ITERATIVE-BATCH AND SEQUENTIAL ALGORITHMS FOR SINGLE MICROPHONE SPEECH ENHANCEMENT

Sharon Gannot *David Burshtein* *Ehud Weinstein*

Dept. of Electrical Engineering – Systems
Tel-Aviv University
Tel-Aviv 69978, Israel
sharong@eng.tau.ac.il

ABSTRACT

Speech quality and intelligibility might significantly deteriorate in the presence of background noise, especially when the speech signal is subject to subsequent processing. In this paper we represent a class of Kalman-filter based speech enhancement algorithms with some extensions, modifications, and improvements. The first algorithm employs the estimate-maximize (EM) method to iteratively estimate the spectral parameters of the speech and noise parameters. The enhanced speech signal is obtained as a byproduct of the parameter estimation algorithm. The second algorithm is a sequential, computationally efficient, gradient descent algorithm. We discuss various topics concerning the practical implementation of these algorithms. Experimental study, using real speech and noise signals is provided to compare these algorithms with alternative speech enhancement algorithms, and to compare the performance of the iterative and sequential algorithms.

1. INTRODUCTION

Speech quality and intelligibility might significantly deteriorate in the presence of background noise, especially when the speech signal is subject to subsequent processing. Speech enhancement algorithms have therefore attracted a great deal of interest in the past two decades [1], [2], [3], [4], [5], [6], [7], [8].

Lim and Oppenheim [5] have suggested to model the speech signal as a stochastic auto-regressive (AR) model embedded in additive white Gaussian noise, and use this model for speech enhancement. The proposed algorithm is iterative in nature. It consists of estimating the speech AR parameters by solving the Yule-Walker equations using the current estimate of the speech signal, and then apply the (non-causal) Wiener filter to the observed signal to obtain an hopefully improved estimate of the desired speech signal. It can be shown that the version of the algorithm which uses the covariance of the speech signal estimate, given at the output of the Wiener filter, is in fact the estimate-maximize (EM) algorithm for the problem at hand. As such, it is guaranteed to converge to the maximum likelihood (ML) estimate of the AR parameters, or at least to a local maximum of the likelihood function, and to yield the best linear filtered estimate of the speech signal, computed at the ML parameter estimate.

Weinstein *et al.* [7] presented a time-domain formulation to the problem at hand. Their approach consists of representing the signal model using linear dynamic state equation, and apply the EM method. The resulting algorithm is similar in structure to the Lim and Oppenheim [5] algorithm, only that the non-causal Wiener filter is replaced by the Kalman smoothing equations. In addition to that, sequential speech enhancement algorithms are presented in [7]. These sequential algorithms are characterized by a forward Kalman filter whose parameters are continuously updated. In [8] similar methods were proposed for the related problem of multi-sensor signal enhancement. Lee *et al.* [4] extended the sequential single sensor algorithm of Weinstein *et al.* by replacing the white Gaussian excitation of the speech signal with a mixed Gaussian term that may account for the presence of an impulse train in the excitation sequence of voiced speech. Lee *et al.* examined the signal to noise ratio (SNR) improvement of the algorithm when applied to synthetic speech input. They also provide limited results on real speech signals.

The use of Kalman filtering was previously proposed by Paliwal and Basu [6] for speech enhancement, where experimental results reveal its distinct advantage over the Wiener filter, for the case where the estimated speech parameters are obtained from the clean speech signal (before being corrupted by the noise). Gibson *et al.* [2] proposed to extend the use of the Kalman filter by incorporating a colored noise model in order to improve the enhancement performances for certain class of noise sources. A disadvantage of the above mentioned Kalman filtering algorithms is that they do not address the model parameters estimation problem. Koo and Gibson [3] suggested an algorithm that iterates between Kalman filtering of the given corrupted speech measurements, and estimation of the speech parameters given the enhanced speech waveform. The resulting algorithm is in fact an approximated EM algorithm.

In this paper we represent the iterative-batch and sequential algorithms that were presented in [7] with some extensions, modifications, and improvements, and discuss various topics concerning the practical implementation of these algorithms. We also compare the performance of the suggested algorithms to existing algorithms in the literature. This discussion is supported by experimental study using recorded speech signals and actual noise sources.

2. THE SIGNAL MODEL

Let the signal measured by the microphone be given by:

$$z(t) = s(t) + v(t) \tag{1}$$

where $s(t)$ represents the sampled speech signal, and $v(t)$ represents additive background noise.

We shall assume the standard LPC modeling for the speech signal over the analysis frame, in which $s(t)$ is modeled as a stochastic AR process, i.e.

$$s(t) = -\sum_{k=1}^{p} \alpha_k s(t-k) + \sqrt{g_s} u(t) \tag{2}$$

where the excitation $u(t)$ is a normalized (zero mean unit variance) white noise, g_s represents the spectral level, and $\alpha_1, \ldots, \alpha_p$ are the AR coefficients. We may incorporate the more detailed voiced speech model suggested in [9] in which the excitation process is composed of a weighted linear combination of an impulse train and a white noise sequence to represent voiced and unvoiced speech respectively. However, as indicated in [10], this approach did not yield any significant performance improvements over the standard LPC modeling.

The additive noise $v(t)$ is also assumed to be a realization from a zero mean possibly non-white stochastic AR process:

$$v(t) = -\sum_{k=1}^{q} \beta_k v(t-k) + \sqrt{g_v} w(t) \tag{3}$$

where $\beta_1, \ldots, \beta_q$ are are the AR parameters of the noise process, and g_v represents its power level. Many of the actual noise sources may be closely approximated as low order, all-pole (AR) processes, in which case a significant improvement may be achieved by incorporating the noise model into the estimation process as indicated in [2], [10].

Following straight-forward algebra manipulations, Eqs. (1) - (3) may be represented in the following state-space form:

$$\begin{aligned} \mathbf{x}(t) &= \Phi \mathbf{x}(t-1) + G\mathbf{r}(t) \\ \mathbf{z}(t) &= \mathbf{h}^T \mathbf{x}(t) \end{aligned}$$

where the state vector $\mathbf{x}(t)$ is defined by:

$$\mathbf{x}^T(t) = \left[\begin{array}{cccc} \mathbf{s}_{p-1}^T(t-1) & s(t) & \mathbf{v}_{q-1}^T(t-1) & v(t) \end{array} \right]$$

where

$$\begin{aligned} \mathbf{s}_p(t) &= \left[\begin{array}{cccc} s(t-p+1) & s(t-p+2) & \ldots & s(t) \end{array} \right]^T \\ \mathbf{v}_q(t) &= \left[\begin{array}{cccc} v(t-q+1) & v(t-q+2) & \ldots & v(t) \end{array} \right]^T \end{aligned}$$

Φ, $\mathbf{h}$ and G may be expressed in terms of the model parameters.

Assuming that all the signal and noise parameters are known, which implies that Φ, $\mathbf{h}$ and G are known, the optimal (minimum mean square error) linear state estimate, which includes the desired speech signal $s(t)$, is obtained using the Kalman smoothing equations. However, since the signal and noise parameters are not known a-priori, they must also be estimated within the algorithm.

3. EM - BASED ALGORITHM

Applying the EM method to the problem at hand, and following the considerations in [11], [7] (see also [8] that considers the two channel case), we obtain an algorithm that iterates between state estimation (E-step) and parameter re-evaluation (M-step). The E-step is implemented by using the Kalman filtering equations. The M-step is implemented by using a non-standard YW equation set, in which correlations are replaced by their a-posteriori values, that are calculated by using the Kalman smoothing equations. The enhanced speech is obtained as a byproduct of the E-step.

Since the algorithm is based on the EM method, it is guaranteed to converge monotonically to the ML estimate of all unknown parameters (under Gaussian assumptions), or at least to a local maximum of the likelihood function, where each iteration increases the likelihood of the estimate of the parameters. As a byproduct, it yields the optimal linear state (signal) estimate, computed using the estimated parameters.

This algorithm is an extension of the algorithm presented in [7] for the case in which the additive noise is modeled more generally as a colored AR process. Since the signal and the noise parameter estimates are computed separately within the algorithm, the increase in computational complexity is quite moderate. However, the realizable improvement in the enhancement performance may be quite significant, as indicated in [2], [10].

In order to reduce the computations involved, we suggest to replace the full smoothing operation with fixed-lag smoothing (delayed Kalman filter estimate) [6] or even just by filtering. As indicated in [10], the resulting algorithm still maintains its nice monotonic convergence behavior.

A simplified EM algorithm may be obtained by iteratively estimating the speech parameters using the enhanced speech signal (by employing the ordinary YW equation set), and then using these parameters to improve the estimate of the enhanced signal (the noise parameters are estimated, using signal segments at which voice activity is assumed not to be present). This simplified EM algorithm was suggested by Koo *et al.* [3]. We found that unlike the EM algorithm, which is guaranteed to be stable and to monotonically increase the likelihood function, the simplified EM algorithm does not possess such properties. The simplified EM algorithm results in performance degradation, which is very significant at the lower SNR range. Similar behavior was noticed by Lim and Oppenheim [5] in the context of an iterative Wiener filter algorithm for the enhancement of speech in the presence of white Gaussian noise.

4. PARAMETER ESTIMATION USING HIGHER-ORDER STATISTICS

To obtain a reliable estimate of the speech signal, it is essential to have a powerful initialization algorithm for the speech and noise parameters. Otherwise, the algorithm might converge to a local minimum of the likelihood function. When the SNR is high, an initial estimate of the speech parameters may be obtained using standard LPC processing, and an initial estimate of the noise parameters may be obtained by employing a voice activity detector, so that the noise statistics are accumulated during silence periods.

Unfortunately, this initialization procedure breaks down at low SNR conditions, below 5 dB in our experiments. However, if the additive noise $v(t)$ is assumed to be Gaussian, then higher-order statistics (HOS) may be incorporated in order to improve the initial estimate of the speech parameters as follows. It can be shown (by invoking basic cumulant properties and recalling (1), (2)) that

$$\begin{aligned}\mathrm{cum}\,(z(t), z(t-l_1), \ldots, z(t-l_M)) = \\ -\sum_{k=1}^{p} \alpha_k \mathrm{cum}\,(z(t-k), z(t-l_1), \ldots, z(t-l_M))\end{aligned}$$

whenever $M \geq 2$, where $\mathrm{cum}(\cdot,\cdot,\ldots)$ denotes the joint cumulant of the bracketed variables. For $M = 1$ we obtain the standard Yule-Walker equations based on second-order statistics. However, in this case the equations do not hold because of the contribution of the additive noise, and this is why the parameter initialization breaks down at low SNR. For $M \geq 2$ we obtain additional Yule-Walker type equations that are insensitive to the presence of additive Gaussian noise. These equations appear to be very useful if the additive noise is "more Gaussian" than the speech signal in the sense that its higher-order cumulants are relatively small in magnitude.

In practice the cumulants are approximated by substituting the unavailable ensemble averages with sample averages, thus obtaining a set of linear equations that may be used to compute the AR parameters $\alpha_1, \ldots, \alpha_p$ directly from the observed signal $z(t)$.

Experimental results using actual speech signal in several typical noise environments indicated that at low SNR conditions, below 5 dB, using fourth-order cumulants ($M = 3$) one typically obtains a better and more robust initial estimate of the speech parameters as compared with the conventional LPC approach based on second-order statistics. The use of third-order cumulants ($M = 2$), was not that effective.

5. SEQUENTIAL ALGORITHM

The iterative-batch EM algorithm requires the use of an analysis window over which the signal and noise statistics are assumed to be stationary. To avoid this assumption, we used a sequential speech enhancement algorithm which is more computationally efficient than the iterative-batch algorithm. Another benefit of the sequential algorithm is that it is delay-less, unlike the iterative-batch algorithm that has an inherent delay of one processing window frame.

Our sequential algorithm is a gradient based algorithm, similar to the algorithm suggested in [7] (see also [8], that considers the two-channel case). This algorithm consists of a forward Kalman filter whose parameters are continuously up-dated. An improvement in the convergence behavior of the algorithm was obtained by normalizing the step sizes, i.e. using a normalized gradient search algorithm.

6. EXPERIMENTS

In order to evaluate the performance of the proposed algorithms, both objective and subjective tests were conducted. In the experiments that we describe below, the speech signal was degraded by additive computer-fan noise, at various SNRs. This noise source is typical of an office environment. It was found to obey the Gaussian assumption with a good degree of approximation.

Fig. 1 presents the median value of Itakura-Saito (IS) measurements obtained by using two sentences (the duration of the first was 25 seconds; the duration of the second was 5 seconds). The algorithms that were examined were the proposed iterative-batch algorithm, the sequential algorithm and the log spectral amplitude estimator (LSAE) algorithm, suggested by Ephraim and Malah [1] (which is an improvement of the short time spectral amplitude (STSA) estimator algorithm suggested by the same authors). As can be seen, the results indicate that the iterative-batch algorithm is superior both to the sequential algorithm and to the LSAE algorithm, especially at SNRs above 5 dB. At SNRs below -5 dB the performances of the iterative-batch and sequential algorithms are essentially identical. Similar results were obtained when we considered the total and segmental SNR measures.

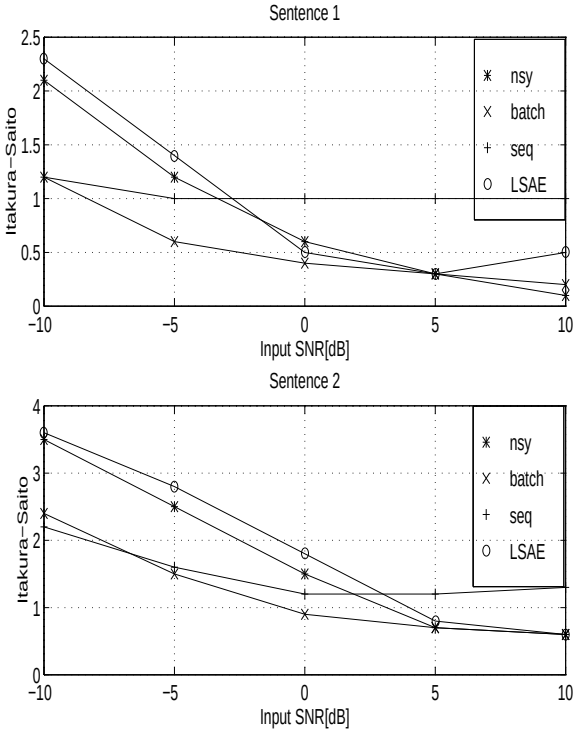

Figure 1. Median Itakura-Saito distortion measure

ASR experiments were conducted using a continuous density hidden Markov model (HMM) based speech recognition system, developed in our laboratory. The speech database was the speaker independent, high quality connected digits recorded at TI (TIDIGITS).

The digit recognition rate of the system when subject to speech signals contaminated by computer fan noise at various SNRs is summarized in Fig. 2. We also show the corresponding recognition rate, when the noisy speech is preprocessed by the iterative-batch enhancement algorithm and by the LSAE algorithm. As can be seen, the iterative-batch algorithm improves the performance by between 4 to

8 dB. The iterative-batch algorithm shows superior performance compared to the LSAE algorithm, especially at the very low and very high SNR range. In fact, at the higher SNR range the LSAE algorithm degrades the performance of the recognizer.

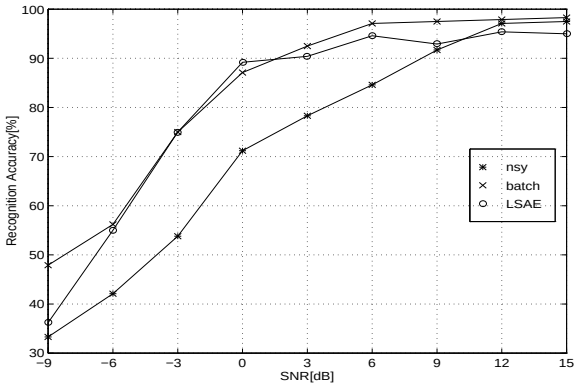

Figure 2. Single digits recognition rate with preprocessing (iterative-batch and LSAE) and without.

The experimental results presented demonstrate the superior performance of the iterative-batch algorithm compared to both the sequential and LSAE algorithms. However, as long as the SNR is not too high, the performance of the sequential algorithm is close to the performance of the iterative-batch algorithm. These conclusions were also supported by informal speech quality tests, and by the assessment of sound spectrograms.

REFERENCES

[1] Y. Ephraim D. Malah. Speech enhancement using a minimum mean square error log-spectral amplituse estimator. *IEEE transactions on Acoustics, Speech and Signal Processing*, Vol. 33(No. 2):443–445, April 1985.

[2] J. D. Gibson B. Koo S.D. Gray. Filtering of colored noise for speech enhancement and coding. *IEEE transactions on Acoustics, Speech and Signal Processing*, Vol. 39:1732–1742, 1991.

[3] B. Koo J.D. Gibson. Filtering of colored noise for speech enhancement and coding. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 349–352, 1989.

[4] K. Y. Lee B. G. Lee and S. Ann. An em-based approach for parameter enhancement with an application to speech signals. *Signal Processing*, Vol. 46:1–14, 1995.

[5] Jae S. Lim and Alan V. Oppenheim. All-pole modeling of degraded speech. *IEEE Transaction on Acoustic, Speech and Signal Processing*, Vol. 26(No. 3):197–210, June 1978.

[6] Anjan Basu K.K. Paliwal. A speech enhancement method based on Kalman filtering. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 177–180, 1987.

[7] E. Weinstein A.V. Oppenheim and M.Feder. Signal Enhancement Using Single and Multi-Sensor Measurement. Technical Report RLE No. 560, M.I.T, Cambridge, MA, M.I.T, November 4 1990.

[8] M. Feder E. Weinstein, A. V. Oppenheim and J. R. Buck. Iterative and sequential algorithms for multisensor signal enhancement. *IEEE Transactions on Signal Processing*, vol. 42:846–859, 1994.

[9] D. Burshtein. Joint modeling and maximum-likelihood estimation of pitch and linear prediction coefficient parameters. *Journal of Acoustical Society of America*, volume 91:1531–1537, March 1992.

[10] S. Gannot. Algorithms for single microphone speech enhancement. Master's thesis, Tel-Aviv University, April 1995.

[11] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, Vol. 3:253–264, 1982.