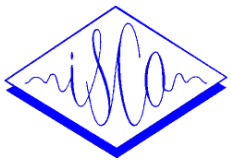

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

JOINT SYSTEM FOR ACOUSTIC ECHO CANCELLATION AND NOISE REDUCTION

G. Faucon - R. Le Bouquin Jeannès

e-mail: Gerard.Faucon@univ-rennes1.fr
Laboratoire de Traitement du Signal et de l'Image - Université de Rennes 1
Bât. 22 - Campus de Beaulieu - 35042 RENNES CEDEX - FRANCE

ABSTRACT

Our concern is the investigation of speech enhancement techniques for hands-free audio terminals, including two major problems: noise reduction and acoustic echo cancellation. The objective is to find a joint structure to get a near-end speech signal with a minimum distortion and low levels of echo and noise. To solve the problems addressed by the basic structures, a new technique is investigated and tested in real conditions.

1. INTRODUCTION

The increasing demand of mobile telephone systems requires speech enhancement techniques particularly adapted in hands-free communications. In this context, two problems appear: first of all, since the near-end speech signal to be transmitted is degraded by ambient noise, a Noise Reduction (NR) system is needed; secondly, the coupling between the loudspeaker and the microphone introduces an echo on the microphone so that an Acoustic Echo Cancellation (AEC) system is required. Efforts have been spread for the development of separate AEC and NR systems; in modern applications involving noisy environments, the requirements for acoustic echo cancellers are more stringent. Only a few studies have been devoted to techniques reducing acoustic echo as well as noise [1,2,3,4]. The dissemination of hands-free audio terminals and the advance in signal processing technology constrain to take the double talk problem into account, to provide a near-end speech signal more pleasant to listen to.

The two basic structures found in the literature are cascaded structures where AEC preceeds or follows NR. These structures are simple to implement but suffer from some defaults. Our goal is to propose joint structures which yield a near-end speech signal only slightly distorted, for a sufficient attenuation of echo and noise. After a recall of the two basic structures, we propose a new structure which is evaluated in a car and in an office environment.

2. AEC AND NR SYSTEMS

First of all, we present briefly the techniques of noise reduction and echo cancellation included in the joint systems.

The acoustic echo canceller is a Generalized Multi-Delay Filter (GMDF) algorithm [5]; it is based on a block frequency-domain adaptive filtering procedure. The two differences with the standard scheme lie in (i) the segmentation of the impulse response into segments, which allows to control independently the overall processing delay, and (ii) the introduction of a parameter controlling the overlap between the successive input blocks to modify the rate at which the filter coefficients are updated.

The noise reduction algorithm is derived from the Minimum Mean-Square Error Short-Time Spectral Amplitude (MMSE STSA) estimator proposed by Ephraim and Malah [6]. It is based on modeling speech and noise spectral components as statistically independent Gaussian random variables. Let $U(f)$ be the spectrum of the NR input $u(t)$, the speech estimator is given by:

$$\hat{S}(f) = G(f)U(f)$$

where $G(f)$ includes a Wiener filtering (simplified version of the MMSE STSA estimator) and a gain function taking the uncertainty of speech presence into account.

3. JOINT STRUCTURES

We assume that only a microphone and a loudspeaker are available. The observation $x(t)$ we receive on the microphone is composed of the near-end speech signal $s(t)$ to be transmitted, an echo $e(t)$ and a noise $n(t)$: $x(t) = s(t) + e(t) + n(t)$. The loudspeaker emits a signal $z(t)$ which is correlated with $e(t)$ and used as a reference to cancel this echo. No reference of noise is available and noise characteristics must be learned during speech pauses to be further used in noise reduction.

3.1. Optimal filtering

In the frequency domain, the optimal filter applied on the observation vector $Y(f) = [X(f) \;\; Z(f)]^T$ in the sense of the minimum mean-square error is:

$$W(f) = \frac{\gamma_{ss}(f)}{\gamma_{ss}(f) + \gamma_{nn}(f)} \begin{bmatrix} 1 \\ -\dfrac{\gamma_{xz}(f)}{\gamma_{zz}(f)} \end{bmatrix}$$

such that $\hat{S}(f) = W^T(f)Y(f)$. $\hat{S}(f)$, $X(f)$, $Z(f)$ represent the spectra of the signals $\hat{s}(t)$, $x(t)$ and $z(t)$ respectively, $\gamma_{ss}(f)$, $\gamma_{nn}(f)$ and $\gamma_{zz}(f)$ are the psd (power spectral densities) of $\hat{s}(t)$, $n(t)$ and $z(t)$. $\gamma_{xz}(f)$ is the cross psd

between the observations $x(t)$ and $z(t)$. There are two steps involved in the optimal structure: in the first step, the echo is estimated by filtering the reference $z(t)$. For an optimal echo canceller, speech and noise are transmitted with no change and the echo is completely cancelled. The echo canceller output is ideally $s(t)+n(t)$. In the second step, the noise is reduced by a Wiener filtering whose gain is $\gamma_{ss}(f)/(\gamma_{ss}(f)+\gamma_{nn}(f))$. We note that the optimal structure is composed of the two optimal filters addressed for each topic.

3.2. Study of the structures

3.2.1. Basic structures

The first basic structure corresponds to the implementation of the optimal filtering where the AEC system preceeds the NR system (Figure 1).

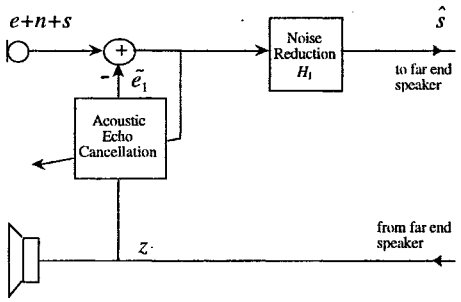

Figure 1. Structure 1

In practice, the AEC system is disturbed by the additive noise and by the near-end speech signal present on the microphone. The noise is omnipresent and the adaptation is necessarily performed in presence of noise. It appears difficult to stop the AEC adaptation in double talk mode, the coefficients of the AEC when speech appears being strongly disturbed. So, to reduce the noise influence on the AEC system, it is suggested to place the NR system before the AEC system (second structure, Figure 2). In this case, stopping the adaptation in double talk mode is possible. Nevertheless, the disturbing noise is less reduced in Single Talk (ST) and Double Talk (DT) modes.

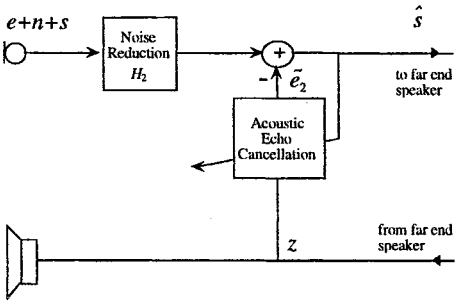

Figure 2. Structure 2

3.2.2. Proposal of a new structure

In the basic structures, $\tilde{e}_1$ and $\tilde{e}_2$ represent the echoes estimated by the AEC system. Figure 3 displays the similarity index (SIM) corresponding to the difference between the original echo and the echo estimated by the AEC computed in ST mode:

$$SIM(k)=10\log\frac{P_e(k)}{P_{e-\tilde{e}_i}(k)}, \quad i=1,2$$

where $P_y(k)$ represents the power of y computed on the kth block of 32 ms. In spite of the distortion brought by the NR system, it is better to reduce first the disturbing noise to obtain a more accurate echo estimate. As a matter of fact, $SIM(k)$ falls on periods when noise is predominant, especially in the first structure. This confirms that the AEC system is dramatically disturbed by the presence of important noise components.

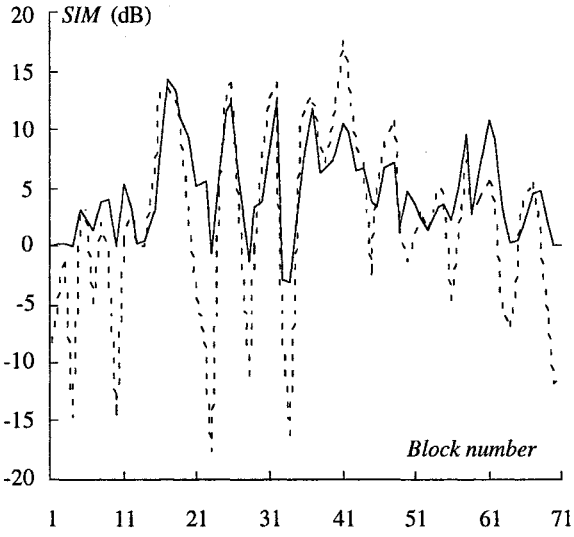

Figure 3. *SIM* vs. block number (block = 32 ms) in ST mode. Solid line: structure 2, dashed line: structure 1

In consequence, we propose a new structure, called structure 3 (Figure 4). The noise influence on the AEC system is first reduced by the introduction of a noise reduction filter H_2 as in structure 2. The echo $\tilde{e}_2$ estimated by the AEC system is subtracted from the observation $x(t)$ to get the signal $v(t)=s(t)+n(t)+e(t)-\tilde{e}_2(t)$. Then, a second noise reduction filter H_3 is applied on $v(t)$ to give the final estimate $\hat{s}(t)$. As in the second structure, the AEC adaptation can be stopped in DT mode.

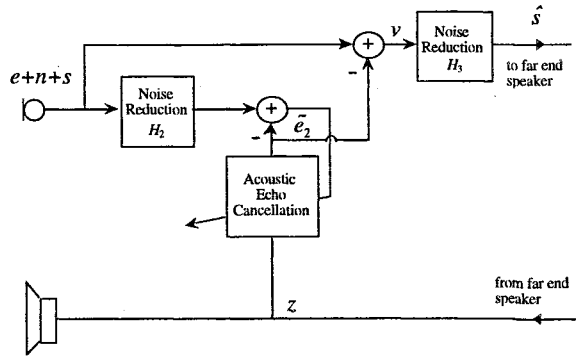

Figure 4. Structure 3

4. RESULTS

An evaluation of performance is realized in a car and in an office environment. The database is obtained by recording the speech signal (near-end speaker), the echo (far-end speaker) and the disturbing noise separately to consider different signal and echo to noise ratios called *SNR* and *ENR* respectively. The sampling rate is 8 kHz. From the recordings, we create files of composite signals as shown Figure 5: the first part consists of a noisy echo (ST mode) and the second part corresponds to speech added to a noisy echo (DT mode).

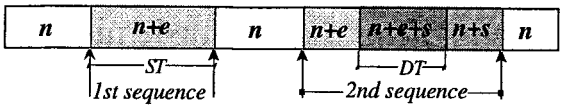

Figure 5. Composite signal

For each file, we define objective measures [7,8] in ST and DT modes:

- the Echo Return Loss Enhancement *ERLE* in both modes,

$$ERLE = \frac{1}{N}\sum_{k=1}^{N} 10\log\frac{P_e(k)}{P_{e_r}(k)}$$

- the noise reduction factor R in both modes,

$$R = \frac{1}{N}\sum_{k=1}^{N} 10\log\frac{P_n(k)}{P_{n_r}(k)}$$

- the speech distortion D in DT mode,

$$D = \frac{1}{N}\sum_{k=1}^{N} 10\log\frac{P_{s-s_f}(k)}{P_s(k)}$$

- the gain G in DT mode,

$$G = \frac{1}{N}\sum_{k=1}^{N} 10\log\frac{P_{e+n}(k)}{P_{s-\hat{s}}(k)}.$$

In these formulas, $P_y(k)$ represents the power of y computed on the kth block of 256 samples, N is the number of blocks corresponding to the estimation performed in ST and DT modes. $\hat{s}$ is the final estimate, n_r and s_f are the filtered versions of n and s using the noise reduction filter (for the third structure, the NR filter is H_3), and e_r is the residual echo obtained as follows:

- in the first structure, e_r is obtained by filtering the difference $(e-\tilde{e}_1)$ using H_1,
- in the second structure, e_r is given by the difference between the echo e filtered by the NR filter H_2 and the estimated echo $\tilde{e}_2$,
- in the third structure, e_r is obtained by filtering the difference $(e-\tilde{e}_2)$ using H_3.

The gain G is a pertinent measure since it takes the speech distortion, the residual noise and the residual echo into account.

For the car environment, the noise is stationary and due to the car moving at 130 km/h and for the office environment, the noise is due to people chatting.

To confirm the previous remarks concerning the AEC adaptation in DT mode, each structure is tested with or without adaptation.

Tables 1 to 3 present the objective measures, averaged on a set of ten files for the car environment, the *SNR* and *ENR* being equal to 3 dB.

	$ERLE_{mean}$	R_{mean}
st. 1	29.24 dB	32.70 dB
st. 2	12.61 dB	21.93 dB
st.3	31.29 dB	31.58 dB

Table 1. Car environment, ST mode

	$ERLE_{mean}$	R_{mean}	D_{mean}	G_{mean}
st. 1	6.37 dB	18.07 dB	-3.89 dB	6.20 dB
st. 2	0.22 dB	16.42 dB	-5.29 dB	5.91 dB
st. 3	10.62 dB	19.39 dB	-4.15 dB	7.86 dB

Table 2. Car environment, DT mode (adaptation continued)

	$ERLE_{mean}$	R_{mean}	D_{mean}	G_{mean}
st. 1	-12.84 dB	9.24 dB	-7.09 dB	-5.02 dB
st. 2	1.35 dB	16.42 dB	-5.29 dB	3.04 dB
st. 3	9.36 dB	16.39 dB	-5.31 dB	7.02 dB

Table 3. Car environment, DT mode (adaptation stopped)

Table 1 corresponds to single talk mode. Structures 1 and 3 are the most performant and give similar results. In double talk mode (Tables 2 and 3), it is confirmed that we'd better to continue the AEC adaptation for the structure 1. For structures 2 and 3 and for this configuration, the gain increases when the adaptation is continued. Among the basic structures, the first one (adaptation continued) is the most performant in DT mode. If we examine all results in both modes, the best performance is obtained with the proposed structure (adaptation continued).

In the same way, Tables 4 to 6 present the objective measures averaged on a set of ten files for the office environment, the *SNR* and the *ENR* being equal to 7 dB.

The results obtained in the office environment show that structures 1 and 3 are still the most performant in ST mode and for all structures, it is better to continue the adaptation in DT mode. The same conclusion holds when the *SNR* and *ENR* are equal to 3 dB. Structures 1 and 3 are the most performant, the structure 3 being slightly the best. Nevertheless, the performance in DT mode remains poor in this environment.

	$ERLE_{mean}$	R_{mean}
st. 1	33.44 dB	27.82 dB
st. 2	14.01 dB	10.75 dB
st.3	33.58 dB	25.95 dB

Table 4. Office environment, ST mode

	$ERLE_{mean}$	R_{mean}	D_{mean}	G_{mean}
st. 1	5.78 dB	12.29 dB	-3.55 dB	0.54 dB
st. 2	1.14 dB	6.80 dB	-5.98 dB	0.75 dB
st. 3	6.96 dB	12.53 dB	-3.83 dB	1.07 dB

Table 5. Office environment, DT mode (adaptation continued)

	$ERLE_{mean}$	R_{mean}	D_{mean}	G_{mean}
st. 1	-6.05 dB	5.41 dB	-6.28 dB	-2.1 dB
st. 2	-1.74 dB	6.80 dB	-5.98 dB	-1.57 dB
st. 3	-0.08 dB	6.89 dB	-5.96 dB	-1.08 dB

Table 6. Office environment, DT mode (adaptation stopped)

Then, we compare the gain of each structure for the car environment when the *ENR* is equal to 0 dB and the *SNR* varies from 0 dB to 9 dB (Figure 6). Globally, the third structure is the most performant, without adaptation for low *SNR* and with adaptation for high *SNR*.

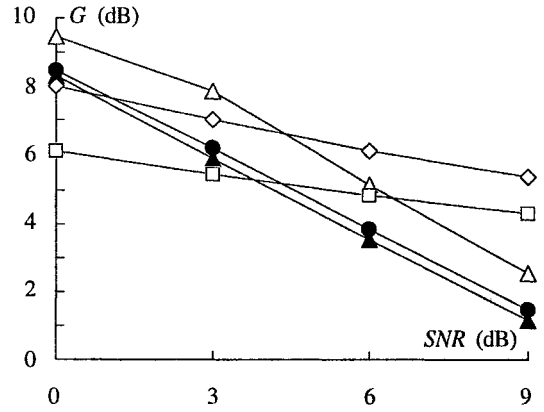

Figure 6. Gain in DT mode
adaptation continued: ● structure 1, ▲ structure 2, △ structure 3
adaptation stopped: □ structure 2, ◇ structure 3

Figure 7 represents the far end and near end speech signals, the microphone signal and the signals estimated by the different structures for a *SNR* equal to 3 dB and an *ENR* equal to 0 dB. It is obvious that the echo is completely suppressed in ST mode using the new structure and in DT mode, the useful signal seems to be well estimated.

5. CONCLUSION

In order to enhance the near-end speech signal in hands-free telecommunications, we investigate a new structure taking advantage of two basic structures. The proposed structure and the first basic structure provide comparable values of *ERLE* in ST mode and our structure yields the upper gain in DT mode. For the car environment and for important *SNR*, it is better to stop the AEC adaptation in the new structure. We plan to optimize the new structure by (i) modifying the AEC adaptation step in DT mode and (ii) using two different noise reduction filters. Other experiments must be carried out on a large variety of signal or echo to noise ratios to confirm the improvement.

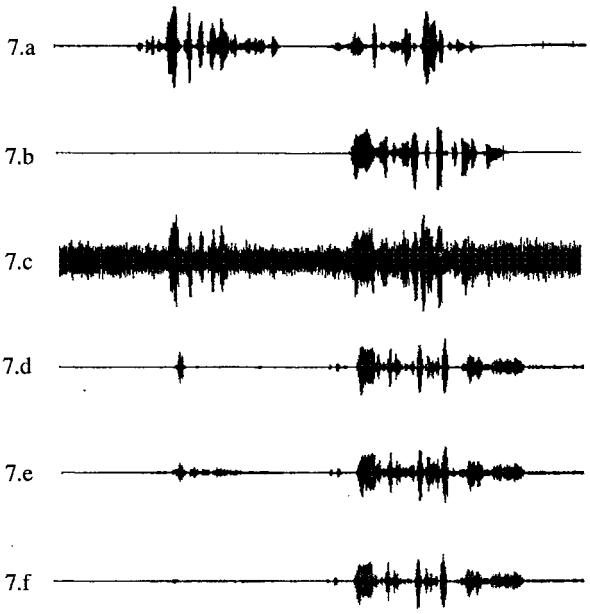

Figure 7. (a) far end speech signal, (b) near end speech signal, (c) microphone signal, (d) signal estimated by structure 1, (e) signal estimated by structure 2, (f) signal estimated by structure 3

ACKNOWLEDGMENTS

The authors wish to thank Telecom Paris for the GMDF algorithm and Matra Communication (Paris) and Page Iberica (Madrid) for the database.

REFERENCES

[1] H. YASUKAWA. *Acoustic Echo Canceller with Sub-band Noise Cancelling*. Electronics Letters, vol. 28, n°15, July 1992, pp. 1403-1404.

[2] R. MARTIN, P. VARY. *Combined Acoustic Echo Cancellation, Dereverberation and Noise Reduction: a Two Microphone Approach*. Annal. des Télécom., tome 49, n°7-8, Juillet-Août 1994, pp. 414-419.

[3] R. MARTIN, J. ALTENHÖNER. *Coupled Adaptive Filters for Acoustic Echo Control and Noise Reduction*. Proc. ICASSP-95, pp. 3043-3046.

[4] B. AYAD, G. FAUCON. *Acoustic Echo and Noise Cancelling for Hands-Free Communication Systems*. 4th International Workshop on Acoustic Echo and Noise Control, Roros, Norway, June 21-23 1995.

[5] E. MOULINES et al. *The Generalized Multidelay Adaptive Filters: Structures and Convergences Analysis*. IEEE Trans. on Signal Processing, vol. 43, n°1, January 1995, pp. 14-28.

[6] Y. EPHRAIM and D. MALAH. *Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator*. IEEE Trans. on ASSP, vol. ASSP-32, n°6, December 1984, pp. 1109-1121.

[7] P. NAYLOR et al. *Enhancement of Hands-Free Telecommunication*. Annal. des Télécom., tome 49, n°7-8, Juillet-Août 1994, pp. 373-379.

[8] R. LE BOUQUIN et al. *Proposal of a Composite Measure for the Evaluation of Noise Cancelling Methods in Speech Processing*. EUROSPEECH'93, Berlin, September 21-23 1993, pp. 227-230.