Speech Reinforcement System for Car Cabin Communications

Alfonso Ortega, Eduardo Lleida, *Member, IEEE*, and Enrique Masgrau, *Member, IEEE*

***Abstract*—A speech reinforcement system is presented to improve communication between the front and the rear passengers in large motor vehicles. This type of communication can be difficult due to a number of factors, including distance between speakers, noise and lack of visual contact. The system described makes use of a set of microphones to pick up the speech of each passenger, then it amplifies these signals and plays them back to the cabin through the car audio loudspeaker system. The two main problems are noise amplification and electro-acoustic coupling between loudspeakers and microphones. To overcome these problems the system uses a set of acoustic echo cancellers, echo suppression filters and noise reduction stages. In this paper, the stability of a speech reinforcement system is studied. We propose a solution based on echo cancellers and residual echo suppression filters. The spectral estimation method for the power spectral density of the residual echo existing after the echo canceller is presented along with the derivation of the optimal residual echo suppression filter. Some results about the performance of the proposed system are also provided.**

***Index Terms*—Acoustic echo cancellation, noise reduction, post-filtering, speech enhancement, speech reinforcement.**

I. Introduction

SPEECH communication among passengers in large motor vehicles can be difficult due to the high noise level present inside the car, the distance among passengers, the lack of visual contact between speaker and listener, the use of sound absorbing materials to quiet the cabin, and many other factors. All of these barriers make passengers raise their voices, move out of their normal seating positions, and even more dangerous, distract the driver.

A speech reinforcement system can be used to facilitate communication among passengers inside a car. The goal of a speech reinforcement system is to deliver the voice of a speaker to listeners with sufficient clarity to be understood. The system proposed in this work uses a set of microphones mounted overhead in the cabin to pick up the speech of each passenger. Afterwards, those signals are amplified and played back into the cabin through the car audio loudspeaker system [1]–[3].

There are two main problems with this system.

1) As the distance between loudspeakers and microphones is relatively small, microphones pick up the signal radiated by the loudspeakers which is the origin of acoustic echo. Due to the amplification stage between the microphones and the loudspeakers, there will be acoustic feedback paths that can become unstable. The electro-acoustic coupling between loudspeakers and microphones causes the system to emit annoying high intensity oscillations (or howling) before becoming unstable.
2) The microphones of the system pick up the speech of each passenger as well as the noise inside the cabin. The noise from the engine, wind or road is amplified by the system along with the speech signal and played back into the cabin, thereby increasing the noise level inside the cabin.

Acoustic echo cancellers (AEC) [4] are widely used to overcome the electro-acoustic coupling between loudspeakers and microphones. They consist of an adaptive filter parallel to the loudspeaker-enclosure-microphone (LEM) path and many algorithms can be used to update the filter taps [5].

Nevertheless, to obtain sufficient echo attenuation and quality of the output signal, the system presented in this paper uses an echo suppression filter (ESF). Several techniques have been proposed for further echo attenuation using residual echo reduction filters [6]–[8]. Unlike these methods, the ESF must ensure stability in this closed-loop reinforcement system in addition to further residual echo attenuation. Hearing aid systems are also examples of closed-loop reinforcement systems. However, some of the solutions used to reduce acoustic feedback in hearing aids are not useful for speech reinforcement inside vehicles. Methods such as adaptive notch filters are only effective if the feedback path is relatively narrowband [9]. Noncontinuous adaptive methods [10] can make the system become unstable since the LEM path in the speech reinforcement system changes faster than the feedback path in hearing aids. Other authors have proposed the use of probe noise [11] but this has been known to annoy passengers.

There are many techniques based on one or two channels that are used to reduce noise in the microphone signal. These include power spectral subtraction, spectral amplitude subtraction, and Wiener filtering or adaptive noise cancellation [12]. However single microphone solutions are preferred by manufacturers in order to reduce costs. Thus, a single microphone method is used here to increase the SNR of the output signal based on the optimal Wiener solution.

Recent studies attempt to combine the tasks of echo control and noise reduction. Several structures have been proposed for this purpose [13], [14]. The cascaded structure derived from optimal filtering is used in this system along with a post-filter to reduce the residual echo before the noise reduction stage.

Manuscript received July 7, 2003; revised September 28, 2004. This work was supported under Project TIC2002-04103-C03-01 from MCyT of the Spanish Government, and by the European Technological Center of LEAR Automotive (EED), Valls, Spain. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

The authors are with the Communication Technologies Group of the Aragon Institute of Engineering Research, University of Zaragoza, 50018 Zaragoza, Spain (e-mail: ortega@unizar.es; lleida@unizar.es; masgrau@unizar.es).

Digital Object Identifier 10.1109/TSA.2005.853006

The solution proposed in this paper makes use of a set of echo cancelers (one for each loudspeaker-microphone pair) and one echo suppression filter for each channel that performs further echo attenuation to ensure the stability of the system. A study about the stability of the system and the discussion about the optimal echo suppression filter are presented in this work. In order to avoid noise amplification, the system uses a noise reduction filter based on Wiener optimal filtering theory.

There are two reasons to minimize the overall delay of the system. The first one is due to the Hass effect [15]. The sound coming from the direct path and the reinforced speech coming from the loudspeaker must be fully integrated to maintain intelligibility. The second reason to minimize the delay of the system is because the speaker would hear his or her own voice delayed and amplified. Even if there was a perfect solution to the echo problem, the gain of the system would be limited due to this effect, depending on the delay of the system and the size of the car.

The paper is organized as follows. A brief description of the system is given in Section II. The electro-acoustic coupling problem is considered in Section III and in the Appendix. A discussion about the noise reduction filter (NRF) is included in Section IV. In Section V, implementation issues relating to the echo suppression filter and noise reduction filter are discussed. Performance measures and results will be presented in Section VI. Finally, in Section VII we present the conclusions along with a summary of the paper.

II. System Overview

Full communication inside a car requires at least a two-channel system. One channel carries the speech signal of the rear passengers to the front part of the car and the other channel carries the speech signal of the front passengers to the rear seats. In a three-row vehicle, a three-channel system would be the most appropriate, with one set of microphones and one set of loudspeakers for each row of seats. A block diagram of a two-channel reinforcement system is shown in Fig. 1.

In a two-channel car cabin communication system, each channel must have two acoustic echo cancellers, an echo suppression filter, a noise reduction filter and the amplification stage. Each channel in a car cabin communication system is a mono channel although the car could have more than one loudspeaker for each row of seats. The same signal is played through several loudspeakers to avoid the problem of stereophonic echo cancellation where there is no unique solution for the echo canceller [16].

To find a practical solution to the electro-acoustic coupling problem, a one-channel system is analyzed and the results are applied to a two-channel reinforcement system.

III. Acoustic Echo and Feedback

A. Problem of the Acoustic Echo and the Electro-Acoustic Feedback

Acoustic echo occurs whenever a microphone picks up the far-end signal radiated by a loudspeaker and transmits it back to the far-end. It is common in hands-free telecommunication systems where loudspeakers and microphones are close together

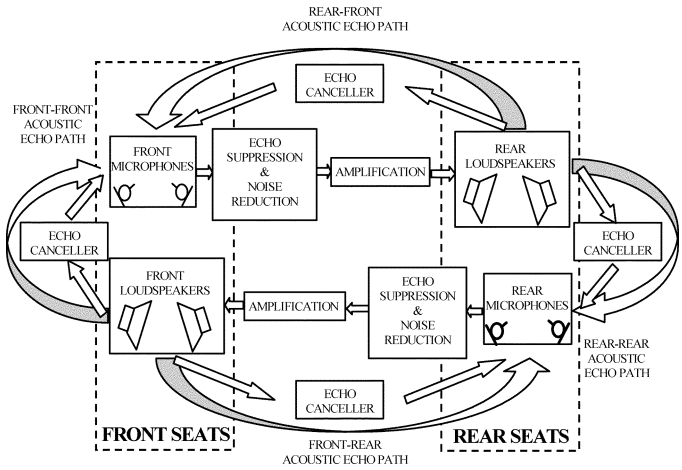

Fig. 1. Block diagram of a two-channel car cabin communication system.

in the same room. An acoustic echo canceller (AEC) can be used to remove the undesired echo signal. The AEC must model the LEM path impulse response with a finite impulse response (FIR) filter in parallel with the LEM path and subtract an echo replica from the microphone signal. A generic AEC is shown in Fig. 2, where $x(n)$ is the far-end speech that goes through the LEM path with impulse response $h(n)$, creating the echo signal $v(n)$. This acoustic echo adds to the near-end speech $s(n)$ and the background noise b(n) to form the microphone signal $d(n)$. The AEC is used to obtain a replica of the acoustic echo $\hat{v}(n)$ from the far-end speech $x(n)$ by filtering it with an adaptive FIR filter $\hat{h}(n)$.

There are many ways of updating the taps of the adaptive filter $\hat{h}(n)$. The choice of method depends on the convergence behavior or computational complexity of the algorithm, among other factors.

When both speakers talk at the same time (near-end and far-end speakers), speech from the local-end can cause the algorithm to produce inaccurate estimates of the impulse response of the LEM path [17], [18]. As a result, the acoustic echo passes through to the far end, a situation known as double-talk. A classical solution is to detect it and freeze the coefficients of the adaptive filter during double-talk periods.

This double-talk situation always exists in the car cabin communication system because the far-end signal is an amplified version of the near-end speech as seen in the block diagram of a one-channel car cabin communication system shown in Fig. 3. Since double-talk is always present, there is little benefit from improving the cancellation precision if adaptation must be performed in the presence of a signal with larger power than the minimum mean square error [9]. For this reason we use reduced complexity adaptive algorithms such as LMS to update the filter weights.

Echo control by means of a traditional AEC in this system is not enough. The estimation of the LEM path carried out by the adaptive filter is inaccurate due to the double talk, and acoustic echo will pass through to the amplification stage. Freezing the coefficients of the adaptive filter can not be the solution. As the system is always in a double talk situation, the coefficients of the adaptive filter would always be frozen.

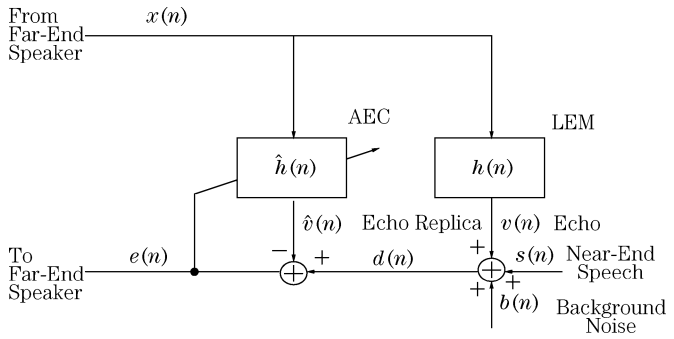

Fig. 2. Diagram of a generic acoustic echo canceller.

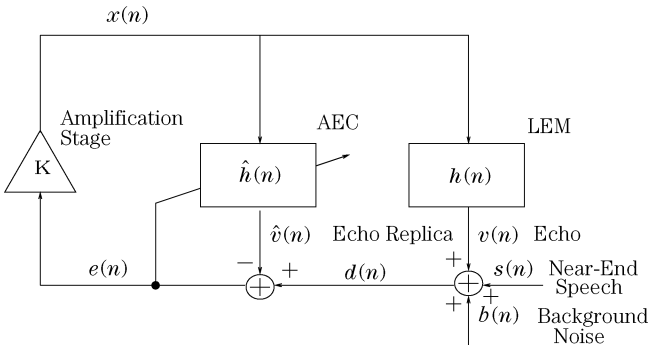

Fig. 3. Diagram of a one-channel CCCS.

This misestimation of the LEM path carried out by the adaptive filter can make the system become unstable since the transfer function of the system in Fig. 3 between the input signal and the output signal in absence of background noise is

$$P_1(e^{j\omega}) = \frac{X(e^{j\omega})}{S(e^{j\omega})} = \frac{K}{1 - K\left[H(e^{j\omega}) - \hat{H}(e^{j\omega})\right]}. \quad (1)$$

The difference between the transfer function of the LEM path $H(e^{j\omega})$ and the transfer function of the adaptive filter $\hat{H}(e^{j\omega})$ can be significant due to double talk. Depending on the value of the gain factor K this difference can make the denominator in (1) approach zero.

As a result, further echo attenuation must be achieved using a filter after the acoustic echo canceller to avoid the system becoming unstable.

B. The Echo Suppression Filter

Besides acoustic echo control by means of a traditional AEC, the car cabin communication system makes use of an echo suppression filter to achieve further echo attenuation. The block diagram of the system with the ESF and the AEC can be seen in Fig. 4. With this ESF and with the noise reduction filter, that will be described in Section IV, the transfer function of the system between the input signal $s(n)+b(n)$ and the output signal $x(n)$ is

$$P_2(e^{j\omega}) = \frac{KW_e(e^{j\omega})W_n(e^{j\omega})}{1 - KW_e(e^{j\omega})W_n(e^{j\omega})\tilde{H}(e^{j\omega})} \quad (2)$$

where $W_e(e^{j\omega})$ is the transfer function of the ESF, $W_n(e^{j\omega})$ is the transfer function of the NRF, and $\tilde{H}(e^{j\omega})$ corresponds to the misadjustment, the difference between the echo path transfer function $H(e^{j\omega})$ and the transfer function of the adaptive filter $\hat{H}(e^{j\omega})$.

To ensure the stability of the system without increasing the noise level inside the cabin, the optimal solution for the ESF can be found by forcing the transfer function of the system to be

$$P_3(e^{j\omega}) = K \cdot W_n(e^{j\omega}). \quad (3)$$

So the optimal expression for the ESF is

$$W_e^o(e^{j\omega}) = \frac{1}{1 + K \cdot W_n(e^{j\omega}) \cdot \tilde{H}(e^{j\omega})}. \quad (4)$$

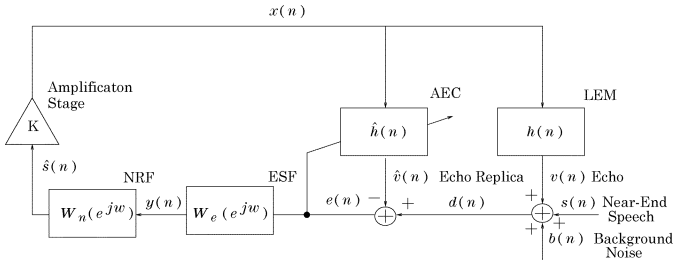

Fig. 4. Block diagram of a CCCS with the echo suppression filter and noise reduction filter.

Although the NRF, $W_n(e^{j\omega})$, obtained from the Wiener filtering theory, is a real valued function, using the true misadjustment function in (4) leads to noncausal ESF. Thus, we use the magnitude of the misadjustment function. This ensures stability (see Appendix) and gives real valued functions for the ESF that follow this expression

$$W_e(e^{j\omega}) = \frac{1}{1 + K \cdot W_n(e^{j\omega}) \cdot \left|\tilde{H}(e^{j\omega})\right|}. \quad (5)$$

According to (5), the optimal ESF depends on the misadjustment function $\tilde{H}(e^{j\omega})$ which is unknown. We estimate the magnitude of this function by using an estimation of the power spectral density (psd) of the residual echo.

The residual echo $r(n)$ can be expressed as the output of a linear system with impulse response

$$\tilde{h}(n) = h(n) - \hat{h}(n) \quad (6)$$

when the input signal is $x(n)$, the output of the reinforcement system.

Thus, the psd of the residual echo is

$$S_r(e^{j\omega}) = S_x(e^{j\omega})\left|\tilde{H}(e^{j\omega})\right|^2 \quad (7)$$

where $S_x(e^{j\omega})$ is the psd of the output signal $x(n)$ and $\tilde{H}(e^{j\omega})$ is the misadjustment function.

According to Fig. 4, we can obtain the psd of the output signal from the psd of the error signal $e(n)$ by using

$$S_x(e^{j\omega}) = S_e(e^{j\omega})\left|KW_e(e^{j\omega})W_n(e^{j\omega})\right|^2. \quad (8)$$

Therefore, the squared magnitude of the misadjustment function is

$$\left|\tilde{H}(e^{j\omega})\right|^2 = \frac{S_r(e^{j\omega})}{\left|KW_e(e^{j\omega})W_n(e^{j\omega})\right|^2 S_e(e^{j\omega})}. \quad (9)$$

Substituting this value of the magnitude of the misadjustment in (5), provides the optimal ESF

$$W_e(e^{j\omega}) = 1 - \sqrt{\frac{S_r(e^{j\omega})}{S_e(e^{j\omega})}} \quad (10)$$

which depends on the psd of the residual echo $S_r(e^{j\omega})$ and the psd of the error signal $S_e(e^{j\omega})$. We can obtain an estimate of $S_e(e^{j\omega})$ since the error signal is directly accessible. However, the procedure to estimate $S_r(e^{j\omega})$ is more elaborate and will be discussed in Section V.

IV. Wiener Noise Reduction Filter

Since the microphones placed on the ceiling of the cabin pick up speech, acoustic echo, and the noise present in the cabin, the reinforcement system must avoid increasing the noise inside the car. In order to minimize the power of the noise before the amplification stage, a Wiener filter, $W_n(e^{j\omega})$, is placed after the ESF as can be seen in Fig. 4.

Assuming that the ESF is working properly

$$y(n) \approx s(n) + b(n) \quad (11)$$

we can express the NRF as follows:

$$W_n(e^{j\omega}) = \frac{S_s(e^{j\omega})}{S_y(e^{j\omega})} \quad (12)$$

where $S_s(e^{j\omega})$ is the psd of the speech signal $s(n)$ and $S_y(e^{j\omega})$ is the psd of the ESF output signal $y(n)$.

Since the speech signal $s(n)$ and the background noise $b(n)$ are statistically independent, the NRF can be defined in terms of the psd of the background noise $S_b(e^{j\omega})$ as follows:

$$W_n(e^{j\omega}) = 1 - \frac{S_b(e^{j\omega})}{S_y(e^{j\omega})}. \quad (13)$$

The discussion about the estimation methods for both power spectral densities can be found in the next section.

V. Power Spectral Densities Estimations

In order to obtain the expressions for the ESF, $W_e(e^{j\omega})$, and the NRF, $W_n(e^{j\omega})$, we need to know the psd of the error signal $e(n)$, the output signal of the ESF $y(n)$, the residual echo $r(n)$ and the background noise $b(n)$. Nevertheless, only the error signal $e(n)$ is directly accessible. This section includes a discussion about the methods followed to obtain the psd needed to calculate $W_e(e^{j\omega})$ and $W_n(e^{j\omega})$.

A. Error Signal Power Spectral Density

An estimate of the short-time psd of the k-th segment of the error signal, $\hat{S}_e(e^{j\omega};k)$, is obtained using the periodogram. In order to reduce the appearance of "musical tones" a Mel scale based frequency smoothing is performed using

$$\tilde{\Gamma}\left(e^{j\frac{2\pi}{N}l};k\right) = \frac{\sum_{i=-M(l)}^{M(l)} \Gamma\left(e^{j\frac{2\pi}{N}(l+i)};k\right)}{2M(l)+1} \quad (14)$$

where $\tilde{\Gamma}(e^{j(2\pi/N)l};k)$ is the smoothed psd, $\Gamma(e^{j(2\pi/N)l};k)$ the psd to be smoothed and $M(l)$ is a frequency dependent index

$$\begin{aligned} M(l) &= m_o \; if \; \frac{l}{N} f_s < 1000 \text{ Hz} \\ M(l) &= 2m_o + 1 \; if \; 1000 \text{ Hz} < \frac{l}{N} f_s < 2500 \text{ Hz} \\ M(l) &= 4m_o + 1 \; if \; \frac{l}{N} f_s > 2500 \text{ Hz} \end{aligned} \quad (15)$$

being N the FFT length and f_s the sampling frequency in Hertz. The number of bins used for the averaging process in each band depends on the index m_o which must be selected to be large enough to reduce the effect of "musical tones" and small enough to maintain adequate frequency resolution. The appearance of musical noise is due to the existence of energy in narrow and isolated bands. This energy can be spread over broader bands by performing this frequency smoothing.

B. Residual Echo Power Spectral Density

The procedure for estimating the psd of the kth frame of residual echo $\hat{S}_r(e^{j\omega};k)$ is quite elaborate since this signal is not directly accessible.

An estimation of the psd of $r(n)$ can be obtained from the estimation of the power spectral densities of the error signal, the microphone signal, and the estimated echo as described in [6], [7]. In this paper, we present a recursive method for estimating the psd of the residual echo from the error signal $e(n)$.

The LEM path can be modeled as a delay block of Δ samples followed by a linear filter $h'(n)$. The delay block will model 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 adaptive filter are set to zero.

Fig. 5 shows a 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 filter composed of the ESF and the NRF.

Assuming stationarity over short time intervals, we want to estimate the short-time psd of the k-th segment of L samples of residual echo $S_r(e^{j\omega};k)$. For this we use the optimal Wiener solution

$$H_r(e^{j\omega};k) = \frac{S_{re}(e^{j\omega};k)}{S_e(e^{j\omega};k)} \quad (16)$$

where $S_e(e^{j\omega};k)$ is the psd of the error signal and $S_{re}(e^{j\omega};k)$ is the cross-power spectral density of the residual echo and the error signal for the k-th segment. We can express this short-time cross-power spectral density as the Fourier transform of the short-time cross-correlation function

$$R_{re}(m;kD) = E\left[r(n;kD)e(n-m;kD)\right] \quad (17)$$

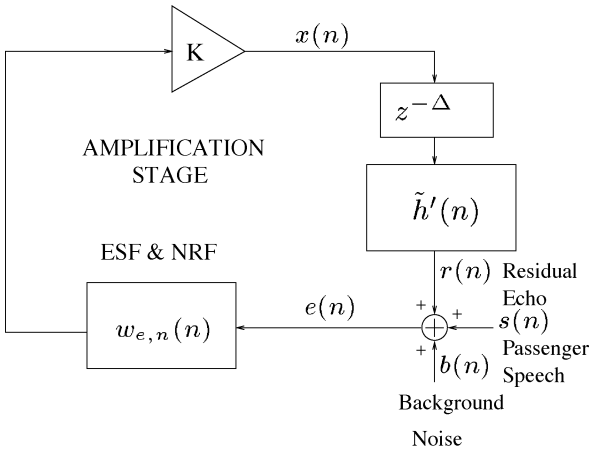

Fig. 5. Simplified diagram of a one-channel speech reinforcement system.

where $E[\cdot]$ denotes expectation of the quantity within the brackets and $(n;kD)$ denotes the n-th sample of the frame starting at sample kD.

The error signal, as can be seen in Fig. 5, 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 function of the residual echo and the error signal, can be expressed as

$$R_{re}(m;kD) = E\left[r(n;kD)r(n-m;kD)\right] + E\left[r(n;kD)s(n-m;kD)\right]. \quad (18)$$

According to Fig. 5, we can express the residual echo as

$$r(n) = K \cdot e(n-\Delta) * \tilde{h}''(n) \quad (19)$$

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

Without loss of generality, consider the delay Δ to be equal to p times D, where D is the time shift between consecutive frames, and p an integer greater than one. Therefore, the k-th frame of residual echo $r(n;kD)$ depends on a previous frame of error signal, $e(n;(k-p)D)$. Thus, the short-time cross-correlation function of the k-th frame of the residual echo and the error signal is

$$R_{re}(m;kD) = E\left[r(n;kD)r(n-m;kD)\right] + KE\left[e\left(n;(k-p)D\right)s(n-m;kD)\right] * \tilde{h}''(n). \quad (20)$$

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

$$\begin{aligned} R_{re}(m;kD) = {} & E\left[r(n;kD)r(n-m;kD)\right] \\ & + KE\left[s\left(n;(k-p)D\right)s(n-m;kD)\right] * \tilde{h}''(n) \\ & + KE\left[r\left(n;(k-p)D\right)s(n-m;kD)\right] * \tilde{h}''(n). \end{aligned} \quad (21)$$

The second and third terms depend on the correlation between different frames, which is expected to be small since speech is not stationary. As the time of separation is long enough these two terms can be considered negligible compared to the first term. Therefore, the short-time cross-power spectral density of the residual echo and the error signal, can be considered to be $S_{re}(e^{j\omega};k) = S_r(e^{j\omega};k)$. Thus, an estimation of the Wiener filter can be obtained by using

$$\hat{H}_r(e^{j\omega};k) = \frac{\hat{S}_r(e^{j\omega};k)}{\hat{S}_e(e^{j\omega};k)} \quad (22)$$

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 kth frame.

An instantaneous estimation of the psd of the residual echo for the next frame can be obtained from the psd of the error signal by using the filter $\hat{H}_r(e^{j\omega};k)$

$$\tilde{S}_r(e^{j\omega};k+1) = \left[\lambda_e + (1-\lambda_e)\hat{H}_r(e^{j\omega};k)\right]^2 \hat{S}_e(e^{j\omega};k) \quad (23)$$

where $0 \le \lambda_e \le 1$ is a bias term that avoids the clipping of any frequency to zero during the estimation of the psd of the residual echo. Afterwards, we perform an exponential time averaging using a forgetting factor δ_e

$$\hat{S}_r(e^{j\omega};k) = \delta_e \cdot \hat{S}_r(e^{j\omega};k-1) + (1-\delta_e) \cdot \tilde{S}_r(e^{j\omega};k). \quad (24)$$

Finally, according to (10), this estimated psd of the k-th segment of residual echo is used to compute the ESF using

$$W_e(e^{j\omega};k) = 1 - \sqrt{\frac{\hat{S}_r(e^{j\omega};k)}{\hat{S}_e(e^{j\omega};k)}}. \quad (25)$$

C. Background Noise Power Spectral Density

The psd of the background noise is estimated to obtain the NRF in a similar way as the residual echo estimation. The NRF in (13) for the kth segment can be expressed as

$$W_n(e^{j\omega};k) = 1 - \hat{H}_n(e^{j\omega};k) \quad (26)$$

where $\hat{H}_n(e^{j\omega};k)$ is a Wiener filter that estimates the psd of the background noise, $S_b(e^{j\omega};k)$, from the output signal of the ESF, $y(n)$, and can be expressed as follows:

$$\hat{H}_n(e^{j\omega};k) = \frac{\hat{S}_b(e^{j\omega};k)}{\hat{S}_y(e^{j\omega};k)} \quad (27)$$

where $\hat{S}_b(e^{j\omega};k)$ is an estimation of the psd of the background noise and $\hat{S}_y(e^{j\omega};k)$ is the estimated psd of the output signal of the ESF, $y(n)$.

The estimation of the psd of $y(n)$ is obtained from the instantaneous estimation of the psd of the residual echo, $\tilde{S}_r(e^{j\omega};k)$, defined in (23) using the following relationship:

$$\tilde{S}_y(e^{j\omega};k) = \hat{S}_e(e^{j\omega};k) - \tilde{S}_r(e^{j\omega};k). \quad (28)$$

Afterwards, a time average is performed in order to reduce the variance of the estimation

$$\hat{S}_y(e^{j\omega};k) = \delta_y \cdot \hat{S}_y(e^{j\omega};k-1) + (1-\delta_y) \cdot \tilde{S}_y(e^{j\omega};k). \quad (29)$$

An instantaneous estimate of the psd of the background noise is obtained from the psd of the signal $y(n)$ using the Wiener filter in (27)

$$\tilde{S}_b(e^{j\omega};k) = \left[\lambda_n + (1-\lambda_n)\cdot \hat{H}_n(e^{j\omega};k)\right]^2 \hat{S}_y(e^{j\omega};k) \tag{30}$$

where $0 \leq \lambda_n \leq 1$ is the bias term that avoids the clipping of any frequency to zero.

Finally, because the statistics of the background noise change very slowly over time, noise can be considered fairly stationary compared to speech [6]. Thus, an exponential time average is performed using a forgetting factor δ_n close to one

$$\hat{S}_b(e^{j\omega};k) = \delta_n \cdot \hat{S}_b(e^{j\omega};k-1) + (1-\delta_n)\cdot \tilde{S}_b(e^{j\omega};k). \tag{31}$$

After estimating the psd for the k-th frame of background noise, it is used to compute the NRF

$$W_n(e^{j\omega};k) = 1 - \frac{\hat{S}_b(e^{j\omega};k)}{\hat{S}_y(e^{j\omega};k)}. \tag{32}$$

VI. Performance Measures and Results

To evaluate the performance of the reinforcement system presented here, a one channel system was implemented and tested using a real loudspeaker-enclosure-microphone path impulse response measured in a car. The sampling rate was 8 kHz and the length of the impulse response was 650 coefficients. An adaptive filter of 350 coefficients was used with a delay line of 50 samples ($\Delta = 50$ samples according to Fig. 5). The length of each frame was $L = 128$ samples and a time shift of $D = 32$ samples was used to keep the overall delay below a few milliseconds. For the Mel scale based frequency smoothing in (14) and (15), $m_o = 2$ and the FFT length, N, was 128 points.

The simulation setup is shown in Fig. 6. In this diagram, $s(n)$ is the speech signal to be amplified, $b(n)$ is the background noise, $d(n)$ is the microphone signal (composed of speech, noise and echo) and $\hat{s}(n)$ is the estimated speech signal that will be amplified by the system becoming $x(n)$, the output signal.

Intermediate signals are also generated to be able to perform useful measures. These include $v(n)$, the acoustic echo signal, and $r(n)$, the residual echo signal which is the difference between the echo signal, $v(n)$, and the estimated echo.

These intermediate signals along with the speech signal, $s(n)$, and the background noise, $b(n)$, are filtered with several copies of the post-filter. The post-filter is composed of the echo suppression filter and the noise reduction filter with an equivalent transfer function $W(z)$. This allows us to compute performance measures such as the speech reinforcement, the echo attenuation, or the signal to noise ratio improvement. Thus, each component of the microphone signal, the acoustic echo, and the residual echo can be processed separately.

The indexes presented here were obtained during double-talk periods, detected by a simple energy voice activity detector (VAD in Fig. 6) operating on the noise free speech signal $s(n)$ using 30 ms long frames.

Several noise free phonetically balanced sentences were used for the simulations. These sentences belong to the ALBAYZIN speech database [19]. This corpus is a Spanish spoken database designed for speech recognition purposes. An equal number of male and female speakers were chosen.

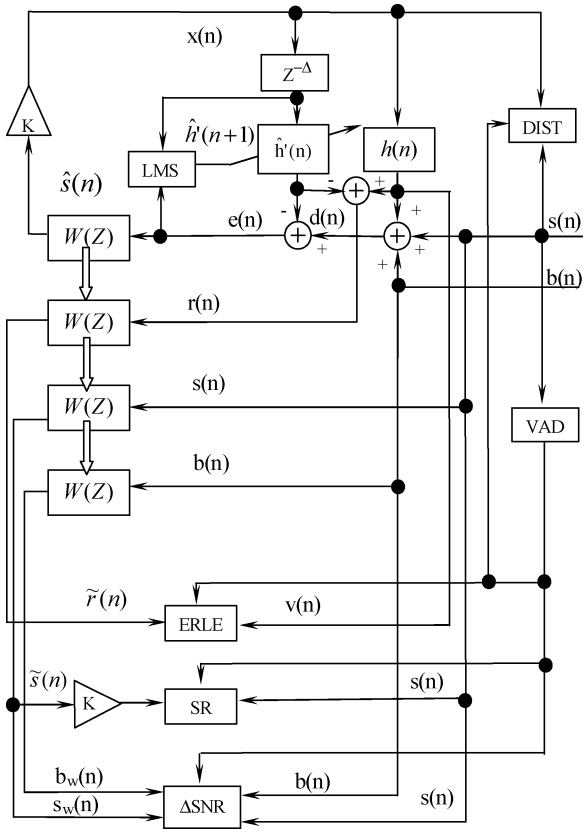

Fig. 6. Simulation setup for performance measure.

The acoustic echo power reduction is measured using the echo-return loss enhancement (ERLE) defined as

$$ERLE = 10\cdot \log_{10}\left(\frac{1}{N}\sum_{k=0}^{N-1}\frac{E\left[v^2(n;k)\right]}{E\left[\tilde{r}^2(n;k)\right]}\right) \tag{33}$$

where $v(n)$ is the acoustic echo signal and $\tilde{r}(n)$ is the residual echo signal (the output of the filter composed of the ESF and the NRF when the input is the residual echo existing after the AEC, $r(n)$).

The feedback stability margin (FSM) was used to calculate how close the reinforcement system is to becoming unstable. To compute this index, the open loop echo gain (OLEG) must be obtained first. The OLEG is the gain of the system in open-loop operation considering only the echo signal

$$OLEG = 10\cdot \log_{10}\left(\frac{E\left[K^2\tilde{r}^2(n)\right]}{E\left[v^2(n)\right]}\right). \tag{34}$$

Without the acoustic echo canceller and the ESF, the maximum attainable gain value is $K = 0.4$ for the LEM path impulse response used in the simulations. Higher values of K make the system become unstable. The maximum value of OLEG is

$$OLEG_{max} = 20\cdot \log_{10}(0.4) = -7.9588. \tag{35}$$

Thus, the feedback stability margin is defined as

$$FSM = OLEG_{max} - OLEG(K). \tag{36}$$

As this value approaches zero, the system is closer to instability.

We define the speech reinforcement (SR) as the ratio of the speech signal power gain and the maximum value of the reinforcement that could be obtained without the AEC and the ESF.

The distortion was evaluated by means of the symmetric Itakura distance [12] between the input speech signal, $s(n)$, and the output signal $x(n)$ with a linear prediction model of 10 coefficients.

A. Echo Suppression Filter Performance

According to (23) and (24) the estimation of the psd of the residual echo depends on two important parameters, δ_e and λ_e.

These parameters can be selected by maximizing ERLE and SR and minimizing distortion. It would be useful to obtain vehicle specific values for these parameters along with the delay Δ that is used before the adaptive filter. That is, it is expected that the best performance of the speech reinforcement system would be obtained after tuning the system for each car. Nevertheless it can be useful to show some simulation results corresponding to different situations (different types of speech, noise and cars) in this section.

Four different impulse responses have been used for these simulations. Two impulse responses corresponding to a medium size car (the first one, from the front loudspeakers to the rear microphones and the second one, from the rear loudspeakers to the front microphones), and the other two, corresponding to a large size car (front-rear and rear-front).

The first parameter to be studied is the forgetting factor (δ_e) used for the exponential time average made over the instantaneous estimates of the psd of the residual echo.

Fig. 7 shows the SR and the ERLE along with the symmetric Itakura distance between $s(n)$ and $x(n)$ as a function of the time constant τ.

Defining the time constant as $\tau(ms) = -(4/\ln(\delta_e))$, values of δ_e equivalent to time constants between 10 ms and 30 ms are the best ones to maximize echo reduction and speech reinforcement minimizing distortion as can be seen in Fig. 7.

The second parameter, λ_e, is the bias term used in the instantaneous estimation of the psd of the residual echo, in order to avoid the clipping of any frequency to zero. Maximizing ERLE and SR and minimizing distortion, we set this parameter around 0.3 as can be seen in Fig. 8.

In Fig. 9, the ERLE is plotted as function of the gain factor K with the ESF and NRF, without the NRF and without NRF and ESF. The ERLE is more than 10 dB higher with the ESF and NRF than only with the AEC which implies an improvement in the FSM. The FSM decreases as the gain factor K increases, unlike the SR that increases with K. Fig. 10 shows the evolution of the FSM and the SR as a function of the gain factor K. As expected, the system is closer to instability as SR increases (as indicated by the FSM). From the simulation experiments, we found that it is possible to obtain SR values around 6 dB with a FSM greater than 6 dB, and SR around 12 dB with a FSM of 1 dB. To obtain these values of SR and FSM the gain factor K must be between 2 and 5.

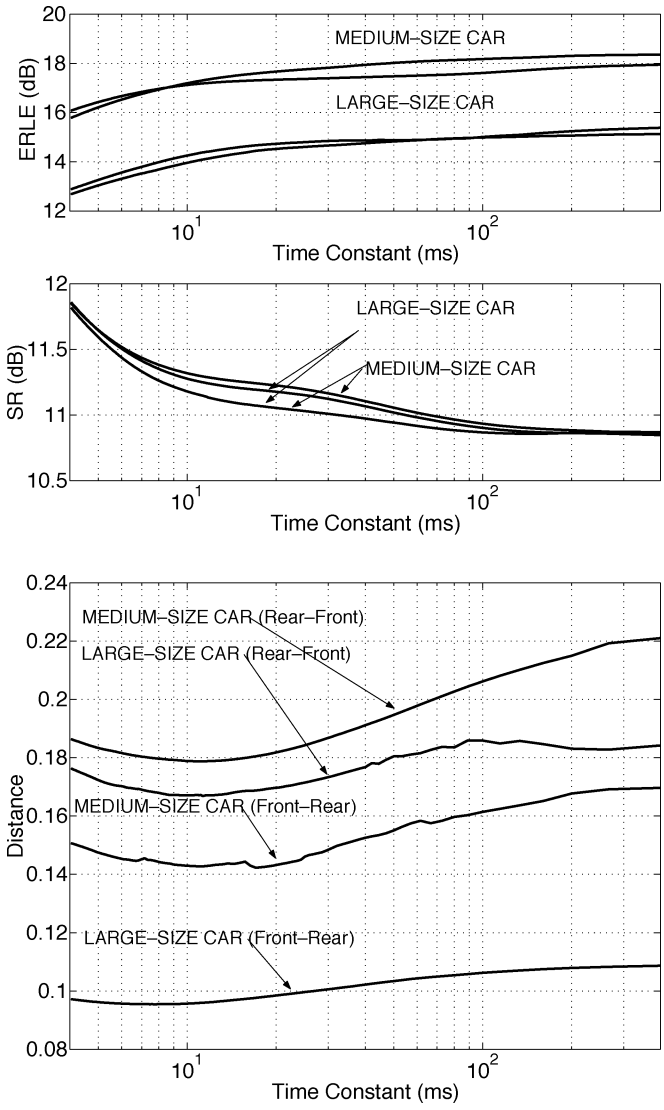

Fig. 7. ERLE, SR, and symmetric Itakura distance over different values of δ_e, evaluated with four different impulse responses.

B. Noise Reduction Performance

The parameter λ_n (bias term) used for this simulation was 0.3 and the forgetting factor δ_n as mentioned in Section V-C was chosen close to 1 (0.9995). Several indexes were used to analyze how the NRF performs enough noise reduction to avoid increasing the noise level inside the cabin.

- The noise attenuation (NA) defined as the ratio of the input noise power to the output noise power

$$NA = \frac{1}{N}\sum_{k=0}^{N-1} 10\log_{10}\left(\frac{E\left[b^2(n;k)\right]}{E\left[b_w^2(n;k)\right]}\right). \tag{37}$$

Two different measures were made using this index, the noise attenuation during silent frames NA_{sil}, and noise attenuation during speech frames NA_{sp}.

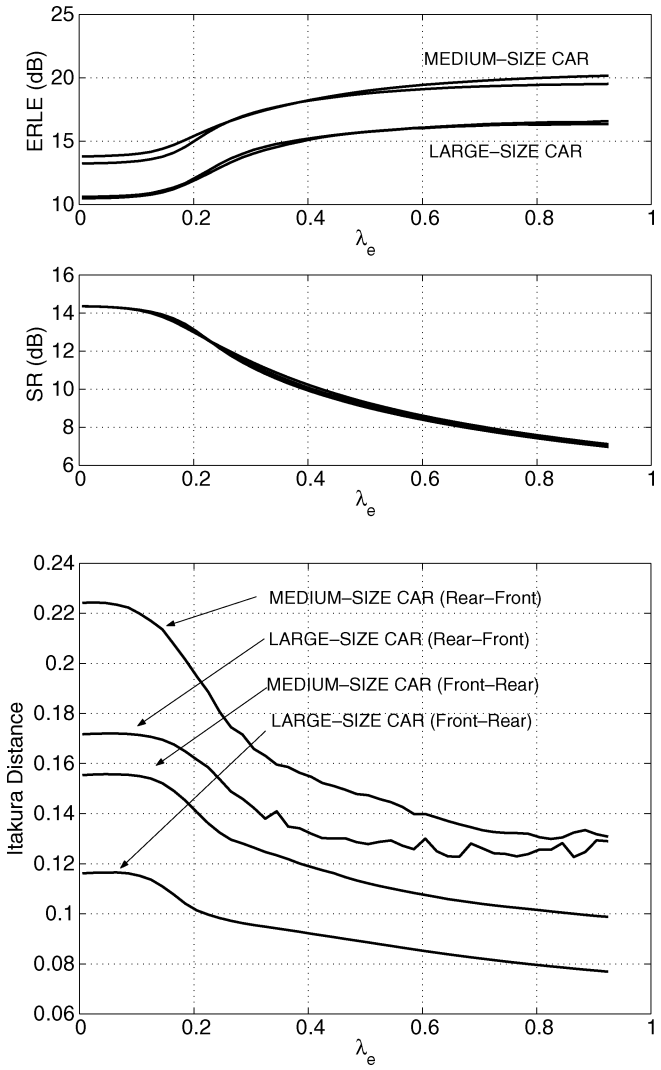

Fig. 8. ERLE, SR, and Itakura distance over different values of λ_e, evaluated with four different impulse responses.

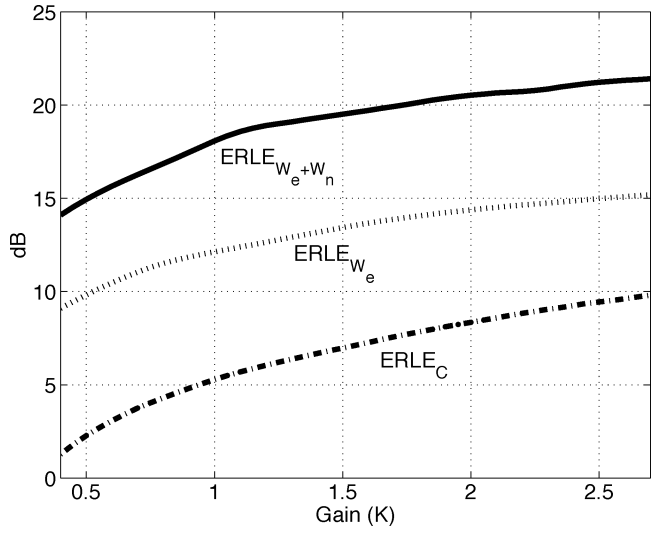

Fig. 9. ERLE evolution over different values of the gain factor K, with the ESF and the NRF (solid line), with (dotted line) the ESF and without the ESF and NRF (dash-dot line).

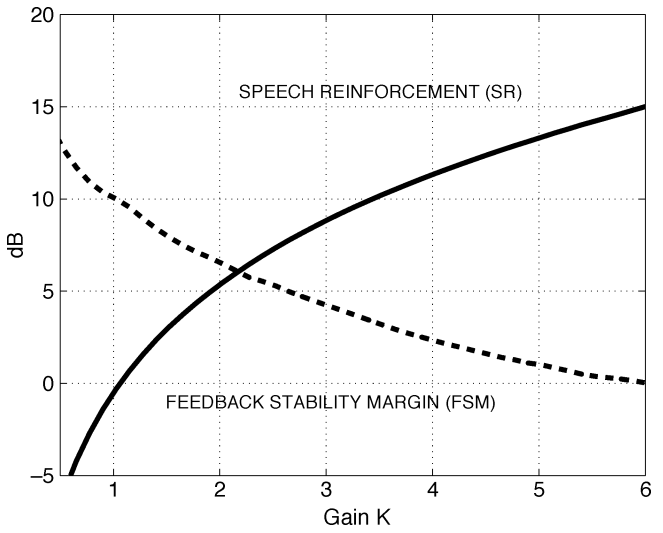

Fig. 10. Speech reinforcement and feedback stability margin.

- The speech attenuation (SA), defined as the ratio of the input speech power to the output speech power

$$SA = \frac{1}{N}\sum_{k=0}^{N-1} 10\log_{10}\left(\frac{E\left[s^2(n;k)\right]}{E\left[s_w^2(n;k)\right]}\right). \tag{38}$$

- The improvement of the segmental signal to noise ratio defined as the difference between the output SNR and the input SNR (both measured as segmental SNR)

$$\Delta SEGSNR = SEGSNR_o - SEGSNR_i \tag{39}$$

where $SEGSNR_o$ is the output segmental SNR

$$SEGSNR_o = \frac{1}{N}\sum_{k=0}^{N-1} 10\log_{10}\left(\frac{E\left[s_w^2(n;k)\right]}{E\left[b_w^2(n;k)\right]}\right) \tag{40}$$

and $SEGSNR_i$ is the input segmental SNR

$$SEGSNR_i = \frac{1}{N}\sum_{k=0}^{N-1} 10\log_{10}\left(\frac{E\left[s^2(n;k)\right]}{E\left[b^2(n;k)\right]}\right). \tag{41}$$

Several noises were added to noise free speech signals to evaluate the performance of the noise reduction filter. Three different situations were considered, car stopped with the engine on, town traffic and car noise recorded while driving on a highway.

In Fig. 11, the mean values of NA_{sil}, NA_{sp}, SA and $\Delta SEGSNR$ are plotted as functions of the input SNR. These measures do not vary significantly for different values of the input SNR except NA_{sil} that ranges from 25 dB to 30 dB for input SNR ranging from 0 dB to 20 dB. During speech frames the NRF presents a NA around 17 dB while the SA is around 10 dB. This increases the signal to noise ratio around 6 dB as can be seen in the segmental signal to noise ratio improvement shown in Fig. 11. As also seen in Fig. 11, during silent periods, the NA is above 27 dB when the input SNR is above 10 dB. Thus, the speech reinforcement system will not increase the noise level inside the cabin what could be very annoying for the passengers. Considering different noise conditions (car

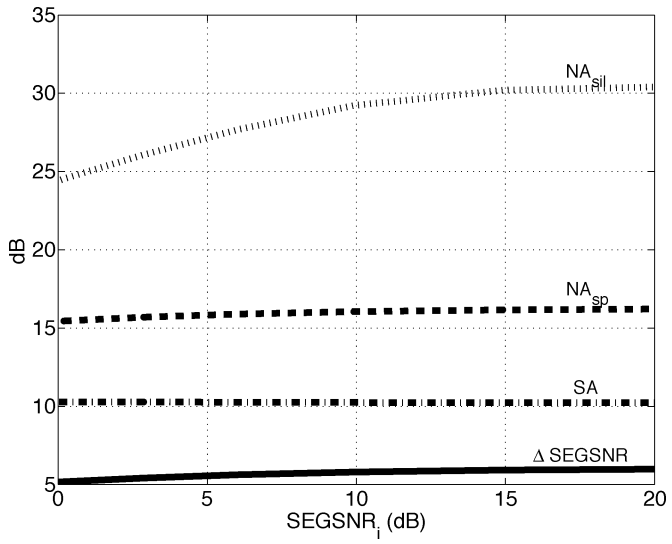

Fig. 11. Mean noise attenuation during silent frames (dotted line), noise attenuation during speech frames (dashed line), speech attenuation (dashed-dotted line) and segmental signal to noise ration improvement (solid line) as functions of the input SNR.

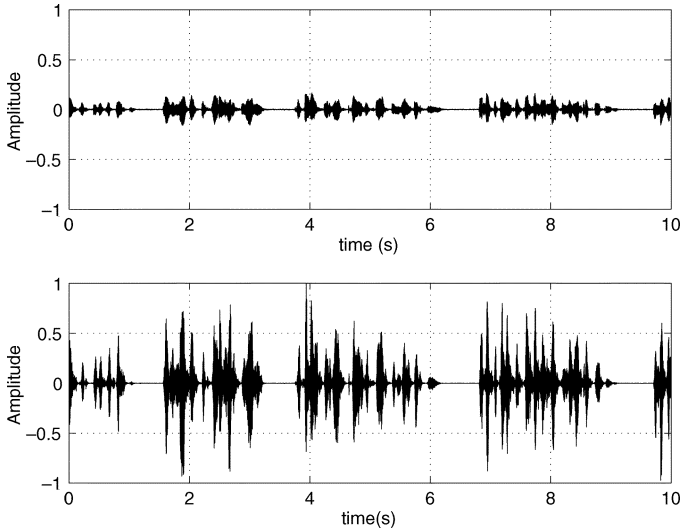

Fig. 12. Example of output speech signals. Reinforcement system without processing and maximum gain ($K = 0.35$) (top) and proposed speech reinforcement system with $K = 4.0$ (bottom).

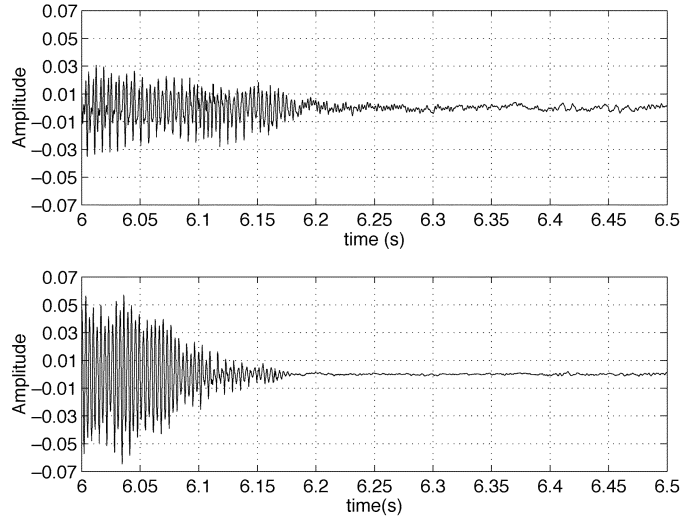

Fig. 13. Example of output speech signals during silent intervals. Reinforcement system without processing and maximum gain ($K = 0.35$) (top) and proposed speech reinforcement system with $K = 4.0$ (bottom).

stopped with the engine on, town traffic and highway noise) with different noise and speech samples, the standard deviation of the indexes presented in Fig. 11 is less than 2 dB.

C. Speech Reinforcement Evaluation

Several simulations were carried out to evaluate the speech reinforcement achieved with the car cabin communication system proposed in this paper.

We considered two situations and compared the output signals. The first was a speech reinforcement system without processing, only an amplification stage between the microphone and the loudspeaker. In this situation, the maximum gain factor K that ensures stability is around 0.4, although the quality of the output speech is very poor due to the appearance of strong tonal components (howling). Smaller values for the gain factor K must be used in order to improve the quality of the output signal. A gain factor $K = 0.35$ was used. The second situation was the speech reinforcement system presented in this work, with the AEC, ESF, and NRF. In this case, the gain factor K used to obtain the output signal was 4.0 for the set-up considered. Greater values of K can be used but distortion with gain factors greater than 5.0 can make the output signal sound uncomfortable, the system is about to become unstable and howling is present, the speech sounds as being said into a rain barrel with a lot of reverberation. In both situations, an input SNR around 12 dB was used by adding real car noise recorded inside a car while driving on a highway to a noise free speech signal.

In Fig. 12, two segments of speech are plotted. The first one was obtained at the output of a speech reinforcement system without echo control and noise reduction (top). The second one was obtained at the output of the proposed reinforcement system (bottom). Fig. 13 shows the end of a speech segment and the beginning of a silent interval. The improvement in speech reinforcement using the AEC, the ESF and the NRF is quite significant with no amplification of the noise in the silent intervals.

To show the spectral alteration of the original signal after passing through the reinforcement system, Fig. 14 illustrates the ratio between an estimate of the psd of the output signal of the proposed system with $K = 4.0$, and the psd of the original signal. The SNR of the original signal was around 12 dB and the time duration of the input sentence was 10 s. This ratio is around 15 dB in most of the system bandwidth. The gain of the system in the lower bands is negative due to the level of noise present in this frequency region and the upper bands are also attenuated because of the anti-aliasing filters of the system falling near 3.7 kHz. This plot shows the frequency dependence of the SR index defined in Section VI for $K = 4.0$.

D. Intelligibility Test

The main purpose of the speech reinforcement system for cars is to improve communications among passengers. Therefore, the intelligibility inside the car must be increased by using this

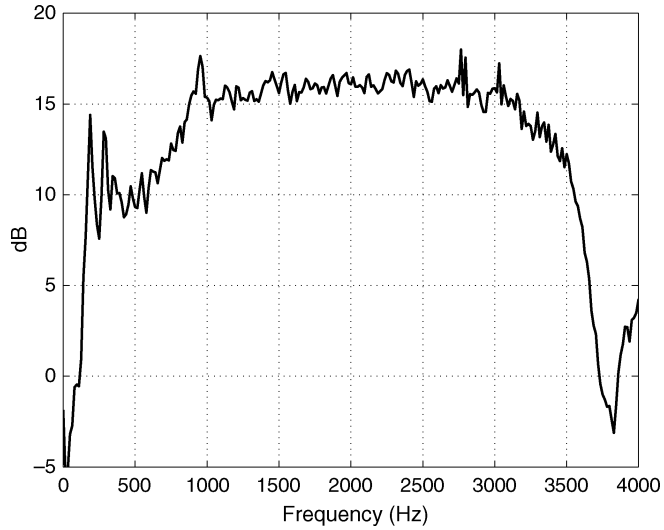

Fig. 14. Frequency dependence of the speech reinforcement index for $K = 4.0$.

car cabin communication system. Speech intelligibility depends on many factors such as the characteristics of the talkers and listeners, the speech material and the properties of the transmission room (reverberation time, signal to noise ratio, distance from the source, etc.) [20]. In order to measure the intelligibility inside a car for different situations, the Speech Transmission Index (STI) has been considered [21]. This index ranges from 0 to 1 and values below 0.6 can be considered as fair (down to 0.45), poor (from 0.3 to 0.45) or bad (below 0.3), meanwhile values from 0.6 to 0.75 can be considered good and above 0.75 excellent [22].

The STI is a physical characteristic of a sound transmission channel which is based on the Modulation Transfer Function (MTF) concept [21]. The MTF has been successfully applied to measure the speech intelligibility. It reflects the effect of reverberation and also the presence of ambient noise. The concept behind MTF is to measure the reduction of the modulation index of the intensity envelope of a signal recorded at the listener's position relative to that of the original signal. A synthetic signal has been used which consists of a speech shaped noise modulated by low frequency sine waves.

To evaluate the STI in the context of the proposed speech reinforcement system, different situations have been considered corresponding to four different setups defined in Table I. The delay between de direct path and the reinforced speech is around 7 ms.

The evaluation has been carried out for different values of the gain factor K and for different signal to noise ratios. The values of STI used as reference are the ones that would be obtained for a communication inside a car without a speech reinforcement system.

Table II shows the values of the STI obtained for different values of K, for all the defined setups with a segmental signal to noise ratio of 10 dB at the microphone position. The gain factor K ranges from 0.35 to 3.0. Higher values of K do not increase STI and values above 4.0 can decrease it because of the appearance of howling in the reinforced speech. Car noise recorded while driving on a highway was added to the test signal

TABLE I
SETUP DEFINITION FOR THE INTELLIGIBILITY TEST

	AEC	ESF	NRF
SETUP A	No	No	No
SETUP B	Yes	No	No
SETUP C	Yes	Yes	No
SETUP D	Yes	Yes	Yes

TABLE II
STI FOR $SEGSNR_i = 10$ dB AND 20 dB OF SNR DEGRADATION BETWEEN MICROPHONE POSITION AND LISTENER POSITION

K	SETUP A	SETUP B	SETUP C	SETUP D
0.35	0.71	0.74	0.71	0.71
1.0	Unstable	0.81	0.81	0.82
3.0	Unstable	0.73	0.82	0.86
Without Speech Reinforcement System				0.56

TABLE III
STI VARIATION WITH THE INPUT SNR FOR THE PROPOSED SYSTEM ($K = 3.0$)

$SEGSNR_i$	-5 dB	0 dB	5 dB	10 dB	15 dB
STI	0.76	0.82	0.84	0.86	0.89

to perform the evaluation. The degradation of the signal to noise ratio between the microphone position and the listener position was around 20 dB. The reference value of the STI for this configuration was 0.56.

It can be observed how STI increases when a speech reinforcement system is used. The need for the whole system can be seen in the last row of Table II when high levels of reinforcement are required.

The values of the STI achieved with the speech reinforcement system proposed in this paper are shown in Table III for different values of the input segmental signal to noise ratio measured at the microphone position. The gain parameter of the system was $K = 3.0$ and setup D was used.

The evolution of the intelligibility improvement achieved with the proposed system ($K = 3.0$) as a function of the segmental SNR at the microphone position for different values of the SNR degradation between the microphone position and the listener position, is shown in Fig. 15. This improvement is obtained as the difference between the STI measured using the speech reinforcement system (STI_o) and STI without the reinforcement system (STI_i), divided by STI_i: $\Delta STI(\%) = (STI_o - STI_i)/(STI_i)100$.

From Fig. 15, it can be concluded that the improvement of the STI achieved with the speech reinforcement system is very important when the degradation of the sound coming from direct path is high.

E. Two-Channel Real-Time Implementation

In addition to evaluating the system by means of simulation, a two channel real time system was built in a dual high-performance floating point DSP board according to the block diagram shown in Fig. 1. It was tested in medium and large size cars. In this real time system, the bandwidth of the speech signal was extended to 8 kHz, so we used a sampling frequency of 16 kHz.

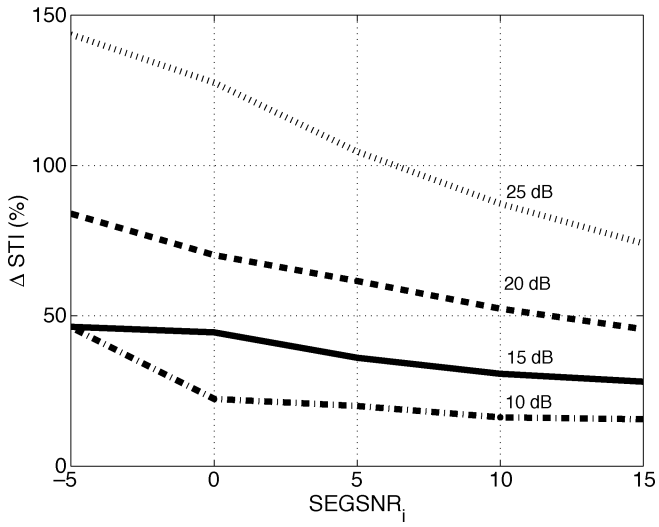

Fig. 15. Evolution of the STI improvement as a function of the input segmental SNR at the microphone position for different values of the SNR degradation between the microphone position and the listener position.

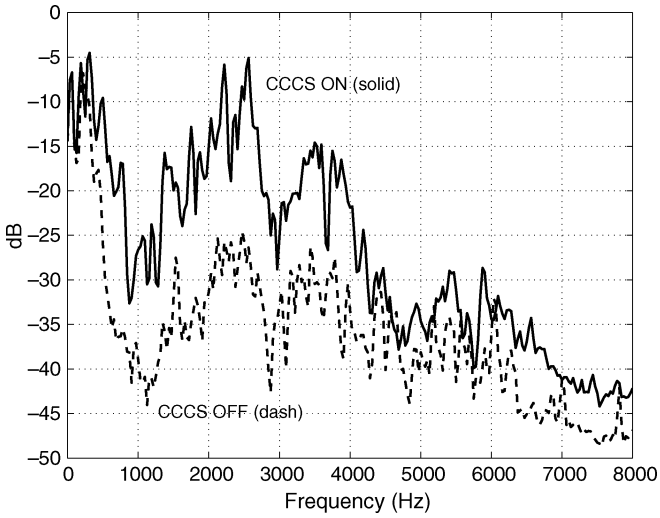

Fig. 16. Average psd of the recorded signal with (solid line) and without (dashed line) the reinforcement system. The car was stopped and the engine off.

The gains (K_R and K_F) of each channel can be controlled manually by the passengers but they also depend on the speed of the car and are automatically modified by the system.

A speed dependent variable ($\Delta G(k)$) is added to the gain factor selected by the user. The value of this variable varies linearly with speed up to 80 km/h and ranges from 0 (when the car is stopped) to 0.32 (when the speed of the car is 80 km/h). From 80 km/h it becomes constant and equal to 0.32. This increase in the gain factor is not instantaneous but time averaged in order to avoid sudden changes in the gain factor according to

$$\Delta G(k) = \alpha \Delta G(k-1) + (1-\alpha) 0.004 \text{ Current Speed(km/h)}. \tag{42}$$

Four directional microphones were mounted on the overhead of the cabin to pick up the speech of each passenger (two in the front and two in the rear). The microphone signals from the front are added and played back by the rear loudspeakers and the rear

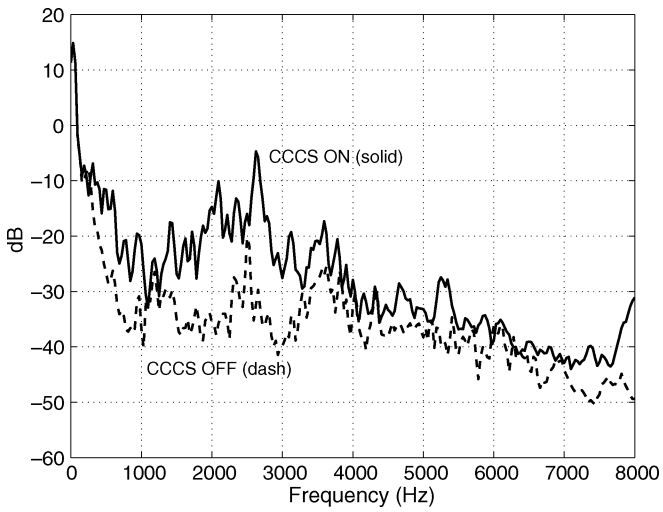

Fig. 17. Average psd of the recorded speech signal with (solid line) and without (dashed line) with the car stopped and the engine on.

microphone signals are also added and played back by the front loudspeakers. To measure the speech reinforcement achieved with this two-channel real-time system, several recordings were made in a large size vehicle under different test conditions. We recorded the output of the rear loudspeakers with a microphone placed 50 cm away from one of the rear loudspeakers, while several sentences were played with a loudspeaker placed in the position of the driver's head. For these recorded signals, the average psd was computed using the Welch method. These estimates were computed using a Hanning window, a 512 point FFT and an overlap of 50% on speech segments of 5 s. Fig. 16 shows the average psd of the speech reinforcement system output when the system is OFF (dashed line) and when the system is ON (solid line). The car was stopped and the engine was off. These are the worst conditions for the stability of the system because during silent segments of the speech input, there is no signal to drive the system.

Fig. 17 shows the same experiment with the engine ON. There is no sound reinforcement in the low frequency region, which contains most of the engine noise power. The reinforcement in both situations (engine ON and OFF) ranges from 10 to 15 dB in the frequency region between 300 Hz and 4 kHz, where most of the speech power is allocated.

Estimates of the speech reinforcement obtained with and without the engine ON are plotted in Fig. 18. The difference between both situations is not significant. In the frequency region from 300 Hz to 4 kHz, the average speech reinforcement is around 15 dB. Thus, these measures show that the two-channel real time implementation and the one-channel simulator have similar performance in the frequency region between 0 to 4000 Hz.

We performed informal tests under several conditions, different speeds ranging from 50 km/h to 120 km/h, different road types, etc. At high speeds and poor road conditions, speech intelligibility is highly degraded and the use of this reinforcement system really improves the communication among the passengers without amplifying the noise and avoiding the appearance of howling.

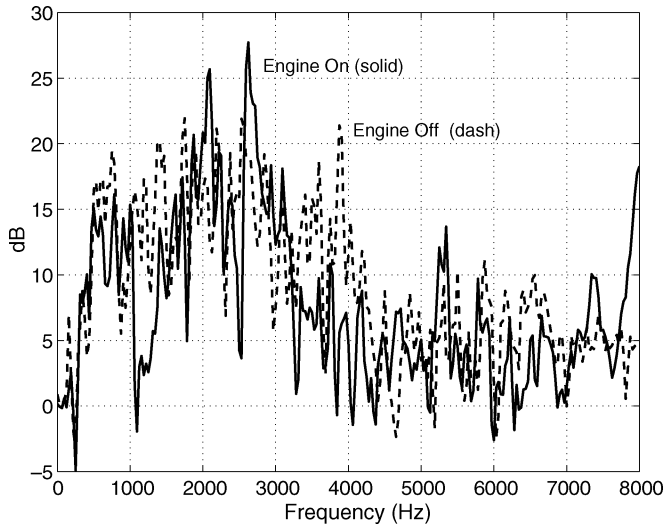

Fig. 18. Estimated speech reinforcement with the engine on (solid line) and the engine off (dashed line).

VII. Conclusion

A speech reinforcement system is presented which improves communication among the passengers in vehicles by picking up the speech of each passenger, amplifying it and playing it back into the cabin through the loudspeakers of the car. Acoustic echo cancellation is needed to avoid howling and ensure the stability of this feedback acoustic system. Continuous double-talk degrades the performance of the acoustic echo cancellers so further echo attenuation is needed to avoid instability. The echo attenuation is performed by using a filter placed after the acoustic echo canceller that is designed to ensure stability avoiding distortion. A noise reduction stage is also needed to avoid increasing the noise level inside the car. Since the noise is picked up by the microphones, amplified by the system and played back into the cabin, a noise reduction filter based on the optimal filter theory is used. The system is able to maintain speech intelligibility levels of good and excellent for a wide range of noise levels and speech attenuation.

In this paper, we derive the optimal expression of the echo suppression filter. We also show that a real valued function should be used to maintain stability. The estimation of the echo suppression filter is performed by using the estimations of the psd of the residual echo.

Simulation results show that speech reinforcement around 15 dB can be achieved, ensuring stability. A noise reduction of 7 dB is obtained, measured over voiced speech segments and a noise attenuation greater than 25 dB in silent segments.

A real-time two-channel implementation was also presented. Several tests on medium and large size cars shows that it is possible to achieve speech reinforcement over 15 dB with low distortion and without howling even under unexpected changes in the acoustic environment such us passengers movements or doors openings.

Appendix

In this Appendix we show that the best option to obtain a real valued function for the ESF from (4) that ensures stability is substituting $\tilde{H}(e^{j\omega})$ by its modulus. The misadjustment function used in (4) is unknown and must be replaced by its estimate in a practical system. First, we will assume that we are able to get a good estimate of the modulus of the misadjustment function but there can be unbounded errors in the estimated phase of the misadjustment. Thus, the estimated misadjustment is

$$\tilde{H}_{est}(e^{j\omega}) = \left|\tilde{H}(e^{j\omega})\right| \cdot e^{j\psi(\omega)} \tag{A1}$$

while the true misadjustment is

$$\tilde{H}(e^{j\omega}) = \left|\tilde{H}(e^{j\omega})\right| \cdot e^{j\phi(\omega)}. \tag{A2}$$

Substituting (A1) and (A2) in (2) and (4) we obtain

$$P_4(e^{j\omega}) = \frac{KW_n(e^{j\omega})}{1 + KW_n(e^{j\omega})\left|\tilde{H}(e^{j\omega})\right|\left(e^{j\psi(\omega)} - e^{j\phi(\omega)}\right)}. \tag{A3}$$

Denoting the denominator in (A3) as

$$D(e^{j\omega}) = 1 + KW_n(e^{j\omega})\left|\tilde{H}(e^{j\omega})\right|\left(e^{j\psi(\omega)} - e^{j\phi(\omega)}\right) \tag{A4}$$

it must satisfy $D(e^{j\omega}) = 1$ to ensure stability avoiding distortion. Nevertheless, this can only be met if we are able to obtain $\psi(\omega) = \phi(\omega)$ what is not possible in practice. In order to avoid the system to become unstable we must ensure that the squared modulus of the denominator of the transfer function of the system

$$\begin{aligned}\left|D(e^{j\omega})\right|^2 = {} & 1 + 2KW_n(e^{j\omega})\left|\tilde{H}(e^{j\omega})\right|\left[\cos\psi(\omega) - \cos\phi(\omega)\right] \\ & + 2\left|KW_n(e^{j\omega})\tilde{H}(e^{j\omega})\right|^2\left\{1 - \cos\left[\psi(\omega) - \phi(\omega)\right]\right\}\end{aligned} \tag{A5}$$

is never equal to zero. The third term in (A5) is never smaller than zero, so $D(e^{j\omega})$ can only approach zero if the second term in (A5) is negative, and this never happens if $\cos\psi(\omega) > \cos\phi(\omega)$. To avoid the denominator in (A3) approach zero we choose $\psi(\omega) = 0$ what makes the second term in (A5) always positive and the squared modulus of $D(e^{j\omega})$ always greater than 1 what ensures the stability of the system.

Finally, if we assume that the magnitude of the misadjustment estimation is inaccurate and we choose $\psi(\omega) = 0$, that is

$$\tilde{H}_{est}(e^{j\omega}) = \left|\tilde{H}(e^{j\omega})\right| + \Delta(e^{j\omega}) \tag{A6}$$

the transfer function of the systems becomes

$$P_5(e^{j\omega}) = \frac{KW_n(e^{j\omega})1}{1 + KW_n(e^{j\omega})\left|\tilde{H}(e^{j\omega})\right|\left(1 + \frac{\Delta(e^{j\omega})}{\left|\tilde{H}(e^{j\omega})\right|} - e^{j\phi(\omega)}\right)} \tag{A7}$$

Thus, to ensure the stability of the system, we must avoid the denominator in (A7) become equal to zero. This sets an upper limit for the gain factor K depending on the magnitude of the misadjustment, $|\tilde{H}(e^{j\omega})|$, and the error in the estimation of the magnitude of the misadjustment, $\Delta(e^{j\omega})$.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for providing constructive comments that improved this paper.

REFERENCES

[1] F. Gallego, E. Lleida, E. Masgrau, and A. Ortega, "Method and system for suppressing echoes and noise in environments under variable acoustic and highly feedback conditions," Patent WO 02/101728 A1.

[2] E. Lleida, E. Masgrau, and A. Ortega, "Acoustic echo and noise reduction for cabin car communication," in *Proc. Eurospeech*, vol. 3, Sep. 2001, pp. 1585–1588.

[3] A. Ortega, E. Lleida, E. Masgrau, and F. Gallego, "Cabin car communication system to improve communication inside a car," in *Proc. ICASSP'02*, vol. 4, May 2002, pp. 3836–3839.

[4] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. 46, no. 3, pp. 497–510, Mar. 1967.

[5] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control, an application of very-high-order adaptive filters," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 42–69, Jul. 1999.

[6] S. Gustanfsson and R. Martin, "Combined acoustic echo control and noise reduction for mobile communications," in *Proc. Eurospeech*, Sep. 1997.

[7] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Process.*, no. 64, pp. 21–32, Jan. 1998.

[8] E. Hänsler and G. U. Schmidt, "Hands-free telephones—joint control of echo cancellation and postfiltering," *Signal Process.*, no. 80, pp. 2295–2305, Jan. 2000.

[9] J. A. Maxwell and P. M. Zurek, "Reducing acoustic feedback in hearing aids," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 304–313, Jul. 1995.

[10] J. M. Kates, "Feedback cancelation in hearing aids: results from a computer simulation," *IEEE Trans. Signal Process.*, vol. 39, no. 3, pp. 553–562, Mar. 1991.

[11] A. M. Engebretson, M. P. O'Connel, and F. Gong, "An adaptive feedback equalization algorithm for the cid digital hearing aid," in *Proc. Annu. Int. Conf. IEEE EMB Soc.*, vol. 12, 1990, pp. 2286–2287.

[12] J. R. Deller Jr., J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[13] G. Faucon, R. L. Bouquin, and R. L. Jeannès, "Joint system for acoustic echo cancellation and noise reduction," in *Proc. Eurospeech*, Sep. 1995, pp. 1525–1528.

[14] R. L. B. Jeannès, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 808–820, Nov. 2001.

[15] H. Hass, "The influence of a single echo on the audibility of speech," *Acustica*, vol. 1, no. 2, 1951.

[16] P. Eneroth, "Stereophonic Acoustic Echo Cancellation: Theory and Implementation," Ph.D. dissertation, Lund Univ., Lund, Sweden, Jan. 2001.

[17] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Trans. Commun.*, vol. 39, no. 11, pp. 1542–1545, Nov. 1991.

[18] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 718–724, Nov. 1999.

[19] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. B. Mariño, and C. Nadeu, "Albayzin speech database: design of the phonetic corpus," in *Proc. Eurospeech*, vol. 1, Sep. 1993, pp. 175–178.

[20] D. Davis and C. Davis, *Handbook for Sound Engineers*, 2nd ed. New York: SAMS, Macmillan, 1991, ch. 32.

[21] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Amer.*, vol. 77, no. 3, pp. 1069–1077, Mar. 1985.

[22] H. J. Steeneken, "The measurement of speech intelligibility," in *Proc. Inst. Acoust.*, 2001.

Alfonso Ortega was born in Teruel, Spain, in 1976. He received the M.Sc. degree in telecommunication engineering from the University of Zaragoza (UZ), Zaragoza, Spain, in 2000.

In 1999, he joined, under a research grant, the Communications Technologies Group, UZ, where he has been an Assistant Professor since 2001. He is also involved as Researcher with the Aragon Institute of Engineering Research (I3A). At present, his research interest lies in the field of adaptive signal processing applied to speech technologies.

Eduardo Lleida (M'91) was born in Spain in 1961. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in signal processing from the Polytechnic University of Catalonia (UPC), Spain, in 1985 and 1990, respectively.

From 1986 to 1988, he was involved in his doctoral work at the Department of Signal Theory and Communications, UPC. From 1989 to 1990, he was Assistant Professor and from 1991 to 1993, he was Associate Professor in the Department of Signal Theory and Communications, UPC. From February 1995 to January 1996, he was with AT&T Bell Laboratories, Murray Hill, NJ, as a Consultant in speech recognition. Currently, he is an Associate Professor of signal theory and communications in the Department of Electronic Engineering and Communications, Universidad de Zaragoza, Spain. He is also member of the Aragon Institute of Engineering Research. His current research interests are in digital signal processing, in particular applied to speech enhancement and recognition in adverse acoustic environments.

Enrique Masgrau (M'84) received the M.S. and Ph.D. degrees in electrical engineering from the Polytechnic University of Catalonia (UPC), Catalonia, Spain, in 1978 and 1983, respectively.

He was an Assistant Professor (1978 to 1992) at UPC. He joined the University of Zaragoza, Spain, in 1992, as a Full Professor with the Department of Electronic Engineering and Communications, where he is currently Manager. He is also a Member of the Aragon Institute of Engineering Research (I3A) where he is Manager of the Communications Technologies Group. His research interests include speech processing, acoustic noise cancellation, MIMO communication techniques, and ICT applications in automotive ("telematics"). In these areas, he has published over 100 technical papers in various international journals and conferences. He has also been serving as Reviewer of several international conferences and journals. He holds three international patents.