OPTIMIZATION OF A NOISE REDUCTION PREPROCESSING IN AN ACOUSTIC ECHO AND NOISE CONTROLLER

Beghdad Ayad, Gérard Faucon and Régine Le Bouquin-Jeannès

Laboratoire de Traitement du Signal et de l’Image - Université de Rennes 1
Bât. 22 - Campus de Beaulieu - 35042 RENNES CEDEX - FRANCE
Beghdad.Ayad@univ-rennes1.fr

ABSTRACT

In hands-free communications, the speech signal to be transmitted is disturbed by ambient noise and acoustic echo. So, a processing to reduce these disturbances must be performed before transmission. Classical solutions are cascaded structures where the acoustic echo canceller proceeds or follows the noise reduction system. Recently, we have proposed a new joint system where a noise reduction preprocessing allows to improve the performance of the acoustic echo canceller. This preprocessing reduces the noise but distorts the original echo. This paper presents an optimization of the preprocessing. Objective results in terms of Echo Return Loss Enhancement and gain are presented.

1. INTRODUCTION

For some applications such as teleconferencing and hands-free telephone, the near-end speech signal to be transmitted is disturbed by ambient noise and by an echo due to the coupling between the microphone and the loudspeaker. The dissemination of hands-free communication systems requires to provide users with some comfort. So, both problems have to be solved to obtain a good quality speech signal. If a number of studies have been done separately on each problem, we can easily derive and compare the results. Recent studies concern joint systems including both processings [1,2]. Our objective is to optimize a joint structure to get a near-end speech signal only slightly distorted and low levels of echo and noise.

The microphone observation \( x(t) \) is composed of the near-end speech signal \( s(t) \), an echo \( e(t) \), a noise \( n(t) \) and the loudspeaker emits a signal \( z(t) \) correlated with \( e(t) \). The optimal structure in the sense of the minimum mean square error to be applied on \( x(t) \) and \( z(t) \) can be easily derived and is composed of two steps. In the first one, the echo is estimated by applying on the reference \( z(t) \) a filtering whose transfer function is

\[
\gamma_{xe}(f)/\gamma_{zz}(f),
\]

where \( \gamma_{xe}(f) \) is the cross power spectral density between signals \( x \) and \( z \); \( \gamma_{zz}(f) \) is the power spectral density of signal \( z \). The filter output is subtracted from the microphone observation. For an ideal echo canceller, speech and noise are transmitted with no change and the echo is completely removed. Then, in the second step, noise is reduced by a Wiener filtering whose gain is

\[
\gamma_{sn}(f)/\gamma_{ss}(f) + \gamma_{en}(f).
\]

So, the optimal structure is composed of the two cascaded optimal filters, where the Acoustic Echo Cancellation (AEC) system precedes the Noise Reduction (NR) system. This structure is called AEC+NR (Figure 1).

In practice, the AEC system is adaptive. The coefficients of the AEC are disturbed by ambient noise which is omnipresent and it appears difficult to stop the AEC adaptation when speech occurs. To reduce the noise influence on the AEC system, the place of the NR system and the AEC system may be exchanged, so that the adaptation can be stopped in Double Talk (DT) mode (speech and echo present). Unfortunately, the disturbing noise is less reduced and this implementation differs from the optimal structure. Nevertheless, the echo estimated by the AEC is closer to the original echo when NR preceeds AEC. In [3], experiments reveal that, in spite of the distortion on the echo due to the NR system, it is better to reduce first the disturbing noise to obtain a more accurate echo estimate. Therefore, a new structure [3,4], called AEC+2NR (Figure 2), has been investigated. The noise influence on the AEC system is first reduced by applying a noise reduction filter \( H_2 \) on the microphone, then an AEC is performed. The echo \( \bar{e}_2 \) estimated by the AEC system is subtracted from the microphone observation \( x(t) \) to get the signal \( v(t) = s(t) + n(t) + e(t) - \bar{e}_2(t) \). Then, a second noise reduction filter \( H_3 \) is applied on \( v(t) \) to get the final estimate. In this way, the AEC adaptation can be stopped in DT mode and \( v(t) \) contains the unchanged speech signal.

In Figure 1, Structure AEC+NR.
2. AEC AND NR 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 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. This algorithm used as a preprocessing (filter \( H_2 \)) in the structure AEC+2NR will be optimized and we give hereafter a more detailed description of this technique to understand where the optimization occurs. Let \( Y(f) \) be the spectrum of a NR input \( y(t) \) which is composed of a signal \( w(t) \) and a noise \( n(t) \). The signal estimate is given by

\[
\hat{w}(f) = G_1(f) - G_2(f) \cdot Y(f)
\]

where \( G_1(f) \) is a Wiener filtering and \( G_2(f) \) represents the gain function taking the uncertainty of speech presence into account [7]. This estimator depends on the \( a\) priori SNR (Signal-to-Noise Ratio), \( R_{\text{prior}} \), the \( a\) posteriori SNR, \( R_{\text{post}} \), defined respectively as:

\[
R_{\text{prior}}(f) = \frac{E[W(f)\hat{W}(f)]}{E[N(f)\hat{N}(f)]}, \quad R_{\text{post}}(f) = -\frac{E[V(f)\hat{V}(f)]}{E[N(f)\hat{N}(f)]}
\]

and the probability of signal absence \( q(f) \). \( E[N(f)\hat{N}(f)] \) is the noise power learned during speech pauses. Ephraim and Malah proposed to estimate the \( a\) priori SNR according to a decision-directed approach [6]:

\[
R_{\text{prior}}^l(f, n) = \lambda R_{\text{post}}^l(f, n-1) + (1 - \lambda) q(n-1) R_{\text{post}}(f, n-1)
\]

where \( n \) is the current block number, \( R_{\text{post}}^l(f, n-1) \) is the amplitude of the signal estimated on the block \((n-1)\), \( \lambda \) is a weighting factor, \( q(n) \) is an operator defined by \( \text{Max}(u, 0) \). \( R_{\text{post}}(f, n) \) is directly given by the ratio of the squared magnitude of the observation on the block \( n \) to the noise power \( E[N(f)\hat{N}(f)] \).

3. OPTIMIZATION OF THE NOISE REDUCTION PREPROCESSING

In the structure AEC+2NR, what is the best noise reduction filter \( H_2 \) to be applied on the microphone observation? A way to modify \( H_2 \) is to change the value of the weighting factor \( \lambda \) in the estimation of the \( a\) priori SNR. We compute a noise reduction factor \( R \) and a distortion factor \( D \) brought by the filter \( H_2 \):

\[
R = \frac{1}{M} \sum_{k=1}^{M} \log \frac{P_e(k)}{P_q(k)}, \quad D = \frac{1}{M} \sum_{k=1}^{M} \log \frac{P_{e-f}(k)}{P_q(k)}
\]

\( e_f \) and \( n_f \) represent the echo and the noise filtered by the noise reduction filter \( H_2 \) respectively, \( P_q(k) \) is the power of \( u \) computed on the \( k \)th echo and noise are present together (Single Talk (ST) mode). Figure 3 shows the distortion \( D \) versus the noise reduction factor \( R \) for different values of \( \lambda \) and different Echo-to-Noise Ratios (ENR). The ENR is given by the ratio of the echo power to the noise power on the \( M \) blocks. When \( \lambda \) tends to 1, the noise reduction and the echo distortion increase, which corresponds to a lower gain of the filter \( H_2 \).

![Figure 2. Structure AEC+2NR](image)

4. RESULTS

The influence of the noise reduction filter \( H_2 \) on the performance of the structure AEC+2NR is evaluated in terms of objective measures. For comparison, the structure AEC+N is also studied.

a) Methodology of evaluation

The database is obtained by recording the speech signal, the echo and the disturbing noise separately to consider various...
SNR and ENR. These signals are recorded in a car, and noise is due to the car moving at 130 km/h. From these recordings, we create files of composite signals (Figure 4), the first part is a noisy echo (ST mode) and the second part corresponds to speech added to a noisy echo (DT mode).

**Figure 4. Composite signal**

Only three objective measures [3,8] are presented:
- the similarity index $SIM$ in ST mode
  \[ SIM = \frac{1}{N} \sum_{k=1}^{N} 10 \log \frac{P_x(k)}{P_{r_i}(k)} \]
  \[ i = 1, 2 \]
- the Echo Return Loss Enhancement $ERLE$ in both modes
  \[ ERLE = \frac{1}{N} \sum_{k=1}^{N} 10 \log \frac{P_x(k)}{P_{r_i}(k)} \]
- the gain $G$ in DT mode
  \[ G = \frac{1}{N} \sum_{k=1}^{N} 10 \log \frac{P_{x+n}(k)}{P_{s}(k)} \]

$k$ is a block index and $N$ is the number of blocks corresponding to the estimation performed in ST or DT mode, $\theta$ is the final estimate of the near-end speech signal, $e_r$ represents a residual echo computed as follows:
- in the structure AEC+NR, $e_r$ is obtained by filtering the difference $e - \theta_i$ using $H_1$,
- in the structure AEC+2NR, $e_r$ is obtained by filtering the difference $e - \theta_i$ using $H_3$.

The optimization only concerns the noise reduction filter $H_2$ in the structure AEC+2NR. The parameters of the AEC system and filters $H_i$ and $H_3$ are fixed. We choose the following parameters: for the GMDF algorithm, the length of the impulse response is 256, it is divided into $L = 2$ segments and the overlapping between successive blocks is (256-32) samples, the adaptation step is 0.33; for the noise reduction filters, $H_i$ and $H_3$, the weighting factor $\lambda$ is 0.98, the probability of signal absence $q(f)$ is 0.5. The noise power is learned on ten blocks of 256 samples where only noise is present.

Figures 5 to 10 present the objective measures, averaged on a set of ten files, where • corresponds to the structure AEC+NR (adaptation continued) and -○- corresponds to the structure AEC+2NR.

In ST mode, the ENR varies from -3 dB to 12 dB and in DT mode, the ENR and SNR are identical and these ratios vary from -3 dB to 12 dB.

**b) Influence of the weighting factor $\lambda$**

In ST mode, the AEC is only disturbed by noise. Figures 5 and 6 represent $SIM$ and $ERLE$ for different values of $\lambda$.

Values of $\lambda$ in the range [0.5; 0.8] lead to comparable results for high ENR; as for low ENR, performance falls because of the imperfect noise reduction and these results are not presented. $\lambda = 0.80$ gives the best $SIM$ and $ERLE$, which corresponds to a noise reduction by $H_2$ around 10 dB (Figure 3). When $\lambda$ tends to 1, the echo is more distorted and the filtering is less efficient.

In DT mode, the question is: is there any advantage to stop the AEC adaptation in the structure AEC+2NR? To this end, two cases are considered:

1. *adaptation continued*. We can see that, for $0.80 \leq \lambda \leq 0.98$, the $ERLE$ (Figure 7) and the gain $G$ (Figure 8) are quite similar; for $\lambda < 0.80$, we note some degradation and results are not presented. Since we obtain interesting results in each mode for $\lambda = 0.80$, this value may be kept in both situations.
To conclude on these experiments, for low ENR (and SNR), it seems better to continue the adaptation and, for high ENR (and SNR), it is better to stop it. Informal listening tests confirm these remarks.

5. CONCLUSION

An optimization of a preprocessing included in a new acoustic echo and noise controller is proposed. Different weighting factors have been considered in the preprocessing in ST and DT modes when the AEC adaptation is continued or stopped. In this last case, the near-end speech signal is less distorted. A complete subjective evaluation has to be conducted to validate the objective measures.

REFERENCES