MODEL-BASED EIGENSPECTRUM ESTIMATION FOR SPEECH ENHANCEMENT

[1]*Vinesh Bhunjun,* [1]*Mike Brookes and* [1]*Patrick Naylor*

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

ABSTRACT

The enhancement of noisy speech in the eigenspectral domain requires an estimate of the clean speech eigenspectrum. In this paper, we develop a model for clean speech eigenspectra for a speaker and use it to constrain the estimation process. As a result, the estimate is contrained by the acoustic space defined by the model and is thus robust even for high levels of noise.

1. INTRODUCTION

Many methods of speech enhancement explicitly incorporate a spectral model of speech sounds. A much-used model for speech denoising is a time-varying autoregressive (AR) model of speech production, the parameters of which are estimated from the noisy speech [1]. Ephraim [2] uses a Hidden Markov Model (HMM) where each Gaussian class is an AR model that represents statistically similar speech sounds. Deng et al [3] propose a statistical model for the static and dynamic cepstral coefficients of the noisy speech and use a Bayesian framework for enhancement. Attias et al [4] choose to model the noisy signal power spectral coefficients instead because they are linearly related to the clean speech and noise coefficients.
In eigendomain-based speech enhancement [5] noisy speech vectors are projected onto the signal subspace, the identification of which requires an estimate of the clean speech eigenspectrum [6]. The eigenspectral estimate is usually obtained by noise energy subtraction, but this may allow it to have values that are not likely to occur for clean speech sounds. In this paper we propose a model for the clean speech eigenspectra for a speaker and use this to apply a soft constraint on the estimate.

2. MODEL-BASED EIGENSPECTRUM ESTIMATION

The enhancement of speech corrupted by additive noise in the eigenspectral domain requires an estimate of the clean speech eigenspectrum. 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, $\mathbf{R}_z$ with eigendecomposition:

$$\mathbf{R}_z = \mathbf{R}_y + \mathbf{R}_w = \mathbf{V}\mathbf{\Lambda}_z\mathbf{V}^T = \mathbf{V}(\tilde{\mathbf{\Lambda}}_y + \tilde{\mathbf{\Lambda}}_w)\mathbf{V}^T \quad (1)$$

where

$$\tilde{\mathbf{\Lambda}}_y = \mathbf{V}^T\mathbf{R}_y\mathbf{V} \quad \tilde{\mathbf{\Lambda}}_w = \mathbf{V}^T\mathbf{R}_w\mathbf{V} \quad (2)$$

We assume that $\tilde{\mathbf{\Lambda}}_y$ and $\tilde{\mathbf{\Lambda}}_w$ are approximately diagonal as in [6]. In this paper, we denote the vectors of diagonal elements of the eigenvalue matrices as follows

$$\mathbf{z} = \mathrm{diag}(\mathbf{\Lambda}_z) \quad \mathbf{y} = \mathrm{diag}(\tilde{\mathbf{\Lambda}}_y) \quad \mathbf{w} = \mathrm{diag}(\tilde{\mathbf{\Lambda}}_w) \quad (3)$$

where $\mathrm{diag}(.)$ gives a vector of the diagonal elements of the input matrix. In practice we observe $\mathbf{z}$ and want to form an estimate $\hat{\mathbf{y}}$ that incorporates prior knowledge of the acoustic space for a speaker.

2.1. Clean speech eigenspectrum model

We obtain feature vectors to train our speech model using clean speech extracts for a single speaker. The covariance matrix $\mathbf{R}_y$ is calculated over frames of 16ms duration. From the eigendecomposition of $\mathbf{R}_y$, we obtain the vector of clean speech eigenvalues, $\mathbf{y}$. We normalize $\mathbf{y}$ in a frame with respect to the energy, $c = \|\mathbf{y}\|_1$, in that frame to make the feature vectors independent of signal amplitude. We then multiply by the DCT matrix, $\mathbf{D}$, to decorrelate the coefficients giving $\mathbf{y}^D_{norm}$:

$$\mathbf{y}^D_{norm} = \mathbf{D}\left(\frac{\mathbf{y}}{c}\right) = \frac{1}{c}\mathbf{D}(\mathbf{y}) = \frac{1}{c}\mathbf{y}^D \quad (4)$$

We use a Gaussian Mixture Model (GMM) to model the energy distribution in the eigenspectrum for different frames of clean speech from the same speaker. The premise is that the clean speech eigenspectra can be clustered in shape-groups according to the spectral characteristics of speech sounds. We perform training on about 600 seconds of speech, sampled at 16kHz, for a single speaker. By excluding periods of silence and pauses, we restrict the training to regions of speech only; the training feature vectors are used to estimate the parameters of the GMM using the Expectation-Maximization (EM) algorithm. Each Gaussian class, C_m, of the GMM is defined by the mean, μ_m, and the covariance matrix, $\mathbf{\Sigma}_m$, with its contribution weighted by α_m:

$$C_m \triangleq \mathcal{N}(\mu_m, \mathbf{\Sigma}_m) \quad \text{for} \quad m \in [1, M] \quad (5)$$

We use $M = 15$ different classes in our implementation. The probability model for the normalized vector of DCT coefficients, $\mathbf{y}^D_{norm}$, is given by

$$P(\mathbf{y}^D_{norm}|GMM) = \sum_{m=1}^{M=15} \alpha_m P(\mathbf{y}^D_{norm}|C_m) \tag{6}$$

2.2. Noise energy distribution

We look at the distribution of the noise energy along an eigenvector and show that it can be approximated with a normal distribution. We assume that the noise signal is stationary over a frame of N samples. Let the frame $\mathbf{w}^{(t)}$ be partitioned into non-overlapping vectors of K samples, $\mathbf{w}^{(t)}(j)$ for $j \in [1, N/K]$, where ${}^{(t)}$ denotes the time-domain signal. Denote the $K \times K$ noise signal covariance matrix for the frame of N samples as $\mathbf{R}_w$. The energy along a normalized eigenvector $\mathbf{v}$ is given by

$$\lambda = \mathbf{v}^T \mathbf{R}_w \mathbf{v} \tag{7}$$

Now, by definition, $\mathbf{R}_w = E\{\mathbf{w}^{(t)}(j)\mathbf{w}^{(t)}(j)^T\} \quad \forall j$, so that

$$\begin{array}{rcl} \lambda & = & \mathbf{v}^T E\{\mathbf{w}^{(t)}(j)\mathbf{w}^{(t)}(j)^T\}\mathbf{v} \\ & = & E\{\mathbf{v}^T \mathbf{w}^{(t)}(j)\mathbf{w}^{(t)}(j)^T \mathbf{v}\} \text{ for a fixed } \mathbf{v} \\ & = & E\{p(j)^2\} \\ \hat{\lambda} & = & \frac{1}{N/K} \sum_{j=1}^{N/K} p(j)^2 \end{array} \tag{8}$$

$p(j) = \mathbf{v}^T \mathbf{w}^{(t)}(j)$ is the coefficient of projection of $\mathbf{w}(j)$ onto $\mathbf{v}$ and its distribution for the different noise types is shown in Figure 1.

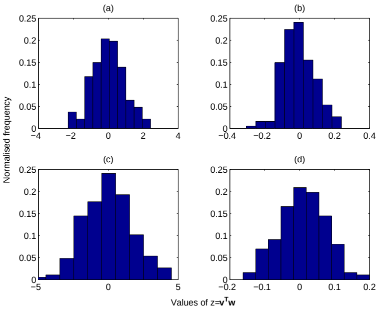

Figure 1: *Histogram of projection coefficients, p, of different noise vectors - (a) white , (b) speech-like, (c) car, (d)operations room - onto a fixed eigenvector*

We select $\mathbf{v}$ to be the eigenvector corresponding to the highest energy in the first frame from 4 s extracts of different approximately stationary noise types from the NOISEX database [7]. For each noise type, we calculate the values of $p(j)$ for $K = 64$ over the whole extract. The $p(j)$ variables are not in general independent, except for white noise. We note from Figure 1 that the distribution plots for speech-like noise, car noise and operations room noise (from NOISEX [7]) are similar to that for white noise. If we assume that $p(j) \sim \mathcal{N}(0, \sigma^2)$ for some σ^2 and that the $p(j)$ variables are independent, λ has a χ^2 distribution with N/K degrees of freedom with mean μ_w and variance σ^2_w:

$$\mu_w = (N/K)\sigma^2 \quad \sigma^2_w/\mu^2_w = 2/(N/K) \triangleq k \tag{9}$$

In this work we approximate the distribution of the noise energy along an eigenvector as $\mathcal{N}(\mu_w, k\mu^2_w)$ even if the $p(j)$s are not independent.

2.3. Noisy signal model

We augment the clean speech eigenspectrum model with the estimated noise statistics to model the noisy speech feature vector, $\mathbf{z}^D$. We assume that each noisy speech vector belongs to a single class in the GMM. Our model for $\mathbf{z}^D$ is given in (10) where both the clean speech and noise vectors are Gaussian random variables: $\mathbf{y}^D_m \sim \mathcal{N}(\mu_m, \mathbf{\Sigma}_m)$ for the m^{th} class of the GMM and $\mathbf{w}^D \sim \mathcal{N}(\mu_w, \mathbf{\Sigma}_w)$ from the previous section with $\mu_w = \mathbf{D}(\text{diag}(\mathbf{V}^T\mathbf{R}_w\mathbf{V}))$.

$$\begin{array}{rcl} \mathbf{z}^D & = & c\mathbf{y}^D_m + \mathbf{w}^D \\ & = & (c_m\mu_m + \epsilon_y) + (\mu_w + \epsilon_w) \\ & = & c_m\mu_m + \mu_w + \epsilon \quad \text{ where } \epsilon = \epsilon_y + \epsilon_w \end{array} \tag{10}$$

where $\epsilon_y \sim \mathcal{N}(\mathbf{0}, c^2_m\mathbf{\Sigma}_m)$, $\epsilon_w \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_w)$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, c^2_m\mathbf{\Sigma}_m + \mathbf{\Sigma}_w)$. In (10), we add the subscript m to c to show the dependence of the scaling factor on the chosen GMM class m. The maximum-likelihood (ML) estimate for this c_m is obtained by maximizing the likelihood function $P(\mathbf{z}^D|\mathbf{y}^D_m, \mathbf{w}^D)$. With the estimate of c_m, we can calculate the noise- and scaling factor-adjusted vector, $\mathbf{z}^D_m$.

$$\mathbf{z}^D_m = \frac{\mathbf{z}^D - \mu_w}{c_m} = \mu_m + \epsilon_m \quad \epsilon_m \sim \mathcal{N}\left(\mathbf{0}, \frac{c^2_m\mathbf{\Sigma}_m + \mathbf{\Sigma}_w}{c^2_m}\right) \tag{11}$$

Since we do not know the class m^* to which the noisy speech vector belongs, we calculate the ML estimate of c_m for each mixture component C_m with $m \in [1, M]$. We also calculate the corresponding likelihood value $P(\mathbf{z}^D_m|GMM)$ using the distribution in (6) and select m^* corresponding to the maximum likelihood value. c_{m^*} is the corresponding scaling factor estimate or equivalently the estimate of the clean speech energy.

In Figure 2 we compare our estimate of the clean speech energy values with the true values for test speech extracts (not used for training) corrupted with noise from the NOISEX database [7]. We observe that our estimate of the clean speech energy from the model, c_m, closely follows the true energy for the different noisy speech extracts.

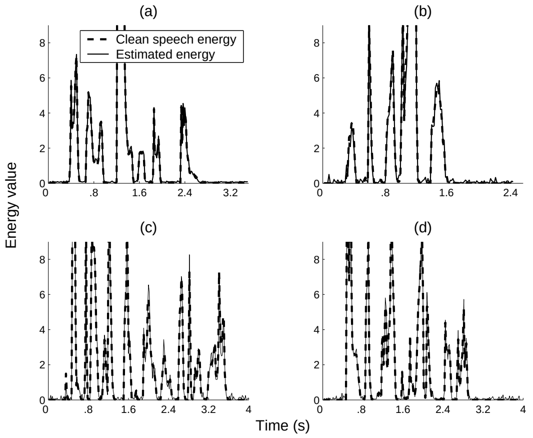

Figure 2: *True clean speech energy and its estimate from the model for (a) speech 1 + white noise, (b) speech 2 + speech-like noise, (c) speech 3 + helicopter noise and (d) speech 4 + aircraft noise all at an input SNR of 5dB.*

2.4. Estimation of clean speech eigenspectrum

Given the acoustic class m^* and scaling factor estimate c_{m^*} for a frame, we want to estimate the clean speech eigenspectrum $\hat{\mathbf{y}}$ for that frame. We replace $(c_{m^*}, \mu_{m^*}, \mathbf{\Sigma}_{m^*})$ by $(c, \mu_y, \mathbf{\Sigma}_y)$ in the text that follows for clarity. We write

$$\mathbf{z}^D = \mathbf{y}^D + \mathbf{w}^D = c\mathbf{y}^D_{norm} + \mathbf{w}^D \qquad (12)$$

where $\mathbf{y}^D \sim \mathcal{N}(c\mu_y, c^2\mathbf{\Sigma}_y)$ and $\mathbf{w}^D \sim \mathcal{N}(\mu_w, \mathbf{\Sigma}_w)$. We choose $\hat{\mathbf{y}}^D$ to maximize the joint probability density function (pdf), $\phi(\mathbf{y}^D, \mathbf{w}^D)$, conditional on the observation $\mathbf{z}^D$ which is the sum of two variables from Gaussian processes.

$$\begin{array}{l} \log(\phi(\mathbf{y}^D, \mathbf{w}^D)) = \\ -\frac{1}{2}(\mathbf{y}^D - c\mu_y)^T(c^2\mathbf{\Sigma}_y)^{-1}(\mathbf{y}^D - c\mu_y) \\ -\frac{1}{2}(\mathbf{w}^D - \mu_w)^T(\mathbf{\Sigma}_w)^{-1}(\mathbf{w}^D - \mu_w) \quad . \end{array} \qquad (13)$$

We substitute $\mathbf{w}^D = \mathbf{z}^D - \mathbf{y}^D$ in (13), differentiate with respect to $\mathbf{y}^D$ and set to zero to give

$$\begin{array}{l} (c^2\mathbf{\Sigma}_y)^{-1}\mathbf{y}^D - (c^2\mathbf{\Sigma}_y)^{-1}c\mu_y = \\ (\mathbf{\Sigma}_w)^{-1}(\mathbf{z}^D - \mu_w) - (\mathbf{\Sigma}_w)^{-1}\mathbf{y}^D \end{array} \qquad (14)$$

which can be simplified to

$$\begin{array}{rcl} \hat{\mathbf{y}}^D & = & \{c^2\mathbf{\Sigma}_y(c^2\mathbf{\Sigma}_y + \mathbf{\Sigma}_w)^{-1}\}(\mathbf{z}^D - \mu_w) + \\ & & \{\mathbf{\Sigma}_w(c^2\mathbf{\Sigma}_y + \mathbf{\Sigma}_w)^{-1}\}c\mu_y \end{array} \qquad (15)$$

and consequently

$$\begin{array}{rcl} \hat{\mathbf{y}} & = & \mathbf{D}^{-1}\left(\hat{\mathbf{y}}^D\right) \\ & = & \mathbf{D}^{-1}\Big(c^2\mathbf{\Sigma}_y\left(\mathbf{\Sigma}_w + c^2\mathbf{\Sigma}_y\right)^{-1}(\mathbf{z}^D - \mu_w) \\ & & +\mathbf{\Sigma}_w\left(\mathbf{\Sigma}_w + c^2\mathbf{\Sigma}_y\right)^{-1}c\mu_y\Big) \quad . \end{array} \qquad (16)$$

The diagonal elements of $\mathbf{\Sigma}_y$ and $\mathbf{\Sigma}_w$ are estimated as $\mathbf{D}^{-1}(k\mu_y^2)$ and $\mathbf{D}^{-1}(k\mu_w^2)$ to give $\mathbf{\Sigma}_y^u$ and $\mathbf{\Sigma}_w^u$ respectively. The updated equation (17) shows that the estimated eigenspectrum is a combination of two possible eigenspectra, $\hat{\mathbf{y}}_s$ and $\hat{\mathbf{y}}_m$, the first obtained by subtracting the noise energy and the second by weighting the class mean. $\hat{\mathbf{y}}$ is closer to $\hat{\mathbf{y}}_m$ for noisy speech with a high noise energy level.

$$\begin{array}{rcl} \hat{\mathbf{y}} & = & c^2\mathbf{\Sigma}_y^u\left(\mathbf{\Sigma}_w^u + c^2\mathbf{\Sigma}_y^u\right)^{-1}\mathbf{D}^{-1}(\mathbf{z}^D - \mu_w) \\ & & +\mathbf{\Sigma}_w^u\left(\mathbf{\Sigma}_w^u + c^2\mathbf{\Sigma}_y^u\right)^{-1}\mathbf{D}^{-1}(c\mu_y) \\ & = & \mathbf{A}_s \times \mathbf{D}^{-1}(\mathbf{z}^D - \mu_w) + (\mathbf{I} - \mathbf{A}_s) \times \mathbf{D}^{-1}(c_{m^*}\mu_{m^*}) \\ & = & \mathbf{A}_s\hat{\mathbf{y}}_s + (\mathbf{I} - \mathbf{A}_s)\hat{\mathbf{y}}_m \end{array} \qquad (17)$$

3. ESTIMATION RESULTS

We analyse the properties of $\hat{\mathbf{y}}_s$ and $\hat{\mathbf{y}}_m$ from (17) by using them to perform speech enhancement [6]. We write

$$\hat{\mathbf{y}}^{(t)} = \mathbf{V}\mathbf{G}\mathbf{V}^T\mathbf{z}^{(t)} \qquad (18)$$

where $\mathbf{G} = \mathrm{f}(\hat{\mathbf{y}}, \mathbf{z})$ with gain transfer function f as in [6]. The noisy test speech was not used in the training of the GMM and it is contaminated with white noise for an input SNR of 5dB. $\mathbf{R}_w$ is estimated from the noisy speech signal for example as in [8]. The model parameters - noise vector μ_m, most likely class m^* and scaling factor c_{m^*} - are obtained as in Section 2.3. We also obtain the true values for the class and scaling factor from the clean speech eigenspectra. In all, we estimate the clean speech signal, $\mathbf{y}^{(t)}$, using (a) the subtraction-based eigenspectrum, $\hat{\mathbf{y}}_s$, giving $\hat{\mathbf{y}}_s^{(t)}$, (b) the model-based eigenspectrum, $\hat{\mathbf{y}}_{\hat{m}}$, with estimated parameters $\hat{m}^*$ and $\hat{c}_{m^*}$ to give $\hat{\mathbf{y}}_{\hat{m}}^{(t)}$, and (c) the model-based eigenspectrum, $\hat{\mathbf{y}}_m$, with true parameters m^* and c_{m^*} to give $\hat{\mathbf{y}}_m^{(t)}$. The spectrograms of the enhanced speech from the three cases, $\hat{\mathbf{y}}_s^{(t)}$, $\hat{\mathbf{y}}_{\hat{m}}^{(t)}$ and $\hat{\mathbf{y}}_m^{(t)}$, are plotted in Figure 3.

Both model-based speech estimates, $\hat{\mathbf{y}}_{\hat{m}}^{(t)}$ and $\hat{\mathbf{y}}_m^{(t)}$, sound more distorted for weak speech sounds compared to $\hat{\mathbf{y}}_s^{(t)}$. This can be seen for example close to time instant 0.7s and 1.1s in Figure 3. Nevertheless, our model successfully synthesizes eigenspectra for speech not encountered during training using only the GMM class and the scaling factor. When these parameters are estimated from the noisy speech, the resulting signal, $\hat{\mathbf{y}}_{\hat{m}}^{(t)}$ (Figure 3(d)), shows a slight degradation compared to $\hat{\mathbf{y}}_m^{(t)}$ (Figure 3(e)), for example close to time instant 1.35s, but the results are still encouraging. The subtraction-based speech estimate, $\hat{\mathbf{y}}_s^{(t)}$, has the least distortion for weak speech sounds for example close to time instant .2s and .7s. However, the level of residual noise is higher as can be seen during silent intervals. This analysis reinforces our choice of estimation equation (17) with $\hat{\mathbf{y}}_s^{(t)}$ chosen when the speech energy dominates and $\hat{\mathbf{y}}_m^{(t)}$ when the noise energy is high.

We analyse further the difference in performance of the two estimates, $\hat{\mathbf{y}}_s$ and $\hat{\mathbf{y}}_{\hat{m}}$. We calculate the normalized error for a frame by subtracting the true speech eigenspectrum, $\mathbf{y}$, from the subtraction-based estimate, $\hat{\mathbf{y}}_s$, and the model-based one, $\hat{\mathbf{y}}_{\hat{m}}$, and normalizing as follows:

$$e_s = \|\hat{\mathbf{y}}_s - \mathbf{y}\|/\|\mathbf{z}\| \quad e_m = \|\hat{\mathbf{y}}_{\hat{m}} - \mathbf{y}\|/\|\mathbf{z}\| \quad . \tag{19}$$

The normalized error terms are averaged over all speech frames for a noisy test speech, e.g. test speech 1 and white noise for different values of the input SNR. The relationship between the average error terms and the input SNR value for this particular speech/noise combination is shown in Figure 4(a) as a solid line for $\hat{\mathbf{y}}_w$ and a dotted line for $\hat{\mathbf{y}}_{\hat{m}}$. The plots for three further speech/noise combinations from the NOISEX [7] database are shown in Figures 4(b), (c) and (d).

For values of the input SNR above -3 dB, the normalized error average and hence speech distortion energy are lower with the subtraction-based estimator, as expected. As the input SNR decreases to very low values, the error average is relatively lower with the model-based estimator confirming that the performance for our model decreases to a lesser extent as input SNR decreases. One way to interpret this result is that for the model, the estimate is selected from a template of different possible fixed speech eigenspectra. With the subtraction approach, the variance of the estimate involves the noise variance which is high for low values of input SNR.

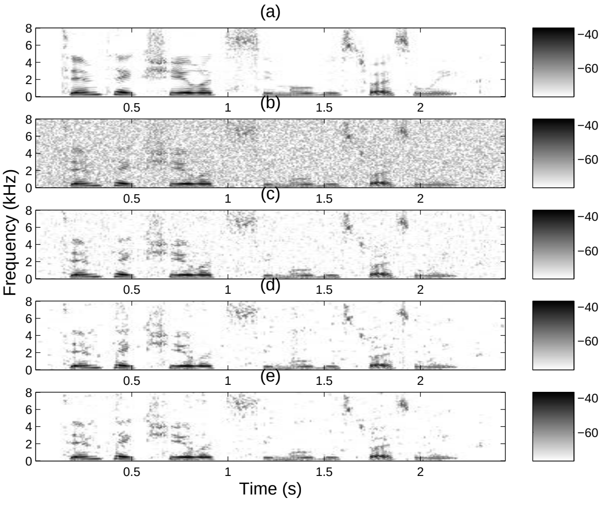

Figure 3: *Spectrograms of (a) clean speech, $\mathbf{y}^{(t)}$, (b) noisy speech, $\mathbf{z}^{(t)}$, (c) subtraction-based estimate, $\hat{\mathbf{y}}_s^{(t)}$, (d) model-based estimate with estimated parameters, $\hat{\mathbf{y}}_{\hat{m}}^{(t)}$ and (e) model-based estimate with true parameters, $\hat{\mathbf{y}}_m^{(t)}$*

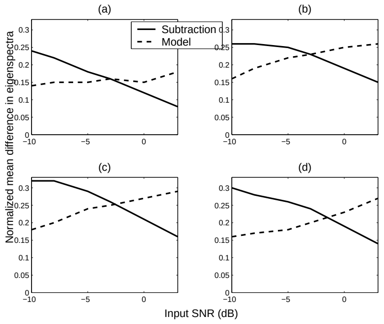

Figure 4: *Mean difference in eigenspectra for (a) test speech 1 + white noise, (b) test speech 2 + speech-like noise, (c) test speech 3 + lynx helicopter noise and (d) test speech 4 + phantom aircraft noise at various input SNR values.*

4. CONCLUSION

In this paper, we propose a model for the eigenspectra of clean speech sounds for a speaker and investigate its use for the enhancement of noisy signal eigenspectra. The estimation equation for the clean speech eigenspectrum combines a noise subtraction-based estimate with a model-based one. The latter is shown to be robust for noisy speech at low input SNR values because it is selected from known clean speech eigenspectrum clusters represented by the proposed model.

5. REFERENCES

[1] J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[2] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992, keywords: Review / survey of state of art, Model-based approach, Hidden Markov models.

[3] L. Deng, J. Droppo, and A. Acero, “A Bayesian approach to speech feature enhancement using the dynamic cepstral prior,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2002, vol. 1, pp. 829–832.

[4] H. Attias, L. Deng, A. Acero, and J.C. Platt, “A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for noise,” in *Proceedings of the 7th Eurospeech Conference*, 2001, pp. 1903–1906.

[5] 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.

[6] A. Rezayee and S. Gazor, “An Adaptive KLT Approach for Speech Enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 87–95, 2001.

[7] 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.

[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.