NOISE POWER SPECTRAL DENSITY ESTIMATION FROM NOISY SPEECH USING ON-LINE TRAINED HIDDEN MARKOV MODELS

Karsten Vandborg Sørensen and Søren Vang Andersen

Department of Communication Technology, Aalborg University
Fredrik Bajers Vej 7, 9220 Aalborg Ø, Denmark
{kvs,sva}@kom.aau.dk

ABSTRACT

In this paper we describe a method for estimation of noise power spectral densities from a noisy speech signal. The method is used in conjunction with a time-frequency domain speech presence detection method that provides connected time-frequency regions of each decision type. In speech absence regions hidden Markov models are trained on-line and in speech presence regions the trained models are used for MMSE optimum estimation. Both types of speech presence regions can be present in each frame and on-line training of the models in speech absence can be conducted while the models in speech presence are used for estimation. Experiments show that the proposed noise PSD estimation method consistently performs better than three state-of-the-art reference methods. For real-life noise types the special case of the hidden Markov model where it reduces to a Gaussian mixture model is shown to be nearly as good as the hidden Markov model.

1. INTRODUCTION

Statistical noise estimation methods very often relies on an assumption of stationarity; the parameters of the noise PDF are estimated during speech absence and kept constant during speech presence. In this paper we investigate if dynamic modeling of the noise PDF using hidden Markov models (HMM's) results in better performance than when using static Gaussian mixture models (GMM's). We evaluate the performance for real-life noise types. Transition probabilities in the HMM's will, to a certain degree, capture the dynamic behavior of non-stationary noise. The general method of HMM's have already proven its worth for modeling different classes of noise. In particular, Sameti et. al. [1] and Gaunard et. al. [2] have used off-line trained noise models for noise classification. In their approach only a predefined set of noise classes are modeled and the off-line training is supervised. Similarly trained models have been suggested by Ghoreishi and Sheikhzadeh [3] for speech pause detection. A subband approach has been proposed by Hosoki et. al. [4] for noise detection. If a subband is unlikely to contain clean speech they classify it as being noise contaminated.

In previous work [5] we have proposed a connected time-frequency domain region speech presence detector and applied it to find bias compensation factors for minimum statistics based noise estimation. We have shown that this approach results in a less spectrally distorted noise estimate than the original minimum statistics based noise estimation [6]. In this paper, we apply our new approach to train subband HMM's on-line while speech is absent. When speech is present we use the most recently trained noise models for MMSE optimum noise power spectral density (PSD) estimation. This way, instead of choosing from a finite set of predefined noise class models, the proposed method models the local behavior of the noise to more accurately adapt to the actual noise environment.

The remainder of this paper is organized as follows. Section 2 describes the signal model, the speech presence hypotheses, the spectral smoothing method, and the fundamental noise PSD estimation approach. Section 3 provides the details of the applied statistical model. In Section 4 the methods for estimation of state probabilities and unknown observations are described. Section 5 contains the experiments and Section 6 provides a discussion of the proposed method and the obtained results.

2. SIGNAL MODEL

We assume that noisy speech $y(i)$ at sampling time index i consists of speech $s(i)$ and additive noise $n(i)$. For joint time-frequency analysis of $y(i)$ we apply the L-point discrete Short-Time Fourier Transform (STFT), i.e.

$$Y(\tau,\omega)=\sum_{\mu=0}^{L-1} y(\tau R+\mu)h(\mu)\exp\left(-j2\pi\omega\mu/L\right), \quad (1)$$

where $\tau\in\mathbb{Z}$ is the (sub-sampled) time index, $\omega\in\{0,1,\ldots,L-1\}$ is the frequency index, and L is the STFT size, which in this paper equals the window length. R is the skip between frames and $h(\mu)$ is a unit energy window function, i.e. $\sum_{\mu=0}^{L-1} h^2(\mu)=1$. From the linearity of (1) we have that $Y(\tau,\omega)=S(\tau,\omega)+N(\tau,\omega)$, where $S(\tau,\omega)$ and $N(\tau,\omega)$ are the STFT coefficients of speech $s(i)$ and additive noise $n(i)$, respectively. We further assume that $s(i)$ and $n(i)$ are zero mean and statistically independent, which leads to a power relation where the noise is additive. Now, let the hypotheses H_0 and H_1 for speech absence and speech presence, respectively, be defined by two power relations, i.e.

$$H_0:\quad E\{|Y(\tau,\omega)|^2\}=E\{|N(\tau,\omega)|^2\} \quad (2)$$

$$H_1:\quad E\{|Y(\tau,\omega)|^2\}=E\{|S(\tau,\omega)|^2\}+E\{|N(\tau,\omega)|^2\}. \quad (3)$$

The decision of which hypothesis to believe is true is done by means of the connected region speech presence detection method, which we have proposed in previous work [5]. This speech presence detection method provides individual decisions at each time-frequency location. At the same time it ensures that decisions of the same type are connected in larger time-frequency regions. In the regions where no speech is detected we can directly observe what we assume to be the realizations of the stochastic noise process. The approach taken in this paper is to exploit this property to train dynamic statistical noise models in connected regions of speech absence and use it for noise PSD estimation in regions of speech presence. Initially, we apply a spectral window of size $2D+1$, centered at ω, to reduce the fluctuations of the noisy speech periodogram bins, i.e.

$$B(\tau,\omega)=\frac{1}{2D+1}\sum_{\widetilde{\omega}=\omega-D}^{\omega+D}|Y(\tau,\widetilde{\omega})|^2. \quad (4)$$

At each time-frequency location (τ,ω) we let the spectrally averaged $B(\tau,\omega)$ constitute the noise PSD estimate if speech is absent in all the noisy speech periodogram bins within the spectral window in (4). If speech is present in any of these bins we turn to HMM based MMSE estimation.

This work was supported by The Danish National Centre for IT Research, Grant No. 329, and Microsound A/S.

3. TRAINING THE STATISTICAL MODEL

We model the spectrally averaged noisy periodogram bins $B(\tau,\omega)$ at each ω using a continuous density HMM with a Gaussian mixture model in each state of the HMM modeling the observation PDF. At current time, say $\mathscr{T}$, we consider the T' most recent spectrally smoothed periodograms, i.e. $B(\tau,\omega)$ for $\mathscr{T} - T' < \tau \leq \mathscr{T}$. If no noisy speech periodogram bins with speech presence was used in (4) to calculate any of these, we denote the case $D(\mathscr{T},\omega) = 0$ and otherwise we denote it $D(\mathscr{T},\omega) = 1$. Binary speech presence detection methods generally needs a certain amount of speech power to detect speech presence. Therefore, the last few noisy speech periodogram bins leading up to a speech presence region will most likely contain enough speech power to contaminate the HMM training. To avoid this, we train the HMM on the training set that consist of the first T spectrally smoothed periodograms within the sliding window of length T'. We train the model on the training set only if $D(\mathscr{T},\omega) = 0$. If $D(\mathscr{T},\omega) = 1$, the model parameters from $\mathscr{T} - 1$ are preserved except for the forward likelihoods, which are estimated by prediction from the forward likelihoods and state transition probabilities of the model at $\mathscr{T} - 1$. The set of T spectrally smoothed bins $B(\tau,\omega)$ for $\mathscr{T} - T' < \tau \leq \mathscr{T} - T' + T$ that constitutes a training set of T scalar observations at ω will cause the means and variances of the GMM's to be scalars. For easy generalization to the vector case we will, however, describe the theory using vector/matrix notation. As the training procedure is the same for all sets of training vectors we use the same notation for all $\mathscr{T}$ and ω, i.e. $\mathbf{X}_1^T \triangleq [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T-1}, \mathbf{x}_T]$. We want to find model parameters $\mathbf{\Phi}$ that maximize the joint likelihood[1] conditioned on the training set $\mathbf{X}_1^T$ of T observation vectors by adjusting the model parameters, i.e.

$$\widehat{\mathbf{\Phi}}_{ML} = \arg\max_{\mathbf{\Phi}} p(\mathbf{X}_1^T|\mathbf{\Phi}). \tag{5}$$

This optimization, however, of the (generally) non-convex objective function $p(\mathbf{X}_1^T|\mathbf{\Phi})$ requires knowledge of the hidden states and mixture components and is therefore not feasible. Instead we use the Baum-Welch algorithm [7], which by alternating maximization will converge to a model parameter estimate corresponding to a local maximum of the likelihood function. To ensure numerical stability we use the lower limit $\epsilon_1 = 10^{-6}$ on the multivariate Gaussian mixture components c_{jk}, the single element of the 1-by-1 covariance matrices $\mathbf{\Sigma}_{jk}$, and the sampled individual Gaussians $b_{jk}(\mathbf{x}_t)$ for all mixture numbers $1 \leq k \leq K$ and states $1 \leq j \leq N$. Also, we use a lower limit $\epsilon_2 = 10^{-2}$ on all entries in the state transition probability matrix $\mathbf{A} = \{a_{ij}\} = P(s_t = j|s_{t-1} = i)$ [8, pp. 381-382] for the states $1 \leq i,j \leq N$.

Let $\alpha_t(i)$ denote the forward likelihoods, defined as the likelihoods for being in state i at t while having produced the observations $\mathbf{X}_1^t$, conditioned on the model, i.e.

$$\alpha_t(i) \triangleq p(\mathbf{X}_1^t, s_t = i|\mathbf{\Phi}). \tag{6}$$

The forward likelihoods $\alpha_0(i)$ initializing the HMM training for $t = 0$ at each $\mathscr{T}$ are initialized as uniform distributions. After initialization at $t = 0$ the forward likelihoods for $1 \leq t \leq T$ in the Baum-Welch training algorithm are induced from the forward likelihoods at $t - 1$, i.e.

$$\alpha_t(j) = \overset{N}{\underset{i=1}{}} \alpha_{t-1}(i)a_{ij}b_j(\mathbf{x}_t), \tag{7}$$

where the sampled observation PDF $b_j(\mathbf{x}_t)$ from state j is given by the weighted mixture of sampled Gaussians $b_{jk}(\mathbf{x}_t)$, i.e.

$$b_j(\mathbf{x}_t) = \overset{M}{\underset{k=1}{}} c_{jk}b_{jk}(\mathbf{x}_t) = \overset{M}{\underset{k=1}{}} c_{jk}\mathscr{N}(\mathbf{x}_t; \boldsymbol{\mu}_{jk}, \mathbf{\Sigma}_{jk}). \tag{8}$$

To avoid numerical instability in the implementation all forward likelihood vectors are scaled to unity l_1-norm [8], i.e.

$$\widetilde{\boldsymbol{\alpha}}_t = \boldsymbol{\alpha}_t / \|\boldsymbol{\alpha}_t\|_1. \tag{9}$$

This scaling does not affect training nor does it affect HMM based estimation.

4. HIDDEN MARKOV MODEL BASED ESTIMATION

When speech presence is detected in a noisy speech periodogram bin it will cause $D(\tau,\omega) = 1$ for T' successive values of τ at all $2D + 1$ frequency indices ω where it is located within the spectral window in (4). While $D(\tau,\omega) = 1$ the forward likelihoods $\boldsymbol{\alpha}_{T'}$ at ω, which are required for HMM based estimation, are predicted from the most recently trained model at ω, i.e. the model at the most recent τ for which $D(\tau,\omega) = 0$. We denote the last observable training set, on which this model has been trained, $\widetilde{\mathbf{X}}_1^T$. When $D(\mathscr{T},\omega) = 1$ we set the sampled observation PDF equal to one for all states j, i.e. $b_j(\mathbf{x}_t) = 1$. This corresponds to an observation $\mathbf{x}_t$ that does not affect the forward likelihoods and therefore it leads to a simplified version of the forward likelihood induction equation in (6), i.e. $\alpha_t(j) = {}_{i=1}^{N} \alpha_{t-1}(i)a_{ij}$ for $j \in \{1, \ldots, N\}$. The simplified equation can be compactly written in vector/matrix notation, i.e.

$$\boldsymbol{\alpha}_t = \mathbf{A}^T \boldsymbol{\alpha}_{t-1}. \tag{10}$$

From (10) it follows that the F'th successively estimated forward likelihood vector for T' (at current time $\mathscr{T}$) is given by

$$\boldsymbol{\alpha}_{T'} = (\mathbf{A}^T)^F \boldsymbol{\alpha}_{T'-F}. \tag{11}$$

In the above $\boldsymbol{\alpha}_{T'-F}$ is relative to current time $\mathscr{T}$, i.e. it equals $\boldsymbol{\alpha}_T$ from time $\mathscr{T} - F + (T' - T)$. We now investigate whether or not this estimate will convergence to a "steady state". Suppose $\mathbf{A}^T \in \mathbb{R}^{N \times N}$ has N linearly independent eigenvectors. Then, by means of the eigenvalue decomposition we have that $\mathbf{A}^T = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$, where $\mathbf{\Lambda}$ is a diagonal matrix with non-increasingly sorted eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1} \geq \lambda_N$ on the diagonal and $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N]$ is a matrix of associated eigenvectors. We then have that $(\mathbf{A}^T)^F = \mathbf{S}\mathbf{\Lambda}^F\mathbf{S}^{-1}$. For the strictly positive, hence primitive, Markov matrix $\mathbf{A}^T$ the Perron-Frobenius theorem for primitive matrices [9, Theorem 1.1] states that there will be a unique dominant eigenvalue $\lambda_1 > |\lambda_i|$ for any eigenvalue $\lambda_i \neq \lambda_1$, which can be associated a strictly positive (or strictly negative) dominant eigenvector $\mathbf{s}_1$. For a positive Markov matrix the dominant eigenvalue will be $\lambda_1 = 1$ [9, p.118]. Now, if we let $F \rightarrow$ we have that the dominant eigenvalue remains one while the rest goes to zero. Therefore, $\lim_{F\rightarrow} \mathbf{S}\mathbf{\Lambda}^F\mathbf{S}^{-1}$ becomes a rank-1 projection matrix that projects onto the subspace $\mathscr{C}(\mathbf{s}_1) \in \mathbb{R}^N$, i.e. the line, spanned by a dominant eigenvector $\mathbf{s}_1$ of $\mathbf{A}^T$. Since we have that $\boldsymbol{\alpha}_{T'-F}$ is strictly positive it is not possible that $\mathbf{s}_1 \perp \boldsymbol{\alpha}_{T'-F}$. In effect,

$$\lim_{P\rightarrow} \mathbf{S}\mathbf{\Lambda}^F\mathbf{S}^{-1}\boldsymbol{\alpha}_{T'-F} = c\mathbf{s}_1, \tag{12}$$

where $c \in \mathbb{R} \setminus 0$ is a constant. In the practical implementation, we have that $\boldsymbol{\alpha}_{T'-F}$ is scaled by (9) to unity l_1-norm. Since left multiplication with the Markov matrix $\mathbf{A}^T$ does not affect the l_1-norm the limit value remains $c\mathbf{s}_1$, but readily scaled to unity l_1-norm. Thus, $c = sign(s_1(i))/||\mathbf{s}_1||_1$ for any $i \in \{1, \ldots, N\}$. To summarize, forward likelihood vectors predicted successively by (10) converge to a dominant eigenvector of $\mathbf{A}^T$. In fact, this implies that the HMM converges to a special case where it reduces to a GMM.

Once the forward likelihoods have been estimated we obtain the noise PSD estimate $\widehat{\mathbf{x}}_T$ as the conditional MMSE optimum observation vector $\widehat{\mathbf{x}}_T^\star$. i.e.

$$\widehat{\mathbf{x}}_T^\star = \arg\min_{\widehat{\mathbf{x}}_T} E\{(\widehat{\mathbf{x}}_T - \mathbf{x}_T)^T(\widehat{\mathbf{x}}_T - \mathbf{x}_T)|\widetilde{\mathbf{X}}_1^T, \mathbf{\Phi}\}. \tag{13}$$

[1] We use $p(\cdot)$ to denote probability density functions and $P(\cdot)$ to denote probability mass functions.

We assume that the unknown observation vector $\mathbf{x}_T$ is drawn from a continuous density HMM with observation PDF's modeled by a GMM in each state, i.e.

$$\mathbf{x}_T \sim \sum_{j=1}^{N}\sum_{k=1}^{K} P(s_T = j|\widetilde{\mathbf{X}}_1^T, \mathbf{\Phi}) c_{jk} \mathscr{N}(\mathbf{x}_T; \boldsymbol{\mu}_{jk}, \mathbf{\Sigma}_{jk}). \quad (14)$$

The conditional MMSE estimate of the observation parameter vector is therefore given by

$$\widehat{\mathbf{x}}_T^{\star} = \sum_{j=1}^{N}\sum_{k=1}^{K} P(s_T = j|\widetilde{\mathbf{X}}_1^T, \mathbf{\Phi}) c_{jk} \boldsymbol{\mu}_{jk}, \quad (15)$$

where the j'th conditional state probability

$$P(s_T = j|\widetilde{\mathbf{X}}_1^T, \mathbf{\Phi}) = p(s_T = j, \widetilde{\mathbf{X}}_1^T|\mathbf{\Phi})/p(\widetilde{\mathbf{X}}_1^T|\mathbf{\Phi}) \quad (16)$$

is obtained as the j'th entry in $\widetilde{\boldsymbol{\alpha}}_T$, given by (9).

5. EXPERIMENTS

For the experiments we use signals sampled at 8 kHz sampling frequency. Both frame and FFT size are $L = 256$ and the frame skip is $R = 128$ samples. The analysis window is a square-root Hanning scaled to unit energy. $D = 2$ defines the size of the spectral window and the number of observations in a training set is $T = 16$. These are taken from a sliding window of length $T' = 20$. In the experiments we evaluate the performance in subbands consisting of individual frequency tracks, i.e. scalar observations for all HMM's. The following methods are compared:

- *HMM(5,1)*: The proposed HMM based noise estimation method with $M = 1$ Gaussian in each state and $N = 5$ states.
- *GMM(5)*: The proposed method with $M = 5$ and $N = 1$. With $N = 1$ the HMM reduces to a GMM with M Gaussians.
- *CR-SPD*: Connected time-frequency region speech presence detection based smooth noise estimation [5].
- *MCRA*[2]: Minima controlled recursive averaging [10].
- *MS*: Minimum statistics noise estimation [6].

The first three methods are based on the connected time-frequency region speech presence detector proposed in [5]. The last two methods, MCRA and MS, serve as well known reference methods, which both feature independence from explicit speech presence detection. The performance of these methods are evaluated by means of their spectral distortion, which we measure as average segmental noise-to-error ratios (SegNER's). We calculate the SegNER directly in the time-frequency domain, as the ratio in dB between the noise energy and the noise estimation error energy. These values are upper and lower limited by 35 and 0 dB [11, p.586], respectively, and averaged over all $\mathscr{T}'$ frames, i.e.

$$SegNER = \frac{1}{\mathscr{T}'} \sum_{\tau=1}^{\mathscr{T}'} \min(\max(NER(\tau), 0), 35), \quad (17)$$

where the noise-to-error ratio, in dB, at τ is given by

$$NER(\tau) = 10\log_{10}\left(\frac{\sum_{\omega=0}^{L-1} |N(\tau,\omega)|^2}{\sum_{\omega=0}^{L-1}(|N(\tau,\omega)| - |\widehat{N}(\tau,\omega)|)^2}\right). \quad (18)$$

We evaluate for noisy speech with 4 different noise types, i.e. highway traffic, car interior, white, and helicopter noise, which (with zero mean) are added to the speech in -5, 0, 5, and 10 dB SNR. In each combination of SNR and noise type we average the SegNER's over speech from 3 male and 3 female speakers from the TIMIT database [12]. Initially, we illustrate the distinct dynamic modeling abilities and the "steady state" estimate convergence of HMM(5,1) with an example where the model is trained on the first 16 observations. For illustrative purposes, as well as for implementation verification, we have in this example used artificially created noise with dynamics that are well captured by the applied model. The example is shown in Fig. 1. We conclude that the initialization and subsequent Baum-Welch training lead to model parameters where the dynamics of the noise have been well captured by the model. The average SegNER's for the five noise estimation methods when applied on the noisy test set are listed in Table 1. The best average SegNER in each combination of noise type and input SNR is emphasized using bold letters. For -5 dB highway traffic noise HMM(5,1) and GMM(5) are equally good, so we emphasize the average SegNER of the method that was best before the SegNER's were rounded. From the table we see that HMM(5,1) has the highest average SegNER's for highway traffic, car interior, and helicopter noise and GMM(5) has the highest average SegNER's for white noise. That GMM(5) is better than HMM(5,1) in white noise is to be expected since the additional degrees of freedom in the HMM(5,1) relative to the GMM(5) tend to cause over-modeling. This is particularly the case when the training sets are small. In case of the HMM(5,1) the local variations in the realizations of stationary stochastic noise processes are, generally, captured by a dynamic model and in case of the GMM(5) by a static model. In all the tested combinations of noise type and SNR the proposed method is considerably better than both MCRA [10] and MS [6]. This is the case in regions of speech presence as well as in regions of speech absence. We note that in all cases the SegNER's for HMM(5,1) and GMM(5) have only minor differences. In Figure 2 we show the original noisy speech and the HMM(5,1) noise PSD estimates for one of the noisy speech signals from the test set.

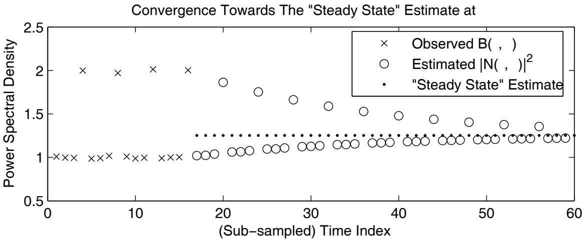

Figure 1: Convergence of the HMM(5,1) MMSE noise estimate towards the "steady state" estimate. Estimated values begin at $\tau = 17$.

6. DISCUSSION

We have proposed a HMM based method for noise PSD estimation that depends on an external connected time-frequency region speech presence detector. The method is trained on-line in speech absence and applied on-line for noise PSD estimation in connected regions of speech presence. Estimates from the proposed method are consistently less spectrally distorted than estimates from any of the three reference methods.

We have shown that for the tested real-life noise types there are only minor differences in performance between GMM(5) and HMM(5,1), i.e. the static model leads to similar performance as the dynamic model. For noise environments with clearly dynamically changing noise PSD PDF's the HMM would be advantageous, though, as it, like we show in Fig. 1, does have the ability to model dynamic behavior of the PDF, that is, non-stationary noise. However, in order for the HMM to be a significantly better model than the GMM the number of states and Gaussians in each GMM must be adequately chosen. Also, the size of the training sets needs to be large enough for the dynamics to be captured during the Baum-Welch training.

At the cost of increased computational complexity and memory requirements the number of states and Gaussians in the HMM

[2] MCRA is implemented for 8 kHz sampling frequency and 16 ms frame skip. Filter coefficients have time constants equal to the coefficients proposed by Cohen and Berdugo [10].

Noise Type	Highway Traffic				Car Interior				White				Helicopter			
Input SNR [dB]	-5	0	5	10	-5	0	5	10	-5	0	5	10	-5	0	5	10
HMM(5,1)	**6.08**	**5.77**	**5.31**	**4.77**	**5.79**	**5.72**	**5.69**	**5.66**	7.20	7.08	6.91	6.75	**6.26**	**5.79**	**5.29**	**4.99**
GMM(5)	6.08	5.75	5.25	4.69	5.74	5.68	5.62	5.60	**7.24**	**7.12**	**6.96**	**6.80**	6.23	5.76	5.28	4.94
CR-SPD	4.85	4.71	4.69	4.42	3.49	3.50	3.50	3.41	5.95	5.94	5.91	5.87	4.94	4.87	4.84	4.76
MCRA^2	0.00	0.05	0.51	2.73	0.03	0.09	0.53	2.67	0.45	5.22	4.92	3.28	0.02	0.09	0.93	2.79
MS	0.37	1.00	1.82	2.44	1.49	1.89	2.14	2.64	0.12	1.93	2.95	2.43	0.21	0.50	0.99	1.87

Table 1: Average segmental noise-to-error ratios in dB when compared to noise realization.

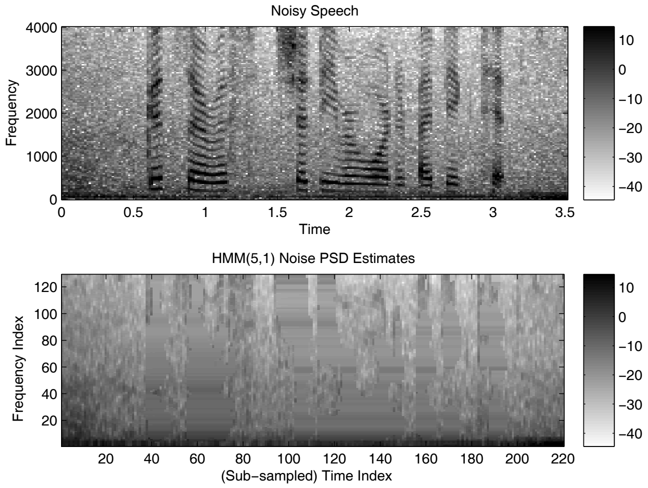

Figure 2: Spectrograms of noisy speech (top) and the HMM(5,1) noise PSD estimates (bottom) for 5 dB SNR highway traffic noise. The female speaker is uttering: *"Good service should be rewarded by big tips."* "Steady state" estimates appear to be dominating.

could be estimated on-line during speech pauses. At each ω various models could be trained and the best choice could be made from measuring the distortion between estimated and observed noise.

We have shown that the MMSE estimate under certain conditions converges to a "steady state" estimate determined by a dominant eigenvector of the transposed transition matrix. In case that the columns of the transposed state transition probability matrix $\mathbf{A}^T$ after Baum-Welch training remains uniformly distributed the trace, which equals the sum of the eigenvalues, will be 1. Therefore the dominant eigenvalue of the symmetric and positive semi-definite matrix, implying non-negative eigenvalues [13, p.269], will be the only eigenvalue different from zero. This will cause immediate convergence to the "steady state" estimate where the HMM reduces to a GMM. For stationary noise we consider this to be desired behavior. More generally, the gap in magnitude between the dominant eigenvalue and the remaining eigenvalues affects the rate of convergence. There will also be an impact from the angle between the forward likelihood vector and the individual subspaces spanned by associated eigenvectors. Experiments have shown that the rate of convergence for a number of models trained on the same training set with different initializations differs for parameters associated with each of the local minima of the likelihood function. There is no direct relationship, however, between the likelihood of a local minimum and the rate of convergence.

For environments with increasing or decreasing noise levels delta parameters will be better suited for the proposed noise estimation method. They will give the "steady state" estimate the ability to follow the level of the noise. Using non-delta representation is the most conservative approach and will unlike the delta representation ensure a stable MMSE estimate. It will, however, not be able to follow increasing nor decreasing noise during speech presence.

If the statistical models are trained with no spectral smoothing of the training sets, i.e. with $D = 0$, the method proposed in this paper could easily be modified to provide estimated noise PSD PDF's at each frequency. This makes the proposed method applicable, and very well suited, for statistical speech enhancement.

The method described in this paper can be applied on subbands of any width. For the model to benefit from any inter-frequency dependencies the vectors in each subband should be modeled using full (non-diagonal) covariance matrices. Better performance could very well be a consequence of the ability to model inter-frequency dependencies in the noise. HMM modeling employing parametric descriptions of larger time-frequency regions is a topic of current research.

REFERENCES

[1] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6(5), pp. 445–455, Sept. 1998.

[2] P. Gaunard, C. G. Mubikangiey, C. Couvreur, and V. Fontaine, "Automatic classification of environmental noise events by hidden Markov models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, May 1998, pp. 3609–3612.

[3] M. H. Ghoreishi and H. Sheikhzadeh, "A hybrid speech enhancement system based on HMM and spectral subtraction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, June 2000, pp. 1855–1858.

[4] M. Hosoki, T. Nagai, and A. Kurematsu, "Speech signal band width extension and noise removal using subband HMM," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2002, pp. I–245–I–248.

[5] K. V. Sørensen and S. V. Andersen, "Speech Enhancement with Natural Sounding Residual Noise based on Connected Time-Frequency Speech Presence Regions," To appear in EURASIP Journal on Applied Signal Processing, 2005.

[6] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9(5), pp. 504–512, July 2001.

[7] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[8] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[9] E. Seneta, *Non-negative Matrices and Markov Chains*. Springer-Verlag, 1981.

[10] I. Cohen and B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Letters*, vol. 9(1), pp. 12–15, Jan. 2002.

[11] John R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Wiley-Interscience, 2000.

[12] *DARPA TIMIT Acoustic-Phonetic Speech Database*, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA, CD-ROM.

[13] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Prentice-Hall, 1997.