REMOVAL OF NOISE FROM SPEECH USING THE DUAL EKF ALGORITHM

Eric A. Wan and Alex T. Nelson

Oregon Graduate Institute of Science & Technology
Dept. of Electrical and Computer Engineering, P.O. Box 91000, Portland, OR 97291

ABSTRACT

Noise reduction for speech signals has applications ranging from speech enhancement for cellular communications, to front ends for speech recognition systems. A neural network based time-domain method called *Dual Extended Kalman Filtering* (Dual EKF) is presented for removing nonstationary and colored noise from speech. This paper describes the algorithm and provides a set of experimental results.

1. INTRODUCTION

While there exists a broad range of traditional speech enhancement techniques (*e.g.*, spectral subtraction, signal-subspace embedding, time-domain iterative approaches, etc. [4]), such methods frequently result in audible distortion of the signal, and are far from satisfactory in real-world noisy environments. Recent neural network based filtering methods utilize data sets where the clean speech is available as a target signal for training. These methods are often effective within the training set, but tend to generalize poorly for actual sources with varying signal and noise levels (a review of neural based approaches can be found in [16]). Furthermore, the network models in these methods do not fully take into account the nonstationary nature of speech. In the approach presented here, we assume the availability of only the noisy signal. Effectively, a *sequence* of neural networks is trained on the specific noisy speech signal of interest, resulting in a nonstationary model which can be used to remove noise from the given signal.

1.1. Nonlinear Speech Model

A noisy speech signal $y(k)$ can be accurately modeled as a nonlinear autoregression with both process and additive observation noise:

$$\begin{aligned} x(k) &= f(x(k-1),\ldots,x(k-M),\mathbf{w}) + v(k) \quad (1)\\ y(k) &= x(k) + n(k), \quad (2)\end{aligned}$$

where $x(k)$ corresponds to the true underlying speech signal driven by process noise $v(k)$, and $f(\cdot)$ is a nonlinear function of past values of $x(k)$ parameterized by $\mathbf{w}$. The speech is only assumed to be stationary over short segments, with each segment having a different model. The available observation is $y(k)$, which contains additive noise $n(k)$. If $f(\cdot)$ is linear, this reduces to the classic Linear Predictive Coding (LPC) model of speech.

The optimal *estimator* given the noisy observations $\mathbf{y}(k) = \{y(k), y(k-1), \ldots, y(0)\}$ is $E[x(k)|\mathbf{y}(k)]$. The most direct way to approximate this conditional expectation would be to train on a set of clean data in which the true $x(k)$ may be used as the target to a neural network. Our assumption, however, is that the clean speech is never available; the goal is to estimate $x(k)$ itself from the noisy measurements $y(k)$ alone.

In order to solve this problem, we assume that $f(\cdot,\cdot)$ is in the class of feedforward neural network models, and compute the dual estimation of both states $\hat{x}$ and weights $\hat{\mathbf{w}}$ based on a Kalman filtering approach. In this paper we provide a basic description of the algorithm, followed by a discussion of experimental results.

This work was sponsored by the NSF under grant ECS-9410823 and grant IRI-9712346

2. DUAL EXTENDED KALMAN FILTERING

By formulating the dual estimation problem in a state-space framework, we can use Kalman filtering methods to perform the estimation in an efficient, recursive manner. At each time point, the Kalman filter provides an optimal estimation by combining a prior prediction with a new observation. Connor *et al.*[3] proposed using an extended Kalman filter with a neural network to perform state estimation alone. Puskorious and Feldkamp [13] and others have posed the weight estimation in a state-space framework to allow for efficient Kalman training of a neural network. In prior work, we extended these ideas to include the dual Kalman estimation of both states and weights for efficient maximum-likelihood optimization for robust nonlinear prediction, estimation, and smoothing [14]. The work presented here develops these ideas in the context of speech processing.

To apply the EKF, we first put the autoregression of Equation 1 and 2 in state-space form:

$$\begin{aligned} \mathbf{x}(k) &= F[\mathbf{x}(k-1)] + Bv(k) \quad (3)\\ y(k) &= C\mathbf{x}(k) + n(k), \quad (4)\end{aligned}$$

where

$$\mathbf{x}(k) = \begin{bmatrix} x(k) \\ x(k-1) \\ \vdots \\ x(k-M+1) \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (5)$$

$$F[\mathbf{x}(k)] = \begin{bmatrix} f(x(k),\ldots,x(k-M+1),\mathbf{w}) \\ x(k) \\ \vdots \\ x(k-M+2) \end{bmatrix},$$

and $C = B^T$. If the model is linear, then $f(\mathbf{x}(k))$ takes the form $\mathbf{w}^T\mathbf{x}(k)$, and $F[\mathbf{x}(k)]$ can be written as $A\mathbf{x}(k)$, where A is a matrix in controllable canonical form. We initially assume the noise terms $v(k)$ and $n(k)$ are white with known variances σ_v^2 and σ_n^2, respectively.

2.1. State Estimation

For a *linear* model with *known* parameters, the Kalman filter (KF) algorithm can be readily used to estimate the states [8]. At each time step, the filter computes the linear least squares estimate $\hat{x}(k)$ and prediction $\hat{x}^-(k)$, as well as their error covariances, $P_{\hat{\mathbf{x}}}(k)$ and $P^-_{\hat{\mathbf{x}}}(k)$. In the linear case with Gaussian statistics, the estimates are the minimum mean square estimates. With no prior information on x, they reduce to the maximum-likelihood estimates.

When the model is nonlinear, the KF cannot be applied directly, but requires a linearization of the nonlinear model at the each time step. The resulting algorithm is called the extended Kalman filter (EKF), and effectively approximates the nonlinear function with a time-varying linear one. The EKF algorithm is as follows:

$$\hat{\mathbf{x}}^-(k) = F[\hat{\mathbf{x}}(k-1), \mathbf{w}] \tag{6}$$

$$P^-_{\hat{\mathbf{x}}}(k) = A(k) P_{\hat{\mathbf{x}}}(k-1) A^T(k) + B\sigma_v^2 B^T \tag{7}$$

$$\text{where} \qquad A(k) = \frac{\partial F[\hat{\mathbf{x}}, \mathbf{w}]}{\partial \hat{\mathbf{x}}(k-1)} \tag{8}$$

$$K(k) = P^-_{\hat{\mathbf{x}}}(k) C^T (C P^-_{\hat{\mathbf{x}}}(k) C^T + \sigma_n^2)^{-1} \tag{9}$$

$$P_{\hat{\mathbf{x}}}(k) = (I - K(k)C) P^-_{\hat{\mathbf{x}}}(k) \tag{10}$$

$$\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}^-(k) + K(k)(y(k) - C\hat{\mathbf{x}}^-(k)). \tag{11}$$

Note that the derivative in Equation 8 corresponds to the linearization of the neural network at the current operation point. This can be found by a single application of standard backpropagation.

When the weights $\mathbf{w}$ are not available, they must be replaced by an estimate, $\hat{\mathbf{w}}$.

2.2. Weight Estimation

Because the model for the speech is not known, the standard EKF algorithm cannot be applied directly. We approach this problem by constructing a separate state-space formulation for the underlying weights as follows:

$$\mathbf{w}(k) = \mathbf{w}(k-1) \tag{12}$$

$$y(k) = f(\mathbf{x}(k-1), \mathbf{w}(k)) + v(k) + n(k), \tag{13}$$

where the state transition is simply an identity matrix, and the neural network $f(\mathbf{x}(k-1), \mathbf{w}(k))$ plays the role of a time-varying nonlinear observation on $\mathbf{w}$. These state-space equations for the weights allow us to estimate them with a second EKF:

$$\hat{\mathbf{w}}^-(k) = \hat{\mathbf{w}}(k-1) \tag{14}$$

$$P^-_{\hat{\mathbf{w}}}(k) = P_{\hat{\mathbf{w}}}(k-1) \tag{15}$$

$$K_{\hat{\mathbf{w}}}(k) = P^-_{\hat{\mathbf{w}}}(k) H^T(k) [H(k) P^-_{\hat{\mathbf{w}}}(k) H^T(k) + \sigma_n^2 + \sigma_v^2]^{-1} \tag{16}$$

$$P_{\hat{\mathbf{w}}}(k) = (I - K_{\hat{\mathbf{w}}}(k) H(k)) P^-_{\hat{\mathbf{w}}}(k) \tag{17}$$

$$\text{where} \quad H(k) = \frac{\partial x^-(k)}{\partial \hat{\mathbf{w}}} \tag{18}$$

$$\hat{\mathbf{w}}(k) = \hat{\mathbf{w}}^-(k) + K_{\hat{\mathbf{w}}}(k)(y(k) - x^-(k)). \tag{19}$$

The use of the EKF for weight estimation can be related to Recursive Least Squares (RLS), and thus represents an efficient second-order on-line optimization method. Note that when x is not available, it must be replaced in the weight filter by an estimate, $\hat{x}$. A maximum-likelihood interpretation of the EKF detailing the implications on the use of x versus $\hat{x}$ is given in [12].

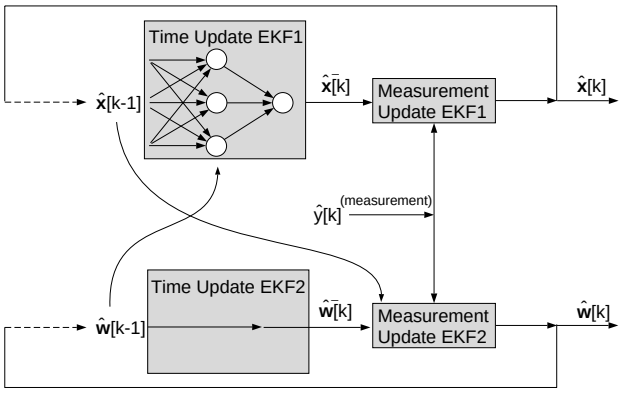

Figure 1: The Dual Extend Kalman Filter. EKF1 and EKF2 represent the filters for the states and the weights, respectively.

The linearization of the network in Equation 18 can be computed as a *dynamic derivative* (or full partial derivative) [17] to account for the recurrent nature of the state-estimation filter, including the dependence of the Kalman gain $K(k)$ on the weights. Unfortunately, the calculation of these derivatives is computationally expensive. Alternatively, this can be avoided completely by ignoring the dependence of $\hat{\mathbf{x}}(k-1)$ on $\hat{\mathbf{w}}$ in Equation 6, resulting in a static linearization of the network. Early results on dynamic derivatives do not indicate performance advantages over the static derivative. Thus, we report results only for static linearization in this paper.

2.3. Dual State and Weight Estimation

The essence of the Dual EKF algorithm is to run the state EKF and weight EKF in parallel (see Figure 1), simultaneously updating estimates of $x(k)$ and $\mathbf{w}$. At each time step, the current estimate of $\mathbf{x}$ is used by the weight filter, and the current estimate of $\mathbf{w}$ is used by the state filter. For finite data sets, the algorithm is run iteratively over the data until the weights converge.

This approach to dual estimation can be justified within a maximum-likelihood framework and can also be related to the Expectation Maximization (EM) algorithm. The approach is also related to work done by Nelson [11] in the linear case, and to Matthews' neural approach [10] to the recursive prediction error algorithm [6]. In the speech literature, the method is most closely related to Lim and Oppenheim's approach to fitting LPC models to degraded speech [9]. It also relates to Ephraim's model-based approach [5], but uses nonlinear estimation to fit the given data instead of using a fixed number of prespecified linear models.

3. EXPERIMENTS

3.1. Nonstationary White Noise

To process noisy speech, the method is applied to successive 64ms windows of the signal (512 points at 8kHz sampling), with a new window starting every 8ms (64 points). A normalized Hamming window is used to emphasize data in the center of the window, and deemphasize data in the periphery. The standard EKF equations are also modified to reflect this windowing in the weight estimation. The result of applying the Dual EKF to a speech signal (se-

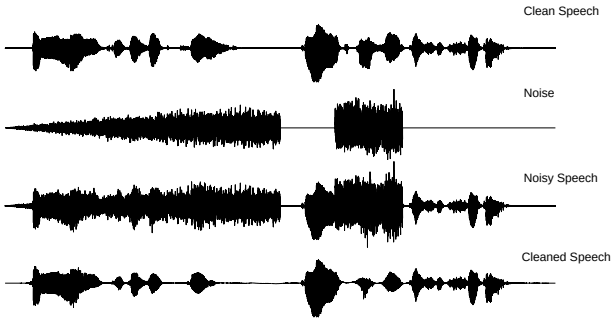

Figure 2: Cleaning noisy speech with the Dual EKF. The TIMIT sentence was approximately 33,000 points (4 sec.) long. Nonstationary white noise was generated artificially and added to the speech to create the noisy signal y.

lected from the TIMIT database) corrupted with simulated nonstationary bursting noise is shown in Figure 2. Feedforward networks with 10 inputs, 4 hidden units, and 1 output were used. Weights typically converged in less than 20 epochs. The results in the figure were computed assuming both σ_v^2 and σ_n^2 were known. The average SNR is improved by 9.94 dB, with little resultant distortion. When σ_n^2 and σ_v^2 are estimated using only the noisy signal, an SNR improvement of 8.50 dB is achieved. In comparison, the "state-of-the-art" technique of *spectral subtraction* [1] achieves an SNR improvement of only 1.26 dB.

3.2. Colored Noise

For most real-world speech applications, we cannot assume the noise is white. For colored noise, the state-space equations 3 and 4 need to be adjusted before Kalman filtering techniques can be employed. Specifically, the measurement noise process is given its own state-space equations,

$$\mathbf{n}(k) = A_n \mathbf{n}(k-1) + B_n v_n(k) \quad (20)$$
$$n(k) = C_n \mathbf{n}(k), \quad (21)$$

where $\mathbf{n}(k)$ is a vector of lagged values of $n(k)$, $v_n(k)$ is white noise, A_n is a simple state transition matrix in controllable canonical form, and B_n and C_n are of the same form as B and C given in Equation 5. Note that this is equivalent to an autoregressive model of the colored noise, which may be fit from a small section of the noisy signal where speech is not present.

With this formulation for the colored noise, it is straightforward to augment both the state $\mathbf{x}(k)$ and the weight $\mathbf{w}(k)$ with $\mathbf{n}(k)$, and write down combined state equations. Specifically, Equations 3 and 4 are replaced by:

$$\begin{bmatrix} \mathbf{x}(k) \\ \mathbf{n}(k) \end{bmatrix} = \begin{bmatrix} F[\mathbf{x}(k-1)] \\ A_n \mathbf{n}(k-1) \end{bmatrix} + \begin{bmatrix} B & 0 \\ 0 & B_n \end{bmatrix} \begin{bmatrix} v(k) \\ v_n(k) \end{bmatrix},$$
$$y(k) = [C \quad C_n] \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{n}(k) \end{bmatrix}, \quad (22)$$

and Equations 12 and 13 are replaced by:

$$\begin{bmatrix} \mathbf{w}(k) \\ \mathbf{n}(k) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & A_n \end{bmatrix} \begin{bmatrix} \mathbf{w}(k-1) \\ \mathbf{n}(k-1) \end{bmatrix} + \begin{bmatrix} 0 \\ B_n \end{bmatrix} v_n(k),$$
$$y(k) = f(\mathbf{x}(k-1), \mathbf{w}(k)) + C_n \mathbf{n}(k) + v(k). \quad (23)$$

The noise processes in these state-equations are now white, and the Dual EKF algorithm can be used to estimate the signal. Note that the colored noise explicitly affects not only the state estimation, but also the weight estimation.

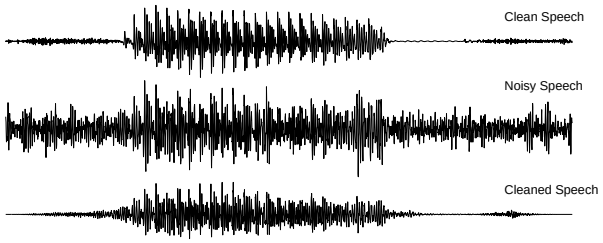

Figure 3: Removing cellular colored noise with the Dual EKF. Initial SNR is -0.16 dB, final SNR is 5.60 dB.

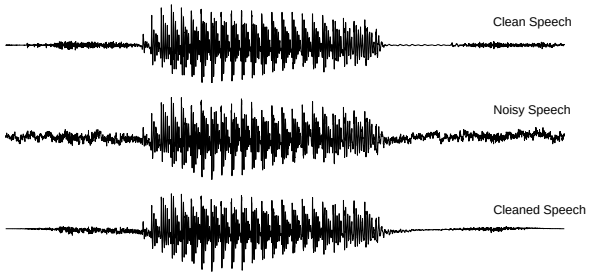

Figure 4: Removing pink noise with the Dual EKF. Initial SNR is 10 dB, final SNR is 13.87 dB.

An actual recording of highway noise through a cellular phone was added to a speech signal to produce the data shown in Figure 3 (3,500 points). Figure 4 shows a similar experiment with pink noise added (spectrograms shown in Figure 5). In both cases, the noise model A_n and process noise variances σ_v^2 and $\sigma_{v_n}^2$ were assumed known (i.e., modeled using knowledge of $x(k)$ and $n(k)$). Experiments were also run with estimated values using only the noisy speech, as described in the next section. Table 1 summarizes the results for several different initial SNR levels. Spectral subtraction results are included for comparison[1].

3.3. Estimating Noise Variances

In the implementation of the Dual EKF, it is assumed that the variances of $v(k)$ and $n(k)$ (or the SNR) are known quantities. Assuming stationarity of the additive noise, the noise variance σ_n^2 (or its full autocorrelation for determining A_n) may be estimated from segments of the data $y(k)$ that do not contain speech. Alternative methods for tracking nonstationary noise are given in [7, 2, 15].

To estimate the process noise variance σ_v^2 (assuming an LPC model for the signal), Lim and Oppenheim [9] used an expression for the inverse Fourier transform of the signal power (which is a function of σ_v^2). We have developed an alternative approach by noting that the process noise variance σ_v^2 can be estimated directly by considering the relationship between the residual AR prediction error for clean and noisy speech [15].

All these approaches, however, are relatively "ad-hoc", and estimating the noise variances remains a critical area for future work. Our current direction is to treat σ_v^2 and σ_n^2 as additional parameters which may be optimized within the Kalman and maximum-likelihood framework.

[1] The authors would like to thank Rick Peterson for his assistance with the spectral subtraction simulations.

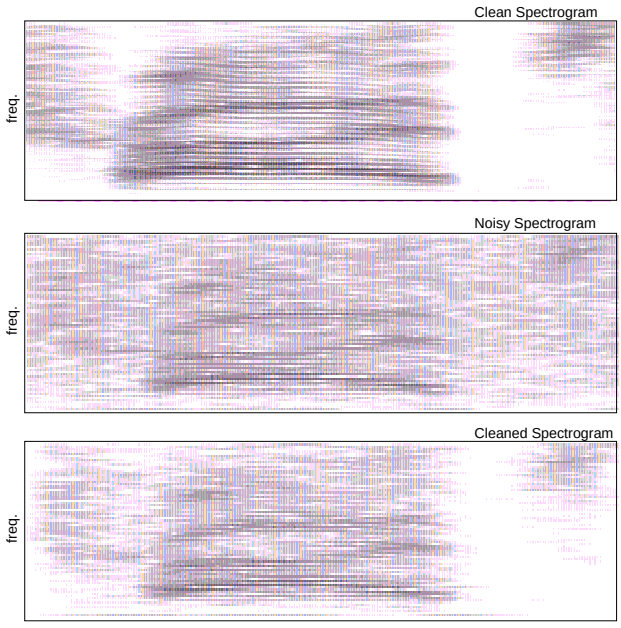

Figure 5: Spectrograms illustrating removal of pink noise with the Dual EKF.

Cellular Noise	Dual EKF(k)	Dual EKF(e)	Spec. Sub.
Init. SNR (dB)	-0.16	-0.16	-0.16
Final SNR	5.60	5.34	2.48
Init. SNR (dB)	5	5	5
Final SNR	9.78	8.01	3.14
Init. SNR (dB)	10	10	10
Final SNR	13.99	12.64	3.54

Pink Noise	Dual EKF(k)	Dual EKF(e)	Spec. Sub.
Init. SNR (dB)	0	0	0
Final SNR	5.52	4.81	2.63
Init. SNR (dB)	5	5	5
Final SNR	9.17	8.88	3.24
Init. SNR (dB)	10	10	10
Final SNR	13.87	12.84	3.56

Table 1: Comparison of methods on colored noise. Results for known noise statistics are indicated by (k), estimated statistics by (e). Spectral subtraction results were obtained using the Duke University Matlab Speech Processing Toolkit. While spectral subtraction was able to suppress noise, distortion of the speech signal resulted in poor SNR values.

4. CONCLUSION AND FUTURE WORK

We have presented the Dual EKF algorithm with preliminary results on its application to speech enhancement in the presence of both nonstationary and colored noise. Initial results compare favorably to current state-of-the-art techniques. However, future work must involve more substantial evaluations based on both objective and subjective criteria. In addition, future algorithmic work will include alternative approaches to variance estimation, as well as the coupling of error statistics, windowing aspects, recurrent training implications, forward-backward methods for *smoothing*, and issues relating to maximum-likelihood estimation and the EM approach.

5. REFERENCES

[1] S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE ASSP-27*, pp. 113-120, April 1979.

[2] J. Cohen. Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America*, vol. 85, no. 6, 1989.

[3] J. Connor, R. Martin, L. Atlas. Recurrent neural networks and robust time series prediction. *IEEE Tr. on Neural Networks*, March 1994.

[4] J. Deller, J. Praokis, J. Hansen. Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, NY, 1993.

[5] Y. Ephraim. Statistical model-based speech enhancement systems. *Proceedings of the IEEE,* Vol. 80, October 1992.

[6] G. Goodwin, K.S. Sin. *Adaptive Filtering Prediction and Control*. Prentice-Hall, Inc., Englewood Cliffs, NJ. 1994.

[7] H.G. Hirsch. Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. *Technical Report TR-93-012, International Computer Science Institute*. 1993.

[8] F. Lewis. *Optimal Estimation*. John Wiley & Sons, Inc. New York, 1986.

[9] J. Lim, A. Oppenheim. All-pole modeling of degraded speech. *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 26, June 1978.

[10] M. B. Matthews and G. S. Moschytz. The identification of nonlinear discrete-time fading-memory systems using neural network models. *IEEE Transactions on Circuits and Systems-II*, 41(11):740–51, November 1994.

[11] L. Nelson, E. Stear. The simultaneous on-line estimation of parameters and states in linear systems. *IEEE Tr. on Automatic Control*, February 1976.

[12] A. Nelson and E. Wan. A Two-observation Kalman framework for maximum-likelihood modeling of noisy time series, submitted to IJCNN'98.

[13] G. Puskorious, L. Feldkamp. Neural control of nonlinear dynamic systems with Kalman filter trained recurrent networks. *IEEE Trn. on NN*, vol. 5, no. 2, 1994.

[14] E. Wan, A. Nelson. Dual Kalman filtering methods for nonlinear prediction, estimation, and smoothing. In *NIPS96 Proceedings*, 1997.

[15] E. Wan and A. Nelson. Neural dual extended Kalman filtering: applications in speech enhancement and monaural blind signal separation. *Neural Networks for Signal Processing VII: Proceedings of the 1997 IEEE Workshop*, ed. Principe, Morgan, Giles, Wilson, 1997.

[16] E. Wan and A. Nelson. Networks for Speech Enhancement, in *Handbook of Neural Networks for Speech Processing*, Edited by Shigeru Katagiri, Artech House, Boston, 1998.

[17] P. Werbos. Neural networks, system identification, and control in the chemical process industries. In *Handbook of Intelligent Control: Fuzzy, Neural, and Adaptive Approaches*, edited by D. White, D. Sofge, Van Nostrand Reinhold, 1992.