ACOUSTIC ECHO CANCELLATION FOR NATURAL SPEECH COMMUNICATION AT THE HUMAN-MACHINE INTERFACE

Walter Kellermann

Telecommunication Laboratory, University Erlangen–Nuremberg
Cauerstr.7, 91058 Erlangen, Germany
wk@LNT.de

ABSTRACT

For hands-free, untethered full-duplex speech communication at the human-machine interface, the acoustic feedback from the loudspeaker(s) to the microphone(s) has to be suppressed. Following the analysis of the problem, the basic concept of acoustic echo cancellation is reviewed. Typical adaptation algorithms are discussed and adaptation control as well as computationally efficient filtering schemes are addressed. Moreover, properties and problems of extensions to multi-channel signal acquisition using microphone arrays, and to multi-channel sound reproduction are outlined.

1. INTRODUCTION

Ideally, for natural speech communication at the human-machine interface, the human should not need to wear or hold any special device and should be allowed to move freely. This implies that acoustic feedback from the loudspeaker(s) into the microphone(s) must be expected.

In full-duplex telecommunication systems, e.g., hands-free telephony or teleconferencing, the remote talker will hear an echoed version of his own speech, with the perceptual annoyance increasing along with signal delay in the network [1, 2, 3]. Standard requirements ask here for 45dB of echo suppression during single-talk and 30dB during double talk [4]. For speech recognition systems with simultaneous speech capture and loudspeaker output, as e.g., desirable for speech-controlled TV sets or game stations, the impairment of recognition rates by acoustic echoes depends on the speech recognizer’s ability to separate interfering echoes from the desired speech.

Disregarding supplementary echo suppression methods, we focus here on acoustic echo cancellation (AEC) as the only known method which, ideally, could remove the echo without impairing other signals [1,2,3]. After reviewing the principles for the single-loudspeaker/single-microphone case, we consider AEC for beamforming microphone arrays, which are attractive because of their spatial filtering capability allowing both dereverberation and noise suppression [5]. Finally, the problems arising for AEC with stereo reproduction are outlined as a basis for AEC with multi-channel reproduction.

2. SINGLE-CHANNEL ACOUSTIC ECHO CANCELLATION (AEC)

The basic scenario for AEC is depicted in Fig. 1 for the case of a single loudspeaker and a single microphone, where \( v(n) \) – as an echoed version of the loudspeaker signal \( u(n) \) – adds to the desired speech \( s(n) \) and to the interfering local noise \( r(n) \) in the microphone signal \( x(n) \) (n denotes discrete time). To remove the echo from the microphone signal \( x(n) \), AEC subtracts an echo estimate \( \hat{v}(n) \) from \( x(n) \) to produce the residual echo

\[
e(n) = u(n) - \hat{v}(n).
\]  

Then, the estimate for the desired signal \( \hat{s}(n) \) is given by:

\[
\hat{s}(n) = x(n) - \hat{v}(n) = s(n) + e(n) + r(n).
\]  

The amount of echo attenuation achieved by AEC is often expressed by the echo return loss enhancement (ERLE) [2], which is the estimated ratio of the powers of \( v \) and \( e \), respectively.

Disregarding possible nonlinearities, the loudspeaker-enclosure-microphone (LEM) system is characterized by its generally time-varying impulse response \( h(k, n) \) [2] and is commonly modelled by a digital FIR filter structure \( \hat{h}(k, n) \) of length \( L_{AEC} \), so that the estimated echo \( \hat{v}(n) \) is given by

\[
\hat{v}(n) = \hat{h}^T(n) \cdot u(n)
\]  

where \( T \) denotes transposition and

\[
\hat{h}(n) = [\hat{h}(k, n), \ldots, \hat{h}(k + L_{AEC} - 1, n)]^T,
\]  

\[
u(n) = [u(n), \ldots, u(n - L_{AEC} + 1)]^T
\]  

The misalignment between the FIR model \( \hat{h}(n) \) and the real LEM system \( h(n) \) is described by the logarithmic system error norm \( D_{log}(n) \):

\[
D_{log}(n) = 10 \cdot \log_{10} \frac{|h(n) - \hat{h}(n)|_2^2}{|h(n)|_2^2}.
\]
with \( || \cdot ||_2 \) denoting the \( L_2 \)-norm. The number of impulse response samples that must be perfectly matched for \( D_{\text{log}} = x \) dB is estimated by [2, 6]

\[
L_{\text{ABC}} \approx \frac{x}{60} \cdot f_s \cdot T_{\text{d0}},
\]

(7)

where \( f_s \) denotes the sampling frequency, and \( T_{\text{d0}} \) is the reverberation time [7]. Accordingly, more than \( L_{\text{ABC}} = 1000 \) impulse response coefficients must be perfectly matched to assure 20dB of \( D_{\text{log}} \) for a typical office with \( T_{\text{d0}} = 400\text{msec} \) and an echo canceller operating at \( f_s = 8\text{kHz} \).

2.1. Adaptation algorithms

For identifying \( \hat{h}(n) \), adaptive filtering algorithms minimize a mean square error criterion based on the input \( u(n) \) and the estimation error \( e(n) \) (assuming \( s(n) = r(n) = 0 \)). In the following, two fundamental algorithms are introduced (for more see, e.g., [8, 9]). Adaptation control in the context of AEC is addressed and frequency domain implementations are outlined briefly.

2.1.1. Fundamental algorithms

Minimizing the mean squared error \( E \{ |e(n)|^2 \} \) for at least wide-sense stationary signals and a time-invariant echo path \( h(k, n) = h(k) \) leads to the Wiener-Hopf equation for the optimum echo canceller \( \hat{h}_{\text{opt}} \) [9]

\[
\hat{h}_{\text{opt}} = \mathbf{R}_{uu}^{-1} \cdot \mathbf{r}_{uv}
\]

(8)

with the time-invariant correlation matrix \( \mathbf{R}_{uu} = E \{ u(n)u^H(n) \} \) and the crosscorrelation vector \( \mathbf{r}_{uv} = E \{ u(n)v^*(n) \} \) (\( \cdot^* \) denotes complex conjugation and \( u^H = (u^*)^T \)). For nonstationary environments, iterative or recursive algorithms are required to approach the Wiener solution Eq.8. As the most popular adaptation algorithm, the NLMS (Normalized Least Mean Square) algorithm [8, 9] updates the filter according to

\[
\hat{h}(n + 1) = \hat{h}(n) + \alpha \frac{u(n)}{u^H(n)u(n)} e^*(n)
\]

(9)

with \( u(n) \) approximating the negative gradient vector, and a step-size parameter \( \alpha \) (0 < \( \alpha < 2 \)). For correlated signals such as speech, \( u(n) \) will not uniformly cover the \( L_{\text{ABC}} \)-dimensional vector space, which implies that the convergence to minimum system error \( D_{\text{log}}(n) \) (Eq.6) is slow [9]. The attractivity of the NLMS algorithm results from its robust convergence behavior [2] and its low computational complexity (about \( 2L_{\text{ABC}} \) multiplications per sampling interval \( T \) are needed for implementing Eqs.1,3,9).

On the other hand, the most powerful and computationally costly adaptation method, the RLS (Recursive least squares) algorithm directly minimizes a weighted sum of previous error samples

\[
J(\hat{h}, n) = \sum_{k=1}^{n} \beta(k)|e(k)|^2, \quad 0 < \beta \leq 1.
\]

(10)

The solution has the form of Eq.8, however with time-dependent estimates for \( \mathbf{R}_{uu}(n), \mathbf{r}_{uv}(n) \). The update equation reads here

\[
\hat{h}(n + 1) = \hat{h}(n) + \mathbf{R}_{uu}^{-1}(n)u(n)e^*(n),
\]

(11)

with

\[
\hat{R}_{uu}(n) = \sum_{k=1}^{n} \beta(k)u(k)u^H(k).
\]

(12)

If an exponential window \( \beta(k) = \lambda^{n-k} \) with the forgetting factor \( 0 < \lambda < 1 \) is used, the inversion of \( \hat{R}_{uu}(n) \) can be replaced by recursively updating the inverse [9]. Then, the resulting complexity of the order of \( L_{\text{ABC}}^2 \) can be further reduced to \( 7L_{\text{ABC}} \) using Fast RLS algorithms, which, however, require extra efforts to assure stable convergence [8].

2.1.2. Adaptation control

Adaptation control has to satisfy two contradicting requirements: For one, changes of the echo path \( h(k, n) \) should be tracked as fast as possible. This asks for a large stepsize \( \alpha \) for the NLMS algorithm (see Eq.9), and a rapidly decaying \( \beta \) for the RLS algorithm (Eq.11), respectively. On the other hand, the adaptation must be robust to interfering local sources \( s(n) \) and noise \( r(n) \), which requires a small stepsize \( \alpha \) and a slowly decaying \( \beta \), respectively [2, 9]. Thus, for robust practical implementations, detection of local source activity and estimation of background noise levels must be fast and reliable, which implies significant extra computational cost [2]. Especially for fast-converging algorithms, the RLS algorithm, a double-filter structure [10] is often preferred, where a nonadaptive foreground filter is replaced by an adaptive background filter only when the background filter performs better.

2.1.3. Frequency subband and transform domain structures

To reduce computational load and to speed up convergence of adaptation algorithms which do not inherently decorrelate \( u(n) \) (e.g., the NLMS algorithm), frequency subband and frequency domain adaptive filtering (FDAF) structures have been developed [3, 11]. Subband structures decompose the fullband signals \( u(n) \) and \( x(n) \) into \( M \) subbands which are usually downsampled by \( R < M \) [1, 12]. The adaptive subband filters operate at a reduced sampling rate and require less coefficients which leads to overall computational savings by a factor of close to \( R^2/M \) compared to fullband adaptive filtering. While the additional complexity for the analysis/synthesis filter banks is relatively small for large \( L_{\text{ABC}} \), the introduced signal delay for \( \hat{e}(n) \) is objectionable in some applications [2, 4].

FDAF structures draw their computational advantage over time-domain filtering from the fast Fourier transform (FFT) and its use for fast convolution [3, 8, 11, 13]. For balancing computational efficiency and signal delay due to the transforms, the system model \( h(k, n) \) is often partitioned into shorter subsystems to allow shorter transform lengths.

2.2. Examples for AEC adaptation

In Fig.2 the convergence for several adaptation schemes is depicted for speech input using the misalignment Eq.6 as criterion (SNR=35dB, \( f_s = 11025\text{Hz}, L_{\text{ABC}} = 2048 \)). The fullband NLMS algorithm is compared to a frequency subband version (\( N = 32, R = 24 \), see [12]) for both the NLMS and the RLS algorithm. Moreover, unconstrained FDAF adaptation is shown for both the single-channel case [13] and the stereo case (according to [14, 15], see Section 4) with an overlap of \( 3L_{\text{ABC}}/4 \) of adjacent time frames.
Regarding computational complexity, the fullband NLMS requires about 8200 real-valued floating point operations (FLOPs) per input sample, which is as much as a fast implementation of the subband RLS algorithm would require [16], 6 times more than needed for b and e, and 13 times more than required for d.

3. AEC WITH BEAMFORMING MICROPHONE ARRAYS

A general model of a beamforming microphone array is shown in Fig.3 [17], where $N$ real-valued sensor signals $x^\nu(n), \nu = 0, \ldots, N - 1$, are filtered by linear time-varying systems with impulse responses $g^\nu(k, n)$ and then summed up (Fig.3).

Fig. 3. General structure for a beamforming microphone array

Two generic options for combining AEC with such an array foresee either one AEC for each of the sensor signals $x^\nu(n)$ ('AEC first') or a single AEC which aims at removing the echo from the beamformer output $y(n)$ ('Beamforming first') [6, 18].

Obviously, 'AEC first' requires $N$ adaptive filters, which face the same system identification problem as in the single-microphone case. The computational complexity does not necessarily increase proportional to $N$ as all operations for filter adaptation that rely exclusively on the input $u(n)$ have to be performed only once, which is especially advantageous for complex adaptation algorithms (cf. Eq.11,12 and [8, 9, 19]). While for signal-independent beamforming [17] the echo suppression provided by AEC and the directivity of the beamformer simply add up, an adaptive signal-dependent beamformer will even compensate for insufficient echo suppression of AEC, e.g., during its initial convergence [18, 20].

The 'beamforming first' structure is attractive as only a single echo canceller is needed. However, this has to incorporate the beamformer into its echo path model $\hat{h}(k, n)$, ideally given by:

$$\hat{h}(k, n) = \sum_{\nu=0}^{N-1} g^\nu(k, n) * h^\nu(k, n),$$

with $h^\nu(k, n)$ connecting $u(n)$ and $x^\nu(n)$. For time-invariant beamforming, $g^\nu(k, n) = g^\nu(k)$, the AEC problem is identical to the single-channel AEC except for an increased filter length determined by $g^\nu(k)$. For typical adaptive beamforming, however, the 'beamforming first' structure will not be effective, as the time-variance of $g^\nu(k, n)$ cannot be tracked by the adaptation algorithm for $\hat{h}(k, n)$ [6, 18]. Then, a decomposition of the time-variant beamforming into a time-invariant stage followed by a time-variant voting stage is applicable, with the AEC acting on the output of the time-invariant stage [6].

4. AEC FOR MULTI-CHANNEL SOUND REPRODUCTION

For speech communication systems with multi-channel sound reproduction, like videoconferencing or game stations, the AEC problem changes characteristically. This is explained here for the stereo case shown in Fig.4 [21, 22]. Echo resulting from the far-end

source signal $u(n) \neq 0$ is perfectly cancelled if

$$g_1(n) * \left[ h_1(n) - \hat{h}_1(n) \right] + g_2(n) * \left[ h_2(n) - \hat{h}_2(n) \right] \triangleq 0.$$  

(14)

Aside from the desired solution $\hat{h}_1(n) = h_1(n), \hat{h}_2(n) = h_2(n)$, Eq.14 has other solutions which depend on $g_1(n), g_2(n)$. This nonuniqueness prevents any adaptation algorithm from finding a stable solution in practical situations. Defining

$$\tilde{h}(n) := \left[ \tilde{h}^1(n) \hspace{1cm} \tilde{h}^2(n) \right]^T \text{ and } u(n) := \left[ u^1(n) \hspace{1cm} u^2(n) \right]^T,$$

with $\tilde{h}_i(n), u_i(n)$ according to Eqs.4,5, respectively, then Eqs.3,6,8 as well as the update equations Eqs.9,11 remain unchanged. However, as $\tilde{u}^1(n)$ and $\tilde{u}^2(n)$ are correlated, the components of the gradient estimate for the NLMS $\hat{v}(n)$ are also highly correlated so that its convergence behaviour is poor. This also translates to the autocorrelation matrix

$$R_{uu} = \begin{pmatrix} R_{u_1 u_1} & R_{u_1 u_2} \\ R_{u_2 u_1} & R_{u_2 u_2} \end{pmatrix}$$

(15)
which is rank-deficient for linearly dependent $u_1(n)$, $u_2(n)$ [21]. In practice, uncorrelated low-level noise can be expected in $u_1(n)$, $u_2(n)$, which changes the nature of $R_{uw}$ from rank-deficient to ill-conditioned, so that a unique solution exists which, however, requires long convergence time. If the group delay difference and the correlation between $g_1(k,n)$ and $g_2(k,n)$ is large, the unmodelled tails of the impulse responses $h_1(k,n), h_2(k,n)$ impair convergence significantly more than in the monaural case [21].

For improved convergence, the correlation between $u_1(n)$, $u_2(n)$ must be reduced without impairing the perceptual audio quality and the spatial representation. The proposed methods include different nonlinear distortion for $u_1(n)$ and $u_2(n)$ [14], adding noise below the auditory masking threshold [23, 24], and time-varying allpass filtering [25]. However, results still lag far behind monaural AEC (see Fig.2). From the structure of the stereo problem it is obvious that convergence problems increase with increasing number of reproduction channels asking for further new and improved algorithms.

5. SUMMARY AND OUTLOOK

Acoustic echo cancellation is a standard technique for the single-channel acquisition/single-channel reproduction scenario since sufficiently powerful signal processing hardware is available at reasonable cost. Current research is fueled here by the desire for faster and more robust convergence. For the combination of AEC with microphone arrays, some synergies have already been explored, further research should yield even more efficient structures. With current knowledge, the greatest challenge for AEC seems to be the multi-channel reproduction scenario, where fast convergence requires efficient decorrelation of the reproduction channels that, ideally, should not impair reproduction quality.

6. ACKNOWLEDGEMENT

The author wishes to thank Herbert Buchner for providing the simulation results and Susanne Keschuny for preparing the illustrations.

7. REFERENCES


