Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 87 (2007) 739–749
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

A hybrid acoustic echo canceller and suppressor

Fredric Lindstrom[a,*], Christian Schüldt[b], Ingvar Claesson[b]

[a]*Konftel AB, Research and Development, Box 268, SE-90106 Umeå, Sweden*
[b]*Blekinge Institute of Technology, Department of Signal Processing, SE-37225 Ronneby, Sweden*

Received 3 April 2006; received in revised form 19 July 2006; accepted 24 July 2006
Available online 22 August 2006

Abstract

Wideband communication is becoming a desired feature in telephone conferencing systems. This paper proposes a computationally efficient echo suppression control algorithm to be used when increasing the bandwidth of an audio conferencing system, e.g. a conference telephone. The method presented in this paper gives a quality improvement, in the form of increased bandwidth, at a negligible extra computational cost. The increase in bandwidth is obtained through combining a conventional acoustic echo cancellation unit and an acoustic echo suppression unit, i.e. a hybrid echo canceller and suppressor. The proposed solution was implemented in a real-time system. Frequency analysis combined with subjective tests showed that the proposed method extends the bandwidth, while maintaining high quality.

Keywords: Acoustic echo cancellation (AEC); Acoustic echo suppression (AES); Wideband; Hybrid

1. Introduction

The market for audio conferencing continues to grow thanks to the strive to save time and reduce travel costs and environmental pollution. Generally, audio conferencing systems are equipped with hands-free loudspeaking audio communication. This paper presents a robust and computationally efficient method to extend the bandwidth of a hands-free audio conference phone. Conference phones traditionally use a communication bandwidth with an upper frequency limit of approximately 3.4 kHz [1]. With the increasing demands of quality and use of IP-telephony, speech codec-based telephony with communication bandwidths of 7 kHz is becoming a desirable feature [2].

Thus, there is a need to find solutions that can handle a wideband audio signal, i.e. to extend the communication bandwidth of a conventional acoustic echo canceller (AEC) conference phone. This task is not uncomplicated, due to robustness requirements and limits of computational resources. One approach is to obtain the extension in bandwidth by adding an acoustic echo suppression (AES) unit, [3–6].

This paper proposes a low-complexity gain control to be used in an AES unit added in parallel with a conventional AEC. In the proposed solution, no assumptions have been made about the structure of the AEC at hand and no signals from the AEC have been used. Thus, the proposed method can be used with good effect in conjunction with any existing AEC based conference phone.

*Corresponding author. Tel.: +46 90706488.
E-mail address: fli@konftel.com (F. Lindstrom).

0165-1684/$ - see front matter
doi:10.1016/j.sigpro.2006.07.006

The outline of the paper is as follows. Section 2 provides a brief overview of AES and cancellation. In Section 3, the hybrid suppressor/canceler solution is presented. The hybrid solution requires a number of frequency splitting/sample rate conversion filters. An analysis and a simple design approach of these filters are provided in Section 4. The proposed control algorithm is presented in Section 5. Section 6 presents a real-time implementation of the proposed solution. Finally, Section 7 concludes the paper.

2. Echo suppression and echo cancellation

AES, or voice switching techniques, are the first introduced solutions to deal with acoustic echoes, [7,8]. An echo suppressor reduces the echo by damping either or both of the sending or/and the receiving signals. The use of adaptive gain echo suppression for half-duplex audio hands-free systems is today a rather well-developed technique, with applications available on chip [9,10]. Echo might not be present at the entire signal spectrum and damping the full-band signal might, thus, not be an optimal solution. An echo suppression filter can be used to obtain a frequency-dependent damping, [11]. A classical problem for the echo suppression solution is the intrinsic half-duplex character of the system, i.e. during simultaneously near and far-end speech one direction of communication is always damped.

Echo cancellation provides a solution that allows increased full-duplex characteristics, [12]. In a hands-free system, acoustic echo is the result of the transformation of the far-end signal as it passes through the loudspeaker, the room and the microphone. The combined influence from the loudspeaker, the room, and the microphone is denoted the loudspeaker enclosure microphone (LEM) system. The purpose of an AEC unit is to adapt the transfer characteristics of an adaptive filter in order to mimic the LEM. Thereby, a replica of the acoustic echo can be produced and the acoustic echo can be cancelled by subtracting the replica from the microphone signal. The solution thus allows simultaneous two-way communication. Overviews of echo cancellation can be found in [8,13–15]. The core of an AEC is a continuously updating adaptive filter [16]. Examples of updating algorithms suitable for real-time AEC implementations are: the normalized least mean square (NLMS), the affine projection algorithm (APA) and, possibly, the fast transversal filter (FTF) [16]. Of these, the NLMS algorithm is the most popular algorithm thanks to low complexity and its robustness to finite precision errors. The key parameter in the NLMS algorithm is the step-size of the adaptive filter update. Suggestions for proper step-size management are found in [17].

3. Hybrid AEC and AES

The concept of a hybrid AEC and acoustic echo suppressor was introduced in the mid 80's [18,19]. The hybrid solution implies a structure where both speech signals, (i.e. the far-end and the near-end signals), are split in two frequency bands, one that contains the lower frequencies and one that contains the higher frequencies. The two bands are processed in different ways. The low frequency part is processed with a full duplex AEC. Acoustic echoes in the low frequency band will therefore be cancelled and communication will not be interrupted in either direction. The high frequency part will be passed with a level dependent damping, i.e. high frequency echoes are suppressed with an adaptive gain.

The main justification for using the hybrid method is that the limited bandwidth of the lower frequency band allows the low frequency signals to be downsampled, thus reducing the computational demand of the AEC. In this paper, the same idea is explored to allow an extension of the communication bandwidth without any significant increase in computational complexity.

The hybrid solution used in this paper is depicted in Fig. 1, where the loudspeaker signal, i.e. the line-in signal received from the far-end, is denoted $x(k)$, k is sample index. The loudspeaker signal generates output in form of an acoustic echo as it is fed to the LEM system. The acoustic echo (or the desired signal) is denoted $d(k)$. The near-end signal, i.e. the signal picked up by the microphone is denoted $y(k)$. The near-end signal $y(k)$ consist of acoustic echo $d(k)$, near-end speech $s(k)$ and background noise $n(k)$, i.e. $y(k) = d(k) + s(k) + n(k)$. The far-end signal, $x(k)$, is divided into a high frequency part, $x_{\mathrm{H}}(k)$ and a downsampled low frequency part, $x_{\mathrm{L}}(l)$, where l is sample index. Likewise, the near-end signal, $y(k)$, is divided into $y_{\mathrm{H}}(k)$ and $y_{\mathrm{L}}(l)$. Frequency splitting/anti-aliasing filters $h_{x_{\mathrm{H}}}$, $h_{x_{\mathrm{L}}}$, $h_{y_{\mathrm{H}}}$, and $h_{y_{\mathrm{L}}}$ are used for this procedure, as depicted in Fig. 1. The low frequency echo cancelled signal $e(l)$ is obtained by subtracting the acoustic echo

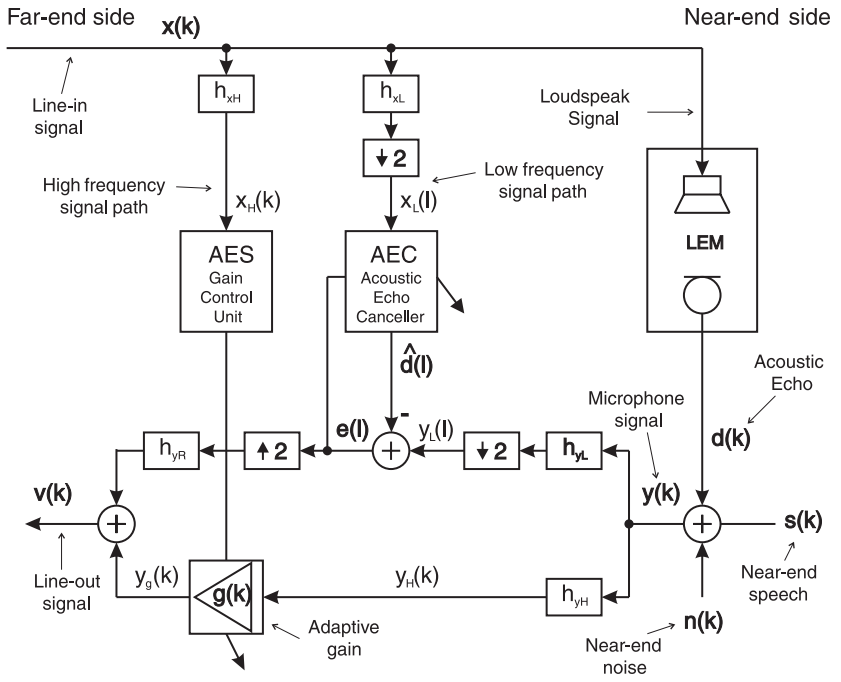

Fig. 1. The scheme of the hybrid solution used in this paper.

estimate $\hat{d}(l)$ from the low frequency microphone signal $y_L(l)$. Real implementations of hands-free systems will almost certainly contain some additional damping in order to maintain system robustness. Such damping is not depicted in Fig. 1. The operation performed on the high frequency signal $y_H(k)$ will be an adaptive attenuation of $y_H(k)$ by a gain factor, $g(k)$, with $g(k)\leqslant 1$, resulting in a possibly damped signal, $y_g(k)$. The adaptation of $g(k)$ is processed by a control unit (CU). The CU sets the value of $g(k)$ depending on the value of some chosen measure of the $x_H(k)$ signal. The line-out signal $v(k)$ is obtained by adding the signal $y_g(k)$ to an upsampled version of $e(l)$, obtained using the anti-image reconstruction filter h_{y_R}.

Several solutions based on the hybrid concept have been proposed, [3–6]. In [4,5] the echo suppression is applied to the output signal $v(k)$, see Fig. 1. A drawback with such a solution is that in a situation where the residual echo is larger in one frequency band, the other band is unnecessarily damped. In [6] this is partly avoided by introducing an attenuation of the upper-band signal, $y_H(k)$, that is equal to the attenuation of the lower-band echo canceller. In [4–6] the processing of the upper-band and lower-band is tightly connected. The aim in this paper is to provide a solution which can be added to an existing lower-band AEC without any assumptions of the processing of that AEC. Such a scheme, i.e. where upper- and lower-band processing are independent was proposed in [3], where the upper-band echo is reduced by using a frequency domain approach. In contrast, the control algorithm proposed in this paper is a low-complexity solution operating in the time domain and implemented in real-time.

Industrial development often relies on extending existing solutions and complexity cost is always an issue. The method proposed in this paper allows an increase of the bandwidth without adding any significant complexity. The independence of the lower- and upper-band processing allows the method to be used with minor effort when extending an existing nonwideband solution.

4. The frequency splitting filters

In this section, the filters $h_{x_H}, h_{x_L}, h_{y_H}, h_{y_L}$ and h_{y_R}, see Fig. 1, used in the hybrid echo canceller/suppressor are discussed. In the following text, a downsampling with a factor 2 is assumed. The treatment of a higher downsampling order is analogous. Upper-case letter versions of introduced signals and filters represent discrete-time Fourier

transforms of their corresponding lower-case letter signal/filter, e.g.

$$X(e^{j\omega}) = \sum_{k=-\infty}^{\infty} x(k)e^{-j\omega k}. \quad (1)$$

The interval of the frequency variable ω is assumed $|\omega| \leqslant \pi$ for all equations.

The signals $x_L(l)$ and $y_L(l)$ are input to the AEC, see Fig. 1. The downsampling and anti-aliasing filtering should not degenerate the performance of the AEC. The following analysis applies:

Assume that the only present input signal is far-end signal with a transform representation $X(e^{j\omega})$ and the LEM is a linear time-invariant system h_{LEM}, then, from Fig. 1, the low frequency part of the microphone signal only consists of low frequency acoustic echo, i.e. $y_L(l) = d_L(l)$. The Fourier transform of the signal $d_L(l)$ is

$$\begin{aligned} D_L(e^{j\omega}) &= 0.5(X(e^{j0.5\omega})H_{LEM}(e^{j0.5\omega})H_{y_L}(e^{j0.5\omega}) \\ &\quad + X(e^{j(0.5\omega-\pi)})H_{LEM}(e^{j(0.5\omega-\pi)})H_{y_L}(e^{j(0.5\omega-\pi)})). \end{aligned} \quad (2)$$

Assume further that $\hat{d}(l)$ is obtained through the filtering of $x_L(l)$ with the filter $\hat{h}_{LEM}$. Then, from Fig. 1, the Fourier transform of the signal $\hat{d}(l)$ is given by

$$\begin{aligned} \hat{D}(e^{j\omega}) &= 0.5(X(e^{j0.5\omega})H_{x_L}(e^{j0.5\omega})\hat{H}_{LEM}(e^{j\omega}) \\ &\quad + X(e^{j(0.5\omega-\pi)})H_{x_L}(e^{j(0.5\omega-\pi)})\hat{H}_{LEM}(e^{j\omega})). \end{aligned} \quad (3)$$

The first terms in Eqs. (2) and (3) correspond to the desired downsampled signals. The second terms in the equations are the aliasing terms. The effect of the aliasing terms on the AEC are analogous to the effects of aliasing in a critically sampled subband AEC [20]. In a critically sampled two-band subband solution, both the upper- and the lower-band are downsampled. This implies that the frequency split has to be done at $\omega = 0.5\pi$. In the solution of this paper, the upper-band is not downsampled, thanks to the low complexity of the upper-band processing. This implies that the frequency split can be at a frequency lower than $\omega = 0.5\pi$, and the design of the frequency splitting filters is thus facilitated.

The portion of the acoustic echo in the lower-band is perfectly cancelled out if

$$\hat{D}(e^{j\omega}) = D_L(e^{j\omega}). \quad (4)$$

Assume that filters h_{x_L} and h_{y_L} provide sufficient damping in the stopband, i.e. for $|\omega| > 0.5\pi$. With sufficient damping we mean that the aliasing terms in Eqs. (2) and (3) become nonsignificant. Then from Eqs. (2) and (3), Eq. (4) is satisfied if the adaptive filter $\hat{H}_{LEM}(e^{j\omega})$ fulfills

$$\hat{H}_{LEM}(e^{j\omega})H_{x_L}(e^{j0.5\omega}) = H_{LEM}(e^{j0.5\omega})H_{y_L}(e^{j0.5\omega}). \quad (5)$$

Eq. (5) demonstrates, that if the filters h_{x_L} and h_{y_L} are selected carelessly the optimal filter characteristics of $\hat{H}_{LEM}(e^{j\omega})$ might be unnecessarily hard or even noncausal. One approach to guarantee that this is avoided, is to choose $h_{x_L} = h_{y_L}$.

The filtering performed should be such that the near-end speech signal is not degenerated. Assume that the only present input signal is a near-end signal with a transform representation $Y(e^{j\omega})$. Then, the scheme in Fig. 1 gives that the Fourier transform of the line-out signal $v(k)$ is

$$\begin{aligned} V(e^{j\omega}) = {} & Y(e^{j\omega})H_{y_H}(e^{j\omega}) \\ & + 0.5(Y(e^{j\omega})H_{y_L}(e^{j\omega})H_{y_R}(e^{j\omega}) \\ & + Y(e^{j(\omega-\pi)})H_{y_L}(e^{j(\omega-\pi)})H_{y_R}(e^{j\omega})). \end{aligned} \quad (6)$$

A perfect reconstruction, i.e.

$$V(e^{j\omega}) = ce^{-j\omega k_0}Y(e^{j\omega}), \quad (7)$$

where c is a nonzero constant and k_0 is a nonnegative integer, thus requires,

$$H_{y_H}(e^{j\omega}) + 0.5H_{y_L}(e^{j\omega})H_{y_R}(e^{j\omega}) = ce^{-j\omega k_0} \quad (8)$$

and

$$H_{y_L}(e^{j(\omega-\pi)})H_{y_R}(e^{j\omega}) = 0. \quad (9)$$

Eq. (8) requires the filter h_{y_H} and the filter operation $0.5h_{y_L} * h_{y_R}$, (where $*$ denotes convolution), to be strictly complimentary. If h_{y_L} and h_{y_R} are TYPE 1 linear phase finite impulse response (FIR) filters a strictly complimentary filter h_{y_H} can be obtained through

$$H_{y_H}(e^{j\omega}) = e^{-0.5j\omega(N_1+N_5)} - 0.5H_{y_L}(e^{j\omega})H_{y_R}(e^{j\omega}), \quad (10)$$

[21,22].

If the strict perfect reconstruction is dropped, a less computationally demanding solution is possible.

The frequency splitting filters will introduce a delay in the signal path. This delay should be as low as possible. The earlier ITU recommendation [23] allows only a 2 ms delay for the signal processing.

In [24], which partly replaces [23], no specific delay is specified for stationary telephones. However, overall delays of 36–52 ms are given as examples of processing delays for mobile hands-free phones. These delays also account for e.g. noise reduction processing.

The filter h_{x_H} is only used to extract information about the power of the high frequency part of $x(k)$. Thus, no hard filter specification requirements are imposed on h_{x_H}.

5. Algorithm for the control unit

In this section an algorithm for the calculation of the gain $g(k)$, (see Fig. 1), is presented. The idea is to find a proper damping of $y_H(k)$ by evaluating the signal $x_H(k)$. If the square of the high frequency acoustic echo, $d_H^2(k)$, is significantly lower than the noise floor in the high frequency band, $f_H(k)$, the acoustic echo is not disturbing. Thus, in order to guarantee sufficient damping the $g(k)$ function should fulfill

$$g(k) \leqslant C_H \frac{f_H(k)}{d_H^2(k)}, \tag{11}$$

where C_H is a constant.

The acoustic echo is not directly measurable. The approach in this paper is to from $x_H(k)$ obtain a signal $\hat{d}_H^2(k)$ that is an estimate of $d_H^2(k)$ and fulfills $\hat{d}_H^2(k) \geqslant d_H^2(k)$.

A noise floor estimate $\hat{f}_H(k)$ can be obtained by measuring the short-time energy during speech pauses, see Section 5.2.

From these estimates the gain function is obtained by

$$g(k) = C_H \frac{\hat{f}_H(k)}{\hat{d}_H^2(k)}. \tag{12}$$

5.1. Estimation of high frequency acoustic echo

The high frequency acoustic echo $d_H(k)$ is generated through the filtering of the loudspeaker signal $x_H(k)$ with the LEM. In this paper it is assumed that the total LEM signal path gain, depicted in Fig. 2, is less than 0 dB for any frequency band. This means that the gain $g(k)$ can be correctly evaluated from $x_H(k)$ and that a fully amplified loudspeaker signal $x(k)$ does not generate an overflowing microphone signal $y(k)$. The acoustic coupling is always less than

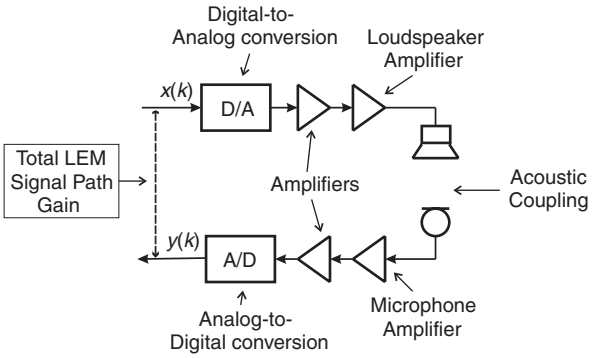

Fig. 2. Schematic illustrating the total LEM signal path gain.

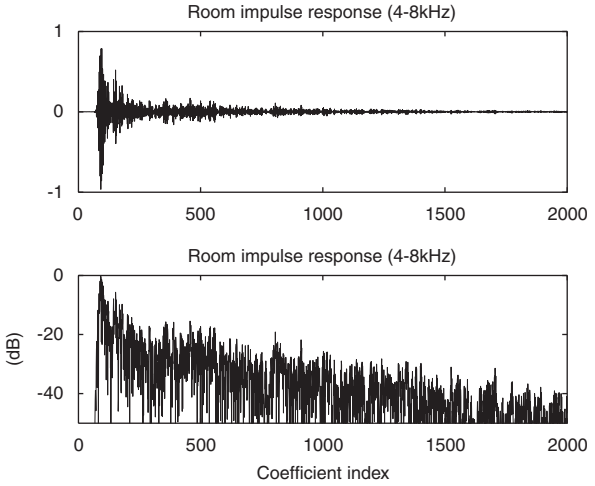

Fig. 3. UPPER PLOT: the impulse response of a typical LEM filter with bandwidth 4–8 kHz, i.e. the impulse response demonstrates the high frequency character of the LEM. LOWER PLOT: the rectified impulse response in dB scale.

0 dB and the amplifier gains are typically known for one piece units, i.e. units without the possibility to connect external microphones/loudspeakers, so the above assumption can generally be fulfilled easily. If any amplifier gain in the LEM signal path is time-variant, e.g. a tunable loudspeaker amplifier, the gain $g(k)$ should be modified so that an increase of the gain in the signal path implies a corresponding decrease of the gain $g(k)$ (or a gain decrease in an amplifier). If the gain in the amplifiers are unknown they need to be adaptively estimated or estimated according to a worst-case scenario. This case is not considered in this paper.

The high frequency part of the first 2000 FIR model coefficients of a typical LEM system is shown in the upper plot in Fig. 3. Other examples of FIR

models depicting the general character of a LEM can be found in [8,15]. The impulse response in Fig. 3 can be divided into three parts: part 1 (index 0–70), part 2 (indices around 80), and part 3 (index >100). The first part consists of zero coefficients. These zeros originate from delays in the LEM system due to D/A and A/D-conversion, sample rate alternation, and the distance between the microphone and the loudspeaker. The second part is the high magnitude "direct" coefficients, i.e. they correspond to a straight signal path directly from the loudspeaker to the microphone (or signal paths that are of the same order as the direct path). The third part consists of the far coefficients, i.e. coefficients that represent the signal path of longer distances between the loudspeaker and the microphone, e.g. a path containing several reflections via the ceiling, the walls, etc. of the enclosure.

Consider a short $x_H(k)$ signal burst. This burst will give rise to an acoustic echo $d_H(k)$. First of all, there is a short delay between the onset of the $x_H(k)$ signal and the emerge of the acoustic echo. Thereafter, there is a fast increase of the acoustic echo. Finally, the acoustic echo will slowly decay after the offset of $x_H(k)$ (Compare with the discussion of the three parts of the LEM in Fig. 3 above). This relation between $x_H(k)$ and $y_H(k)$ is illustrated in Fig. 4. In Fig. 4 the delay between the onset of the loudspeaker signal (dotted line, sample index 1200) and the emerge of the echo (solid line, sample index 1280) can be observed. Further, the slow decay of the echo (solid line, sample index 6800–9000) after the termination of the loudspeaker signal (dotted line, sample index 6800) is shown.

Based on the above observations the following estimate $\hat{d}_H^2(k)$ is proposed:

$$\hat{d}_H^2(k) = \begin{cases} (1-\gamma_f)\hat{d}_H^2(k-1) + \gamma_f x_H^2(k-T) \\ \quad \text{if } x_H^2(k-T) \geqslant \hat{d}_H^2(k), \\ (1-\gamma_s)\hat{d}_H^2(k-1) + \gamma_s x_H^2(k-T) \\ \quad \text{otherwise,} \end{cases} \tag{13}$$

where T is a constant delay determined by the part 1 delay in the LEM, and γ_f and γ_s are two averaging constants with $\gamma_f > \gamma_s$. The constant γ_f yields a "fast increase" and γ_s a "slow decrease". The use of two different averaging constants correspond to the fast increase and slow decrease described in relation to the LEM part 2 and part 3 described above.

In Fig. 5 the square of the acoustic echo, $d_H^2(k)$ (obtained through a real system) is plotted together with the $\hat{d}_H^2(k)$ signal.

5.2. Estimation of noise floor

The estimation $\hat{f}_H(k)$ evaluates the noise floor, i.e. background noise level. The method proposed here is based on comparison of long-term and short-term power averages.

A block-processing method is used in order to reduce computational complexity. For every M

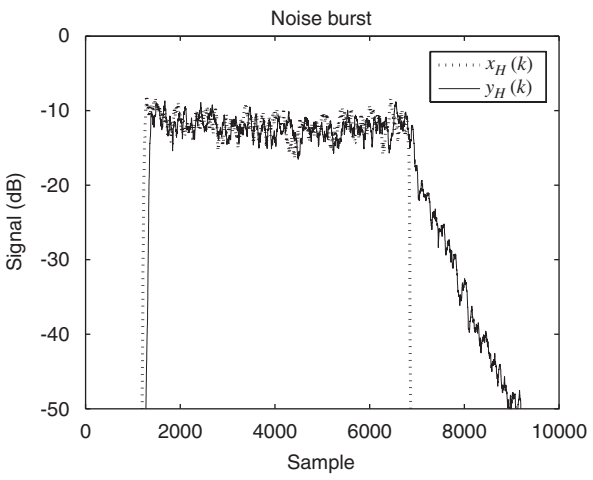

Fig. 4. A $x_H(k)$ noise burst (dotted signal) with corresponding echo, i.e. the $y_H(k)$ signal.

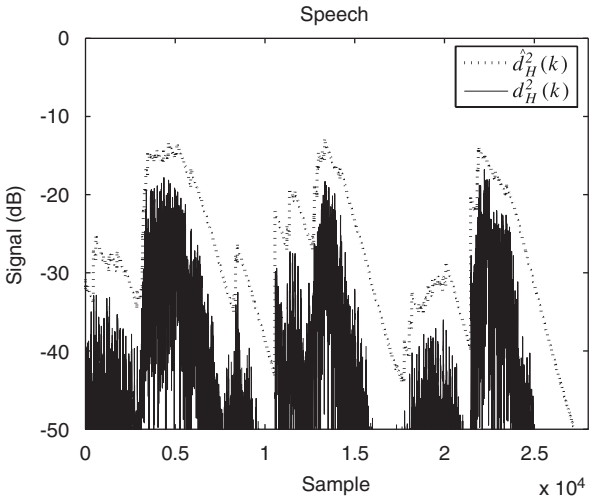

Fig. 5. The momentary high frequency acoustic echo $d_H^2(k)$ and the signal $\hat{d}_H^2(k)$. In this plot it can be seen that the function $\hat{d}_H^2(k)$ fulfills $\hat{d}_H^2(k) \geqslant d_H^2(k)$.

sample, (i.e. $k = M, 2M, 3M, \ldots$), the short-term power $P_y(k)$ for the latest M samples of the high frequency microphone signal $y_H(k)$ is calculated,

$$P_y(k) = \frac{1}{M}\sum_{i=0}^{M-1} y_H^2(k-i). \quad (14)$$

The maximum, $P_{\max}(k)$, and minimum, $P_{\min}(k)$ values for the L latest $P_y(k)$ estimates are given by

$$P_{\max}(k) = \max\{P_y(k), \ldots, P_y(k-(L-1)M)\}, \quad (15)$$

$$P_{\min}(k) = \min\{P_y(k), \ldots, P_y(k-(L-1)M)\}. \quad (16)$$

If the difference between $P_{\max}(k)$ and $P_{\min}(k)$ is less than a constant C_P the long-term and short-term power average of the signal $y_H(k)$ are similar, and the signal $y_H(k)$ is considered to contain only background noise. In this case the estimation of the high frequency near-end background noise floor is updated, i.e.

$$\hat{f}_H(k) = \begin{cases} (1-\gamma_n)\hat{f}_H(k-1) + \gamma_n P_{\min}(k) \\ \quad \text{if } P_{\max}(k) - P_{\min}(k) \leqslant C_P \\ \hat{f}_H(k-1) \\ \quad \text{otherwise,} \end{cases} \quad (17)$$

where γ_n is an averaging constant.

The proposed gain function $g(k)$ is thus defined through Eqs. (12)–(17).

5.3. Complexity discussion

Assume a full-band NLMS-based AEC solution operating with a sampling frequency f_s. With an echo canceling duration of T seconds, the NLMS algorithm will require an adaptive FIR filter of the length $N = Tf_s$. For every sample, a digital signal processor (DSP) capable of multiply–add-and-accumulate and two memory accesses in parallel with arithmetic will require N instructions for the filtering, and $2N$ instructions for the update of the coefficients of the adaptive filter. Thus, the total number of DSP instructions per second for the AEC method, I_{AEC}, is given by

$$I_{AEC} = 3Nf_s = 3T(f_s)^2. \quad (18)$$

If the bandwidth is to be extended by factor 2, the sampling frequency is increased by factor 2 and Eq. (18) shows that the complexity is increased by factor 4.

Assume a sample rate of 8 kHz before the extension and a canceling length of $T = 250$ ms. This gives that the unextended NLMS AEC requires 48 million instructions per second (MIPS), and the extended version 192 MIPS, i.e. a straightforward extension implies a quite large increase in required computational resources.

If the bandwidth is increased by factor 2 using the proposed method the control algorithm as given in Eqs. (12)–(17) only requires a few extra instructions, thanks to the low complexity of Eqs. (12)–(13) and the block implementation of the noise estimation. The number of required instructions I_F for the five filters h_{x_L}, h_{x_H}, h_{y_L}, h_{y_H} and h_{y_R} is given by

$$I_F = (c_{x_L} + c_{x_H} + c_{y_L} + c_{y_H} + c_{y_R})f_s, \quad (19)$$

where c_{x_L}, c_{x_H}, c_{y_L}, c_{y_H} and c_{y_R} are the numbers of coefficients in h_{x_L}, h_{x_H}, h_{y_L}, h_{y_H} and h_{y_R}, respectively. If all filters are assumed to be of FIR type, typical values in an industrial implementation are e.g. $c_{x_L} = c_{y_L} = c_{y_H} = c_{y_R} = 49$ and $c_{x_H} = 13$. Assume $f_s = 16$ kHz and that h_{y_L}, h_{x_L} and h_{y_R} are implemented using a polyphase filters. This, implies that $I_F \approx 2$ MIPS. If all filters are fifth order IIR filters the complexity is given by $I_F \approx 0.8$ MIPS.

The NLMS AEC can be implemented with less complexity, e.g. using sub band/frequency domain implementations. However, the above numbers indicates that the proposed method has a significantly lower complexity as compared with a straightforward extension even in a low-complexity AEC.

6. Real-time implementation

6.1. Implementation

In order to evaluate the proposed method two real-time systems were implemented. The first system, denoted S, is an implementation of an NLMS-based AEC. This implementation include a nonlinear processor for additional damping of residual echo, as indicated in Section 3. (The presentation of this nonlinear processor is out of the scope of this paper.) The second system is an extension of S, denoted S_{EXT}, which uses the method presented in Sections 3–5. The communication bandwidth of system S was (250, 3400 Hz), and the bandwidth of system S_{EXT} was (250, 7000 Hz). These limits were chosen bearing in mind the standards for regular PSTN and the ITU 7 kHz speech coder, respectively, see [1,2] and the limits of the equipment (loudspeaker). The parameter values used in the real-time implementation are given in Table 1.

Table 1
Parameters and corresponding values in the real-time implementation.

Parameter	Value
C_H	0.67
γ_f	0.9980
γ_s	0.25
T	80
M	512
L	8
γ_n	2×10^{-6}
C_p	0.004

The two systems were implemented on a fix-point DSP [25]. Beside the algorithms presented in this paper, noise reduction and comfort noise were implemented in both solutions as well.

6.2. Setup

The near-end speech signal was received through the microphone of a real commercial conference phone, and the near-end output signal was transmitted through the loudspeaker of the same phone. The far-end input signal was fed to a headset, located in another room, in order to provide acoustic isolation. The far-end output signal was obtained by a hand-held microphone, and delayed 100 ms by a delay circuit. The delay was introduced to simulate the delay in telephone wires and switching offices, and to make acoustic echoes clearly audible at the far-end side. The setup was done in an office with a reverberation time of approximately 400 ms expressed by RT60, where RT60 defines the reverberation time required for the sound level in a room to decrease by 60 dB after an impulse. The signal-to-noise-ratio (SNR) in the signal picked up by the near-end side microphone was approximately 40 dB when the near-end speech was produced by a loudspeaker.

6.3. Evaluation

To obtain a set of near-end and far-end speech signals with corresponding phone loudspeaker and phone line-out signals a PC with a 4-channel soundcard was used, see Fig. 6. Channels 1 and 2 recorded the loudspeaker and the phone line-out signals, respectively. Channels 3 and 4 played the near-end speech and far-end speech signals, respectively. The played session consisted of near-end talk, far-end talk, and doubletalk. Recordings were done for both the S and S_{EXT} solutions.

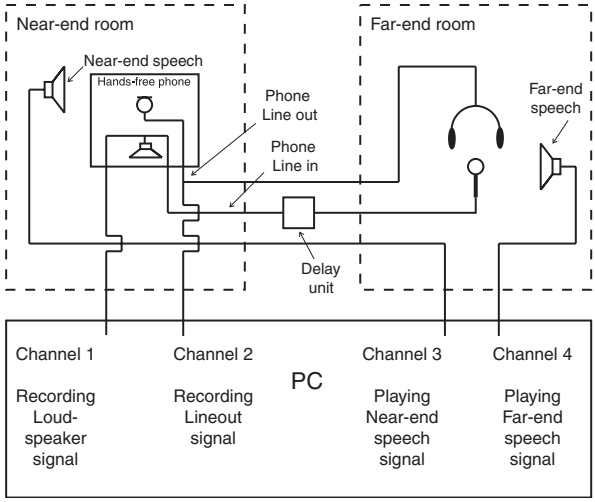

Fig. 6. The measurement setup.

An informal subjective real-time evaluation of both the methods was also performed. One person placed him–herself at the near-end side, and another person placed him-herself at the far-end side. These people carried on a normal conversation, containing sessions of doubletalk. Throughout the test repeated switches between solution S and solution S_{EXT} mode were performed. During the subjective tests other people moved in and out of the room in order to provide nonstationary LEM transfer characteristics.

6.4. Results

In Fig. 7 the short-time average power of the signals $y_L(l)$, $e(l)$, $y_H(k)$ and $y_g(k)$ are shown for a situation where the AEC has converged, a speech signal is present on the loudspeaker signal $x(k)$ and no near-end speech is present, i.e. the signals in Fig. 7 consist of only noise and echo. Fig. 7 demonstrates that the short-time power of the undamped high frequency echo (the power of $y_H(k)$) can be significantly higher than the power of the lower band AEC residual echo, (the power of $e(l)$). Further, Fig. 7 shows that the processed high frequency echo $y_g(k)$ maintains the same (or lower) level as the high frequency background noise. (Background noise level can be seen in Fig. 7 during the plotted first two seconds).

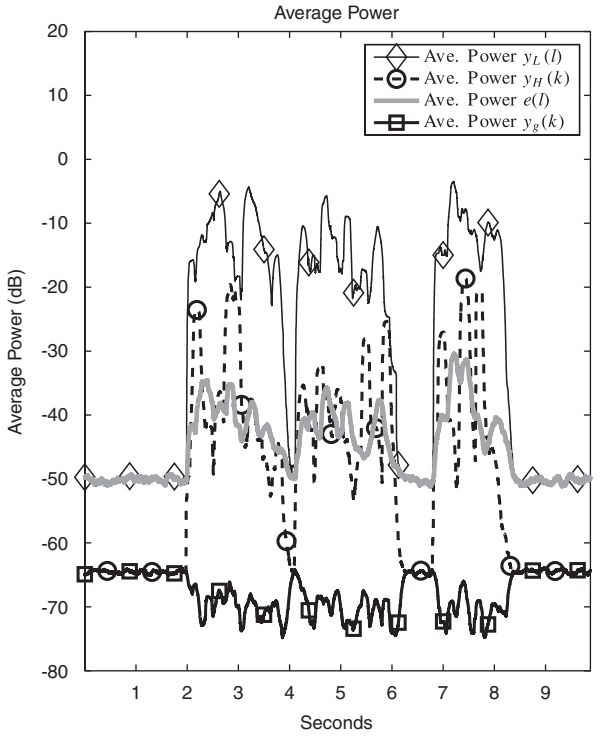

Fig. 7. Short-time average power of the lower-band microphone signal $y_{\mathrm{L}}(l)$, the lower-band residual echo signal $e(l)$, the upper-band microphone signal $y_{\mathrm{H}}(k)$ and the upper-band signal after damping $y_g(k)$ in a single far-end speech situation, with a converged AEC.

Table 2
Long-time power of the signals in Fig. 7.

Parameter	Value (dB)
$P_{y_{\mathrm{L}}}$	−14
P_e	−42
$P_{y_{\mathrm{H}}}$	−31
P_{y_g}	−66

The long-time power $P_{(\cdot)}$ of the signals in Fig. 7 are shown in Table 2. $P_{(\cdot)}$ is defined through

$$P_e = \frac{1}{J}\sum_{j=0}^{J-1} \mathrm{e}^2(l-j), \tag{20}$$

where J and l are set so that the summation is performed over the whole 10 s duration depicted in Fig. 7.

Echo return loss enhancement (ERLE) [13] is defined as

$$\mathrm{ERLE}(l) = \frac{E\{d^2(l)\}}{E\{d^2(l) - \hat{d}^2(l)\}}, \tag{21}$$

where $E\{\cdot\}$ denotes expected value. Since the noise level is relatively low in the experiment setup, average ERLE values after convergence can be estimated from the powers in Table 2. The estimated ERLE of the narrowband S system driven by a [250, 3400 Hz] signal is thus given by $-(P_{y_{\mathrm{L}}} - P_e) = 28\,\mathrm{dB}$. If the narrowband S system is driven by a wideband [250, 7000 Hz] signal it will not be able to cancel the high frequency signal and in this case the estimated ERLE will be $-(P_{y_{\mathrm{L}}} - P_{e+y_{\mathrm{H}}}) = 16\,\mathrm{dB}$.

The adaptive upper-band gain working in system S_{EXT} yields reduction of upper-band echo of $-(P_{y_{\mathrm{L}}} - P_{y_g}) = 35\,\mathrm{dB}$, i.e. sufficient for the residual echo in the upper-band to maintain the same (or lower) level as the background noise, as illustrated in Fig. 7.

Spectrograms of the loudspeaker and line-out signals for the conventional narrowband solution S are presented in Fig. 8, and for the proposed solution S_{EXT} in Fig. 9. The spectrograms of the near-end and far-end input speech signals are shown in Fig. 10, i.e. Fig. 10 presents the ideal, perfect frequency characteristics for the two solutions. By comparing the spectrograms in Figs. 8 and 9, it is clear that the proposed method gives a more natural frequency representation, in that it also contains high frequency components. The subjective real-time tests of the two systems using two-way communication showed that the extended bandwidth of the proposed system significantly increases the perceived quality. The reduction of the line-out

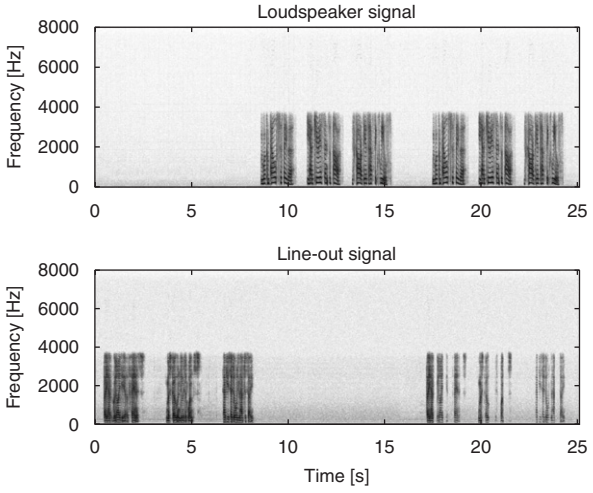

Fig. 8. Spectrograms of the conventional AEC solution, near-end single talk between 0–8.5 s, far-end single talk between 8.5–17 s, doubletalk between 17–25 s.

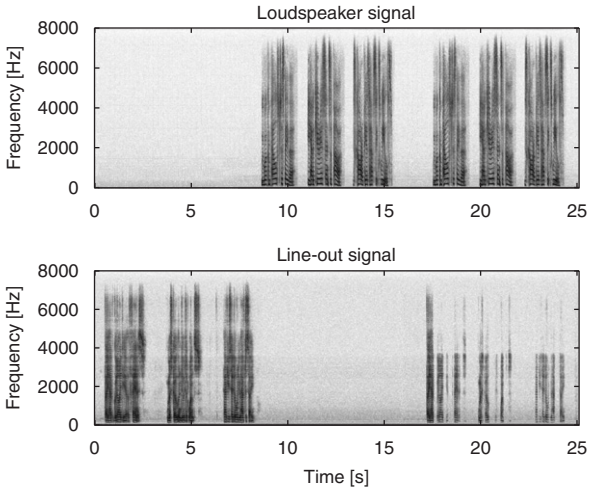

Fig. 9. Spectrograms of the proposed solution, near-end single talk between 0–8.5 s, far-end single talk between 8.5–17 s, doubletalk between 17–25 s.

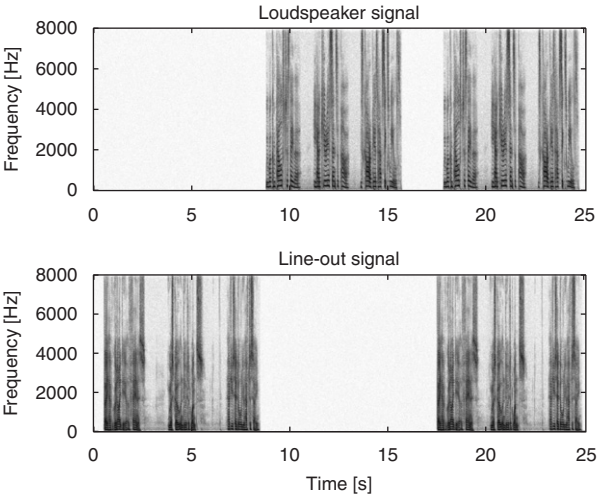

Fig. 10. Spectrograms of an ideal solution, near-end single talk between 0–8.5 s, far-end single talk between 8.5–17 s, doubletalk between 17–25 s.

signal bandwidth during doubletalk was not perceived as disturbing, i.e. it did not render a half-duplex feeling. Further, the subjective tests showed that no audible artifacts such as e.g. click sounds, distortion, or modulation are introduced by the proposed method.

7. Conclusions

A low-complexity method for increasing the bandwidth of an audio conferencing unit based on a hybrid acoustic echo canceller/suppressor solution was presented. A control algorithm for the suppression part was proposed. The algorithm in the suppressor unit was designed to be independent of the canceller unit. This was done in order to be able to use the extension method in conjunction with already existing echo cancellers with minor effort. An analysis of the frequency splitting filters present in the hybrid echo canceller/suppressor was provided and a set of suitable filter designing guidelines were presented. The proposed solution has been implemented and evaluated in real-time for a bandwidth extension from 3.4 to 7 kHz upper frequency limit. Subjective listening tests showed that the proposed solution increases the perceived quality thanks to the extended bandwidth. The extra computational load required by the proposed method was insignificant. Thus, the proposed method is a cost-effective way to increase the performance of an audio conference phone.

Acknowledgments

The above research was supported by the Swedish Knowledge Foundation (KKS). The authors thank the members of the staff at Konftel AB and Blekinge Institute of Technology for their evaluation of the proposed system.

References

[1] TBR21, European Telecommunications Standards Institute, 1998.
[2] ITU-T Recommendation G.722, 7 kHz audio—coding within 64 kbit/s, ITU-T Recommendations, 1998.
[3] F. Wallin, C. Faller, Perceptual quality of hybrid echo canceller/suppressor, Proceedings of IEEE ICASSP'04, vol. 4 (2004) 157–160.
[4] P. Heitkämper, Optimization of an acoustic echo canceller combined with adaptive gain control, in: Proceedings of IEEE ICASSP'95, Detroit, Michigan, 1995, pp. 3047–3050.
[5] P. Heitkämper, M. Walker, Adaptive gain control for speech quality improvement and echo suppression, in: Proceedings of IEEE ISCAS'93, vol. 1, Chicago, IL, 1993, pp. 455–458.
[6] W. Armbrüster, Wideband acoustic echo canceller with two filter structure, in: Proceedings of EUSIPCO 92, vol. 3, Bruxelles, Belgium, 1992, pp. 1611–1617.
[7] W.F. Clemency, F.F. Romanow, A.F. Rose, The Bell system speakerphone, AIEE Trans. 76 (1957) 148–153.
[8] J. Benesty, Y. Huang (Eds.), Adaptive Signal Processing, Springer, Berlin, 2003.
[9] U4082B, Low voltage voice-switched IC for hands-free operation, Atmel, 2001.
[10] IC03b, Semiconductors for wired telecom systems, Philips, 1998.

[11] E. Hänsler, G. Schmidt, Hands-free telephones—joint control of echo cancellation and post filtering, Signal Processing 80 (2000) 2295–2305.
[12] M.M. Sondhi, An adaptive echo canceler, Bell Syst. Tech. J. 46 (1967) 497–510.
[13] E. Hänsler, G. Schmidt, Acoustic Echo and Noise Control a Practical Approach, Wiley, New York, 2004.
[14] S. Gay, J. Benesty, Acoustic Signal Processing for Telecommunication, Kluwer Academic Publishers, Dordrecht, 2000.
[15] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, J. Tilp, Acoustic echo control, IEEE Signal Process. Mag. 16 (4) (1999) 42–69.
[16] S. Haykin, Adaptive Filter Theory, fourth ed., Prentice-Hall, Englewood Cliffs, NJ, 2002.
[17] A. Mader, H. Puder, G.U. Schmidt, Step-size control for acoustic echo cancellation filters—an overview, Signal Processing 80 (2000) 1697–1719.
[18] O.A. Horna, Echo canceller with extended frequency range, US Patent 4,609,787, September 2, 1986.
[19] T. Araseki, K. Ochiai, Echo canceller for attenuation acoustic echo signals on a frequency divisional manner, US Patent 4,670,903, June 2, 1987.
[20] A. Gilloire, M. Vetterli, Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation, IEEE Trans. Signal Process. 40 (8) (1992) 1862–1875.
[21] S.K. Mitra, Digital Signal Processing a Computer-based Approach, McGraw-hill, New York, 1998.
[22] P.P. Vaidyanathan, Multirate Systems and Filterbanks, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[23] ITU-T Recommendation G.167, General characteristics of international telephone connections and international telephone circuits—Acoustic echo controllers, ITU-T Recommendations, 1993.
[24] ITU-T Recommendation P.340, Transmission charactersitics and speech quality parameters of hands-free terminals, ITU-T Recommendations, 2000.
[25] ADSP-BF533 Blackfin processor hardware reference, Analog Devices, 2005.