Methods for Noise Cancellation based on the EM Algorithm [1]

Meir Feder
Dept. of Electrical Engineering and Computer Science
Research Laboratory of Electronics, Room 36-615
Massachusetts Institute of Technology
Cambridge, MA 02139
and
Department of Ocean Engineering
Woods Hole Oceanographic Institution

Alan V. Oppenheim
Dept. of Electrical Engineering and Computer Science
Research Laboratory of Electronics, Room 36-615
Massachusetts Institute of Technology
Cambridge, MA 02139

Ehud Weinstein
Department of Ocean Engineering
Woods Hole Oceanographic Institution
Woods Hole, MA 02543
and
Department of Electronic Systems
Tel-Aviv University, Tel-Aviv, Israel

Abstract

Single microphone speech enhancement systems have typically shown limited performance, while multiple microphone systems based on a least-squares error criterion have shown encouraging results in some contexts. In this paper we formulate a new approach to multiple microphone speech enhancement. Specifically, we formulate a maximum likelihood (ML) problem for estimating the parameters needed for canceling the noise in a two microphone speech enhancement system. This ML problem is solved via the iterative EM (Estimate-Maximize) technique. The resulting algorithm shows encouraging results when applied to the speech enhancement problem.

1 Introduction

The problem of noise cancellation in single and multiple microphone environments has been extensively studied [1]. The performance of the various techniques in the single microphone case seems to be limited. However, in a two or multiple microphone case, the performance of an enhancement system may be improved due to the existence of reference signals.

Widrow et al. [2] suggested an adaptive solution for the two microphone case, based on the LMS algorithm. Adaptive solution based on the RLS algorithm also exist, and these algorithms have been applied in a speech enhancement context. [3], [4].

For the single microphone case, one of the variety of methods that have been suggested is the iterative enhancement method proposed by Lim and Oppenheim [5]. Although not developed from this point of view, this method can be shown to be an instance of a general iterative algorithm for maximum likelihood, introduced by Dempster et al. [6] and referred to as the EM algorithm.

In the EM algorithm, the observations are considered "incomplete" and the algorithm iterates between estimating the sufficient statistics of the "complete data" given the observations and a current estimate of the parameters (the E step) and maximizing the likelihood of the complete data, using the estimated sufficient statistics (the M step).

In [5] the observations are the desired signal with additive noise and the "complete data" is the signal and noise separately. The unknown parameters are some spectral parameters of the signal (LPC parameters, for speech). The algorithm iterates between Wiener filtering applied to the observations using the current spectral parameters of the signal (the E step), and updating the spectral parameters using the results of the Wiener filter (the M step).

In this paper we develop and demonstrate a method, based on the EM algorithm, for noise cancellation, in a two microphone situation. We emphasize the two microphone case, although the results may be extended to the more general multiple microphone case. While the problem may be posed as a maximum likelihood problem, maximizing the likelihood directly is complicated, and consequently the EM algorithm is used. The resulting procedure may be considered as an extension of the method in [5] to two microphones. We also propose an adaptive algorithm based on the EM algorithm, which may be an alternative to Widrow's approach in [2].

This paper is organized as follows: In the next section we describe the EM algorithm. In section 3 we formalize a ML problem whose solution (the estimated parameters) are used for canceling the noise. This ML problem is solved via the EM algorithm. The solution, together with a possible on-line scheme for its implementation, is described in section 4. We conclude, by evaluating the performance of the suggested system, especially compared with the technique of Widrow, in [2].

2 The EM algorithm

Let $\underline{Y}$ denote the data vector with the associated probability density $f_{\underline{Y}}(\underline{y};\underline{\theta})$ indexed by the parameter vector $\underline{\theta} \in \Theta$. Θ is a subset of the Euclidean K-space. Given an observed $\underline{y}$, the ML estimate $\hat{\underline{\theta}}_{ML}$ is the value of $\underline{\theta}$ that maximizes the log-likelihood, that is

$$\max_{\underline{\theta}\in\Theta} \log f_{\underline{Y}}(\underline{y};\underline{\theta}) \Longrightarrow \hat{\underline{\theta}}_{ML} \qquad (1)$$

Suppose that the data vector $\underline{Y}$ can be viewed as being incomplete, and we can specify some "complete" data $\underline{X}$ related to $\underline{Y}$ by

$$H(\underline{X}) = \underline{Y} \qquad (2)$$

[1] This work has been supported in part by the Advanced Research Projects Agency monitored by ONR under contract N00014-81-K-0742 and the National Science Foundation under grant ECS-8407285 at MIT, and in part by ONR under contract N00014-85-K-0272 at WHOI. M. Feder acknowledges the support of Woods Hole Oceanographic Institution.

CH2396-0/87/0000-0201 $1.00 © 1987 IEEE

where $H(\cdot)$ is a non-invertible (many to one) transformation.

The EM algorithm is directed at finding the solution to (1); however it does so by making an essential use of the complete data specification. The algorithm is basically an iterative method. It starts with an initial guess $\underline{\theta}^{(0)}$, and $\underline{\theta}^{(n+1)}$ is defined inductively by

$$\max_{\underline{\theta}\in\Theta} E\left\{\log f_{\underline{X}}(\underline{x};\underline{\theta})/\underline{y};\underline{\theta}^{(n)}\right\} \Longrightarrow \underline{\theta}^{(n+1)} \qquad (3)$$

where $f_{\underline{X}}(\underline{x};\underline{\theta})$ is the probability density of $\underline{X}$, and $E\left\{\cdot/\underline{y};\underline{\theta}^{(n)}\right\}$ denotes the conditional expectation given $\underline{y}$, computed using the parameter value $\underline{\theta}^{(n)}$. The intuitive idea is that we would like to choose $\underline{\theta}$ that maximizes $\log f_{\underline{X}}(\underline{x};\underline{\theta})$, the log-likelihood of the complete data. However, since $\log f_{\underline{X}}(\underline{x};\underline{\theta})$ is not available to us (because the complete data is not available), we maximize instead its expectation, given the observed data $\underline{y}$. Since we used the current estimate $\underline{\theta}^{(n)}$ rather than the actual value of $\underline{\theta}$ which is unknown, the conditional expectation is not exact. Thus the algorithm iterates, using each new parameter estimate to improve the conditional expectation on the next iteration cycle (the E step) and then uses this conditional estimate to improve the next parameter estimate (the M step).

The EM algorithm was first presented by Dempster et al. in [6]. The algorithm was suggested before, however not in its general form, by several authors e.g. [7], [8], [9].

3 The two-microphone ML problem

Suppose that in a two microphone situation, one microphone measures the desired (speech) signal with additive noise, while the second microphone measures a reference noise signal, which is correlated to the "noise" component of the signal measured in the first microphone.

We assume that we observe $y_1(t)$ and $y_2(t)$ as indicated in Figure 1, where $A(z)$ is an FIR filter, $e(t)$ is Gaussian white noise, and $s(t)$ is the desired signal.

Specifically, then

$$y_1(t) = s(t) + n(t)$$

$$n(t) = \sum_{k=0}^{q} a_k y_2(t-k) + e(t) \qquad (4)$$

or,

$$y_1(t) = s(t) + \sum_{k=0}^{q} a_k y_2(t-k) + e(t) \qquad (5)$$

The desired signal $s(t)$, is a speech signal. In formulating the ML problem, we assume that $s(t)$ is a sample function from a stationary Gaussian process whose spectrum is known up to some parameters. The unknown parameters $\underline{\theta}$, are $\{a_k\}$, the spectral parameters of $s(t)$ (which will be denoted $\underline{\phi}$), and σ^2.

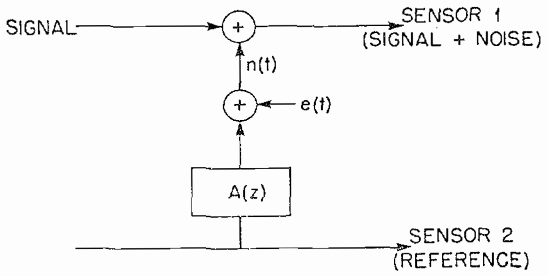

Figure 1: The observations

For the rest of the paper we assume that all the signals are discrete. Assuming that the observation window, $0 \le t \le T-1$, is long enough so that the Fourier coefficients are uncorrelated, the likelihood function is easily expressed in the frequency domain as follows:

$$\log p\left(y_1(t), y_2(t), \;;\underline{\theta}\right) = \sum_{\omega} \log p\left(Y_1(\omega), Y_2(\omega)\;;\underline{\theta}\right) \qquad (6)$$

where

$$Y_i(\omega) = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} y_i(t) e^{-\jmath\omega t}$$

Now, as shown in appendix A, maximizing (6) is equivalent to minimizing

$$\sum_{\omega}\left[\log\left(P_s(\omega)+\sigma^2\right) - \frac{|Y_1(\omega) - A(\omega)\cdot Y_2(\omega)|^2}{P_s(\omega)+\sigma^2}\right] \qquad (7)$$

with respect to σ^2 and the coefficients of $P_s(\omega)$ and $A(\omega)$, where $A(\omega)$ is the frequency response of the FIR filter i.e

$$A(\omega) = \sum_{k=0}^{q} a_k e^{-\jmath\omega k}$$

and $P_s(\omega)$ is the power spectrum of $s(t)$, e.g if $s(t)$ is an AR process of order p, with coefficients $\{b_i\}_{i=1}^{p}$ and gain G,

$$P_s(\omega) = \frac{G}{|1 - \sum_{i=1}^{p} b_i e^{-\jmath\omega i}|^2}$$

We recall that the objective is to estimate the signal $s(t)$ and/or its parameters. Solving the ML problem will provide us the spectral parameters of the signal. Having the $\{a_k\}$ parameters, we may cancel the noise, and have a signal estimate.

4 Solution via the EM algorithm

A direct maximum likelihood solution (i.e minimizing (7)) is complicated and therefore the use of the EM algorithm is suggested. In this approach the complete data is chosen to be $\{s(t), n(t), y_2(t)\}$.

The choice of this "complete data" is motivated by the simple maximum likelihood solution available if indeed $s(t), n(t)$ and $y_2(t)$ are observed separately. The maximum likelihood estimate of $\{a_k\}$ and σ^2 is achieved by least squares fitting of $y_2(t)$ to $n(t)$. The spectral parameters of $s(t)$ are also easily estimated, e.g by solving the normal equation resulting from the sample covariance of $s(t)$ for the LPC parameters.

More specifically, observing the expression for the likelihood of the "complete data" (appendix B), the parameters are estimated by

$$\begin{aligned} \sigma^2, \{a_k\} \quad &\Longleftarrow \quad \max_{\sigma^2,\{a_k\}} \log p(n(t)/y_2(t)) \qquad (8) \\ &\Longleftarrow \quad \min_{\sigma^2,\{a_k\}} \sum_{t=0}^{T-1}\left(n(t) - \sum_{k=0}^{q} a_k y_2(t-k)\right)^2 + T\cdot\log\sigma^2 \end{aligned}$$

and $\underline{\phi}$, the spectral parameters of P_s, by

$$\underline{\phi} \Longleftarrow \max_{\underline{\phi}} \log p(s(t)) = \min_{\underline{\phi}} \sum_{\omega} \log P_s(\omega;\underline{\phi}) + \frac{|S(\omega)|^2}{P_s(\omega;\underline{\phi})} \qquad (9)$$

where $S(\omega)$ is the Fourier transform of $s(t)$, i.e

$$S(\omega) = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} s(t) e^{-\jmath\omega t}$$

In some special cases, (9) is simpler; e.g when $s(t)$ is assumed to be an AR process it reduces to solving the Yule-Walker equation, using the sample autocorrelation of $s(t)$.

Note, observing eqs. (8), (9), that the sufficient statistics of the complete data is $n(t)$, and $|S(\omega)|^2$. The sufficient statistics is linear for the noise part, and quadratic for the signal part. Thus The E step of the algorithm requires the following expectations;

$$\hat{n}(t) = E\left\{n(t)/s(t) + n(t) = y_1(t), y_2(t); \underline{\theta}^{(n)}\right\} \qquad (10)$$

and

$$|\widehat{S(\omega)}|^2 = E\left\{|S(\omega)|^2/S(\omega) + N(\omega) = Y_1(t), Y_2(t); \underline{\theta}^{(n)}\right\} \quad (11)$$

where $\underline{\theta}^{(n)}$ denotes the parameters $\{a_k\}$,σ^2 and $\underline{\phi}$ in the n^{th} iteration.

In the E step, the above conditional expectations are calculated, using the current estimate of the parameters $\underline{\theta}^{(n)}$. The resulting procedure is,

- Generate a signal $x(t)$

$$x(t) = y_1(t) - \sum_{k=0}^{q} a_k y_2(t-k) \qquad (12)$$

(Note that given $\{a_k\}$, $x(t) = s(t) + e(t)$.)

- Apply a Wiener filter to $x(t)$. Get an estimate of

$$\hat{S}(\omega) = \frac{P_s}{P_s + \sigma^2} \cdot X(\omega) \qquad (13)$$

$$\hat{E}(\omega) = X(\omega) - \widehat{S(\omega)} \qquad (14)$$

$$|\widehat{S(\omega)}|^2 = |\hat{S}(\omega)|^2 + \frac{P_s \cdot \sigma^2}{P_s + \sigma^2} \qquad (15)$$

where $E(\omega)$ is the Fourier transform of $e(t)$ and $X(\omega)$ is the Fourier transform of $x(t)$.

- The estimate of $n(t)$ is

$$\hat{n}(t) = \sum_{k=0}^{q} a_k y_2(t-k) + \hat{e}(t) \qquad (16)$$

For the M step, we substitute the expectation above in (8),(9) instead of the measured statistics. The resulting M step is,

- Solve the following least squares problem

$$\min_{\{a_k\}} \sum_t (\hat{e}(t) - \sum_{k=0}^{p} a_k y_2(t-k))^2$$

- Add the result to the previous estimate of $\{a_k\}$ to get a new estimate of $\{a_k\}$.
- Update the signal spectral parameter
For LPC parameters, solve the normal equation using the estimated sample covariance matrix, (the inverse Fourier transform of $|\widehat{S(\omega)}|^2$).

The EM algorithm is summarized in Figure 2.

This procedure can be implemented either on the entire data on each iteration or adaptively, so that on each iteration an updated segment of data is produced. The resulting adaptive algorithm will be an alternative to the LMS and RLS algorithms suggested for solving the least-squares problem that arises in Widrow's approach in [2].

5 Conclusion

The suggested algorithm has been implemented, with $s(t)$ a speech signal. The signal $y_2(t)$ was band limited noise with a flat spectrum from zero to 3 KHz. The FIR filter $A(z)$, was of order 10. $y_1(t)$ was generated according to Figure 1, and the SNR in $y_1(t)$ was approximately -20 db. The results were compared with a "batch" version of the least-squares algorithm, corresponding to estimating the $\{a_k\}$'s via the following least-square

$$\min_{\{a_k\}} \sum_t \left(y_1(t) - \sum_{k=1}^{q} a_k y_2(t-k)\right)^2$$

and then canceling the "noise" and estimating the signal by

$$s(t) = y_1(t) - \sum_{k=1}^{q} a_k y_2(t-k)$$

Both algorithms produced good enhancement of the speech signal, and although there were perceptible differences, the overall quality of both was similar.

The direct least-square approach assumes that $y_2(t)$ and $s(t)$ are uncorrelated, and this assumption is critical. Our algorithm do not require this assumption. In a second experiment, $y_2(t)$ included a delayed version of the speech signal. The direct least squares approach canceled part of the signal, together with the noise, resulting in poor quality. In comparison, the performance of our algorithm, was still good.

Further experiments will include a comparison of adaptive algorithms. Another important project is the evaluation of the EM algorithm and its comparison with direct least-squares method, when both noise and signal couple into both microphones.

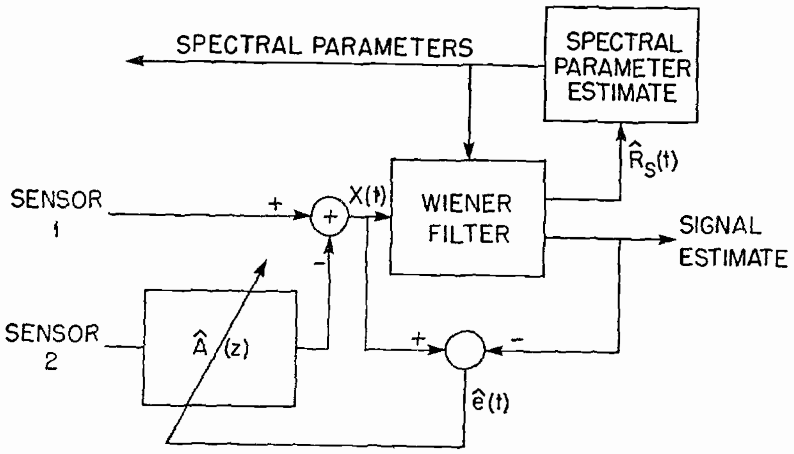

Figure 2: The suggested algorithm

A Derivation of (7)

It is easier to express the joint distribution of the signals $y_1(t)$ and $y_2(t)$ in the frequency domain, since the Fourier transform of the signals generate in each frequency random variables that are independent of the variables in other frequencies. Thus

$$\log p\left(y_1(t), y_2(t)\ ;\underline{\theta}\right) = \sum_{\omega} \log p\left(Y_1(\omega), Y_2(\omega)\ ;\underline{\theta}\right) \qquad (17)$$

In each frequency,

$$\log p\left(Y_1(\omega), Y_2(\omega)\right) = \log p\left(Y_1(\omega)/Y_2(\omega)\right) + \log p\left(Y_2(\omega)\right) \quad (18)$$

where $Y_1(\omega), Y_2(\omega)$ are the Fourier transforms of the signals $y_1(t), y_2(t)$.

However, $\log p\left(Y_2(\omega)\right)$ is independent of $\underline{\theta}$. Given $Y_2(\omega)$ and the parameters $\underline{\theta}$, (i.e $A(\omega), P_s(\omega)$ and σ^2,

$$\begin{aligned}\log p\left(Y_1(\omega)/Y_2(\omega)\ \underline{\theta}\right) = \quad & -\log \pi \left(P_s(\omega) + \sigma^2\right) - \\ & - \tfrac{|Y_1(\omega) - A(\omega)\cdot Y_2(\omega)|^2}{P_s(\omega)+\sigma^2} \qquad (19)\end{aligned}$$

So, maximizing the likelihood, is equivalent to minimizing (7).

B The likelihood of the "complete data"

The likelihood of the "complete data", $L_c(\underline{\theta})$, satisfy

$$\begin{aligned} L_c(\underline{\theta}) &= \log p(s(t), n(t), y_2(t); \underline{\theta}) \\ &= \log p(s(t), n(t)/y_2(t); \underline{\theta}) + \log p(y_2(t); \underline{\theta}) \qquad (20)\end{aligned}$$

However, $\log p(y_2(t))$ is independent of $\underline{\theta}$. Also, given $y_2(t)$, $s(t)$ and $n(t)$ are independent, thus

$$L_c(\underline{\theta}) = \log p(n(t)/y_2(t); \underline{\theta}) + \log p(s(t)/y_2(t); \underline{\theta}) \qquad (21)$$

The term $\log p(n(t)/y_2(t); \underline{\theta})$ depends only on $\{a_k\}$ and σ^2. Maximizing this term is equivalent to minimizing (8).

Under our assumptions, $y_2(t)$ may be related to $s(t)$. However, this relation is arbitrary, and unknown. Thus, we assume that the probability distribution of $s(t)$ given $y_2(t)$ will be the a-priori distribution of $s(t)$. This distribution depends only on the parameters of $P_s(\omega)$, and it is the probability of a stationary random process with power spectrum $P_s(\omega)$,

$$\log p\left(s(t); \underline{\phi}\right) = -\sum_{\omega} \log P_s(\omega; \underline{\phi}) + \frac{|S(\omega)|^2}{P_s(\omega; \underline{\phi})} \qquad (22)$$

where $S(\omega)$ is the Fourier transform of $s(t)$.

Thus, maximizing this term, is equivalent to minimizing (9).

References

[1] J. S. Lim (Editor). *Speech Enhancement.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[2] B. Widrow et al. Adaptive noise canceling: principles and applications. *Proc. IEEE*, 63:1692–1716, 1975.

[3] S. F. Boll and D. C. Pulsipher. Suppression af acoustic noise in speech using two microphone adaptive noise cancellation. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-28:752–753, 1980.

[4] W. A. Harrison, J. S. Lim, and E. Singer. A new application of adaptive noise cancellation. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-34:21–27, 1986.

[5] J. S. Lim and A. V. Oppenheim. All pole modeling of degraded speech. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-26:197–210, 1978.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Annales of the Royal Statistical Society*, 1–38, 1977.

[7] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *The Annals of MathematicalStatistics*, 41:164–171, 1970.

[8] H. O. Hartley and R. R. Hocking. The analysis of incomplete data. *Biometrics*, 27:783–808, 1971.

[9] T. Orchard and M. A. Woodsbury. A missing information principle: theory and applications. In *Proceedings of 6th Berkley Symposium on math. stat. and prob.*, pages 697–715, 1972.