# AN EM-BASED PROBABILISTIC APPROACH FOR ACOUSTIC ECHO SUPPRESSION

*Nilesh Madhu*[*]

Institute of Communication Acoutics (IKA)
Ruhr-Universität Bochum, Germany
nilesh.madhu@rub.de

*Ivan Tashev and Alex Acero*

Microsoft Research,
One Microsoft Way, Redmond, WA 98052,USA
{ivantash,alexac}@microsoft.com

**ABSTRACT**

This paper introduces a new Acoustic Echo Suppression (AES) algorithm for suppressing the residual echo after the Acoustic Echo Canceller (AEC). By temporally segmenting the frequency bins of the residual signal spectrum into blocks and modelling the data in each block and each frequency bin as realizations of a random variable, we can compute the probability of presence of residual echo and derive an appropriate ML suppression rule based on this probability. The computation of the probabilities is based on the Expectation Maximization algorithm. The proposed method shows better performance as compared to state of the art methods for residual echo suppression while producing no audible degradation in the near end signal and no musical noise. Test results indicate that the proposed approach provides an increase in the ERLE of up to 3 dB more than the state of the art echo suppressor while yielding a comparable mean opinion score (MOS) for the near end speech quality. Furthermore, the proposed method is independent of the double talk detector – which makes it robust to misclassifications on the part of the AEC algorithm.

***Index Terms***— Echo suppression, EM learning, Probabilistic suppression rule, Suppression rule, Adaptive filters

## 1. INTRODUCTION

Acoustic Echo Cancellation (AEC) algorithms often do not provide sufficient echo reduction. In part, this is because they model room impulse responses by short, finite-length filters, which cannot completely cancel echoes and, in part, because the reverberation tail is stochastic in nature – making it impossible to model by linear filters. This mismatch between the true impulse response and the estimated response of the AEC leads to what is termed as residual echo.

Over the past years, many methods have been developed to suppress the residual echo. These approaches, generically called Acoustic Echo Suppressors (AES), treat the echo signal as an uncorrelated interference that must be suppressed, and use methods from the area of single-channel signal enhancement for this purpose. Echo cancellation and suppression is generally done in the short-time–frequency domain (the spectral representation of an acoustic signal obtained from overlapped, windowed, discrete frequency transforms). AES algorithms then weight each short-time–frequency (T-F) point according to some optimization criterion that assigns a high gain to T-F points containing near-end signals and suppresses T-F points that predominantly contain echo.

The simplest of all techniques is center-clipping, which produces significant suppression – albeit with equally drastic near-end speech distortion, and is, furthermore, heavily dependent on the doubletalk detector. More sophisticated techniques estimate the power spectral density (PSD) of the residual echo for Weiner filtering [1, 2, 3] or spectral subtraction [4]. However, all these approaches bring with them the disadvantages common to single-channel noise suppression algorithms: distortions and musical noise.

The recent approach of [5] models the magnitude of the residual echo, in each frequency bin and in each time frame, as a linear combination of the loudspeaker signals of the previous time-frames, for that frequency bin. This leads to the estimation of a temporal, linear 'filter' for each frequency that minimizes the mean square error between the current microphone input frame and the loudspeaker outputs for the current and previous $L$ frames. For this approach, it is assumed that the AEC removes most of the phase information in the echo signal (corresponding to the direct path and the early reflections) leaving only the reverberation tail to be suppressed by the AES. The adaptation of the regression coefficients is done in the absence of doubletalk, necessitating a good doubletalk detector. Similar approaches are proposed in [6, 7]. As these methods also rely on spectral subtraction, they are sensitive to musical noise.

This paper is motivated by the approaches in [3, 5]. The spectrum of the residual is temporally segmented into short blocks (of the order of 0.3 – 0.5 s). For each block and each bin, we consider two mutually exclusive cases: (a) The block contains, mainly, near end signals (a mixture of near end speech and noise) or (b) it contains, mainly, the residual echo. Depending upon the hypothesis, we write the corresponding probability density function of the observed data and compute the probabilities of the respective hypothesis. Based on these estimates, we develop a new rule for residual suppression. The performance of the approach is compared with the state of the art [5] using the ERLE measure and the objective perceptual evaluation of near end speech quality [8, 9].

## 2. SIGNAL MODEL

The block diagram of the classic AEC+AES system is indicated in Figure 1. $z\left(t\right)$ denotes the far end signal; $x\left(t\right)$, the microphone input; $s\left(t\right)$, the near end speech and $\tilde{v}\left(t\right)$, the near end noise. $y\left(t\right)$ is the output from the AEC, containing the residual and the near end signals, and $\hat{y}\left(t\right)$ represents the output of the AES. We then have the following relations between these signals, in the time domain:

$$\begin{aligned} x\left(t\right) &= h\left(t\right) * z\left(t\right) + s\left(t\right) + \tilde{v}\left(t\right) \\ y\left(t\right) &= \left(h\left(t\right) - \hat{h}\left(t\right)\right) * z\left(t\right) + s\left(t\right) + \tilde{v}\left(t\right) \\ \hat{y}\left(t\right) &\approx s\left(t\right) + \tilde{v}\left(t\right), \end{aligned} \tag{1}$$

where $\hat{h}\left(t\right)$ is the estimate of the room impulse response from the AEC and $*$ indicates convolution. However, as the AEC and AES

[*]Work done while at Microsoft Research.

generally perform in the short-time–frequency domain, we may rewrite the above equation in this domain as:

$$\begin{aligned} X(k,n) &= H(k,n)\,Z(k,n) + S(k,n) + \tilde{V}(k,n) \\ Y(k,n) &= \left(H(k,n) - \hat{H}(k,n)\right) Z(k,n) + S(k,n) + \tilde{V}(k,n) \\ \hat{Y}(k,n) &\approx S(k,n) + \tilde{V}(k,n), \end{aligned} \tag{2}$$

where $k$ indicates the discrete frequency bin index and $n$, the frame. As far as the AES is concerned, we need distinguish only between

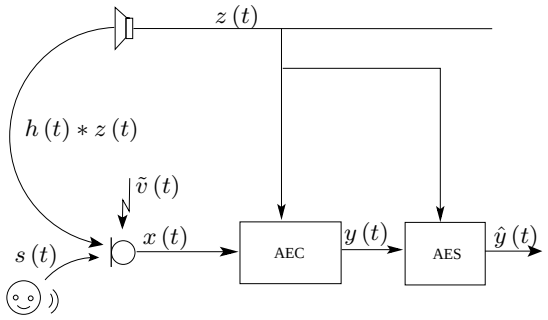


**Fig. 1**. Block diagram of the typical Acoustic Echo Reduction system

near end signals and the residual: we need not be concerned with the individual properties (speech/noise) of the near end signals. This is left to the noise suppressor which usually follows the AES. Subsequently, in a slight abuse of notation, we shall simplify the equations by using $V(k,n)$ to denote the *sum* of the near end signals:

$$V(k,n) = S(k,n) + \tilde{V}(k,n) \tag{3}$$

## 3. EM BASED AES

For each frequency bin $k$, we process the residual in a blockwise manner considering, at a time, a *block* consisting of $N$ frames of the spectrum. Such a block $b$ may be written as:

$$\mathbf{Y}_b(k) = \left(Y(k,bN), \ldots, Y(k,(b+1)N-1)\right)^T, \tag{4}$$

and is depicted in Figure 2, for a frequency bin $k$. Note that the

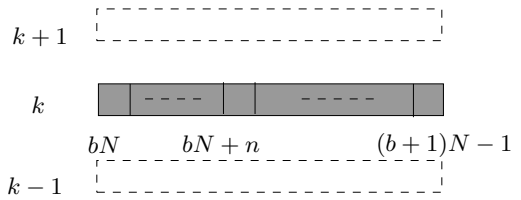


**Fig. 2**. Selection of a block of frequency frames for the proposed approach

selected blocks may or may not overlap.

For each block, and each bin, we shall posit two mutually exclusive hypotheses:

1. $\mathcal{H}_1$: the block contains, predominantly, the residual echo.
2. $\mathcal{H}_0$: the block contains, predominantly, signals from the near end,

and aim to find the probability of each hypothesis for each block and each bin. Once we have these probabilities, we compute the output of the AES in a manner analogous to the maximum likelihood estimate of McAulay and Malpass [10]:

$$\hat{\mathbf{Y}}_b(k) = P\left(\mathcal{H}_0^{k,b}\right) \mathbf{Y}_b(k). \tag{5}$$

Note that we have explicitly specified the block and frequency bin indices to indicate that the probabilities of the hypotheses vary with time and frequency. For the subsequent development of the method, however, they shall be dropped for purposes of convenience, and reintroduced where necessary. We next proceed by assuming that the signals are realizations of random variables, which allows us to write the probability of each observation of $Y(n) = Y(bN+n)$ given the speaker (far end) signal $\mathbf{Z}(n) = \left(Z(bN+n-L), \ldots, Z(bN+n)\right)^T$ as

$$p\left(Y(n)\,|\mathbf{Z}(n)\right) = \sum_{i=0}^{1} P(\mathcal{H}_i)\, p\left(Y(n)\,|\mathcal{H}_i, \mathbf{Z}(n)\right) \tag{6}$$

where we model the probabilities of the observed signal based on each hypothesis as:

$$p\left(Y(n)\,|\mathcal{H}_0, \mathbf{Z}(n)\right) \sim \mathcal{N}\left(0, \Psi_{v_1}\right) \tag{7}$$

$$p\left(Y(n)\,|\mathcal{H}_1, \mathbf{Z}(n)\right) \sim \mathcal{N}\left(\mathbf{W}^H \mathbf{Z}(n), \Psi_{v_2}\right). \tag{8}$$

Equation (8) is inspired by the approach in [5]: the residual is modelled as a linear combination of the current and $L$ previous frames of the speaker signal.

Note that the variances $\Psi_{v_1}$ and $\Psi_{v_2}$ in (7) and (8) are modelled as different variables to take the mismatch of the regression model into account.

The unknown parameters that need to be estimated are: the probabilities $P(\mathcal{H}_i)$, the variances $\Psi_{v_i}$, and the regression coefficients $\mathbf{W}$. These can be written as a vector of parameters:

$$\mathbf{\Theta} = \left(P(\mathcal{H}_0), P(\mathcal{H}_1), \Psi_{v_1}, \Psi_{v_2}, \mathbf{W}^T\right)^T \tag{9}$$

For notational convenience, we shall subsequently replace the $P(\mathcal{H}_i)$ by $\alpha_i$. Our problem, now, is the following: given $N$ observations $Y(n)$ and the corresponding far end signal vectors $\mathbf{Z}(n)$, $n = 0, \ldots, N-1$, estimate the parameter vector $\mathbf{\Theta}$. The maximum likelihood [11] solution of this problem is:

$$\begin{aligned} \mathbf{\Theta}_{\text{opt}} &= \underset{\mathbf{\Theta}}{\text{argmax}}\, p\left(Y(0), \ldots, Y(N-1)|\mathbf{Z}(0), \ldots, \mathbf{Z}(N-1), \mathbf{\Theta}\right) \\ &= \underset{\mathbf{\Theta}}{\text{argmax}} \prod_{n=0}^{N-1} p\left(Y(n)|\mathbf{Z}(n), \mathbf{\Theta}\right), \end{aligned} \tag{10}$$

where the last simplification is obtained by assuming conditional temporal independence among the signals $Y$. Using (7) and (8) in (10), we obtain the following *log* likelihood function:

$$\begin{aligned} \mathbf{\Theta}_{\text{opt}} &= \underset{\mathbf{\Theta}}{\text{argmax}}\, \ln\left(\prod_{n=0}^{N-1} \sum_{i=0}^{1} P(\mathcal{H}_i)\, p\left(Y(n)|\mathcal{H}_i, \mathbf{Z}(n), \mathbf{\Theta}\right)\right) \\ &= \underset{\mathbf{\Theta}}{\text{argmax}} \sum_{n=0}^{N-1} \ln\left(\sum_{i=0}^{1} \alpha_i p\left(Y(n)|\mathcal{H}_i, \mathbf{Z}(n), \mathbf{\Theta}\right)\right). \end{aligned} \tag{11}$$

Maximizing the log likelihood function directly for the optimal $\mathbf{\Theta}$ is not trivial. However, recognizing that the form of (11) is similar

to the estimation of the mixture of densities problem [12], we may proceed towards solving this using the *Expectation Maximization* algorithm. Following [12], we form the complete log likelihood function by positing the existence of unobserved data that indicate which hypothesis "generated" each observed sample $Y(n)$. We then take the expected value over this complete log-likelihood function (the expectation step) and, after some trivial algebraic manipulations, obtain the auxiliary equation $Q\left(\mathbf{\Theta}, \mathbf{\Theta}^{(\ell)}\right)$ as:

$$Q\left(\mathbf{\Theta}, \mathbf{\Theta}^{(\ell)}\right) = \sum_{n=0}^{N-1}\sum_{i=0}^{1} \ln\left(\alpha_i p\left(Y(n)|\mathcal{H}_i, \mathbf{Z}(n), \mathbf{\Theta}\right)\right) \cdot P\left(\mathcal{H}_i|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right), \quad (12)$$

where $\mathbf{\Theta}^{(\ell)}$ is the estimate of the parameter vector at iteration $\ell$. In the above, we may obtain $P\left(\mathcal{H}_i|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right)$ using the Bayes' rule as:

$$P\left(\mathcal{H}_i|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) = \frac{\alpha_i^{(\ell)} p\left(Y(n)|\mathcal{H}_i, \mathbf{Z}(n), \mathbf{\Theta}\right)}{\sum_{i'=0}^{1} \alpha_{i'}^{(\ell)} p\left(Y(n)|\mathcal{H}_{i'}, \mathbf{Z}(n), \mathbf{\Theta}\right)}. \quad (13)$$

From (7), (8) and (12), the auxiliary function may be expanded to:

$$\begin{aligned} Q\left(\mathbf{\Theta}, \mathbf{\Theta}^{(\ell)}\right) = \sum_{n=0}^{N} \Bigg( & \sum_{i=0}^{1} \ln\left(\alpha_i\right) P\left(\mathcal{H}_i|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) \\ & - \left(\ln\left(\Psi_{v_1}\right) + \frac{|Y(n)|^2}{\Psi_{v_1}}\right) P\left(\mathcal{H}_0|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) \\ & - \left(\ln\left(\Psi_{v_2}\right) + \frac{|Y(n) - \mathbf{W}^H \mathbf{Z}(n)|^2}{\Psi_{v_2}}\right) P\left(\mathcal{H}_1|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) \Bigg) \end{aligned} \quad (14)$$

The optimal parameters for the next iteration may then be found by maximizing (14) with respect to the parameters, and using the constraints $\sum_i \alpha_i = 1$. This yields the following update rules:

$$\alpha_i = \frac{1}{N}\sum_{n=0}^{N-1} P\left(\mathcal{H}_i|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) \quad (15)$$

$$\Psi_{v_1} = \frac{\sum_{n=0}^{N-1} |Y(n)|^2 P\left(\mathcal{H}_0|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right)}{\sum_{n=0}^{N-1} P\left(\mathcal{H}_0|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right)} \quad (16)$$

$$\Psi_{v_2} = \frac{\sum_{n=0}^{N-1} |Y(n) - \mathbf{W}^{(\ell)H} \mathbf{Z}(n)|^2 P\left(\mathcal{H}_1|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right)}{\sum_{n=0}^{N-1} P\left(\mathcal{H}_1|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right)} \quad (17)$$

$$\mathbf{W} = \mathbf{R}^{-1}\mathbf{p} \quad (18)$$

where:

$$\mathbf{R} = \left(\sum_{n=0}^{N-1} P\left(\mathcal{H}_1|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) \mathbf{Z}(n)\mathbf{Z}^H(n)\right)$$

and

$$\mathbf{p} = \left(\sum_{n=0}^{N-1} P\left(\mathcal{H}_1|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) Y^*(n)\mathbf{Z}(n)\right).$$

If we choose to neglect the additional variance for $\mathcal{H}_1$ due to the mismatch in modelling, we reduce the system of variables by one and the update rules are modified according to:

$$\begin{aligned} \Psi_v = \frac{1}{N}\Bigg( & \sum_{n=0}^{N-1} |Y(n)|^2 P\left(\mathcal{H}_0|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right) \\ & + \sum_{n=0}^{N-1} |Y(n) - \mathbf{W}^{(\ell)H}\mathbf{Z}(n)|^2 P\left(\mathcal{H}_1|Y(n), \mathbf{Z}(n), \mathbf{\Theta}^{(\ell)}\right)\Bigg). \end{aligned} \quad (19)$$

Note that only the update rule for the variance is affected.

Thus, we can obtain an estimate for the probability of each hypothesis in each block and each bin, and use this to compute the output signal as in (5).

## 4. EXPERIMENTAL RESULTS

To test the proposed approach, signals were recorded at a sampling frequency of 16 kHz, in an office room with a $T_{60}$ of 0.27s. The spectrum was computed using a 512 point FFT and a Hann window with a 50% overlap. The AEC is performed using a classic frequency domain NLMS algorithm. The output of the AEC was then processed using the EM based suppressor and was compared to the magnitude regression based approach of [5]. The parameters for our algorithm were as follows: $N = 24$, $L = 8$. The blocks were selected in a non-overlapping manner. For each frequency, 20 iterations of the EM algorithm were carried out. The quantitative performance of the algorithms was evaluated on the basis of their average Echo Return Loss Enhancement (ERLE) value, which is computed as in [13], for example. We also test the perceptual quality of the near end speech after the AES, using a slightly modified version of the wideband perceptual evaluation of speech quality [9] to obtain the Mean Opinion Score (MOS). The results are presented in Table 1. It may be seen

**Table 1**. Test results for the AEC, the AEC+AES based on the magnitude regression model of [5] and the AEC+AES based on the proposed system

| Test | Baseline | Reg. AES | EM-AES |
|---|---|---|---|
| ERLE (dB) | 12.37 | 17.54 | 21.15 |
| MOS | 3.2 | 3.69 | 3.78 |

that the proposed approach performs very well, yielding an average of 8.7dB of residual echo suppression as compared to the 5dB average of the regression based approach. The MOS values of both approaches are comparable indicating that the achieved ERLE increase is *not* at the cost of near end sound quality. Figures 4 and 5 illustrate the performance of the proposed AES approach in the presence of near end signal. For comparison, the segment containing only the near end signal is presented in Figure 3. Observe that the near end signal is preserved during its passage through the AES – including short segments comprising plosives and fricatives.

## 5. DISCUSSION AND CONCLUSIONS

We have proposed an approach to AES based on a probabilistic model. The probability of presence of residual echo is estimated by means of the EM algorithm and is used to compute an ML estimate of the output. Also, as the gains are limited between [0,1], and the estimation is performed on short blocks, musical noise is absent in the output.

One issue that needs to be addressed is the overlap between adjacent blocks. Decreasing the overlap would increase the latency of the

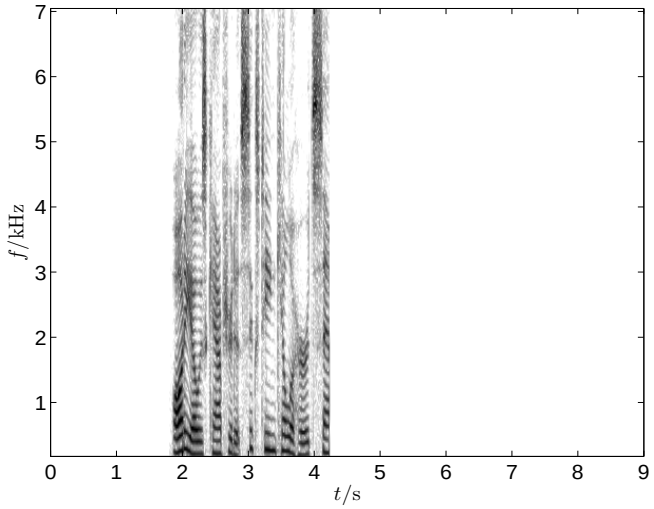


**Fig. 3**. Spectrum of the near end speech only.

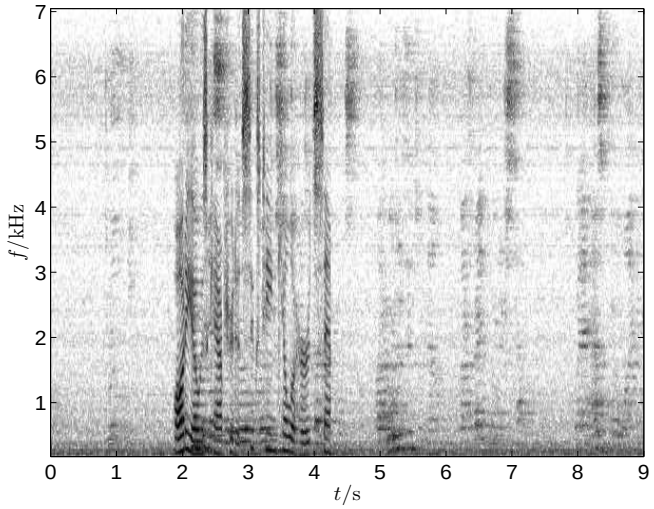


**Fig. 4**. Spectrum showing the residual along with doubletalk. Note the significant amount of residual echo.

system, whereas increasing the overlap increases the computational load. Therefore, this factor must be judiciously selected so that it provides an acceptable trade off between latency and computational expense.

Another issue is the tracking and the use of the regression coefficients and the variances in each hypothesis. Currently these variables are not utilized in the computation of the gain function - save as necessary 'nuisance' parameters. However, they provide useful information regarding the state of the environment. Optimal utilization of these parameters shall be the subject of a future publication.

## 6. REFERENCES

[1] G. Enzner, R. Martin, and P. Vary, "Unbiased residual echo power estimation for hands free telephony," in *Proc. ICASSP*, 2002.

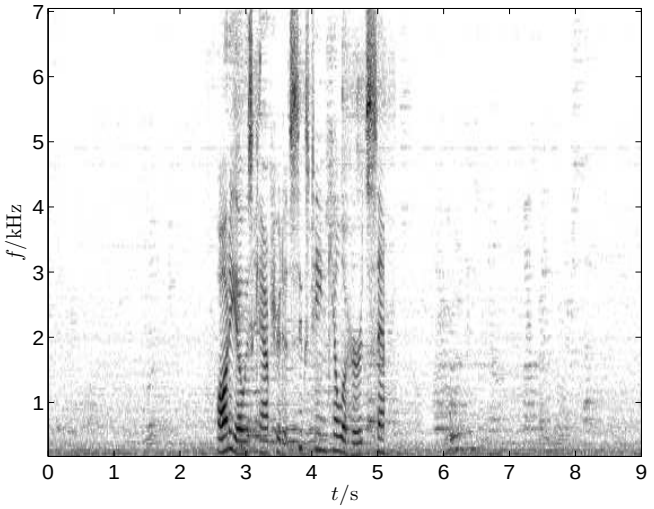


**Fig. 5**. Output of the proposed AES - Near end speech is preserved, residual is suppressed.

[2] M. Kallinger and K. Kammeyer, "Residual echo estimation with the help of minimum statistics," in *IEEE Benelux Signal Processing Symposium*, Mar. 2002.

[3] S. Y. Lee, J. W. Shin, H. S. Yin, and N. S. Kim, "A statistical model based post-filtering algorithm for residual echo suppression," in *Proc. INTERSPEECH*, 2007.

[4] C. Faller and C. Tournery, "Stereo acoustic echo control using a simplified echo path model," in *Proc. IWAENC*, Sept. 2006.

[5] A. S. Chhetri, A. C. Surendran, J. W. Stokes, and J. C. Platt, "Regression based residual acoustic echo suppression," in *Proc. IWAENC*, Sept. 2005.

[6] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. WASPAA*, Oct. 2001.

[7] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 1048 – 1062, Sept. 2005.

[8] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.," 2001.

[9] ITU-T Recommendation P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2005.

[10] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980 1980.

[11] H. L. Van Trees, *Detection, Estimation and Modulation Theory - Part I*, John Wiley & Sons, Ltd., New York, 2001.

[12] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep. ICSI-TR-97-021, University of Berkeley, 1997.

[13] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, John Wiley & Sons, Ltd., New York, 2006.