CODEBOOK-BASED BAYESIAN SPEECH ENHANCEMENT

Sriram Srinivasan, Jonas Samuelsson and W. Bastiaan Kleijn

Dept. Signals, Sensors and Systems
KTH (Royal Institute of Technology), Stockholm, Sweden
{sriram.srinivasan,jonas.samuelsson,bastiaan.kleijn}@s3.kth.se

ABSTRACT

In this paper, we propose a Bayesian approach for the estimation of the short-term predictor parameters of speech and noise, from the noisy observation. The resulting estimates of the speech and noise spectra can be used in a Wiener filter or any state-of-the-art speech enhancement system. We utilize a-priori information about both speech and noise in the form of trained codebooks of linear predictive coefficients. In contrast to current Bayesian estimation approaches that consider the excitation variances as part of the a-priori information, in the proposed method they are computed analytically based on the observation at hand. Consequently, the method performs well in nonstationary noise conditions. Experimental results confirm the superior performance of the proposed method compared to existing Bayesian approaches, such as those based on hidden Markov models.

1. INTRODUCTION

The freedom and flexibility provided by mobile communications introduces new challenges, one of the most prominent being the suppression of background acoustic noise. Mobile users communicate in different environments with varying types and amounts of background noise. Noise reduction remains a challenging problem largely due to the wide variety of background noise types and the difficulty in estimating their statistics. A majority of noise suppression techniques fall into the category of single-channel algorithms that have only a single microphone to obtain the input signal, and are thus attractive in mobile applications due to cost and size factors. Examples of such methods include [1, 2]. A drawback is that noise estimates need to be obtained from the noisy speech observation. This has proved to be a particularly difficult task, especially in nonstationary noise conditions. Most noise estimation methods typically employ a buffer of past noisy data from which estimates are obtained [3, 4]. While on the one hand large buffers result in accurate estimates, on the other hand they make it difficult to deal with changing noise, which is often the case in practice.

A method to estimate the short-term predictor (STP) parameters of speech and noise using a-priori information was presented in [5]. The STP parameters consist of the linear predictive (LP) coefficients and the excitation variance, which is the variance of the prediction error. The use of a-priori information about noise eliminates the dependence on buffers of past data providing good performance in nonstationary noise conditions [6]. The a-priori information consists of trained codebooks of LP coefficients. Each pair of vectors from the speech and noise codebooks, together with the (unknown) excitation variances represent a model of the noisy speech spectrum. The excitation variances are determined by finding the best spectral fit between the observed and the modelled noisy spectrum, with respect to a particular distortion measure. The codebook pair that minimizes the distortion measure, together with the corresponding excitation variances are selected as the best representation of the underlying speech and noise spectra. In [7], maximum-likelihood (ML) estimates of the STP parameters were obtained by using the Itakura-Saito distortion measure. By eliminating the dependence on long-term estimates of the noise spectrum, it is possible to react to quickly changing noise conditions.

In the ML estimation proposed in [7], the LP coefficients were considered to be deterministic parameters. In this paper, we treat them as random variables and obtain minimum mean square error (MMSE) estimates. While in [7], one pair of speech and noise LP vectors was selected from the codebooks, the MMSE estimate of the speech (noise) LP vector is a weighted sum of the speech (noise) codebook vectors. Such a soft-decision estimation approach allows for a proportionate contribution from closely competing candidates.

The HMM based systems [1, 2] also perform Bayesian estimation using a-priori information. In [1], the clean signal is modelled using Gaussian AR HMMs. The minimum mean-squared error (MMSE) estimator of clean speech given the noisy speech is obtained as a weighted sum of MMSE estimators corresponding to each state of the HMM for the clean signal. This approach is generalized in [2] to include noise HMMs as well. However, the HMM based systems treat the excitation variance as part of the a-priori information. The MMSE estimate in [8] also treats the excitation variance as part of the a-priori information. In the method proposed in this paper, in addition to obtaining the MMSE estimate of the speech and noise LP coefficients, we also compute the MMSE estimate of the speech and noise excitation variances based on the observed noisy speech and the codebooks. We eliminate the dependence on conventional noise estimation schemes, which is a fundamental limitation of current single-channel methods. Consequently, the method proposed here can perform well in nonstationary noise conditions.

2. BACKGROUND

In this section, we provide a brief overview of the codebook based ML estimation procedure. Consider an additive noise model where speech and noise are independent:

$$y(n) = x(n) + w(n), \tag{1}$$

This work was partially supported by the European Commission under the ANITA project (IST-2001-34327)

where $y(n), x(n)$ and $w(n)$ represent the noisy speech, clean speech, and noise respectively. We have trained codebooks of speech and noise spectral shapes parameterized as LP coefficients. We consider only the envelope of the spectrum and not its fine structure. The noisy spectrum can be modelled by a combination of speech and noise LP spectral shapes from the respective codebooks, together with their excitation variances. Given the spectral shapes and excitation variances, the modelled noisy spectrum can be written as

$$\hat{P}_y(\omega) = \frac{\sigma_x^2}{|A_x(\omega)|^2} + \frac{\sigma_w^2}{|A_w(\omega)|^2}, \tag{2}$$

where σ_x^2 and σ_w^2 are the excitation variances of clean speech and noise respectively, and

$$A_x(\omega) = \sum_{k=0}^{p} a_{x_k} e^{-j\omega k}, \quad A_w(\omega) = \sum_{k=0}^{q} a_{w_k} e^{-j\omega k}. \tag{3}$$

$\theta_x = (a_{x_0}, \ldots, a_{x_p}), \theta_w = (a_{w_0}, \ldots, a_{w_q})$ are the LP coefficients of clean speech and noise with p, q being the respective LP-model orders and $a_{x_0} = a_{w_0} = 1$. The parameters to be estimated are $\{\sigma_x^2, \sigma_w^2, \theta_x, \theta_w\}$. The codebook indices corresponding to the ML estimate of the speech and noise LP vectors are given by

$$\{i^*, j^*\} = \arg\min_{i,j} \{ \min_{\sigma_x^2, \sigma_w^2} d_{\mathrm{IS}}(P_y, \frac{\sigma_x^2}{|A_x^i|^2} + \frac{\sigma_w^2}{|A_w^j|^2})\}, \tag{4}$$

where $A_x^i(\omega)$ and $A_w^j(\omega)$ are the spectra of the i^{th} and j^{th} speech and noise codebook entries respectively and d_{IS} is the Itakura-Saito measure [9]. For a given $A_x(\omega)$ and $A_w(\omega)$, the excitation variances that minimize the Itakura-Saito distortion between $P_y(\omega)$ and $\hat{P}_y(\omega)$ can be shown to be the solution to the following system of equations [6]:

$$\mathbf{C} \begin{bmatrix} \sigma_x^2 \\ \sigma_w^2 \end{bmatrix} = \mathbf{D}, \tag{5}$$

where $\mathbf{C}, \mathbf{D}$ are given by

$$\mathbf{C} = \begin{bmatrix} \| \frac{1}{P_y^2(\omega)|A_x(\omega)|^4} \| & \| \frac{1}{P_y^2(\omega)|A_x(\omega)|^2|A_w(\omega)|^2} \| \\ \| \frac{1}{P_y^2(\omega)|A_x(\omega)|^2|A_w(\omega)|^2} \| & \| \frac{1}{P_y^2(\omega)|A_w(\omega)|^4} \| \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} \| \frac{1}{P_y(\omega)|A_x(\omega)|^2} \| \\ \| \frac{1}{P_y(\omega)|A_w(\omega)|^2} \| \end{bmatrix}, \tag{6}$$

where $\|f(\omega)\| = \int |f(\omega)| d\omega$.

3. BAYESIAN ESTIMATION

In this section, we first derive the Bayesian MMSE estimates of the speech and noise STP parameters by treating the speech and noise LP coefficients and their excitation variances as random variables. The resulting framework is then used to obtain MMSE estimates of two useful functions of the STP parameters, one of which is shown to result in the MMSE estimate of the clean speech waveform, given the noisy speech.

3.1. MMSE estimation of STP parameters

Let θ_x and θ_w denote the random variables corresponding to the speech and noise LP coefficients respectively. Let σ_x^2 and σ_w^2 denote the random variables corresponding to the speech and noise excitation variances respectively. We wish to jointly estimate the speech and noise LP coefficients and the excitation variances such that the mean-squared error is minimized. Let $\theta = [\theta_x, \theta_w, \sigma_x^2, \sigma_w^2]$. The desired MMSE estimate can be written as

$$\hat{\theta} = \mathrm{E}\{\theta|\mathbf{y}\}, \tag{7}$$

where $\mathbf{y} = [y(1)y(2)\ldots y(N)]$ is the vector of noisy observations for the current frame, with N the frame length. We rewrite (7) as

$$\hat{\theta} = \int_\Theta \theta p(\theta|\mathbf{y}) d\theta = \int_\Theta \theta \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} d\theta, \tag{8}$$

where the integral is over the space $\Theta = \Theta_x \times \Theta_w \times \Sigma_x \times \Sigma_w$, where Θ_x, Θ_w represent the support-space of the vectors of speech and noise LP coefficients and Σ_x, Σ_w represent the support-space for the speech and noise excitation variances. From the independence assumption, we have

$$p(\theta) = p(\theta_x, \sigma_x^2)p(\theta_w, \sigma_w^2). \tag{9}$$

For simplicity, we assume $p(\theta_x, \sigma_x^2) = p(\theta_x)p(\sigma_x^2)$ and likewise for the noise.

We next show that given θ_x, θ_w and the noisy speech $\mathbf{y}$, the likelihood $p(\mathbf{y}|\theta)$ decays rapidly from its maximum value as a function of the deviation from the ML estimate of the variances $\sigma_x^{2,\mathrm{ML}}$ and $\sigma_w^{2,\mathrm{ML}}$ obtained using (5) and (6). We first consider the case where noise is not present. In the absence of background noise, under Gaussianity assumptions, the probability density of the speech samples given the LP parameters can be written as

$$p_{\mathbf{x}}(\mathbf{x}|\mathbf{a}_x, \sigma_x^2) = \frac{1}{(2\pi)^{N/2}|\mathbf{R}_x|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T \mathbf{R}_x^{-1}\mathbf{x}), \tag{10}$$

where $\mathbf{x} = [x(0)x(1)\ldots x(N\text{-}1)]^T$, $\mathbf{a}_x = [1\ a_{x_1}\ a_{x_2}\ldots a_{x_p}]^T$ and $\mathbf{R}_x = \sigma_x^2(\mathbf{A}_x^T\mathbf{A}_x)^{-1}$, where $\mathbf{A}_x$ is the $N \times N$ lower triangular Toeplitz matrix with $[1 a_{x_1} a_{x_2} \ldots a_{x_p} 0 \ldots 0]^T$ as the first column. Assuming that the frame length is large so that the covariance matrix $\mathbf{R}_x$ can be described as circulant and is hence diagonalized by the Fourier transform, we have $\mathbf{R}_x = F^H \sigma_x^2 \Lambda_x F$, where F denotes the Fourier transform matrix, the superscript ^{H} denotes complex conjugate transpose and Λ_x is a diagonal matrix containing the eigenvalues of $\mathbf{R}_x$ scaled down by σ_x^2. We wish to study the effect of a deviation δ_x in the excitation variance σ_x^2 on $p_{\mathbf{x}}(\mathbf{x}|\mathbf{a}_x, \sigma_x^2)$. Let $\mathbf{R}'_x = F^H(\sigma_x^2 + \delta_x)\Lambda_x F$. We have

$$p_{\mathbf{x}}(\mathbf{x}|\mathbf{a}_x, \sigma_x^2 + \delta_x) = \frac{1}{(2\pi)^{N/2}|\mathbf{R}'_x|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T \mathbf{R}_x'^{-1}\mathbf{x})$$

$$= \underbrace{\frac{1}{(2\pi)^{N/2}(\sigma_x^2 + \delta_x)^{\frac{N}{2}} \prod_{i=1}^N \lambda_{x_i}}}_{A} \underbrace{\exp(-\frac{1}{2}\sum_{i=1}^N \frac{|\mathcal{X}_i|^2}{(\sigma_x^2 + \delta_x)\lambda_{x_i}})}_{B}, \tag{11}$$

where $\mathcal{X}_i$, $1 \le i \le N$ are the Fourier transform coefficients of $\mathbf{x}$. We note that δ_x can take values in the range $[-\sigma_x^2, \infty)$. For positive values of δ_x, the exponential in term B converges to one and the behavior of the likelihood is dominated by $(\sigma_x^2 + \delta_x)^{\frac{-N}{2}}$, which indicates a rapid decay. For negative values of δ_x, term B dominates resulting in an exponential decay of the likelihood.

Considering the case where noise is present, we can show in a

similar fashion

$$p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_x, \mathbf{a}_w, \sigma_x^2+\delta_x, \sigma_w^2+\delta_w) \quad (12)$$
$$= \prod_{i=1}^{N} \frac{\exp(-\frac{1}{2}\frac{|\mathcal{Y}_i|^2}{(\sigma_x^2+\delta_x)\lambda_{x_i}+(\sigma_w^2+\delta_w)\lambda_{w_i}})}{(2\pi)^{1/2}[(\sigma_x^2+\delta_x)\lambda_{x_i}+(\sigma_w^2+\delta_w)\lambda_{w_i}]^{1/2}},$$

where $\mathcal{Y}_i,\ 1 \leq i \leq N$ are the Fourier transform coefficients of $\mathbf{y}$ and δ_w, λ_{w_i} are defined analogously to δ_x, λ_{x_i} respectively. It can be seen from (12) that the likelihood exhibits a behavior similar to the clean speech case for both positive and negative values of δ_x and δ_w.

Thus given θ_x, θ_w and the noisy speech $\mathbf{y}$, the likelihood is significant only at the ML estimate of the excitation variances and we can approximate (8) as

$$\hat{\theta} \approx \int_{\Theta_x}\int_{\Theta_w} \theta' \frac{p(\mathbf{y}|\theta_x, \theta_w; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}})p(\theta')}{p(\mathbf{y})} d\theta_x d\theta_w, \quad (13)$$

where $\theta' = [\theta_x,\ \theta_w,\ \sigma_x^{2,\mathrm{ML}},\ \sigma_w^{2,\mathrm{ML}}]$ and $\hat{\theta} = [\hat{\theta}_x,\ \hat{\theta}_w,\ \hat{\sigma}_x^2,\ \hat{\sigma}_w^2]$. Note that we now have only a double integral over the support-space of the LP coefficients. $p(\mathbf{y})$ can be obtained as

$$p(\mathbf{y}) = \int_{\Theta_x}\int_{\Theta_w} p(\mathbf{y}|\theta_x, \theta_w; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}})p(\theta')d\theta_x d\theta_w. \quad (14)$$

In practice, the integrals in (13) and (14) are evaluated using numerical integration:

$$\hat{\theta} = \frac{1}{N_x N_w}\sum_{i,j=1}^{N_x,N_w} \theta'_{ij} \frac{p(\mathbf{y}|\theta_x^i, \theta_w^j; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}})p(\sigma_{x,ij}^{2,\mathrm{ML}})p(\sigma_{w,ij}^{2,\mathrm{ML}})}{p(\mathbf{y})}, \quad (15)$$

$$p(\mathbf{y}) = \frac{1}{N_x N_w}\sum_{i,j=1}^{N_x,N_w} p(\mathbf{y}|\theta_x^i, \theta_w^j; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}})p(\sigma_{x,ij}^{2,\mathrm{ML}})p(\sigma_{w,ij}^{2,\mathrm{ML}}),$$

where $\theta'_{ij} = [\theta_x^i,\ \theta_w^j,\ \sigma_{x,ij}^{2,\mathrm{ML}},\ \sigma_{w,ij}^{2,\mathrm{ML}}]$, θ_x^i and θ_w^j are the i^{th} speech codebook and j^{th} noise codebook entries respectively, $\sigma_{x,ij}^{2,\mathrm{ML}}$, $\sigma_{w,ij}^{2,\mathrm{ML}}$ are the maximum likelihood estimates of the speech and noise excitation variances corresponding to $\mathbf{y}, \theta_x^i$ and θ_w^j, and N_x, N_w are the speech and noise codebook sizes. Here we assume that the codebooks model the probability density of the AR data. For simplicity, we assume that the excitation variances are uniformly distributed in the interval $[0, \sigma_{\max}^2]$. The exact value of $\sigma_{\max}^2$ is irrelevant since, for a uniform distribution, the terms cancel out in the numerator and denominator of (15). Using the equivalence of the log-likelihood and the Itakura-Saito distortion, we can compute

$$p(\mathbf{y}|\theta_x, \theta_w; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}) = C\exp(-d_{\mathrm{IS}}(P_y, \hat{P}_y^{\mathrm{ML}})). \quad (16)$$

The constant C also appears in the expression for $p(\mathbf{y})$, and thus cancels out in the numerator and denominator of (15). The estimate $\hat{\theta}$ can be used to construct a Wiener filter to obtain the enhanced speech:

$$H_1(\omega) = \frac{\hat{\sigma}_x^2}{|\hat{A}_x(\omega)|^2}/\Big(\frac{\hat{\sigma}_x^2}{|\hat{A}_x(\omega)|^2} + \frac{\hat{\sigma}_w^2}{|\hat{A}_w(\omega)|^2}\Big), \quad (17)$$

where $\hat{A}_x(\omega), \hat{A}_w(\omega)$ are the spectra corresponding to $\hat{\theta}_x, \hat{\theta}_w$ respectively.

Since interpolation of LP coefficients can result in unstable filters, alternate representations are often used [10]. Representations that are guaranteed to result in stable synthesis filters include line spectral frequencies (LSF), autocorrelation coefficients (ACR), reflection coefficients and log-area ratios (LAR). Among these, it has been shown that LSFs result in the best performance and thus we perform the MMSE estimation in the LSF domain.

3.2. MMSE estimation of functions of $\theta_x, \theta_w, \mathbf{y}$

The estimation framework represented by (13) can be used to obtain MMSE estimates of different parametric representations based on the LP coefficients. For notational convenience, we define the function

$$f(\theta_x, \theta_w, \mathbf{y}) = \frac{p(\mathbf{y}|\theta_x, \theta_w; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}})p(\theta')}{p(\mathbf{y})}. \quad (18)$$

The MMSE estimate of any function $g(\theta_x, \theta_w, \mathbf{y})$ can be obtained as

$$\hat{g}(\theta_x, \theta_w, \mathbf{y}) = \int_{\Theta_x}\int_{\Theta_w} g(\theta_x, \theta_w, \mathbf{y})f(\theta_x, \theta_w, \mathbf{y})d\theta_x d\theta_w. \quad (19)$$

$g(\cdot)$ depends on $\mathbf{y}$ since $\sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}$ depend on $\mathbf{y}$. For example, let $g(\theta_x, \theta_w, \mathbf{y})$ be the Wiener filter defined as $H(\omega; \theta_x, \theta_w, \mathbf{y}) = \frac{\sigma_x^{2,\mathrm{ML}}}{|A_x(\omega)|^2}/(\frac{\sigma_x^{2,\mathrm{ML}}}{|A_x(\omega)|^2} + \frac{\sigma_w^{2,\mathrm{ML}}}{|A_w(\omega)|^2})$, where $\sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}$ are the ML estimates of the speech and noise excitation variances given θ_x, θ_w, and $\mathbf{y}$ obtained according to (5) and $A_x(\omega), A_w(\omega)$ are the spectra of the speech and noise LP coefficients θ_x, θ_w. The MMSE estimate $H_2(\omega)$ of the Wiener filter is obtained as:

$$H_2(\omega) = \int_{\Theta_x}\int_{\Theta_w} H(\omega; \theta_x, \theta_w, \mathbf{y})f(\theta_x, \theta_w, \mathbf{y})d\theta_x d\theta_w. \quad (20)$$

We note that the enhanced speech obtained by filtering $\mathbf{y}$ with the filter $H_2(\omega)$ is the MMSE estimate of the clean waveform, $\mathrm{E}\{\mathbf{X}|\mathbf{y}\}$, where $\mathbf{X}$ is the random variable corresponding to clean speech. This can be seen if we write

$$\mathrm{E}\{\mathbf{X}|\mathbf{y}\} = \int_{\Theta} p(\theta|\mathbf{y})\mathrm{E}\{\mathbf{X}|\mathbf{y}, \theta\}d\theta \quad (21)$$
$$= \int_{\Theta_x}\int_{\Theta_w} f(\theta_x, \theta_w, \mathbf{y})\mathrm{E}\{\mathbf{X}|\mathbf{y}, \theta_x, \theta_w; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}\}d\theta_x d\theta_w.$$

For Gaussian AR models, $\mathrm{E}\{\mathbf{X}|\mathbf{y}, \theta_x, \theta_w; \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}\}$ can be equivalently evaluated in the frequency domain as $H(\omega; \theta_x, \theta_w, \mathbf{y})\mathcal{Y}(\omega)$, where $\mathcal{Y}(\omega)$ is the Fourier transform of $\mathbf{y}$.

If we are interested in directly obtaining speech or noise spectra (e.g., as a noise estimation scheme), their MMSE estimates are given by

$$\hat{P}_x^{\mathrm{mmse}}(\omega) = \int_{\Theta_x}\int_{\Theta_w} \frac{\sigma_x^{2,\mathrm{ML}}}{|A_x(\omega)|^2} f(\theta_x, \theta_w, \mathbf{y})d\theta_x d\theta_w \quad (22)$$

$$\hat{P}_w^{\mathrm{mmse}}(\omega) = \int_{\Theta_x}\int_{\Theta_w} \frac{\sigma_w^{2,\mathrm{ML}}}{|A_w(\omega)|^2} f(\theta_x, \theta_w, \mathbf{y})d\theta_x d\theta_w, \quad (23)$$

where $\sigma_x^{2,\mathrm{ML}}$ is the ML estimate of the speech excitation variance given θ_x, θ_w and $\mathbf{y}$. A corresponding Wiener filter can be written as

$$H_3(\omega) = \frac{\hat{P}_x^{\mathrm{mmse}}(\omega)}{\hat{P}_x^{\mathrm{mmse}}(\omega) + \hat{P}_w^{\mathrm{mmse}}(\omega)}. \quad (24)$$

4. EXPERIMENTS

In this section, we describe the experiments performed to evaluate the performance of the MMSE estimation scheme. The test set comprised of ten speech utterances, five male and five female, from the TIMIT database, resampled at 8 kHz. A 10-bit speech codebook of dimension 10 was trained using 10 minutes of speech from the TIMIT database with the generalized Lloyd algorithm (GLA) [11]. The test utterances were not included in the training. A frame length of 240 samples was used with 50% overlap between adjacent frames. The frames were windowed using a Hann window. The noise types considered were highway noise (obtained by recording noise on a freeway as perceived by a pedestrian standing at a fixed point), siren noise (a two-tone siren recorded inside an emergency vehicle), speech babble noise (from Noisex-92) and white Gaussian noise. The noise codebooks were trained using the GLA with two minutes of training data. The noise samples used in the training and testing were different. We used the classifier described in [7] to select a particular noise codebook for a given input frame.

Tables 1 and 2 show the segmental signal-to-noise ratio (SSNR) and spectral distortion (SD) values for the noisy input and the enhanced speech obtained using the HMM method [2], the ML approach [7] and the three Wiener filters obtained with the Bayesian approach according to equations (17), (20) and (24) respectively. It can be seen that the Wiener filter (20) which is obtained as a weighted sum of the Wiener filters corresponding to each pair of speech and noise codebook vectors provides the best performance in terms of both SSNR and SD. All the Wiener filters obtained using the Bayesian approach perform better than those obtained using the HMM method and the ML approach. For the two-tone siren noise, the Bayesian approach performs slightly worse than the ML approach. The reason for this is that the siren at any instant is in one of two disjoint states. A weighted sum of these disjoint states thus performs worse than the best state alone (as in the ML approach). In general, the Bayesian MMSE approach provides superior performance.

Noise	Noisy	HMM	ML	H_1	H_2	H_3
Highway	1.9	5.9	7.2	8.6	8.5	8.4
White	0.7	6.1	6.1	7.2	7.6	7.5
Babble	1.3	4.0	5.2	6.3	6.0	5.9
Siren	0.7	6.8	11.1	10.1	10.2	10.2

Table 1. SSNR values (in dB) averaged over ten utterances at 10 dB input SNR for the HMM based method, the ML estimator and the proposed MMSE estimators.

Noise	Noisy	HMM	ML	H_1	H_2	H_3
Highway	3.3	3.0	2.8	2.9	2.5	2.5
White	4.6	3.6	3.8	4.7	3.5	3.6
Babble	3.2	3.1	3.1	3.2	2.7	2.8
Siren	4.7	3.4	2.2	2.8	2.7	2.7

Table 2. SD values (in dB) averaged over ten utterances at 10 dB input SNR for the HMM based method, the ML estimator and the proposed MMSE estimators.

5. CONCLUSIONS

We have presented an MMSE approach for the estimation of the short-term predictor parameters of speech and noise using a-priori information. A distinguishing feature of the proposed technique is that unlike current MMSE techniques, the excitation variances of the speech and noise AR models are computed on a frame-by-frame basis resulting in good performance in nonstationary noise environments. We have also presented MMSE estimation of different LP based parametric representations of the spectrum. In line with intuition, MMSE estimation of the clean waveform, and of the speech and noise spectra, results in better performance than individually obtaining the MMSE estimates of the STP parameters as seen from the performance of H_1, H_2 and H_3. Experimental results show that the proposed MMSE estimation scheme provides superior performance compared to existing methods.

6. REFERENCES

[1] Y. Ephraim, "A minimum mean square error approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1990, pp. 829–832.

[2] H. Sameti, H. Sheikhzadeh, and L. Deng, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 445–455, Sept. 1998.

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

[4] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, June 2000, pp. 1875–1878.

[5] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2001, pp. 669–672.

[6] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, Sept. 2003, pp. 1405–1408.

[7] ——, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Speech Audio Processing*, Accepted.

[8] M. Kuropatwinski and W. B. Kleijn, "Minimum mean square error estimation of speech short-term predictor parameters under noisy conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 96–99.

[9] R. M. Gray, A. Buzo, A. H. Gray Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. 28, no. 4, pp. 367–376, Aug. 1980.

[10] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier Science B.V., 1995, ch. 12, pp. 433–468.

[11] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.