EIGENDOMAIN-BASED NOISE ESTIMATION WITH THE MINIMUM STATISTICS APPROACH

[1]*Vinesh Bhunjun*, [1]*Mike Brookes and* [1]*Jimi Y. C. Wen*

[1]{vinesh.bhunjun , mike.brookes, yung.wen}@imperial.ac.uk
[1]Imperial College London, Dept of Electrical Engineering, London SW7 2AZ, UK

ABSTRACT

One of the challenges for single-channel speech enhancement is to estimate the noise statistics from a signal containing both speech and noise. In this paper, we present a technique for eigendomain-based noise estimation that uses minimum statistics to control the adaptation rate along each eigenvector. We demonstrate that this technique gives robust noise tracking for non-stationary noise.

1. INTRODUCTION

Single-channel speech enhancement algorithms require an accurate estimate of the noise statistics. This estimate must be obtained from a signal containing both the speech and the noise, assumed to be additive and uncorrelated with the speech. One approach is to employ a Voice Activity Detector (VAD) and to estimate the noise statistics during periods of speech absence. Alternatively, continuous noise update schemes, normally using minimum statistics (MS) and/or recursive averaging (RA), can be used to update the noise estimates even in the presence of speech. In this paper, we present a continuous update method which operates in the eigendomain. We show that our algorithm can track rapidly changing noise.

An estimation technique introduced by Martin [1], [2], [3] consists of finding the minimum of smoothed noisy-signal power spectral estimates for each frequency bin over a window of D frames. This method assumes that, over a sufficiently long window, the smoothed estimate will contain troughs that correspond to the noise energy. For an accurate noise estimate, the window length for the minimum value search, D, needs to be long enough to bridge any periods of speech activity. A conflicting desirable property is to have a short window so that the minimum estimate can track changes in the noise statistics and to give low latency. The author uses a window length between 0.5 and 1.5 seconds in his implementation.

Doblinger [4] uses RA to track the minimum value using a computationally more efficient scheme. However, the learning rate for his algorithm has to be low because there is no way to distinguish between a rise in the noise level and the onset of speech. Consequently, his method suffers from a slow update of the noise statistics. Cohen and Berdugo [5], [6] combine the MS and RA approaches thereby benefiting from their strengths without their drawbacks. The minimum estimate is used to derive the probability of speech presence in each frequency bin. The noise estimate is updated using the recursive averaging equation with the learning rate determined by the estimated probability of speech presence. The combined technique has the simplicity and efficiency of RA with the robustness of the MS method. Since the minimum estimates are only used indirectly, their precision is not critical.

2. EIGENDOMAIN-BASED SPEECH PRESENCE PROBABILITY USING MINIMUM STATISTICS

Eigendomain speech enhancement [7] consists of identifying the speech subspace and projecting the speech onto it before estimating the speech energy. In a previous paper [8], we proposed a continuous noise update scheme in the eigendomain which exploits the persistence of noise energy bands in the noisy signal. The eigendomain is particularly suited for noise estimation because it gives a sparse representation for speech signals so that the noise can be estimated from the subspace complementary to that of the speech.

For additive uncorrelated noise, the covariance matrices of the clean speech, $\mathbf{R}_y$, and of the noise, $\mathbf{R}_w$, add up to give the noisy speech covariance matrix, $\mathbf{R}_z \in \mathbb{R}^{K\times K}$ with eigendecomposition:

$$\mathbf{R}_z = \mathbf{R}_y + \mathbf{R}_w = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{1}$$

where λ_k with $k \in [1, K]$ are the diagonal elements of $\mathbf{\Lambda}$ and $\mathbf{R}_z$ is calculated for frame n as in [9]. Our aim is to estimate $\mathbf{R}_w$ using a continuous noise estimation technique. We modify the frequency-based minimum finding algorithm in [5] to operate in the eigendomain and propose a probability model for the presence of speech energy along each eigenvector.

2.1. Minimum values in the eigendomain

In this section, we present an eigendomain-based minimun estimator and investigate the probability distribution for the minimum values. We propose the following recursive equation for finding the minimum along each eigenvector instead of each frequency bin [5]; the diagonal elements of $\mathbf{L}(n)$, $l_{k,k}(n)$, are the minimum energy values along the corresponding eigenvector.

$$\mathbf{L}(n) = \mathbf{B}(n)\mathbf{L}(n-1)\mathbf{B}(n)^T \tag{2}$$

where $\mathbf{B}(n) = \mathbf{M}(n)\mathbf{V}(n)^T\mathbf{V}(n-1)$ and $\mathbf{M}(n)$ is diagonal with elements

$$m_{k,k}(n) = \min\left(\sqrt{\lambda_k(n)/l_{k,k}(n-1)}, 1\right) \tag{3}$$

It is possible to reason about (2) by first assuming that $\mathbf{V}(n) = \mathbf{V}(n-1)$ giving the minimum finding step as

$$\mathbf{L}(n) = \mathbf{M}(n)\mathbf{L}(n-1)\mathbf{M}(n)^T \tag{4}$$

or equivalently $l_{k,k}(n) = m_{k,k}^2 l_{k,k}(n-1)$. The formulation in (2) is a generalization of (4) for the case where $\mathbf{V}(n) \neq \mathbf{V}(n-1)$.
We obtain the ratio of the k^{th} eigenvalue to the minimum energy along the corresponding eigenvector, $l_{k,k}(n)$, as

$$\rho_k(n) = \lambda_k(n)/l_{k,k}(n) \tag{5}$$

which is the eigendomain-based equivalent for the ratio in [5]. To analyse the distribution of the ρ values (5) for speech and noise, we denote by H_y the hypothesis that the speech energy associated with an eigenvector, λ_{y_k}, exceeds that for noise, λ_{w_k}, and H_w for the converse, i.e.

$$H_y : \lambda_{y_k} > \lambda_{w_k} \quad H_w : \lambda_{y_k} \leq \lambda_{w_k} \tag{6}$$

We label as H_y or H_w the eigenvectors in all frames of four noisy speech instances obtained by adding noise instances from the NOISEX database [10] to 4 speech extracts for a global input SNR of 5dB in each case. For each eigenvector of each frame, we also calculate the values of ρ as in (5). We plot in Figure 1 the histogram of the $\log_{10}(\rho)$ values corresponding to the eigenvectors labelled H_y (yellow) and do the same for those labelled H_w (red), with the region of overlap appearing as orange.
Even though the noise instances used are different, the ρ values for H_w are clustered at the low end, typically below 10. In [5], a hard threshold is applied to decide between the two hypotheses. A possible problem with this approach is the degree of overlap in the distribution of the ρ values for H_y and H_w. In particular, this problem shows up as the input SNR decreases. This is illustrated by plotting in Figure 2 the distribution curves for a speech extract corrupted with white noise at four different input SNR values. As the input SNR decreases from 5dB to -10dB, the

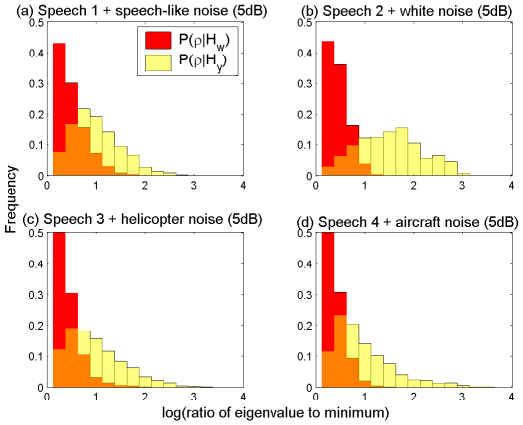

Figure 1: *Histogram of $\log_{10}(\rho)$ values for speech extracts corrupted with 4 different noise types at 5dB input SNR.*

distribution for H_w values (red) remains constant. However, the ρ values for H_y (yellow) shift to the left so that the degree of overlap is more pronounced. This is because the minimum energy values used to calculate the values of ρ increase as the input SNR decreases. In these circumstances, a hard threshold leads to a high error rate in the classification of eigenvectors and a probability-based framework gives greater robustness.

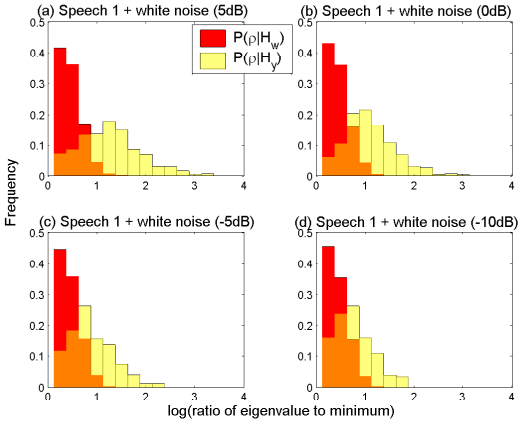

Figure 2: *Histogram of $\log_{10}(\rho)$ values for a speech extract corrupted with white noise at different input SNR values.*

2.2. Probability model

In this section we develop a probability model for the presence of speech energy along an eigenvector, $P(H_y|\rho)$. We use Bayes theorem to obtain an estimate of the conditional probability values, $P(H_y|\rho)$, from the likelihood values, $P(\rho|H_y)$ and $P(\rho|H_w)$, in the previous section.

$$P(H_y|\rho) = \frac{P(\rho|H_y)}{P(\rho|H_y) + P(\rho|H_w)P(H_w)/P(H_y)} \tag{7}$$

For the case of equal prior probabilities, i.e. $P(H_w)/P(H_y)$ =1, we plot in Figure 3(a) the conditional probability curves for the noisy speech extracts used in the previous section. An interesting feature of all the curves is the almost linear dependence of the probability values on the $\log_{10}(\rho)$ values in the range 0 to about 1.3. We also plot in Figure 3(b) the conditional probability curves for the noisy speech extract from the previous section at different input SNR values. The noteworthy feature, for $\log_{10}(\rho)$ values below 1.3, is the higher values of $P(H_y|\rho)$ for a fixed value of ρ as the input SNR decreases, for example for curve (d) compared to curve (a). This reflects the increase in the overlap of the distribution curves for the ρ values as mentioned in the previous section (Figure 2).

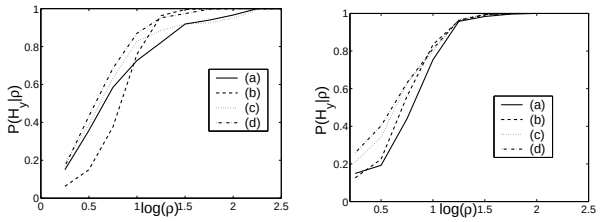

Figure 3: *Conditional probability curves for noisy speech extracts in (a) Figure 1 and (b) Figure 2.*

We model the conditional probability curves in Figure 3(a) as a straight line whose slope depends on $P(H_w)/P(H_y)$ for $0 < \log_{10}(\rho) < 1.3$. To estimate $P(H_w)/P(H_y)$, we exploit one of the general properties of speech energy in the eigenspectrum, namely that the eigenvalues associated with speech typically cluster together among the highest eigenvalues. Consequently, the ρ values in (5) are normally high for adjacent eigenvectors with associated speech energy. If we apply a fixed threshold, $\tau = 1.3$, to some adjacent ρ values in a frame, the fraction, $r_k(n)$, of those that exceed the threshold is given by

$$r_k(n) = \frac{1}{|W|} \sum_{i \in W} (\rho_{k-i}(n) > \tau) \tag{8}$$

where W defines the adjacent eigenvectors and $|W|$ is the number of such eigenvectors, e.g. $W = \{-3, \ldots, 3\}$ and $|W| = 7$. $r_k(n)$ is an estimate of the proportion of eigenvectors with label H_y in that vicinity and can be used to estimate $P(H_w)/P(H_y)$. For example, if the value of ρ for an eigenvector is low but those for adjacent eigenvectors in the same frame are high, the ratio of prior probabilities is skewed more towards H_y than H_w.

The ratio of prior probability values for an eigenvector of a frame is estimated as

$$\log_4(P(H_w)/P(H_y)) = 1 - 2r \tag{9}$$

where the frame index and eigenvector rank have been omitted for clarity. With (9), $P(H_w)/P(H_y)$ is estimated as 4 for $r = 0$ and $1/4$ for $r = 1$, with the limits 4 and $1/4$ chosen to account for possible errors in classification. For $r = 0.5$, the ratio is 1 as required, i.e. H_y and H_w are equally likely.

3. RESULTS

With the probability estimate from the previous section, we use an adaptive noise update scheme to estimate $\mathbf{R}_w(n)$.

$$\begin{aligned}
\hat{\mathbf{R}}_w(n) &= \mathbf{V}(n)\Big(\mathbf{A}(n)\mathbf{\Lambda}_w^{(n)}(n-1)\mathbf{A}(n) \\
&\quad +\mathbf{A}^c(n)\mathbf{\Lambda}(n)\mathbf{A}^c(n)\Big)\mathbf{V}(n)^T \\
\mathbf{\Lambda}_w^{(n)}(n-1) &= \mathbf{V}(n)^T\hat{\mathbf{R}}_w(n-1)\mathbf{V}(n) \\
\mathbf{A}(n) &= \mathrm{diag}(a_1(n), \ldots, a_K(n)) \\
a_k(n) &= \sqrt{\alpha + (1-\alpha)p_k(n)} \\
\mathbf{A}^c(n) &= (\mathbf{I} - \mathbf{A}(n)\mathbf{A}(n))^{1/2}
\end{aligned} \tag{10}$$

where α, set at 0.8 in our implementation, controls the smoothing applied. This formulation is the eigendomain version of the frequency-based one in [5] and guarantees a positive semi-definite $\hat{\mathbf{R}}_w(n)$. The enhancement then proceeds as in [7]. For comparison, we also enhance the speech using an estimate of the noise statistics using an implementation [11] of Martin's MS approach [1]. The metric for comparison between the two techniques is the difference in segmental SNR for the enhanced speech using the proposed technique instead of Martin's MS, calculated over frames containing speech only.

We first assess the performance of the two approaches for different window lengths for the minimum search. In his paper, Martin [1] suggests a window length between 0.5s and 1.5s. The difference in segmental SNR values for different speech and noise combinations is calculated for different window lengths for the minimum search. The results, shown in Figure 4, indicate that for a window length of 0.8s the proposed algorithm typically performs slightly better than Martin's since the segmental SNR improvement for the proposed approach generally exceeds his. As the window length decreases, Martin's technique suffers more than the proposed one and the difference in segmental SNR increases. This is because his minimum energy estimates and hence the noise energy values may be overestimated for very short window lengths which results in excessive attenuation of the speech energy. Even with a window length as short as 0.16s, the proposed method copes well because of the robustness of the proposed probability model and because the minimum estimate is not directly used for signal/noise estimation.

A more challenging test consists of estimating the noise statistics for noise that varies rapidly with time. We concatenate 1s extracts from the following noise instances in the NOISEX database [10]: phantom aircraft noise, white

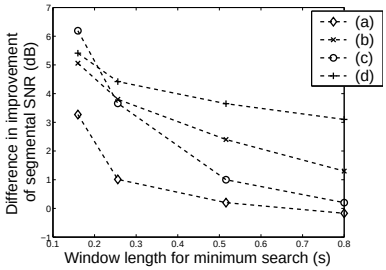

Figure 4: *Difference in segmental SNR improvement between proposed approach and Martin's for different window lengths applied to the noisy speech extracts for Figure 1.*

noise, speech-like noise and lynx helicopter noise. We corrupt a speech extract with the resulting noise samples for an input SNR of 5dB with the noisy speech spectrogram shown in Figure 5(a). The spectrogram for the enhanced speech from the proposed approach and Martin's for a window length of 0.25s are shown in Figures 5(b) and (d). The segmental SNR plot in Figure 5(c) shows that for strong speech regions (corresponding to dark areas in the spectrograms), the segmental SNR for the proposed approach (solid line) is generally higher leading to a lower distortion to speech.

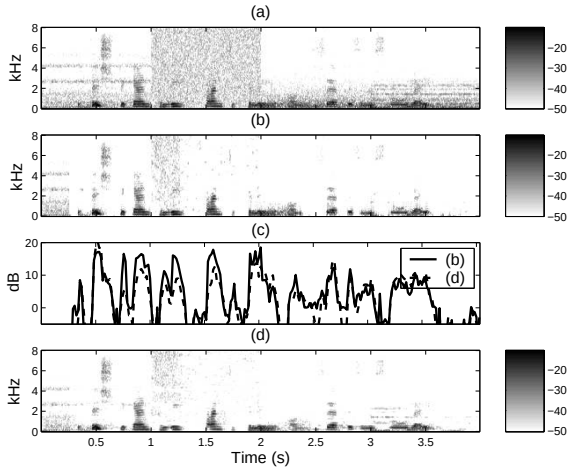

Figure 5: *Spectrogram of (a) noisy speech, (b) noisy speech enhanced using proposed approach and (d) noisy speech enhanced using an implementation [11] of Martin's technique [1]. (c) shows the segmental SNR of the enhanced speech for the proposed (solid) and Martin's (dashed) methods*

4. CONCLUSION

In this paper, we propose an eigendomain-based probability model for speech presence in each eigenvector which leads to robust noise tracking even in the presence of speech in a frame. Tests with real noise indicate that the window for the minimum search can be made quite small without greatly affecting the performance. The proposed technique can thus estimate the noise statistics without excessive speech distortion even when the noise is changing rapidly.

This research project was supported by an ORS award.

5. REFERENCES

[1] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of VII European Signal Processing Conference, EUSIPCO 94*, 1994, pp. 1182–1185.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[3] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006.

[4] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proceedings Eurospeech*, 1995, vol. 2, pp. 1513–1516.

[5] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 393–399, 2003.

[7] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[8] V. Bhunjun and D.M. Brookes, "Narrowband noise estimation in the subspace domain," in *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004, pp. 1–4.

[9] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700–708, 2003.

[10] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 241–246, 1993.

[11] D.M. Brookes, "VOICEBOX: Speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2003.