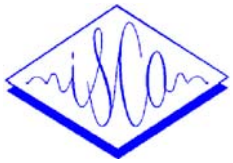

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

NOISE-REGULARIZED ADAPTIVE FILTERING FOR SPEECH ENHANCEMENT

Eric A. Wan and Rudolph van der Merwe

Oregon Graduate Institute of Science & Technology
Dept. of Electrical and Computer Engineering, P.O. Box 91000, Portland, OR 97291
http://www.ece.ogi.edu/˜ericwan

ABSTRACT

The removal of noise from speech signals has applications ranging from speech enhancement for cellular communications to front ends for speech recognition systems. In this paper, we present a new nonlinear time-domain method called *Noise-Regularized Adaptive Filtering*. The approach is based on minimum mean-squared estimation using a modified cost function and allows designing both linear and nonlinear filters using only the observed noisy speech.

1. A GENERAL FRAMEWORK FOR MMSE ESTIMATION

Given a noisy speech signal,

$$y_k = s_k + n_k \ , \tag{1}$$

where s_k is the unobserved clean speech and n_k is additive noise, our goal is to design a nonlinear filter to estimate the clean speech

$$\hat{s}_k = g(\mathbf{y}_k) \ . \tag{2}$$

This filter maps the vector $\mathbf{y}_k = [y_{k-M} \ldots y_k \ldots y_{k+M}]^T$ to an estimate of the speech signal as illustrated in Figure 1 ($L = 2M+1$ is the filter window length). For a minimum mean-square error (MMSE) cost,

$$\min_g E[(s_k - g(\mathbf{y}_k))^2] \ , \tag{3}$$

the corresponding optimal solution is given by the the conditional mean

$$g^*(\mathbf{y}_k) = E[s_k|\mathbf{y}_k] \ . \tag{4}$$

With Gaussian statistics, the conditional mean corresponds to a linear estimator. This simplification has been the basic assumption underlying almost all approaches to speech enhancement (e.g., Wiener and Kalman filtering, spectral subtraction, signal subspace methods [1]). In this paper we introduce a general design approach that encompasses both linear and nonlinear estimation.

This work was sponsored in part by the NSF under grant IRI-9712346

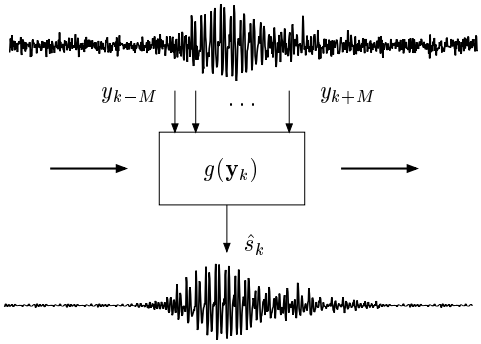

Figure 1: Nonlinear filtering for speech enhancement.

The fundamental reason that the MMSE cost function cannot be minimized directly to find $g(\cdot)$ is that s_k is not available. However, consider the expansion:

$$\begin{aligned} E[(s_k - g(\mathbf{y}_k))^2] \quad = \quad & E[(y_k - g(\mathbf{y}_k))^2] + \\ & 2E[n_k \cdot g(\mathbf{y}_k)] - \\ & 2E[y_k \cdot n_k] + E[n_k^2] \end{aligned} \tag{5}$$

The last two terms are independent of $g(\cdot)$. Thus the optimal $g(\cdot)$ can be found by minimizing the alternative cost,

$$\min_g \left\{ E[(y_k - g(\mathbf{y}_k))^2] + 2E[n_k \cdot g(\mathbf{y}_k)] \right\} \ . \tag{6}$$

The advantage of this formulation is that the first term only depends on the observed noisy speech, whereas the expectation in the second term can be evaluated using only knowledge of the noise statistics. The clean speech is not needed. The first term can also be viewed as the cost associated with filtering the noisy signal to itself, while the second term acts as a noise dependent *regularizer* which prevents the filter $g(\cdot)$ from becoming the identity map. Thus we refer to the resulting estimators as *Noise-Regularized Adaptive Filters* (NRAF).

With speech we cannot assume stationarity. We follow the typical approach of windowing the speech into short overlapping frames, and then design a new filter for each frame. The noisy speech is filtered within each frame, windowed appropriately, and then overlap-added to produce the resulting enhanced speech. Thus we reformulate Equation 6 using a sample expectation for the first term

$$\min_g \left\{ \frac{1}{N} \sum_{k=1}^{N} (y_k - g(\mathbf{y}_k))^2 + 2E[n_k \cdot g(\mathbf{y}_k)] \right\} \ , \tag{7}$$

where N is the frame length. The regularization term still involves the "long-term" statistics of the noise which can be approximated from a preceding non-speech segment. This final form of the cost function will be used to unify a number of different speech enhancement approaches.

2. LINEAR ESTIMATION

We first consider three cases when the estimator is restricted to be linear and indicate their relation to traditional methods.

CASE 1: Linear filter.

For a linear estimator, $\hat{s}_k = \mathbf{w}^T \mathbf{y}_k$, and Equation 7 simplifies to

$$\min_{\mathbf{w}} \left\{ \frac{1}{N} \sum_{k=1}^{N} (y_k - \mathbf{w}^T \mathbf{y}_k)^2 + 2E[n_k \mathbf{w}^T \mathbf{y}_k] \right\} . \quad (8)$$

The regularizer $E[n_k \mathbf{w}^T \mathbf{y}_k] = E[\mathbf{w}^T (n_k(\mathbf{s}_k + \mathbf{n}_k))] = \mathbf{w}^T \mathbf{r}_n$, where $\mathbf{r}_n$ is the autocorrelation of the noise. Taking the gradient with respect to $\mathbf{w}$ yields the closed form solution:

$$\mathbf{w} = \hat{\mathbf{R}}_y^{-1} (\hat{\mathbf{r}}_y - \mathbf{r}_n), \quad (9)$$

where $\hat{\mathbf{R}}_y$ is the sample autocorrelation over the current frame for $\mathbf{y}_k$. This is simply an approximation to the optimal Wiener solution $\mathbf{w}^* = \hat{\mathbf{R}}_y^{-1} \hat{\mathbf{r}}_{yx}$, where the cross correlation of the clean speech and the noise $\hat{\mathbf{r}}_{yx} \approx \hat{\mathbf{r}}_y - \mathbf{r}_\mathbf{n}$ (i.e., a mix between long-term and short-term estimates of the statistics). Note also the relation to spectral-subtraction [2]. By definition the spectrum of the noisy speech and the spectrum of the noise are defined as the Fourier Transforms of $\hat{\mathbf{r}}_y$ and $\mathbf{r}_n$ respectively.

CASE 2: Linear filter with KLT preprocessing.

We again consider a linear estimator $\hat{s}_k = \mathbf{w}^T \tilde{\mathbf{y}}_k$, but with $\tilde{\mathbf{y}}_k = \mathbf{U} \cdot \mathbf{y}_k$, where $\mathbf{U}$ is an $L_e \times L$ matrix corresponding to a linear transformation (L_e is the embedding dimension). We choose $\mathbf{U} = \mathbf{U}_{KLT}$ as the KL Transform (principle eigenvectors of the sample autocorrelation matrix $\hat{\mathbf{R}}_y$) to embed the noise plus signal to a lower dimensional subspace. This decorrelates the signal, helps to separate the speech and noise, and allows the subsequent filter to be designed with a fewer number of inputs. The modified cost to be minimized corresponds to

$$\min_{\mathbf{w}} \left\{ \frac{1}{N} \sum_{k=1}^{N} (y_k - \mathbf{w}^T \mathbf{U} \mathbf{y}_k)^2 + 2\mathbf{w}^T \mathbf{U} \mathbf{r}_n \right\} , \quad (10)$$

which can again be solved in closed form:

$$\mathbf{w} = (\mathbf{U}^T \hat{\mathbf{R}}_y \mathbf{U})^{-1} \mathbf{U}^T (\hat{\mathbf{r}}_y - \mathbf{r}_n) . \quad (11)$$

CASE 3: Signal subspace embedding.

For the final case, we modify the MMSE cost to design a multiple-input-multiple-output (MIMO) filter which provides a simultaneous block estimate of the speech, $\hat{\mathbf{s}}_k = [\hat{s}_{k-M} \ldots \hat{s}_k \ldots \hat{s}_{k+M}]^T$. The corresponding MMSE cost is

$$\min_{\mathbf{W}} E[(\mathbf{s}_k - \mathbf{W}\tilde{\mathbf{y}}_k)^T \mathbf{Q} (\mathbf{s}_k - \mathbf{W}\tilde{\mathbf{y}}_k)] , \quad (12)$$

where $\mathbf{W}$ is now an $L \times L_e$ matrix, $\tilde{\mathbf{y}}_k = \mathbf{U} \cdot \mathbf{y}_k$, and $\mathbf{Q}$ is a $L \times L$ matrix allowing for general weighted MMSE. With $\mathbf{Q} = diag([0 \ldots 0\, 1\, 0 \ldots 0])$ and $\mathbf{U} = \mathbf{I}$, we have *CASE 1*, and for $\mathbf{U} = \mathbf{U}_{KLT}$ we have *CASE 2*.

The modified sampled MMSE cost can be written as

$$\min_{\mathbf{W},\mathbf{U}} \left\{ \frac{1}{N} \sum_{k=1}^{N} \mathbf{e}_{\mathbf{y}_k}^T \mathbf{Q} \mathbf{e}_{\mathbf{y}_k} + 2E[\mathbf{n}_k^T \mathbf{Q}\mathbf{U}\mathbf{W}\mathbf{n}_k] \right\} , \quad (13)$$

with $\mathbf{e}_{\mathbf{y}_k} = \mathbf{y}_k - \mathbf{U}\mathbf{W}\mathbf{y}_k$. We now solve for both the filter weights $\mathbf{W}$ and embedding transformation $\mathbf{U}$ for an arbitrary fixed weighting $\mathbf{Q}$.

If $\mathbf{Q} = \mathbf{I}$, the noise is assumed white, and we restrict the embedding dimension of $\mathbf{U}$ to be fixed and less than that of $\mathbf{y}_k$, then the resulting filter is the same as classical signal subspace embedding (for fixed embedding dimension) which guarantees minimum signal distortion for a fixed noise residual [3]. In this case $\mathbf{U}$ can be expressed analytically as the KLT of $\hat{\mathbf{R}}_s = \hat{\mathbf{R}}_y - \mathbf{R}_\mathbf{n}$, if the matrix remains positive-semi-definite (PSD). The matrix $\mathbf{W}$ involves a set of Lagrange dependent gain factors followed by an inverse KLT. To avoid these restrictions, we take an iterative approach (as will also be the case in nonlinear filtering) in which the cost is optimized by gradient descent. It is straight forward to derive the necessary update rules for $\mathbf{U}$ and $\mathbf{W}$,

$$\begin{aligned} \mathbf{W}_{i+1} &= \mathbf{W}_i + \mu(\frac{1}{N} \mathbf{Q} \mathbf{e}_{\mathbf{y}_k} \mathbf{y}^T \mathbf{U}_i^T) - \mu(\mathbf{Q}\mathbf{R}_n \mathbf{U}_i^T) \\ \mathbf{U}_{i+1} &= \mathbf{U}_i + \mu(\frac{1}{N} \mathbf{W}_i^T \mathbf{Q} \mathbf{e}_{\mathbf{y}_k} \mathbf{y}_i^T) - \mu(\mathbf{W}_i^T \mathbf{Q}\mathbf{R}_n) , \end{aligned}$$

where μ controls the learning rate[1]. For simulations we consider the case where $\mathbf{Q} = \mathbf{I}$ corresponding to classic signal subspace embedding methods, and $\mathbf{Q} = diag([\beta \ldots \beta\, 1\, \beta \ldots \beta])$ with $(0 < \beta < 1)$ allowing a tradeoff between *CASE 2* and classic signal-subspace embedding (this allows the center tap $\hat{s}_k$ to be emphasized during minimization).

3. NON-LINEAR ESTIMATION

For the nonlinear case, we use standard feedforward neural networks to allow non-parametric learning of the conditional mean filter. While similar neural architectures have been proposed for speech enhancement, they have relied on the need for clean targets to train the network

[1] In practice we use batch learning with an adaptive learning rate (see [4]).

off-line (see Wan and Nelson for a review [5]). In our case, the clean signal is never assumed available. Instead, we again minimize the modified cost function.

$$\min_{\mathbf{w}} \left\{ \frac{1}{N} \sum_{k=1}^{N} (y_k - g(\mathbf{w}, \mathbf{y}_k))^2 + 2E[n_k \cdot g(\mathbf{y}_k)] \right\} , \tag{14}$$

where the neural estimator is represented as $g(\mathbf{w}, \mathbf{y}_k)$ and $\mathbf{w}$ are the weights of the network. The first term corresponds to training the network with noisy target data. The second term corresponds to the expected product between the noise and the neural network output and acts to regularize the weights of the network.

The evaluation of the regularization term cannot be performed analytically. Instead, we find an approximate solution using the *Unscented Transformation* (UT), a method for calculating the statistics of a random variable which undergoes a nonlinear transformation [6]. This involves forming the augmented covariance matrix

$$\mathbf{P} = \frac{1}{L+1+\kappa} \begin{bmatrix} \mathbf{U}^T \hat{\mathbf{R}}_y \mathbf{U} & \mathbf{U}^T \mathbf{r}_n \\ \mathbf{r}_n^T \mathbf{U} & \mathbf{r}_n(0) \end{bmatrix} . \tag{15}$$

From this we form $2L+3$ *sigma* vectors corresponding to the mean of $[\tilde{\mathbf{y}}_k \; n_k]$, and the mean (+)plus and (-)minus each column of the matrix square-root of $\mathbf{P}$. These sigma vectors are partitioned as $[\tilde{\mathcal{Y}}_i \; \eta_i]$. The desired expectation is then approximated as

$$\frac{1}{L+1+\kappa} \left\{ \kappa\eta_0 g(\mathbf{w}, \tilde{\mathcal{Y}}_0) + \frac{1}{2} \sum_{i=1}^{2L+2} \kappa\eta_i g(\mathbf{w}, \tilde{\mathcal{Y}}_i) \right\} , \tag{16}$$

where κ is a scaling factor. The process is illustrated in Figure 2. The resulting expectation is accurate to at least the second order (compared to the limited first order accuracy of a truncated Taylor series expansion).

Finally, to minimize the cost in Equation 14 we perform gradient based descent. As the evaluation of the regularization term can be viewed as a sequence of modified forward passes through the network, the gradients can be calculated in a manner similar to that of standard *back-propagation* (see [4]). For each epoch κ is chosen randomly ($-L < \kappa < 36 - L$) and the square-root covariance matrix is rotated by a random unitary matrix to avoid network overfitting to the sigma points. For a complete description of the algorithm see http://ece.ogi.edu/NSEL/.

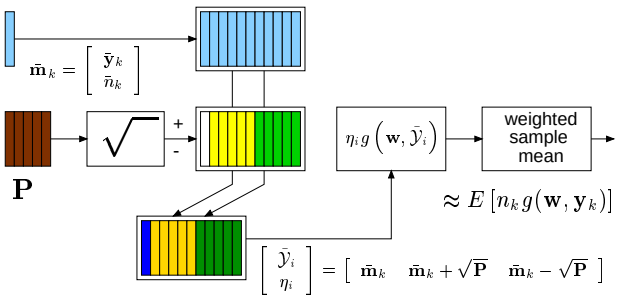

Figure 2: Illustration of the Unscented Transformation for expectation evaluation

The neural network structure can also be modified to have multiple outputs for a block estimation of $\mathbf{s}_k$. In this case, the number of hidden neurons can be made less than the number of outputs forcing an embedding of the input space. This has the potential to provide nonlinear component analysis in contrast to linear KLT embedding (results are not reported in this conference paper).

4. RESULTS

Experiments are performed using samples from the OGI Speech Enhancement Assessment Resource (SpEAR) [7]. All speech and noise sources have been *acoustically* combined to simulate a real noise environment. Synchronous clocking is used to provide an exact time-aligned reference to the clean speech signal. This allows the subsequent use of objective measures such as SNR.

For all experiments we use 8kHz speech, 600 point frames (overlap of 4), filter window $L = 25$, and fixed embedding dimension of $L_e = 19$. These values were found empirically by cross-validation. For the nonlinear estimator we use two-layer feedforward networks with 5 hidden nodes and KLT preprocessing.

Figure 3 summarizes the performance on a sample speech sentence for a number of different noise sources. Table 1 gives performance for an example of actual Lom-

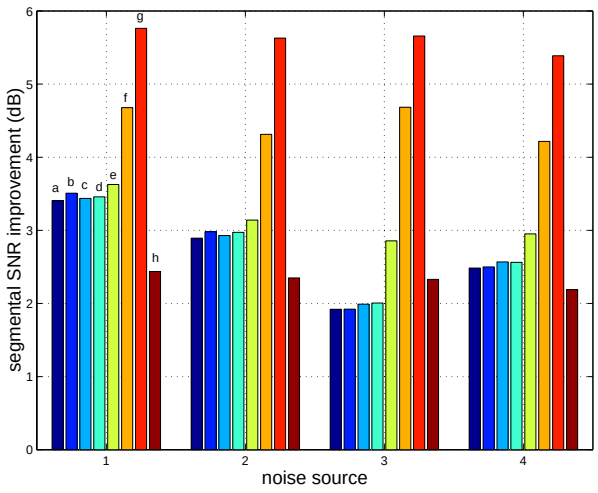

Figure 3: Comparison of segmental SNR performance for different noise sources 1) white (input SNR = 6.08 , segmental SNR = 1.55), 2) pink (input SNR = 4.34, segmental SNR = 0.3) , 3) factory (input SNR = 5.16, segmental SNR = 1.07), 4) F16 (input SNR = 4.61, segmental SNR = .46). Algorithms: a) linear, b) linear with KLT, c) signal subspace (SS), d) weighted signal subspace (SSw), e) nonlinear with KLT, f) linear with clean target, g) nonlinear with clean target, h) standard implementation of spectral subtraction with 256 point frames (Duke Speech Processing Toolkit). Note for these experiments a 3dB improvement in segmental SNR corresponds to approximately 5db improvement in SNR (Segmental SNR has been shown to correlate more closely with subjective quality evaluations).

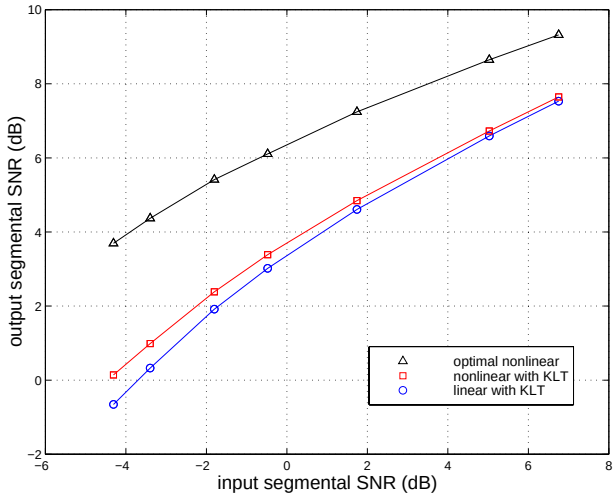

Figure 4: Performance over a range of segmental SNR levels comparing linear NRAF with KLT, nonlinear NRAF with KLT, and nonlinear with clean targets.

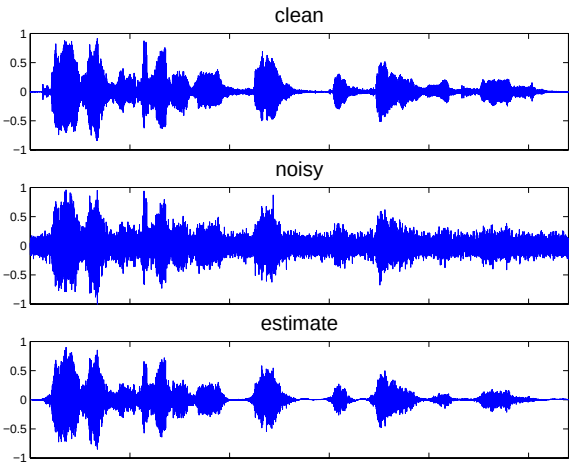

Figure 5: Time domain plots of noisy, clean, and estimated speech using nonlinear NRAF.

bard speech, and finally Figure 4 illustrates the performance over a range of input SNR levels. Figure 5 illustrates performance with a plot in the time domain.

Results also show performance for both the linear and neural estimators which are trained using the known clean signal. While this is unrealistic, it allows a comparison of the potential performance benefits of nonlinear estimation. From these experiments we see that all the linear methods have essentially equivalent performance. The nonlinear NRAF filter shows a clear improvement. As expected, the nonlinear filter trained using the clean speech has the best performance, indicating the advantage of nonlinear estimation.

5. CONCLUSIONS AND FUTURE WORK

To summarize, NRAF provides a general design method for adaptive estimation utilizing only the noisy data. The approach is based on a simple manipulation of the MMSE cost function. Several simple variations lead to a number of popular linear methods which we further extend to the nonlinear case.

While these methods are clearly effective, it is also worth highlighting the sources of error that prevent the filters from achieving optimal performance (known clean signal). In both the linear and nonlinear case the long-term statistic $\mathbf{r}_n$ is used to approximate the true noise statistics over the current frame. The shorter the frame length, the more variation in the actual short-term noise statistics (methods should be investigated which estimate time-varying SNR levels). In the nonlinear case, there is an additional approximation, namely the use of the UT to evaluate $E[n_k g(\mathbf{w}, \mathbf{y}_k)]$. It is our belief that better approaches to estimating this regularization term can lead to significant improvement in performance.

Other work in progress includes the extension of nonlinear estimation to the MIMO case allowing for nonlinear embedding, as well as dynamically determining the appropriate embedding dimension [8]. We also plan more extensive experiments including the use of the filters for front-ends to ASR systems, and finally modification of the MMSE cost to allow incorporation of perceptual based measures.

Δ dB	Lin	Lin KLT	Lin SS	Lin SSw	Nonlin KLT	Opt. Nonlin
SNR	6.58	6.53	6.74	6.71	7.14	10.32
segSNR	2.16	2.12	2.22	2.21	3.05	6.77

Table 1: Performance on Lombard speech. Input SNR = -0.82 dB, input segmental SNR = -2.85 dB.

6. REFERENCES

[1] J. John R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. New York, NY: Macmillan Publishing Company, 1993.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. ASSP-27, pp. 113–20, April 1979.

[3] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 251–66, July 1995.

[4] S. Haykin, *Neural Networks : A Comprehensive Foundation*. Englewood Cliffs, NJ: Macmillan College Publishing Company, Inc, 1994.

[5] E. Wan and A. Nelson, *Handbook of Neural Networks for Speech Processing*, ch. Networks for Speech Enhancement. Artech House, Boston (in press), 1999.

[6] S. J. Julier and J. K. Uhlmann, "A General Method for Approximating Nonlinear Transformations of Probability Distributions," tech. rep., RRG, Dept. of Engineering Science, University of Oxford, Nov 1996. http://www.robots.ox.ac.uk/~siju/work/publications/letter_size/Unscented.zip.

[7] "SpEAR : Speech Enhancement Assessment Resource." http://ece.ogi.edu/NSEL/data.

[8] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1109–1114, 1985.