A MODULAR APPROACH TO SPEECH ENHANCEMENT WITH AN APPLICATION TO SPEECH CODING

Anthony J. Accardi and Richard V. Cox

AT&T Labs – Research, Florham Park, NJ 07932

ABSTRACT

Ephraim and Malah's MMSE-LSA speech enhancement algorithm, while robust and effective, is difficult to tune and adjust for the tradeoff between noise reduction and distortion. We suggest a means of generalizing this design, which allows for other estimators besides the MMSE-LSA to be used within the same supporting framework. When a modified version of Ephraim and Van Trees's spectral domain constrained signal subspace estimator is used in this manner, we obtain a system with greater flexibility and similar performance. We also explore the possibility of using different speech enhancement techniques as pre-processors for different parameter extraction modules of the IS-641 speech coder. We show that such a strategy can increase the quality of the coded speech and lead to a system that is more robust to differing noise types.

1. INTRODUCTION

It is usually the case that for a given speech enhancement scheme, a tradeoff must be made between the amount of noise removed and the distortion introduced as a side effect. If too much noise is removed, listeners could find the resulting distortion objectionable and reject the enhanced waveform in favor of the original noisy speech. The energy of the noise and distortion are not the only factors involved in influencing listener preference; tonal distortions become annoying when just audible, while a certain level of "natural sounding" background noise is well-tolerated. Residual background noise also serves to perceptually mask slight distortions, making its removal even more troublesome.

1.1. The MMSE-LSA Estimator

We say that clean speech $x[n]$ is corrupted by additive noise $d[n]$ to produce a noisy speech signal $y[n]$:

$$y[n] = x[n] + d[n]. \tag{1}$$

We assume that the speech and noise are independent. Let Y_k, X_k, and D_k be the k^{th} DFT coefficients of the noisy speech, clean speech, and noise respectively, where

$$X_k = A_k e^{j\phi_k}, \tag{2}$$

$$Y_k = R_k e^{j\theta_k}. \tag{3}$$

We assume that the DFT coefficients of both the speech and the noise are independent Gaussian random variables.

Ephraim and Malah's minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator [4] is a member of the class of short time spectral amplitude (STSA) estimators that modify the spectral amplitude of the noisy speech and leave the phase untouched. The enhancement is realized as a gain:

$$\hat{A}_k = G(\eta_k, \gamma_k) R_k \tag{4}$$

where η_k and γ_k are called the *a priori* and *a posteriori* SNR's respectively:

$$\eta_k = \frac{E[A_k^2]}{E[|D_k|^2]}, \tag{5}$$

$$\gamma_k = \frac{R_k^2}{E[|D_k|^2]}. \tag{6}$$

Ephraim and Malah's system contains a number of novel features, such as a decision-directed approach to estimating η_k that reduces musical noise [3] [2], and the addition of a gain modification factor, M_k, that implements a soft decision [8] [3]. A modified version of the MMSE-LSA algorithm builds upon this structure [7]. With soft decision in this altered scheme, (4) becomes

$$\hat{A}_k = M_k \cdot G(\xi_k, \gamma_k)\big|_{\xi_k = \frac{\eta_k}{1-q_k}} R_k \tag{7}$$

where q_k is the *a priori* probability of speech absence in the k^{th} frequency bin. This scheme also includes a noise adaptation module that identifies noise-only frames and is capable of tracking non-stationary noise, even when speech is present.

The modified MMSE-LSA exhibits good noise distortion properties, as the residual noise in the enhanced speech usually sounds similar in character to $d[n]$, and is therefore perceived as "natural". However, it is awkward to adjust the tradeoff between the residual noise level and the speech distortion. The most effective way to do this is through the forgetting factor in the decision-directed estimate of η_k, which does not offer enough flexibility [2].

1.2. A Signal Subspace Approach

Ephraim and Van Trees developed a speech enhancement approach based on signal subspace decomposition [5]. In this work it is noted that the vector space containing the noisy speech can be decomposed into a signal-plus-noise subspace and a noise-only subspace. Once identified, the noise-only subspace can be eliminated and the speech can be subsequently estimated from the remaining signal-plus-noise subspace.

In vector notation, we have

$$\mathbf{y} = \mathbf{x} + \mathbf{d} \tag{8}$$

where $\mathbf{d}$ is assumed to be white noise. We consider applying a linear filter $\mathbf{H}$ to the noisy speech to obtain the enhanced speech

$$\hat{\mathbf{x}} = \mathbf{H}\mathbf{y}. \tag{9}$$

We can decompose the residual error into a term solely dependent on the clean speech, called the signal distortion $\mathbf{r_x} = (\mathbf{H} - \mathbf{I})\mathbf{x}$, and a term solely dependent on the noise, called the residual noise $\mathbf{r_d} = \mathbf{H}\mathbf{d}$:

$$\begin{aligned} \mathbf{r} &= \hat{\mathbf{x}} - \mathbf{x} \\ &= (\mathbf{H} - \mathbf{I})\mathbf{x} + \mathbf{H}\mathbf{d} \\ &= \mathbf{r_x} + \mathbf{r_d}. \end{aligned} \tag{10}$$

We consider Ephraim and Van Trees's spectral domain constrained estimator, where the energy of the signal distortion $\bar{\varepsilon}_{\mathbf{x}}^2$ is minimized while each of the eigenvalues of the residual noise is constrained below a constant proportion of the noise variance $\sigma_{\mathbf{d}}^2$:

$$\min_{\mathbf{H}} \bar{\varepsilon}_{\mathbf{x}}^2 \quad \text{subject to} \quad E[|\mathbf{u}_k^{\#}\mathbf{r_d}|^2] \leq \alpha_k \sigma_{\mathbf{d}}^2. \tag{11}$$

Here $\mathbf{u}_k$ is the k^{th} eigenvector of the noisy speech, $(\cdot)^{\#}$ denotes vector conjugate transpose, $0 \leq \alpha_k \leq 1$, and the constraint is applied to each k in the signal-plus-noise subspace. The solution for $\mathbf{H}$ takes on a particularly simple form — a Karhunen-Loeve Transform is applied to the noisy speech signal, the resulting eigenvalues are multiplied by a set of gains, and these products undergo an inverse KLT to yield the enhanced speech. The gains are given by

$$G_k = \sqrt{\alpha_k}. \tag{12}$$

The constraints, the α_k's, still need to be specified. Ideally they would be based on a perceptual hearing model, but Ephraim and Van Trees found that

$$\alpha_k = \exp\left(-\nu \frac{\sigma_{\mathbf{d}}^2}{\lambda_{\mathbf{x}}[k]}\right) \tag{13}$$

yields good results, where $\lambda_{\mathbf{x}}[k]$ is the k^{th} eigenvalue of the clean speech $\mathbf{x}$ and ν is a constant that determines the level of aggression of the enhancement. Both $\sigma_{\mathbf{d}}^2$ and $\lambda_{\mathbf{x}}[k]$ need to be estimated in order to implement the algorithm.

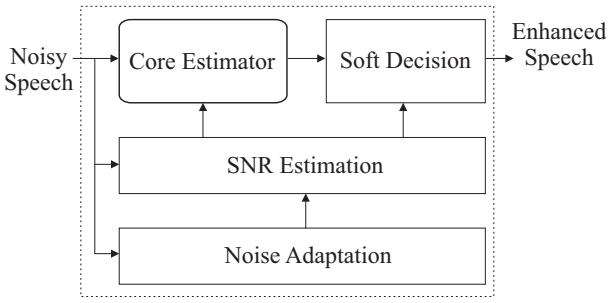

Figure 1: The general structure of the modified MMSE-LSA algorithm.

Although this signal subspace framework provides for an explicit tradeoff between speech distortion and residual noise reduction through the α_k's, there is no accounting for noise distortion, which can have damaging effects on the quality of the enhanced speech.

2. SIGNAL SUBSPACE AS A CORE ESTIMATOR

The general structure of the modified MMSE-LSA algorithm is shown in Figure 1. It consists of a core estimator (the MMSE-LSA) embedded in a supporting framework. We wish to substitute the spectral domain constrained signal subspace estimator for the MMSE-LSA as the core estimator, while making use of the remaining modules. Our goal is to obtain the flexibility and tradeoff control of the signal subspace approach along with the good distortion reduction properties provided by the remainder of the MMSE-LSA framework.

The first difficulty encountered when incorporating signal subspace in the MMSE-LSA framework is that these components operate in different domains — signal subspace computes the KLT while the MMSE-LSA modules calculate the DFT of the noisy speech. We therefore make the stationary process long observation time (SPLOT) assumption, where we approximate the KLT with the DFT and work exclusively in the frequency domain. With this simplification, the eigenvalues become the spectral coefficients of the noisy speech. Now the signal subspace gains G_k are applied to the noisy speech DFT coefficients Y_k.

The removal of the noise-only subspace contributes to unnatural sounding noise structuring. In fact, with the DFT approximation to the KLT, the removal of the noise-only subspace corresponds to a hard decision. We will therefore omit the explicit removal of the noise-only subspace done in [5] and instead rely on the soft decision and noise adaptation modules to take advantage of the uncertainty of speech presence, as these techniques are known to usually lead to

more natural sounding noise [3] [7].

The spectral domain constrained estimator assumes that the noise is white. Therefore, we will first whiten the noise with a whitening filter $\mathbf{W}$, then apply the estimator, and finally invert the whitening operation. Because we are approximating the KLT with the DFT, the changes introduced are subtle. It is straightforward to verify that the form of the estimator does not change. However, the constraint is modified. We now have

$$E[|\mathbf{u}_k^{\#}\tilde{\mathbf{r}}_{\mathbf{d}}|^2] \leq \alpha_k \tilde{\sigma}_{\mathbf{d}}^2 \tag{14}$$

where

$$\tilde{\mathbf{r}}_{\mathbf{d}} = \mathbf{H}\mathbf{W}\mathbf{d} \tag{15}$$

is the residual whitened noise, and

$$\tilde{\sigma}_{\mathbf{d}}^2 = E[|\mathbf{u}_k^{\#}\mathbf{W}\mathbf{d}|^2] \tag{16}$$

is the variance of the whitened noise. The expectations in (14) and (16) are energy spectral coefficients of the residual whitened noise and the whitened noise respectively. Now, dividing the k^{th} constraint in (14) by the magnitude squared of the k^{th} component of the whitening filter in the frequency domain, we obtain our new constraint:

$$S_{r_d r_d}[k] \leq \alpha_k S_{dd}[k]. \tag{17}$$

Here $S_{r_d r_d}[k]$ and $S_{dd}[k]$ are the k^{th} spectral coefficients of the residual noise and original noise, respectively. The final step is to choose the constraints α_k. Following Ephraim and Van Trees, we suggest a form similar to (13):

$$\alpha_k = \exp\left(-\nu \frac{S_{dd}[k]}{S_{xx}[k]}\right) = \exp\left(-\frac{\nu}{\eta_k}\right). \tag{18}$$

In this manner, we heavily base our core estimator on the decision-directed estimate of η_k given by the MMSE-LSA framework, and benefit from the resulting reduction in musical noise. With soft decision, the magnitudes of the DFT coefficients of the enhanced speech for this hybrid algorithm become

$$\hat{A}_k = M_k \cdot \exp\left(-\frac{\nu}{\xi_k}\right)\Big|_{\xi_k = \frac{\eta_k}{1-q_k}} R_k. \tag{19}$$

3. CORE ESTIMATORS FOR THE IS-641

When designing a speech enhancement pre-processor for the IS-641 speech coder (a 7.4 kb/s ACELP codec) [6], it seems natural to use different speech enhancement techniques as pre-processors for different parameter extraction modules of the coder, since different modules make use of different aspects of the input speech in order to code it. By using the MMSE-LSA framework with different core estimators for the different enhancement types, a variety of pre-processors can be implemented with some savings in overall complexity. Here we propose to use different core estimators for the LPC analysis and the residual processing of the IS-641, as shown in Figure 2.

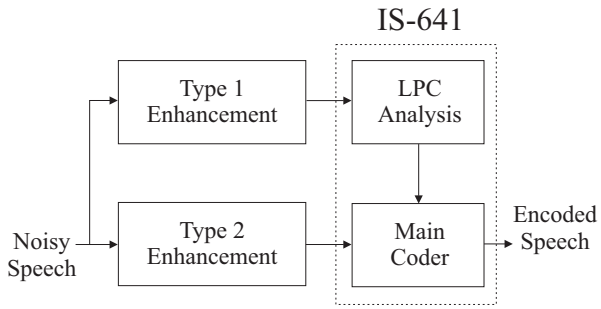

Figure 2: Using two different types of speech enhancement as a pre-processor for the IS-641.

For the "type 1" enhancement for the LPC analysis we will use a magnitude-squared core estimator: $E[|X_k|^2|Y_k]$. Assuming the DFT coefficients of both the noise and clean speech are independent Gaussian random variables, we can easily compute

$$\begin{aligned} E[|X_k|^2 \mid Y_k] &= \left(\frac{S_{xx}[k]}{S_{xx}[k]+S_{dd}[k]}\right)^2 |Y_k|^2 \\ &\quad + \frac{S_{dd}[k]S_{xx}[k]}{S_{xx}[k]+S_{dd}[k]}. \end{aligned} \tag{20}$$

We can express this in terms of parameters in the MMSE-LSA framework as

$$E[|X_k|^2 \mid Y_k] = \left(\frac{\eta_k}{1+\eta_k}\right)^2 |Y_k|^2 + \frac{\eta_k}{1+\eta_k} S_{dd}[k]. \tag{21}$$

Note that the noise adaptation module provides an estimate of the spectral coefficients of the noise $S_{dd}[k]$.

We will use the hybrid algorithm (with the spectral domain constrained estimator as a core estimator) from Section 2 for the "type 2" enhancement for the residual calculations. We will refer to this pre-processor system as "combination enhancement".

4. SUBJECTIVE TEST RESULTS

MOS tests were conducted on a number of enhanced noisy speech samples. Three male and three female speakers each provided a sentence pair sampled at 8 kHz. Both car noise (with SNR's of 10, 15, and 20 dB) and babble (with SNR's of 10 and 20 dB) were then added to this clean speech. A number of enhancement techniques were applied: the hybrid algorithm described in Section 2 (Hyb), the modified MMSE-LSA (LSA) [7], the magnitude-squared estimator from Section 3 (Mag2), the combination enhancement

Enh.	Clean	Babble		Car Noise		
		10 dB	20 dB	10 dB	15 dB	20 dB
None	4.09	3.08	3.74	2.72	3.17	3.48
Hyb	4.12	2.88	3.72	3.25	3.67	3.88
LSA	4.12	3.21	3.79	3.19	3.58	3.79
Mag2	4.13	3.20	3.85	3.15	3.53	3.75
Comb	4.10	2.93	3.84	3.22	3.66	3.93
127	4.11	2.93	3.75	2.93	3.45	3.80

Table 1: MOS scores for different enhancement types (by row) and different noise types and intensities (by column). All samples have been coded with the IS-641.

technique also described in Section 3 (Comb), and the IS-127 noise suppression pre-processor (127) [9]. As a final step, all speech samples were coded with the IS-641 speech coder. The MOS results are shown in Table 1.

All of these enhancement techniques substantially improved the quality of speech contaminated by car noise. The hybrid algorithm generally performed the best, with the combination enhancement at a close second. The modified MMSE-LSA and magnitude-squared algorithms trailed behind slightly. This disparity is primarily due to differences in the residual noise energy — the hybrid and combination enhancement remove more of the car noise than both the MMSE-LSA and magnitude-squared algorithms, without introducing very much additional distortion. All four of these techniques proved superior to the IS-127 pre-processor, indicating the strength of the framework in [7].

The results for babble were more disappointing. Enhancement usually helped in the 20 dB case, with the magnitude-squared and combination enhancement techniques performing the best. For the 10 dB case, however, only the MMSE-LSA and magnitude-squared algorithms were able to increase the quality of the noisy speech. The hybrid algorithm performed the worst in this case. The reason for this performance is that the noise adaptation module had a good deal of trouble properly tracking the highly non-stationary babble. The hybrid algorithm makes a much more aggressive enhancement decision given this noise estimate than the MMSE-LSA or magnitude-squared estimators. When the noise tracking is very accurate, this leads to less residual noise without many distortions, as with car noise. But when the noise tracking is more error-prone, as with babble, the hybrid algorithm causes extra unwanted distortions.

It is interesting to note how the combination enhancement exhibits the strengths of both of its constituent enhancement routines, the hybrid and magnitude-squared algorithms, to become more robust to differing noise types. Better quality is obtained with combination enhancement than with the hybrid algorithm in babble and with the magnitude-squared algorithm in car noise. In fact, with the exception of 10 dB babble, the MOS scores for the speech samples processed with the combination enhancement are within 0.03 of those for the best enhancement algorithm tested for each noise type.

5. CONCLUSIONS

We have described a means of replacing the MMSE-LSA estimator with a more flexible signal subspace based estimator, while retaining the supporting modules and their beneficial contributions to restricting distortion. We have also suggested how multiple core estimators can make use of this same framework as a single pre-processor for a speech coder. Such an approach appears to be robust to differing noise types. A more detailed treatment can be found in [1].

6. ACKNOWLEDGMENTS

We wish to thank Y. Ephraim, D. Malah, J. Allen, and B. Gold for their valuable assistance and advice with this work.

7. REFERENCES

[1] A. J. Accardi, “A Modular Approach to Speech Enhancement with an Application to Speech Coding”, M.Eng. thesis, MIT, 1998.

[2] O. Cappé, “Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor”, IEEE Trans. Speech and Audio Proc., vol. 2, pp. 345–349, 1994.

[3] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator”, IEEE Trans. ASSP, vol. ASSP–32, pp. 1109–1121, 1984.

[4] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator”, IEEE Trans. ASSP, vol. ASSP–33, pp. 443–445, 1985.

[5] Y. Ephraim and H. L. Van Trees, “A Signal Subspace Approach for Speech Enhancement”, IEEE Trans. Speech and Audio Proc., vol. 3, pp. 251–266, 1995.

[6] T. Honkanen *et. al.*, “Enhanced Full Rate Speech Codec for IS-136 Digital Cellular System”, ICASSP '97, Munich, pp. 731–734, 1997.

[7] D. Malah, R. V. Cox, and A. J. Accardi, “Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments”, submitted to ICASSP '99, Phoenix, 1999.

[8] R. J. McAulay and M. L. Malpass, “Speech Enhancement Using a Soft-Decision Noise Suppression Filter”, IEEE Trans. ASSP, vol. ASSP–28, pp. 137–145, 1980.

[9] T. V. Ramabadran, J. P. Ashley, and M. J. McLaughlin, “Background Noise Suppression for Speech Enhancement and Coding”, IEEE Workshop on Speech Coding and Tel., Pocono Manor, PA, pp. 43–44, 1997.