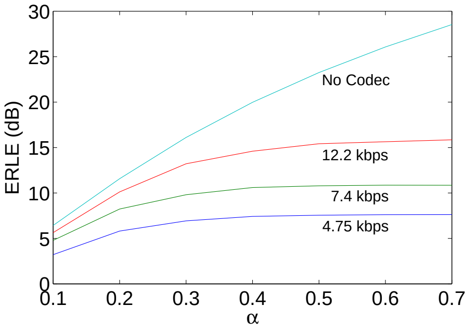

Figure 8: Average ERLE from the network-based AEC (FBLMS) for a female speech as a function of the loudspeaker saturation parameter α and the GSM AMR bit-rate.

Codec Name	Bit-Rate (kbps)	LSS			No LSS		
		Male	Female	WN	Male	Female	WN
G.723.1	5.3	16.73	18.80	27.36	53.51	65.40	72.56
	6.3	17.14	19.02	27.19	60.83	66.28	73.58
G.728	16	17.91	20.03	27.08	59.68	66.09	76.93
G.729	8	18.61	19.91	33.07	53.81	59.81	71.63
GSM AMR	4.75	17.60	18.59	33.47	48.65	59.46	61.64
	5.15	18.16	19.34	33.72	55.11	50.27	61.06
	5.9	18.79	20.27	32.74	64.01	62.25	67.36
	6.7	18.85	20.09	33.28	57.31	63.98	64.01
	7.4	18.69	20.21	30.83	48.53	64.43	71.06
	7.95	18.79	20.34	28.06	51.91	63.48	64.20
	10.2	18.68	20.20	30.45	49.53	63.63	65.34
	12.2	18.71	20.14	29.60	40.50	64.39	51.06
No Speech Coding		**17.89**	**19.97**	**26.60**	**51.99**	**61.08**	**77.04**

Table 2: Average ERLE (dB) from the handset-based AEC (FBLMS) with PF (α = 0.4 for LSS).

5. CONCLUSIONS

The results from the network-based AEC simulations show that unless there is a severe loudspeaker saturation, it is mostly the speech coding nonlinearity applied to the acoustic echo signal and not the loudspeaker nonlinearity that limits the linear AEC performance. While NLMS exhibits robustness against the saturation-type nonlinearity at low volume level, none of the linear AEC algorithms tested (NLMS, FBLMS, RLS) were able to adequately handle the speech coding nonlinearity. The ERLE is strongly correlated with the speech codec bit-rate, where the highest average ERLE obtained by using popular speech codecs such as GSM AMR and G.728 was around 16 dB regardless of the degree of loudspeaker saturation. Also, the sparsity of codebook in speech codecs based on LPC-ABS is likely a major factor in the reduction of the ERLE for unvoiced and silent portions of a speech signal. The post-filter removal makes a measurable difference for some speech codecs, and there are still many parts within a speech codec, such as the perceptual weighting filter, the encoding parameter quantization, and the channel coding bit allocation, that must be analyzed carefully be-

Codec Name	Bit-Rate (kbps)	LSS		No LSS	
		PF	No PF	PF	No PF
G.723.1	5.3	8.64	8.81	9.38	9.66
	6.3	10.05	10.34	10.66	11.02
G.728	16	13.83	15.05	15.79	18.00
G.729	8	10.41	11.35	11.02	12.09
GSM AMR	4.75	7.29	7.61	7.55	7.81
	5.15	7.16	7.43	7.80	8.05
	5.9	8.44	8.77	8.78	9.04
	6.7	9.19	9.46	9.63	9.98
	7.4	10.15	10.77	11.12	11.74
	7.95	10.07	10.54	10.74	11.35
	10.2	13.38	13.69	14.74	14.98
	12.2	14.18	14.58	15.86	16.25
No Speech Coding		**17.89**		**51.99**	

Table 3: Average ERLE (dB) from the network-based AEC (FBLMS) for a male speech (α = 0.4 for LSS).

Codec Name	Bit-Rate (kbps)	LSS		No LSS	
		PF	No PF	PF	No PF
G.723.1	5.3	8.83	9.24	9.54	9.93
	6.3	10.21	10.41	10.59	11.06
G.728	16	14.61	16.38	15.99	19.01
G.729	8	10.67	11.56	11.17	12.23
GSM AMR	4.75	7.42	7.57	7.57	7.77
	5.15	7.79	8.01	8.15	8.36
	5.9	8.50	8.80	8.77	9.15
	6.7	9.38	9.82	9.79	10.28
	7.4	10.60	11.19	11.08	11.77
	7.95	10.44	10.90	10.94	11.54
	10.2	13.72	13.99	14.60	14.92
	12.2	14.59	14.90	16.26	16.67
No Speech Coding		**19.97**		**61.08**	

Table 4: Average ERLE (dB) from the network-based AEC (FBLMS) for a female speech (α = 0.4 for LSS).

Codec Name	Bit-Rate (kbps)	LSS		No LSS	
		PF	No PF	PF	No PF
G.723.1	5.3	4.46	4.48	4.50	4.51
	6.3	5.42	5.42	5.43	5.43
G.728	16	15.84	15.68	16.20	16.02
G.729	8	6.82	6.83	6.88	6.86
GSM AMR	4.75	1.56	1.57	1.54	1.56
	5.15	1.61	1.61	1.56	1.56
	5.9	2.17	2.19	2.13	2.15
	6.7	2.14	2.16	2.17	2.20
	7.4	7.01	6.89	7.09	6.95
	7.95	4.60	4.52	4.62	4.54
	10.2	9.42	9.33	9.45	9.36
	12.2	11.05	10.93	11.09	10.97
No Speech Coding		**26.60**		**77.04**	

Table 5: Average ERLE (dB) from the network-based AEC (FBLMS) for a white noise (α = 0.4 for LSS).

Codec Name	Bit-Rate (kbps)	Male		Female	
		Voiced	Unvoiced	Voiced	Unvoiced
G.723.1	5.3	9.59	6.17	9.61	5.75
	6.3	10.96	7.00	10.65	6.80
G.728	16	15.77	15.95	16.03	14.60
G.729	8	11.45	5.93	11.26	6.09
GSM AMR	4.75	7.88	0.89	7.64	1.82
	5.15	8.34	0.76	8.23	2.56
	5.9	9.37	2.49	8.91	2.21
	6.7	10.32	3.67	9.93	3.57
	7.4	11.56	6.66	11.15	7.12
	7.95	11.38	5.91	11.10	5.29
	10.2	15.07	11.72	14.70	10.28
	12.2	16.36	12.24	16.36	12.55
No Speech Coding		**53.74**	**45.42**	**61.71**	**52.40**

Table 6: Average ERLE (dB) from the network-based AEC (FBLMS) with PF and no LSS for voiced and unvoiced portions of female and male speeches.

fore any new AEC algorithms that work in conjunction with speech coding can be formulated.

6. ACKNOWLEDGEMENTS

The main author is grateful for the help and the insights gained from the third co-author during the internship at Tellabs.

REFERENCES

[1] A. Guérin, G. Faucon, and R. L. Bouquin-Jeannès, “Nonlinear acoustic echo cancellation based on Volterra filters,” *IEEE Trans. on Speech and Audio Processing.*, vol. 11, pp. 672–683, Nov. 2003.

[2] A. N. Birkett and R. A. Gourban, “Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects,” in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 1995, pp. 103–106.

[3] F. Kuech, A. Mitnacht, and W. Kellermann, “Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.*, Mar. 2005, vol. 3, pp. 105–108.

[4] M. Rages and K. C. Ho, “Limits on echo return loss enhancement on a voice coded speech signal,” in *Proc. IEEE Midwest Symp. on Circuits and Systems*, Aug. 2002, vol. 2, pp. 152–155.

[5] G. Ezner, H. Krüger, and P. Vary, “On the problem of acoustic echo control in cellular networks,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Sep. 2005, pp. 213–216.

[6] M. Hasegawa-Johnson and A. Alwan, “Speech coding: Fundamentals and applications,” *Wiley Encyclopedia of Telecommunications*, vol. 5, pp. 2340–59, 2003.

[7] L. Lamel, R. Kassel, and S. Seneff, “Speech database development: Design and analysis of the acoustic-phonetic corpus,” in *Proc. DARPA Speech Recognition Workshop*, Mar. 1987, pp. 26–32.

[8] Int. Telecom. Union, “ITU-T G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s,” 1996.

[9] Int. Telecom. Union, “ITU-T G.728: Coding of speech at 16 kbit/s using low-delay code excited linear prediction,” 1992.

[10] Int. Telecom. Union, “ITU-T G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP),” 1996.

[11] Europ. Telecom. Stand. Inst., “ETSI TS 126 104 V6.1.0: ANSI-C code for the floating-point Adaptive Multi-Rate (AMR) speech codec,” 2004.

[12] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 2002.