

Residual Echo Power Estimation for Speech Reinforcement Systems in Vehicles

Alfonso Ortega, Eduardo Lleida, Enrique Masgrau

Communication Technologies Group
Aragon Institute of Engineering Research (I3A)
University of Zaragoza, Spain
ortega@unizar.es

Abstract

In acoustic echo cancellation systems, some residual echo exists after the acoustic echo canceler (AEC) due to the fact that the adaptive filter does not model exactly the impulse response of the Loudspeaker-Enclosure-Microphone (LEM) path. This is specially important in feedback acoustic environments like speech reinforcement systems for cars where this residual echo can make the system become unstable. In order to suppress this residual echo remaining after the AEC, postfiltering is the most used technique. The optimal filter that ensures stability without attenuating the speech signal depends on the power spectral density (psd) of the residual echo that must be estimated. This paper presents a residual echo psd estimation method needed to obtain the optimal echo suppression filter in speech reinforcement systems for cars.

1. Introduction

There are many factors that can make communications among passengers inside a car difficult: high noise level inside the cabin, distance among passengers, lack of visual contact between speaker and listener, etc. In order to make communications easier a speech reinforcement system is proposed [1, 2]. The speech reinforcement system in vehicles, known as Cabin Car Communication System, makes use of a set of microphones placed on the overhead of the cabin to pick up the speech of each passenger. Afterwards, it amplifies those signals and returns them to the cabin through the car audio system. This solution presents two problems:

1. Microphones pick up the signal radiated by the loudspeakers what is the origin of acoustic echo. Because there is an amplification stage between microphones and loudspeakers, there will be acoustic paths that can make the system become unstable.
2. The noise present inside the cabin coming from the engine, the road or the wind is picked up by the microphones and amplified by the system resulting in an increase of the noise level of the cabin.

An acoustic echo canceller (AEC) is used to overcome the first problem. Nevertheless, in order to achieve enough echo attenuation and sufficient quality in the output signal, the system presented in this paper makes use of an Echo Suppression Filter (ESF) after the AEC. To reduce the noise present in the microphone signal we use a single microphone method based on the Wiener solution.

Another important aspect of this system is that the overall delay must be short enough to achieve full integration of the

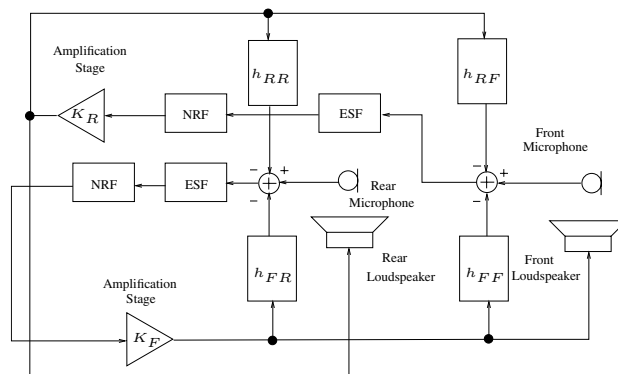


Figure 1: Schematic diagram of a two channel speech reinforcement system.

sound coming from the direct path and the reinforced speech coming from the loudspeaker.

This paper is organized as follows. In section 2 a brief description of the system is presented. A discussion about the optimal expression for the Echo Suppression Filter (ESF) will be studied in section 3. The way how the psd of the residual echo is estimated will be shown in section 4. In section 5 performance measures and results will be presented and in section 6 we present the conclusions along with a summary of the paper.

2. Description of the problem.

In order to make communications among passengers in vehicles easier, a two channel speech reinforcement system is required. One channel must take the speech of the rear passengers to the front seats and the other one must take the speech of the front passengers to the rear seats. A block diagram of the two channel system can be seen in Fig. 1. In a two channel system, for each channel, there must be two echo cancellers, an Echo Suppression Filter, a Noise Reduction Filter and an amplification stage with gain factor K . For the sake of simplicity, a one channel system will be studied here. The block diagram of a one channel system can be seen in Fig. 2.

This one channel system is composed of one acoustic echo canceller (AEC), an Echo Suppression Filter (ESF) a Noise Reduction Filter (NRF) and an amplification stage. The algorithm used to update its coefficients is a Least Mean Square algorithm because real time operation must be achieved and low computational complexity algorithms must be used.

Due to the fact that the microphone signal is always com-

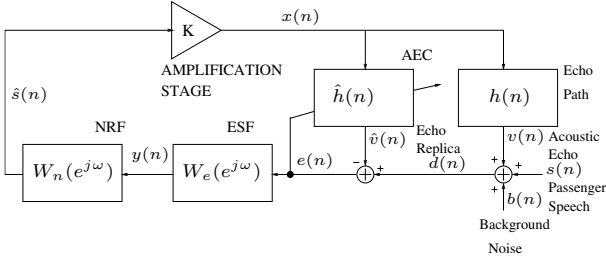


Figure 2: Schematic diagram of a two channel speech reinforcement system.

posed of the acoustic echo, the noise present in the cabin and the speech signal, the adaptive filter can't converge to a good estimate of the LEM path because passenger's speech disturbs the adaptation of the filter. This effect is usually known in telephony as double-talk. The classical solution to deal with it is to detect double-talk and to freeze the coefficients during this double-talk periods. However to freeze the filter taps can not be the solution for the reinforcement system as this double-talk situation is permanent and the filter coefficients would be always frozen. As a result of this bad estimation performed by the adaptive filter, acoustic echo will pass through to the amplification stage and will be played back into the cabin creating acoustic feedback paths that can make the system to become unstable and are the origin of annoying high intensity oscillations known as howling.

To perform further echo attenuation and ensure the stability of the system an Echo Suppression Filter is placed after the AEC. The optimal expression for the ESF to ensure stability will be discussed in section 3.

After the ESF a Noise Reduction Filter is placed in order to avoid increasing the noise inside the car reducing the noise played back into the cabin by the loudspeakers. This filter is performed by means of a Wiener filter as discussed in [1].

3. Echo Suppression Filter

The Echo Suppression Filter placed after the AEC must ensure stability, attenuate the residual acoustic echo and avoid distortion. According to the block diagram of a one channel system shown in Fig. 2 the transfer function of the system between the input signal $s(n) + b(n)$ and the output signal $x(n)$ is

$$P(e^{j\omega}) = \frac{K \cdot W_e(e^{j\omega}) \cdot W_n(e^{j\omega})}{1 - K \cdot W_e(e^{j\omega}) \cdot W_n(e^{j\omega}) \cdot \tilde{H}(e^{j\omega})} \quad (1)$$

where $\tilde{H}(e^{j\omega})$ known as misadjustment, is the difference between the transfer function of the LEM path $H(e^{j\omega})$ and the transfer function of the adaptive filter $\hat{H}(e^{j\omega})$. Due to the permanent double talk situation, this difference can be significant, and the system can become unstable when the denominator in (1) approaches to zero depending on the value of the gain factor K . The optimal solution for the ESF $W_e(e^{j\omega})$ that ensures stability, avoiding howling and minimizing distortion without increasing the noise level inside the car can be found by forcing the transfer function of the system to be

$$P(e^{j\omega}) = K \cdot W_n(e^{j\omega}) \quad (2)$$

Thus the optimal expression for the Echo Suppression Filter is

$$W_e(e^{j\omega}) = \frac{1}{1 + K \cdot W_n(e^{j\omega}) \cdot \tilde{H}(e^{j\omega})} \quad (3)$$

The optimal ESF depends on the misadjustment function $\tilde{H}(e^{j\omega})$ which is unknown. However, an estimation of the misadjustment can be obtained by using the psd of the residual echo.

We assume that the ESF, $W_e(e^{j\omega})$, is a real valued function. This assumption discussed in [3] gives a good approximation in our problem. The Wiener filter NRF, $W_n(e^{j\omega})$, is also a real valued function.

We can express the residual echo $r(n)$ as the output of a linear system with impulse response $\tilde{h}(n) = h(n) - \hat{h}(n)$ when the input signal is $x(n)$, the output of the speech reinforcement system.

Thus, the psd of the residual echo is

$$S_r(e^{j\omega}) = S_x(e^{j\omega}) \cdot |\tilde{H}(e^{j\omega})|^2 \quad (4)$$

which depends on the psd of the output signal $S_x(e^{j\omega})$ and the misadjustment function $\tilde{H}(e^{j\omega})$. According to Fig. 2 we can obtain the psd of the output signal $x(n)$ from the psd of the error signal $e(n)$

$$S_x(e^{j\omega}) = S_e(e^{j\omega}) \cdot K^2 \cdot |W_e(e^{j\omega})|^2 \cdot |W_n(e^{j\omega})|^2 \quad (5)$$

Therefore, the squared modulus of the misadjustment function can be expressed as follows

$$|\tilde{H}(e^{j\omega})|^2 = \frac{S_r(e^{j\omega})}{K^2 \cdot |W_e(e^{j\omega})|^2 \cdot |W_n(e^{j\omega})|^2 \cdot S_e(e^{j\omega})} \quad (6)$$

Using the expression of the squared modulus of the misadjustment in (3) and assuming that $W_e(e^{j\omega})$ is a real valued function, the optimal ESF results

$$W_e(e^{j\omega}) = 1 - \sqrt{\frac{S_r(e^{j\omega})}{S_e(e^{j\omega})}} \quad (7)$$

which depends on the psd of the residual echo $S_r(e^{j\omega})$ and the psd of the error signal $S_e(e^{j\omega})$. The error signal $e(n)$ is directly accessible so we can obtain an estimation of $S_e(e^{j\omega})$ using periodogram methods but the estimation of the psd of the residual echo needs more elaborated procedures that will be discussed in the next section.

4. Residual Echo Power Spectral Density Estimation.

An estimation of the residual echo psd $\hat{S}_r(e^{j\omega})$ can be obtained from the estimation of the psd of the error signal $e(n)$ by using the iterative method described in this section.

The LEM path can be modeled as a delay block of Δ samples followed by a linear filter $h'(n)$. This delay block models the electro-acoustic delay of the loudspeaker to microphone path plus some processing delay.

To compensate for this delay the first Δ coefficients of the AEC are set to zero.

Fig. 3 shows the simplified schematic diagram of the speech reinforcement system with this decomposition where $\tilde{h}'(n)$ is the misadjustment function without the first Δ null samples that are modeled by the block $z^{-\Delta}$ and $w_{e,n}(n)$ is the impulse response of the linear system composed of the ESF and the NRF.

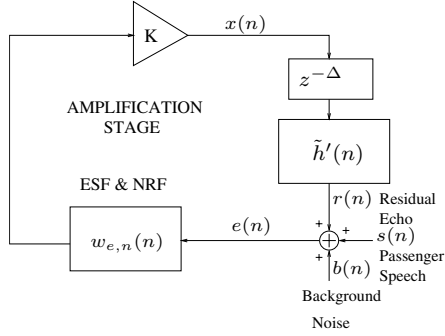


Figure 3: *Simplified Schematic diagram of a speech reinforcement system.*

Assuming stationarity on short periods of time (10-20 ms) we are interested in obtaining an estimation of the short-term psd of the k -th segment of length L samples of residual echo $S_r(e^{j\omega}; k)$. In order to estimate this short term psd, we use the optimal Wiener solution

$$H_r(e^{j\omega}; k) = \frac{S_{re}(e^{j\omega}; k)}{S_e(e^{j\omega}; k)} \quad (8)$$

where $S_e(e^{j\omega}; k)$ is the psd of the error signal $e(n)$ and $S_{re}(e^{j\omega}; k)$ is the cross-power spectral density of the residual echo $r(n)$ and the error signal $e(n)$.

We can express the short term cross-power spectral density of the residual echo and the error signal as the Fourier transform of the short term cross correlation

$$R_{re}(m; kL) = E[r(n; kL)e(n - m; kL)] \quad (9)$$

where $E[\cdot]$ denotes expected value.

The error signal $e(n)$, as can be seen in Fig. 3, is composed of the speech signal $s(n)$, the background noise $b(n)$ and the residual echo $r(n)$.

Assuming statistical independence between the background noise and the rest of the components of $e(n)$, the short time cross correlation of the residual echo and the error signal can be expressed as

$$R_{re}(m; kL) = E[r(n; kL)r(n - m; kL)] + E[r(n; kL)s(n - m; kL)] \quad (10)$$

According to Fig.3 we can express the residual echo as

$$r(n) = K \cdot e(n - \Delta) * \tilde{h}''(n) \quad (11)$$

where $\tilde{h}''(n) = \tilde{h}'(n) * w_{e,n}(n)$.

For the sake of simplicity and better understanding, lets consider that the length of a signal frame L is equal to Δ . Therefore, the k -th frame of the residual echo $r(n; kL)$ depends on the previous frame of the error signal $e(n; (k - 1)L)$. Thus, the k -th frame short time cross correlation of the residual echo and the error signal results

$$R_{re}(m; kL) = E[r(n; kL)r(n - m; kL)] + K \cdot E[e(n, (k - 1)L)s(n - m; kL)] * \tilde{h}''(n) \quad (12)$$

Assuming again statistical independence between the background noise and the rest of the components of $e(n)$

$$R_{re}(m; kL) = E[r(n; kL)r(n - m; kL)] + K \cdot E[s(n, (k - 1)L)s(n - m; kL)] * \tilde{h}''(n) + K \cdot E[r(n, (k - 1)L)s(n - m; kL)] * \tilde{h}''(n) \quad (13)$$

The second and third terms depends on the correlation between consecutive frames. Due to the non stationary nature of speech signal, a low correlation value will be expected. As this value is convolved with $\tilde{h}''(n)$ that is relatively small if the AEC is working, this two terms can be neglected in front of the first term that is the autocorrelation of the residual echo in the actual frame.

Therefore, the short term cross-power spectral density of the residual echo and the error signal can be considered to be $S_{re}(e^{j\omega}; k) = S_r(e^{j\omega}; k)$. In this way, an estimation of the Wiener filter, $H_r(e^{j\omega}; k)$, can be obtained by using

$$\hat{H}_r(e^{j\omega}; k) = \frac{\hat{S}_r(e^{j\omega}; k)}{\hat{S}_e(e^{j\omega}; k)} \quad (14)$$

where $\hat{S}_e(e^{j\omega}; k)$ is the estimation of the psd of the error signal and $\hat{S}_r(e^{j\omega}; k)$ is the estimation of the psd of the residual echo for the k -th frame.

By using the filter $\hat{H}_r(e^{j\omega}; k)$, an instantaneous estimation of the psd of the residual echo for the next frame can be obtained as

$$\tilde{S}_r(e^{j\omega}; k + 1) = \left(\lambda_e + (1 - \lambda_e) \hat{H}_r(e^{j\omega}; k) \right)^2 \hat{S}_e(e^{j\omega}; k) \quad (15)$$

where $0 \leq \lambda_e \leq 1$ is a bias term that avoids the clipping of any frequency to zero during the estimation of the psd of the residual echo. $\hat{S}_e(e^{j\omega}; k)$ is an estimation of the psd of the error signal.

This estimation is consequent with the fact shown in (11) that the psd of the k -th frame of residual echo depends on the psd of the previous frame of the error signal

Afterwards, we perform an exponential time averaging using using a forgetting factor δ_e

$$\hat{S}_r(e^{j\omega}; k) = \delta_e \hat{S}_r(e^{j\omega}; k - 1) + (1 - \delta_e) \tilde{S}_r(e^{j\omega}; k) \quad (16)$$

5. Simulation Results

In this section, an evaluation of the recursive estimation of the psd of the residual echo is presented.

For the simulation, we used a 600 coefficient real acoustic path impulse response measured in a car and the length of the adaptive filter was 350 coefficients. Several noise free speech recordings were used as passenger's speech and real car noise, recorded while driving on a highway, as background noise. The sampling rate used was 8 KHz, giving a delay of around 20 ms in the LEM path. The length of each signal frame was 16 ms and to reduce the overall delay of the reinforcement system, a time overlap of 75% was used. For comparison purposes, Welch's power spectral estimation technique was used to estimate the true psd of the residual echo with a window length of 16 ms.

According to equations (15) and (16) the estimation of the psd of the residual echo depends on two parameters: the bias term λ_e and the forgetting factor δ_e . These values were chosen to minimize the distortion and to achieve the maximum echo attenuation and speech reinforcement. Defining the time constant as $\tau(ms) = \frac{4}{\ln(\delta_e)}$, values of the forgetting factor δ_e equivalent

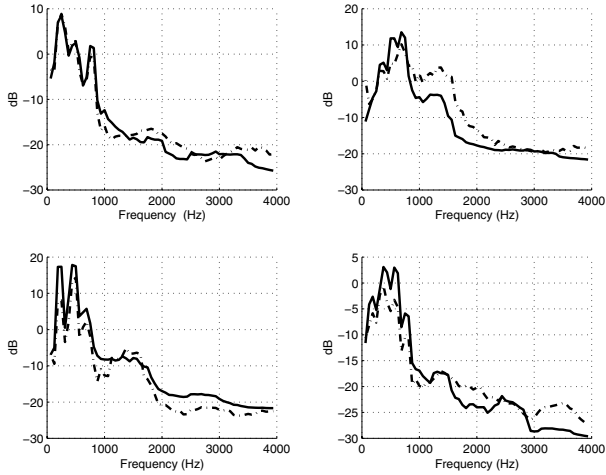


Figure 4: Power spectral density comparison. Real psd (Solid line) and estimated psd (Dashed line).

to time constants between 10 ms and 30 ms are the best ones as can be found in [2]. The variation of the echo attenuation plus the speech reinforcement over different values of λ_e presented in [2] shows that this parameter should be around 0.3.

Figure 4 shows several samples of the psd of the residual echo estimated with the proposed method and the real psd estimated using Welch method. The signal to noise ratio was around 30 dB. As can be seen, the proposed method gives a good approximation of the psd of the residual echo.

In order to evaluate the performance of the estimation we use the mean and the variance of the *Log Estimated-to-True residual echo psd ratio* for each frequency.

$$LETR(e^{j\omega}) = 10 \log \left(\frac{\hat{S}_r(e^{j\omega})}{S_r(e^{j\omega})} \right) \quad (17)$$

A study of this parameters has been carried out with different signal to noise ratios. In Fig. 5 the mean and the variance of the *Log Estimated-to-True residual echo psd ratio* is shown for signal to noise ratios of 10, 20 and 30 dB.

It can be seen that the bias of the estimation is small and practically the same for all the SNR values considered. In the medium and low frequency region, where most of the power of the speech is concentrated, the bias of estimator ranges from -2 dB to 2 dB.

Regarding the variance of this parameter, we can observe from Fig. 5 that it is more sensitive to SNR variations than the mean, although it doesn't present high dependence on the noise power for typical SNR values in a car.

We have also observed that the higher the power of the residual echo, the more accurate the estimation is. This makes our estimation to give a good performance for the reinforcement system.

Several tests have been carried out using the estimated and the true psd of the residual echo for the computation of the ESF. The subjective perception quality is quite similar using the true and estimated psd for values of the speech reinforcement up to 10 dB.

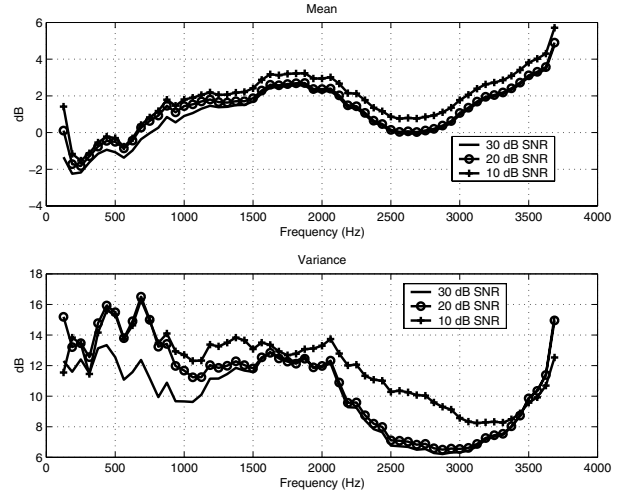


Figure 5: Mean (upper plot) and Variance (lower plot) of the log estimated-to-true residual echo psd ratio for different SNR values from 10 dB to 30 dB

6. Conclusions

In this paper a power spectral density estimator for the residual echo of a speech reinforcement system for vehicles has been presented. This is a challenging task because of the feedback nature of the system. The speech reinforcement system proposed makes use of acoustic echo cancellers and echo suppression filters to cancel the acoustic echo. Due to the permanent double talk situation, the acoustic echo canceller can't converge to a good estimation of the loudspeaker to microphone path. A good estimation of the psd of the residual echo is needed to compute the echo suppression filter that must ensure the stability of the system. Simulation results show that the proposed estimator gives a good approximation to the true psd of the residual echo. The acoustic echo control system allows speech reinforcements from 10 to 15 dB with negligible distortion and echo perception. A full two channel speech reinforcement system has been developed and tested in a medium size car.

7. Acknowledgements

This work has been supported by the project TIC2002-04103-C03-01 from the spanish MCYT.

8. References

- [1] E. Lleida, E. Masgrau, and A. Ortega, "Acoustic echo and noise reduction for cabin car communication," vol. 3, pp. 1585–1588, September 2001.
- [2] A. Ortega, E. Lleida, E. Masgrau, and F. Gallego, "Cabin car communication system to improve communication inside a car," vol. 4, pp. 3836–3839, May 2002.
- [3] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, pp. 21–32, January 1998.