Audio Engineering Society
Convention Paper
Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

Perceptually Motivated Low Complexity Acoustic Echo Control

Christof Faller[1]

[1] *Media Signal Processing Research, Agere Systems, Murray Hill, New Jersey 07974, USA*

Correspondence should be addressed to Christof Faller (cfaller@agere.com)

ABSTRACT

In hands-free two-way communications systems, the loudspeaker signal feeds back to the microphone, resulting in an undesired echo signal component in the microphone signal. Acoustic echo cancelers model the echo path and subtract an estimate of the echo signal from the microphone signal to remove the undesired echo. We present a novel scheme which estimates the echo signal in terms of its spectral envelope. The time and frequency resolution with which this estimation is carried out is chosen according to perceptual criteria. Given the estimated spectral envelope of the echo signal component, speech enhancement and noise suppression algorithms are used to suppress the echo. The presented scheme has low complexity and a high degree of robustness.

1 INTRODUCTION

In full-duplex hands-free telecommunication systems the far-end speech at the loudspeaker feeds back to the microphone. The result is that there is an undesired echo signal component in the microphone signal. An acoustic echo canceler (AEC) [1], which re-

moves the undesired echo signal, is shown in Fig. 1. It is assumed that the echo path can be accurately represented by a linear filter h_e. The far-end talker signal x goes through this echo path and adds to the microphone signal together with the near-end talker signal v and the ambient noise w,

$$y(k) = \mathbf{h}_e^T \mathbf{x}(k) + v(k) + w(k)\,, \tag{1}$$

where

$$\begin{aligned} \mathbf{x}(k) &= [x(k)\, x(k-1) \ldots x(k-M+1)]^T \\ \mathbf{h}_e &= [h_{e,0}\, h_{e,1} \ldots h_{e,M-1}]^T\,, \end{aligned} \tag{2}$$

and M is the length of the echo path response. The AEC uses an adaptive filter $\hat{h}_e$ to estimate to echo path response h_e. The error signal is defined as

$$e_e(k) = y(k) - \hat{\mathbf{h}}_e^T \mathbf{x}(k)\,, \tag{3}$$

where

$$\hat{\mathbf{h}}_e = [\hat{h}_{e,0}\, \hat{h}_{e,1} \ldots \hat{h}_{e,L-1}]^T \tag{4}$$

is the adaptive filter coefficient vector of length L (usually $L < M$).

The sum of the near-end talker and ambient noise, $v + w$, is the noise in the microphone signal with respect to estimation of the echo path. When the near-end talker is silent, i.e. $v(k) = 0$, this noise is only w and the adaptive filter $\hat{h}_e$ converges to a good estimate of the echo path response h_e and the echo is canceled successfully. When doubletalk is present, i.e. the near-end and far-end talkers are active at the same time, the near-end talker signal v acts as high level uncorrelated noise, causing the adaptive filter to diverge, resulting in insufficient echo cancellation. To prevent this from happening, a doubletalk detector [2, 3, 4] is used. Whenever doubletalk is detected, the adaptive filter coefficients are frozen.

Usually an echo path impulse response of 50−300 ms needs to be considered. To achieve even a modest improvement, e.g. a misalignment

$$\boldsymbol{\epsilon}_e = \mathbf{h}_e - \begin{bmatrix} \hat{\mathbf{h}}_e \\ \mathbf{0} \end{bmatrix} \tag{5}$$

of 20 dB below the uncanceled impulse response h_e, a cancellation filter with $L = 500$ taps needs to be considered at 8 kHz sampling rate for a small office. For larger rooms and higher sampling rates the number of taps that need to be considered rises to several thousand. As a result of the high number of taps the computational complexity of echo cancelers is high. Additionally, high order predictors such as are used for AECs are difficult to implement with high enough precision with fixed point arithmetic, as is often necessary for DSP implementations.

Since traditional echo cancelers are based on modeling a physical system (the acoustic echo path), it is hard to incorporate perception. The time-frequency resolution considered is dictated by the properties of the acoustic echo path and has to be chosen such that the estimated echo signal is accurate in magnitude and phase. That is, because the echo canceler simply subtracts the estimated echo from the microphone signal to cancel it. Magnitude or phase differences between the echo and its estimate result in insufficient cancellation.

In this paper, we are proposing a novel approach for echo control that is largely based on perceptual considerations. We call the scheme *Perceptual Echo Control* (PEC). The physical system is not estimated accurately. Instead of estimating the echo signal itself, only its spectral envelope is estimated. The time-frequency resolution in which this estimation is carried out is chosen according to perceptual criteria. Given this estimate, the spectral envelope of the microphone input signal is modified such that the echo is suppressed while maintaining the near-end talker signal, enabling duplex communication.

The paper is organized as follows. Section 2 de-

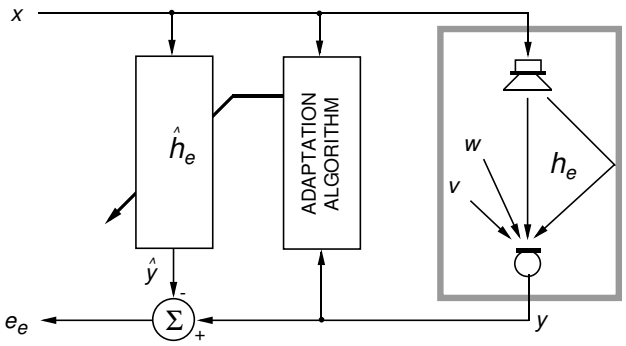

Fig. 1: An acoustic echo canceler removes echoes which arise from the coupling between the loudspeaker and the microphone.

scribes the PEC scheme in detail. Section 3 describes the results obtained from a set of numerical simulations and informal subjective test sessions. Conclusions are drawn in Section 4.

2 PERCEPTUAL ECHO CONTROL (PEC) SCHEME

2.1 Estimation of the Spectral Envelope of the Echo Signal

PEC assumes that the echo path can be modeled in terms of spectral envelopes. The spectral envelope is estimated by estimating the power of the echo signal in a number of subbands. Each of these power values corresponds to one discrete value of the spectral envelope. A filterbank is used with subbands that have similar properties as the spectral decomposition of the auditory system, such that the subband power values are the determining factors for loudness perception [5] of different signal components. This domain is suitable for suppressing signal components such as the echo signal. Also, the number of subbands of this domain is low, e.g. 20, resulting in low complexity of the scheme.

The signal is processed block wise using windowed FFTs with 50 % overlap. We use a sine window for analysis and synthesis. For a sample rate of 16 kHz we use a 256−point FFT. These parameters result in a delay of 16 ms (256 samples). The linearly spaced FFT coefficients are grouped into non-overlapping "subbands" such that each subband is one "critical bandwidth" wide. We found that the frequency resolution was sufficient when choosing the critical bandwidth equal to two ERB (*Equivalent Rectangular Bandwidth* [6]). In the remaining part of this paper *subband* denotes one group of samples corresponding to a critical band.

Figure 2 shows the proposed scheme. The loudspeaker signal x and the microphone signal y are converted to the subband domain. Then the processing is carried out for each subband separately. In the following, this processing is described for one subband. Given the signal of one subband of the loudspeaker signal $x(k)$, the operator P in Fig. 2 computes the power $p_x(n)$, where n is the time index in the (subsampled) subband domain. Similarly, $p_y(n)$ is computed for the corresponding subband of the microphone signal $y(k)$.

The concept of PEC assumes that the echo path can be modeled in the subsampled subband power domain with a linear filter h. The power of the far-end talker p_x "goes through" h resulting in the power of the echo signal component at the microphone,

$$p_e(n) = \mathbf{h}^T \mathbf{p}_x(n)\,, \tag{6}$$

where

$$\begin{aligned} \mathbf{p}_x(n) &= [p_x(n)\, p_x(n-1) \ldots p_x(n-N+1)]^T \\ \mathbf{h} &= [h_0\, h_1 \ldots h_{N-1}]^T\,, \end{aligned} \tag{7}$$

and N is the length of the modeling filter h. The power of the microphone signal is the sum of the powers of the echo signal p_e, near-end talker signal p_v, and ambient noise p_w,

$$p_y(n) = p_e(n) + p_v(n) + p_w(n)\,. \tag{8}$$

Similarly to how an AEC estimates the echo path response h_e, PEC estimates for each subband the modeling filter h with an adaptive filter $\hat{h}$. In this case, the error signal is defined as

$$e(n) = p_y(n) - \hat{p}_y(n)\,, \tag{9}$$

where

$$\hat{p}_y(n) = \hat{\mathbf{h}}^T \mathbf{p}_x(n) \tag{10}$$

is the estimated power of the echo and

$$\hat{\mathbf{h}} = [\hat{h}_0\, \hat{h}_1 \ldots \hat{h}_{N-1}]^T \tag{11}$$

is the adaptive filter coefficient vector of length N. The filter coefficients of $\hat{h}$ are adapted using either *least mean squares* (LMS) [7], *normalised least mean squares* (NLMS) [7], *recursive least square* (RLS) [7], *fast recursive least mean square* FRLS [8], or similar algorithms.

But as expected, (6) is only an approximation of the power of the echo signal with limited accuracy. The mean absolute relative error, averaged over all subbands is about 0.25 for a segment of male speech of 6 s length when h is chosen to be the Wiener solution for the chosen segment of speech. For traditional echo cancellation, the echo path is modeled with much higher accuracy. However, our results suggest that the accuracy achieved here is enough for the desired echo suppression. By taking into account prior knowledge (limitation of the estimated variables to

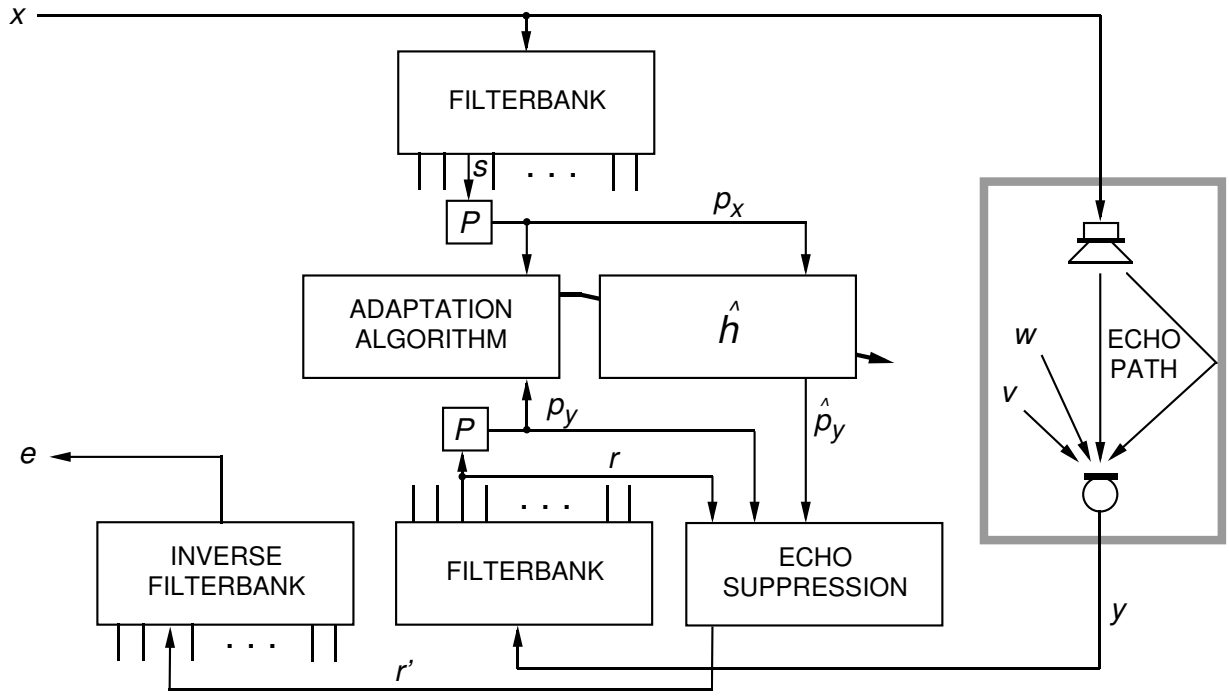

Fig. 2: PEC scheme is shown with the processing for one subband. An adaptive filter $\hat{h}$ estimates the power of the echo signal. The echo is suppressed according to this estimate.

their physically possible ranges) and perception (integrating the estimated variables as much in time as possible) the negative impact of this low precision is limited.

2.2 Suppression of the Echo Signal

For suppression of the echo, we chose to use a technique called *parametric spectral subtraction* [9]. Each subband signal $r(n)$ is scaled according to

$$r'(n) = a(n)r(n)\,, \tag{12}$$

with

$$a(n) = \left[\frac{p_y^{\frac{b}{2}}(n) - \eta\hat{p}_y^{\frac{b}{2}}(n)}{p_y^{\frac{b}{2}}(n)}\right]^{\frac{1}{b}}, \tag{13}$$

where b and η determine the strength and properties of the echo suppression. Before scaling (12), $a(n)$ is limited to the range $[0, 1]$, such that the microphone signal is never amplified. For $b = 2$ and $\eta = 1$, (13) is similar to *spectral power subtraction* [10], i.e. the subband is scaled such that the resulting power is equal to the power of the subband signal without the echo signal component.

After scaling (12), each subband of the microphone signal is converted back to the time-domain with the inverse of the filterbank used.

With the specific time-frequency resolution as defined by the FFT and subband-width parameters, the quality is considerably improved when $a(n)$ is smoothed over frequency and time. For smoothing over time a single-pole recursive average is used,

$$a'(n) = \alpha a(n) + (1-\alpha)a'(n-1)\,, \tag{14}$$

where we use $\alpha = 0.3$. For smoothing over frequency, $a(n)$ is averaged with a three tap filter between subbands, with filter coefficients $\{0.3, 1.0, 0.3\}$.

2.3 Doubletalk Control

For preventing that the filter coefficients diverge during doubletalk, we implemented a *doublepath* structure [11]. That is, additional memory is used for a second complete set of filter coefficients. One set is denoted the *foreground process* and is used for carrying out the echo suppression. The foreground filters are not adapted. The other set is called the

background process and the corresponding filter coefficients are adapted similarly to the regular PEC implementation. The coefficients from the background process are copied to the foreground process whenever the background process performs well and better than the background process.

2.4 Future Additions

Future versions of PEC may use a more sophisticated perceptual model for the suppression by taking into account masking [5], loudness [12], and continuity illusions [13] explicitly. Also, features such as noise-suppression [10, 14] or comfort noise [15] may be added to PEC.

3 SIMULATIONS AND RESULTS

The aim of this section is to assess the performance of the presented scheme. It is not trivial to compare PEC to conventional echo cancelers since measures such as the misalignment have a different meaning for PEC. The performance of PEC can not be measured just by means of SNR or mean square error since PEC is optimized taking perception into consideration. Therefore, we not only assessed PEC in terms of numerical simulations but also integrated PEC into a real-time *voice over IP* (VoIP) system and subjectively evaluated its performance.

3.1 Numerical Simulations

All the simulations were carried out using an NLMS algorithm for adapting the adaptive filter coefficients. For the simulations we used noisy wideband speech signals with a sample rate of 16 kHz. We used Gaussian noise for w for an SNR of 20 dB when the near-end talker is silent. We used a measured room impulse response of length 4096 samples (256 ms). As mentioned before, the FFT size used is 256 samples with sine windows with 50 % overlap. The number of the 2 ERB wide subbands is 17 and $M = 4$ filter coefficients are used for each subband. We experimented with different number of filter coefficients and it turned out that $M = 4$ was sufficient.

For optimizing and assessing the adaptation algorithm of PEC, the misalignment for PEC is normalized and expressed in dB,

$$m(n) = 10 \log_{10} \left(\frac{\| \mathbf{h} - \hat{\mathbf{h}}(n) \|^2}{\| \mathbf{h} \|^2} \right), \qquad (15)$$

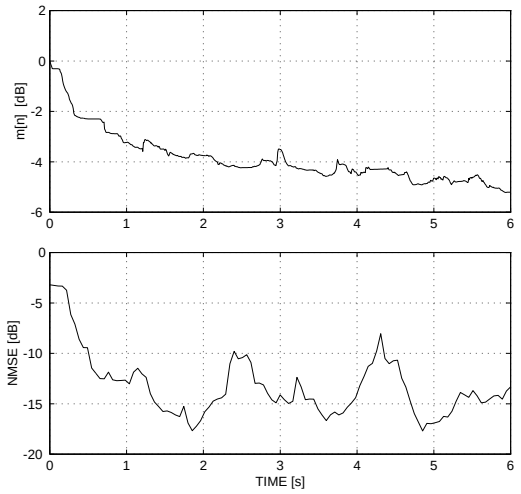

Fig. 3: The convergence $m(n)$ for a segment of speech (top) and the corresponding NMSE (bottom).

where $\mathbf{h}$ is the vector of the desired filter coefficients (6) (we used the least square solution over the whole time span of the simulation as the desired coefficients) and $\hat{\mathbf{h}}(n)$ are the NLMS obtained filter coefficients at time index n. The NLMS step size can be optimized for each subband by using (15). A more global measure can be obtained by using (15) with $\mathbf{h}$ and $\hat{\mathbf{h}}(n)$ being the concatenation of the filter coefficients of each subband. The following figures show this measure and it is called *convergence* in the remaining part of this paper. For all simulations the initial filter coefficients were set to zero, $\hat{\mathbf{h}}[0] = \mathbf{0}$. The suppression parameters (13) are chosen to be $b = 2$ and $\eta = 1$.

The top of Fig. 3 shows an example of $m(n)$ for 6 s of male speech. The *normalized mean square error* (NMSE),

$$NMSE = 10 \log_{10} \frac{\sigma_e^2}{\sigma_y^2}, \qquad (16)$$

is used as a measure for the echo suppression. An estimate of the NMSE (.5 s rectangular estimation window) is shown on the bottom of Fig. 3. Among other factors, the correspondence of $m(n)$ and the NMSE depends on the SNR, b, and η.

To asses how sensitive PEC is to phase changes, we periodically inverted the phase of the microphone signal $y(k)$ as shown in the top of Fig. 4. The bottom

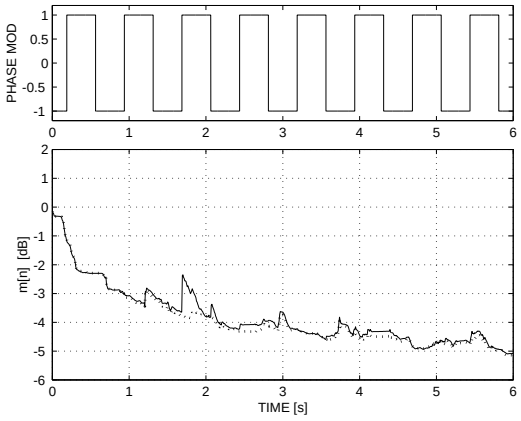

Fig. 4: The phase of the echo signal is toggled between 0 (1) and π (-1) (top). The convergence of the NLMS obtained filter coefficients with (bottom, solid) and without (bottom, dotted) phase toggling.

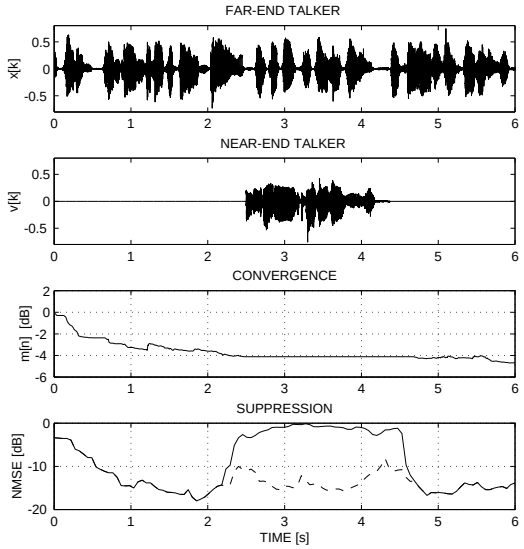

Fig. 5: Simulation with an active far-end and near-end talker with ideal doubletalk detection.

of Fig. 4 shows that the adaptation performance of PEC is virtually the same for both cases, with and without phase modulation. This is in contrast to traditional echo cancelers which are very sensitive to such phase changes.

Figure 5 shows the results of a simulation with doubletalk and ideal doubletalk detection. The estimation filter stays converged during the doubletalk and the NMSE (bottom of Fig. 5, solid) suggests that the doubletalk is not suppressed. To assess if during doubletalk the echo is still being suppressed, we rerun the simulation with identical adaptive filter coefficients at each time instance. The result (bottom of Fig. 5, dashed) suggests that the echo is suppressed, also in the period when doubletalk occurs.

3.2 Subjective Evaluation

We integrated PEC into a conferencing system consisting of two laptops connected over IP. Loudspeakers were connected to both laptop computers and their internal microphones were used. We tested the system with about 10 sessions where two people talked to each other for $15-60$ min. The untrained subjects generally were pleased with the performance of the system, which never run instable. Only one experienced subject, was able to make the system temporarily instable by moving a loudspeaker very close to the microphone.

These informal test sessions were carried out with three types of loudspeakers (small, medium, large). PEC proved to be robust for all of them, without a need for re-tuning parameters.

During doubletalk both talkers mostly get through simultaneously, however there often arise some artifacts. If one of the talkers is much louder, it may be that the other talker gets mostly suppressed. Therefore, it is also important to balance the system in the sense that different clients operate at similar levels.

4 CONCLUSIONS

We presented a novel scheme for echo control: perceptual echo control (PEC). Not the waveform of the echo signal is estimated, but only the spectral envelope thereof. The number of parameters that need to be estimated is significantly reduced, such that the computational complexity of the scheme is significantly lower than that of traditional acoustic echo cancelers (AECs).

Advantages of PEC over traditional AECs is reduced computational complexity and insensitivity for sudden changes in the phase or spectral fine structure of the echo path. The latter result in that when minor echo path changes occur (minor head and body movements) no residual echo appears. Thus in contrast to conventional AECs, there is no need for an additional suppression scheme for residual echoes.

Also, PEC is insensitive to time-jitter and phase changes making it particularly suitable for implementation on personal computers which often can not guarantee jitter-free audio input and output.

During doubletalk PEC introduces some artifacts into the speech. When the loudness of near-end and far-end is not balanced it may occur that one of the talker signals is partially suppressed. Despite of these drawbacks, the subjects of informal testing with a real-time VoIP conferencing system graded the system as performing well.

5 REFERENCES

[1] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. 46, pp. 497–510, Mar. 1967.

[2] D. L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Commun.*, vol. 26, pp. 647–653, May 1978.

[3] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE trans. Commun.*, vol. 39, pp. 1542–1545, Nov. 1991.

[4] T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE trans. Commun.*, vol. 44, pp. 1421–1427, Nov. 1996.

[5] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, New York, 1999, 1999.

[6] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[7] S. Haykin, *Adaptive Filter Theory (third edition)*, Prentice Hall, 1996.

[8] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer, 2001.

[9] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[10] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE trans. Acoust. Speech Sig. Processing*, vol. 27, no. 2, pp. 113–120, Nov. 1979.

[11] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceler with two echo path models," *IEEE trans. on Communications*, vol. 25, no. 6, pp. 589–595, June 1977.

[12] J. J. Zwislocki, "Temporal summation of loudness: an analysis," *J. Acoust. Soc. Am.*, vol. 46 (2), no. 2, pp. 431–441, 1969.

[13] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, Cambridge, Massachusetts, 1990.

[14] Y. Ephrain and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Processing (Boston)*, 1983, pp. 1118–1121.

[15] International Telecommunication Union, "Digital network echo cancellers," Recommendation ITU-T G.168, 2000.