On-line Noise Estimation Using Stochastic-Gain HMM for Speech Enhancement

David Y. Zhao, *Student Member, IEEE*, W. Bastiaan Kleijn, *Fellow, IEEE*, Alexander Ypma, and Bert de Vries

***Abstract*— We propose a noise estimation algorithm for single-channel noise suppression in dynamic noisy environments. A stochastic-gain hidden Markov model (SG-HMM) is used to model the statistics of non-stationary noise with time-varying energy. The noise model is adaptive and the model parameters are estimated on-line from noisy observations using a recursive estimation algorithm. The parameter estimation is derived for the maximum likelihood criterion and the algorithm is based on the recursive expectation maximization (EM) framework. The proposed method facilitates continuous adaptation to changes of both noise spectral shapes and noise energy levels, e.g., due to movement of the noise source. Using the estimated noise model, we also develop an estimator of the noise power spectral density (PSD) based on recursive averaging of estimated noise sample spectra. We demonstrate that the proposed scheme achieves more accurate estimates of the noise model and noise PSD, and as part of a speech enhancement system facilitates a lower level of residual noise.**

***Index Terms*— noise suppression, noise estimation, SG-HMM, gain modeling, noise model adaptation**

I. Introduction

Enhancement of noisy speech in mobile devices is a challenging task due to the large diversity and variability in the interfering acoustic noise. It is desirable for such systems to operate in both stationary and non-stationary noisy environments. However, effective design of noise estimation, a key component, has proven to be particularly challenging for non-stationary noise sources.

Traditional noise estimation techniques are based on recursive averaging of past noisy spectra, using frames that are likely to be dominated by noise. The update of the noise estimate is commonly controlled by a voice-activity detector (VAD), e.g., [1], a speech presence probability estimate [2], or order statistics [3, 4]. The relatively long window length for averaging inherently limits the performance of the noise estimation algorithms for non-stationary noises.

Consider a dynamic noisy environment with multiple moving noise sources. The movement of the noise source and the recording device leads to large variations in the noise energy. Further, the noise sources may have unique and time-varying spectral content. Such noise can be modeled reasonably well using noise models with multiple states and time-varying noise gains [5–7]. In combination with an elaborate data-driven speech model, the noise energy level can be estimated instantaneously using the noisy observation. Such schemes have been shown to perform well in many real-world non-stationary noisy environments, such as traffic noise. Nevertheless, the methods assume prior knowledge of the noise type, such that the proper noise model, obtained through off-line training, is used.

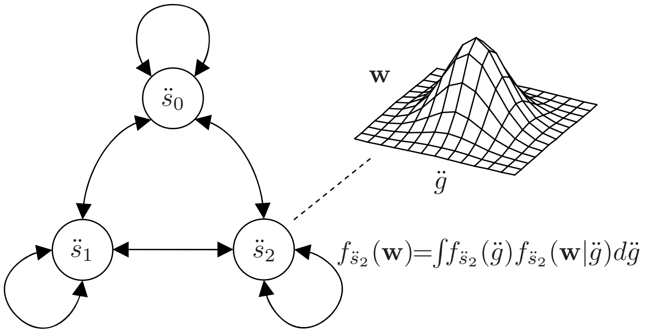
$\mathbf{w}$
$\ddot{g}$
$\ddot{s}_0$
$\ddot{s}_1$
$\ddot{s}_2$
$f_{\ddot{s}_2}(\mathbf{w})=\int f_{\ddot{s}_2}(\ddot{g})f_{\ddot{s}_2}(\mathbf{w}|\ddot{g})d\ddot{g}$

Fig. 1. Schematic diagram of a stochastic-gain HMM based noise model. $\mathbf{w}$ denotes a noise signal vector, s denotes an HMM state and g denotes a gain variable. The double dots $\ddot{}$ is used for the noise model to differentiate from the corresponding speech model (denoted overbar $\bar{}$).

It is commonly accepted that the overall characteristics of speech can be learned reasonably well from a (sufficiently rich) database of speech [5–10]. However, noise can vary greatly in real-world situations. Hence, it is in many cases impractical to capture all of this variation in off-line trained noise models, and on-line learning of changing noise characteristics is necessary.

In this work, we propose an extension of [5] using an adaptive noise model. We assume that the noise environment is unknown, and the noise model parameters are estimated on-line using the noisy observations. The system is based on stochastic-gain hidden Markov models (SG-HMM) [5] for modeling statistics of both speech and noise. A distinguishing feature of SG-HMM is that the gain is modeled explicitly as a random process with state-dependent distributions. The schematic diagram of an SG-HMM is shown in Fig. 1. Such a model is suitable for both speech and non-stationary noise with time-varying energy [5].

Estimation of the noise model parameters is optimized to maximize the likelihood of the noisy model. The proposed implementation is based on the recursive expectation maximization (EM) framework [11, 12]. The noise model is re-estimated/updated continuously from the noisy observa-

D. Y. Zhao and W. B. Kleijn are with the School of Electrical Engineering, (KTH) Royal Institute of Technology, Stockholm, Sweden (e-mail: david.zhao@ee.kth.se; bastiaan.kleijn@ee.kth.se).

A. Ypma and B. de Vries are with GN ReSound A/S, Algorithm R & D, Eindhoven, The Netherlands (e-mail: aypma@gnresound.com; bdevries@gnresound.com).

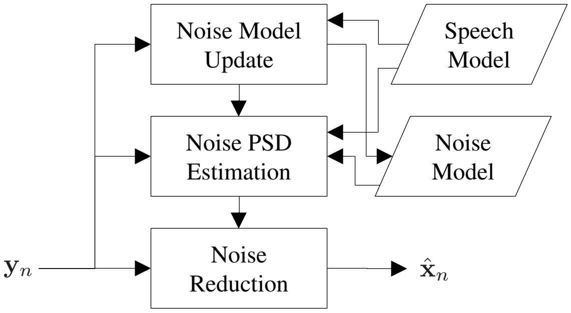

Fig. 2. Overview of a speech enhancement system using the proposed noise estimation algorithm. $\mathbf{y}_n$ denotes the noisy frame of time index n, and $\hat{\mathbf{x}}_n$ denotes the enhanced speech.

tions without using a VAD. Therefore, the scheme facilitates adaptation and correction to changing noise characteristics. Furthermore, we propose a safety-net state strategy to improve the robustness of the method and to avoid convergence to a locally optimal solution.

A number of estimators can be derived using the speech and noise models. In this work, we consider a system structure as shown in Fig. 2. The noise model is first re-estimated and updated using the current noisy observation. The noise power spectral density (PSD) is then estimated using the speech and noise models. The estimated noise PSD can be combined with any short-time spectral attenuation based speech enhancement systems.

In the remaining of this paper, on-line noise model parameter estimation is derived in section II. The noise PSD estimation algorithm is discussed in section III. Finally, experiments and results are presented in section IV.

II. Noise model parameter estimation

In this section, we present the on-line estimation algorithm based on SG-HMM modeling of speech and noise. The speech model is obtained off-line as described in [5]. The signal model is presented in section II-A, and the on-line model-parameter estimation of the noise model is presented in section II-B. A safety-net state strategy for improving the robustness of the method is presented in section II-C.

A. Signal model

We assume that the clean speech is contaminated by independent additive noise. The noisy signal is processed in frames of K samples, typically 20-32 ms, within which we can assume the stationarity of the speech and noise. The n'th noisy speech signal frame is modeled as

$$\mathbf{Y}_n = \mathbf{X}_n + \mathbf{W}_n, \qquad (1)$$

where $\mathbf{Y}_n = [Y_n[0], \ldots, Y_n[K-1]]^T$, $\mathbf{X}_n = [X_n[0], \ldots, X_n[K-1]]^T$ and $\mathbf{W}_n$= $[W_n[0], \ldots, W_n[K-1]]^T$ are random vectors of the noisy speech signal, clean speech and noise, respectively. We use uppercase letters to represent random variables, and lowercase letters to represent realizations of these variables. For simplicity, we use the notation $f(\mathbf{y}_n)$ for the probability density function of $\mathbf{Y}_n = \mathbf{y}_n$. The speech model and the estimation of the model parameters are identical to [5]. We use the notation overbar $^-$ to denote the parameters of the speech HMM and double dots $\ddot{}$ for the noise HMM.

The statistics of noise is modeled using an SG-HMM with explicit gain models in each state. Let $\mathbf{w}_0^n = \{\mathbf{w}_0, \ldots, \mathbf{w}_n\}$ denote the sequence of the noise frame realizations from 0 to n. The probability density function (PDF) of $\mathbf{w}_0^n$ is then modeled as

$$f(\mathbf{w}_0^n) \;=\; \sum_{\ddot{\mathbf{s}} \in \ddot{\mathbf{S}}} \prod_{t=0}^{n} \ddot{a}_{\ddot{s}_{t-1}\ddot{s}_t} f_{\ddot{s}_t}(\mathbf{w}_t), \qquad (2)$$

where the summation is over the set of all possible state sequences $\ddot{\mathbf{S}}$, and for each realization of the state sequence $\ddot{\mathbf{s}} = [\ddot{s}_0, \ddot{s}_1, \ldots, \ddot{s}_n]$, $\ddot{s}_n$ denotes the state of frame n, $f_{\ddot{s}_n}(\mathbf{w}_n)$ denotes the state dependent probability of $\mathbf{w}_n$ at state $\ddot{s}_n$, and $\ddot{a}_{\ddot{s}_{n-1}\ddot{s}_n}$ denotes the transition probability from state $\ddot{s}_{n-1}$ to $\ddot{s}_n$ with $\ddot{a}_{\ddot{s}_{-1}\ddot{s}_0}$ being the initial state probability. The time index n is sometimes dropped when the information is clear from the context.

Using an SG-HMM, the state-dependent PDF incorporates explicit gain models. Let $\ddot{g}'_n = \log \ddot{g}_n$ denote the noise gain in the logarithmic domain, the noise gain of state $\ddot{s}$ has the PDF

$$f_{\ddot{s}}(\ddot{g}'_n) \;=\; \frac{1}{\sqrt{2\pi\ddot{\sigma}^2_{\ddot{s}}}} \exp\!\left(-\frac{1}{2\ddot{\sigma}^2_{\ddot{s}}} \left(\ddot{g}'_n - \ddot{\mu}_{\ddot{s}}\right)^2\right), \qquad (3)$$

where $\ddot{\mu}_{\ddot{s}}$ and $\ddot{\sigma}^2_{\ddot{s}}$ are the mean and the variance. The log-normal PDF can be motivated by its simplicity and the non-negativity of the gain. Such a model has demonstrated good performance for modeling speech and noise gains [5]. Due to the one-to-one mapping of $\ddot{g}'_n$ and $\ddot{g}_n$, we use an appropriate notation depending on the context in the remainder of the paper. We note that (3) is more general than the noise gain model of [5]. In (3), a separate gain model is used for each noise state, while in [5], the same noise gain model is shared across multiple noise states.

The state-dependent PDF of the noise SG-HMM is defined by the integral over the noise gain variable,

$$f_{\ddot{s}}(\mathbf{w}_n) \;=\; \int_{-\infty}^{\infty} f_{\ddot{s}}(\ddot{g}'_n) f_{\ddot{s}}(\mathbf{w}_n|\ddot{g}'_n) d\ddot{g}'_n, \qquad (4)$$

$$f_{\ddot{s}}(\mathbf{w}_n|\ddot{g}'_n) \;=\; \frac{1}{(2\pi\ddot{g}_n)^{\frac{K}{2}}} \exp\!\left(-\frac{1}{2\ddot{g}_n}\mathbf{w}_n^{\sharp}\ddot{\mathbf{D}}_{\ddot{s}}^{-1}\mathbf{w}_n\right), \qquad (5)$$

where $\ddot{g}_n = \exp(\ddot{g}'_n)$, $\sharp$ denotes the Hermitian transpose and the covariance matrix $\ddot{\mathbf{D}}_{\ddot{s}} = (\mathbf{A}^{\sharp}_{\ddot{s}}\mathbf{A}_{\ddot{s}})^{-1}$, where $\mathbf{A}_{\ddot{s}}$ is a $K \times K$ lower triangular Toeplitz matrix with the first $\ddot{p}+1$ elements of the first column consisting of the AR coefficients $[\ddot{\alpha}_{\ddot{s}}[0], \ddot{\alpha}_{\ddot{s}}[1], \ddot{\alpha}_{\ddot{s}}[2], \ldots, \ddot{\alpha}_{\ddot{s}}[\ddot{p}]]^T$ with $\ddot{\alpha}_{\ddot{s}}[0] = 1$. Note that the covariance matrix $\ddot{\mathbf{D}}_{\ddot{s}}$ has determinant one. For a given noise gain $\ddot{g}_n$, the PDF $f_{\ddot{s}}(\mathbf{w}_n|\ddot{g}'_n)$ is considered to be a $\ddot{p}$-th order zero-mean Gaussian AR density function, equivalent to white Gaussian noise of variance $\ddot{g}_n$ filtered by the all-pole AR model filter.

The Gaussian AR density function can be simplified under the assumption of large K. $\mathbf{A}^T_{\ddot{s}}$ is then well approximated by a circulant matrix, which is diagonalized by a Fourier transformation matrix. Applying the Fourier transformation followed by Parseval's theorem, the density function is approximately [13],

$$f_{\ddot{s}}(\mathbf{w}_n|\ddot{g}'_n) \;\approx\; \frac{1}{(2\pi\ddot{g}_n)^{\frac{K}{2}}}\exp\!\left(-\frac{1}{2\ddot{g}_n}\sum_{i=0}^{\ddot{p}} C_r(i)\ddot{r}_{\ddot{s}}[i]r_w[i]\right), \quad (6)$$

where $C_r(i) = 1$ for $i = 0$, $C_r(i) = 2$ for $i > 0$, $\ddot{r}_{\ddot{s}}$ and r_w are the autocorrelations of the AR coefficients and the observations respectively

$$\ddot{r}_{\ddot{s}}[i] \;=\; \sum_{j=0}^{\ddot{p}-i}\ddot{\alpha}_{\ddot{s}}[j]\ddot{\alpha}_{\ddot{s}}[j+i], \quad (7)$$

$$r_w[i] \;=\; \sum_{j=0}^{K-i-1} w_n[j]w_n[j+i]. \quad (8)$$

B. On-line parameter estimation

The noise model parameters are obtained on-line from degraded speech using the maximum likelihood (ML) parameter estimation approach. The parameters consist of $\theta = \{\ddot{a}_{\ddot{s}'\ddot{s}}, \ddot{\mu}_{\ddot{s}}, \ddot{\sigma}^2_{\ddot{s}}, \ddot{\alpha}_{\ddot{s}}[i]\}$, which are the transition probabilities, means and variances of the logarithmic noise gain, and autoregressive model parameters. For notational convenience, let s denote a composite state of the noisy HMM, consisting of combination of the state $\bar{s}$ of the speech model component and the state $\ddot{s}$ of the noise model component. The summation over a function of the composite state corresponds to summation over both the speech and noise states, e.g., $\sum_s f(s) = \sum_{\bar{s}}\sum_{\ddot{s}} f(\bar{s},\ddot{s})$.

The estimation problem involves incomplete observations, and a direct optimization of the ML criterion is not straightforward. For this type of estimation problems, the recursive EM approach [11, 12] provides a convenient framework that often leads to practical solutions. Applied to our estimation problem, the hidden variables at frame n consist of $\mathbf{z}_n = \{s_n, \ddot{g}_n, \bar{g}_n, \mathbf{x}_n\}$, which are the composite state, noise and speech gain, and clean speech vector. The on-line parameter estimation can then be formulated as [11, 12]

$$\hat{\theta}_n \;=\; \arg\max_{\theta}\mathcal{Q}_n(\theta|\hat{\theta}_0^{n-1}), \quad (9)$$

where $\hat{\theta}_0^{n-1} = \{\hat{\theta}_j\}_{j=0..n-1}$ denotes the estimated parameters from the first frame to frame $n-1$ and the auxiliary $\mathcal{Q}_n(\cdot)$ function is defined as

$$\mathcal{Q}_n(\theta|\hat{\theta}_0^{n-1}) \;=\; \int_{\mathbf{z}_0^n} f(\mathbf{z}_0^n|\mathbf{y}_0^n,\hat{\theta}_0^{n-1})\log f(\mathbf{z}_0^n,\mathbf{y}_0^n|\theta)d\mathbf{z}_0^n \quad (10)$$

where the integral is over the hidden variables $\mathbf{z}_0,\ldots,\mathbf{z}_n$.

The auxiliary function (10) can be simplified using the approximations proposed in [5, 14]. We make use of the most probable gains $\{\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t}\}$ for each composite state s_t at frame t, defined as

$$\{\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t}\} \;=\; \arg\max_{\bar{g}_t,\ddot{g}_t}\log f_{s_t}(\ddot{g}_t,\bar{g}_t,\mathbf{y}_t). \quad (11)$$

Applying the approximations (derivation are given in appendix I), we approximate the unscaled state probabilities $\omega_t(s_t)$, transition probabilities $\omega'_t(s_{t-1},s_t)$ and the scaling factor Ω_t using

$$\begin{aligned}\omega_t(s_t) &= f(s_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})f_{s_t}(\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{y}_t|\hat{\theta}_{t-1}) && (12)\\ \omega'_t(s_{t-1},s_t) &= f(s_{t-1}|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})f(s_t|s_{t-1},\hat{\theta}_{t-1})\\ &\quad f_{s_t}(\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{y}_t|\hat{\theta}_{t-1}) && (13)\\ \Omega_t &= \sum_s\omega_t(s) = \sum_{s'}\sum_s\omega'_t(s',s). && (14)\end{aligned}$$

The auxiliary $\mathcal{Q}_n(\cdot)$ function (10) is then simplified to

$$\begin{aligned}\mathcal{Q}_n(\theta|\hat{\theta}_0^{n-1}) &\approx \sum_{t=0}^{n}\mathcal{L}_t(\theta|\hat{\theta}_0^{t-1}) && (15)\\ \mathcal{L}_t(\theta|\hat{\theta}_0^{t-1}) &= \sum_s\frac{\omega_t(s)}{\Omega_t}\int f_s(\mathbf{x}_t|\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{y}_t,\hat{\theta}_{t-1})\\ &\quad \log f_s(\mathbf{y}_t|\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{x}_t,\theta)d\mathbf{x}_t\\ &\quad +\sum_{s'}\sum_s\frac{\omega'_t(s',s)}{\Omega_t}\log\ddot{a}_{\ddot{s}'\ddot{s}}\\ &\quad +\sum_s\frac{\omega_t(s)}{\Omega_t}\log f_s(\hat{\ddot{g}}_{s_t}|\theta)\\ &= \mathcal{L}_{t_1}+\mathcal{L}_{t_2}+\mathcal{L}_{t_3}. && (16)\end{aligned}$$

The three additive terms are relevant for estimation of the AR parameters, transition probabilities and gain model parameters, respectively. The update equations for these parameters are given in the following paragraphs.

By a change of variable, $\mathbf{y}_t = \mathbf{x}_t + \mathbf{w}_t$, the auxiliary function with respect to the AR parameters becomes

$$\begin{aligned}\sum_{t=0}^{n}\mathcal{L}_{t_1} &= \sum_{t=0}^{n}\sum_s\frac{\omega_t(s)}{\Omega_t}\int f_s(\mathbf{w}_t|\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{y}_t,\hat{\theta}_{t-1})\\ &\quad \log f_s(\mathbf{w}_t|\hat{\ddot{g}}_{s_t},\theta)d\mathbf{w}_t\\ &\sim \sum_{\ddot{s}}\sum_{i=0}^{\ddot{p}}C_r(i)\ddot{r}_{\ddot{s}}[i]\left(\sum_{t=0}^{n}\sum_{\bar{s}}\frac{\omega_t(s)}{\Omega_t}\right.\\ &\quad \left.\frac{\int f_s(\mathbf{w}_t|\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{y}_t,\hat{\theta}_{t-1})r_w[i]d\mathbf{w}_t}{\hat{\ddot{g}}_{s_t}}\right), && (17)\end{aligned}$$

where the last step is due to (6) after grouping relevant terms and neglecting the constant terms. Taking the first derivative with respect to the noise AR parameters for state $\ddot{s}$ at frame n and setting to zero, we obtain the Levinson-Durbin recursive equations with estimated autocorrelation sequence. The estimated autocorrelation $\hat{\ddot{r}}_{\ddot{s}}[i]_n$ can be formulated as a recursive algorithm to facilitate on-line learning (with no additional delay),

$$\begin{aligned}\hat{\ddot{r}}_{\ddot{s}}[i]_n &= \left(\sum_{t=0}^{n}\sum_{\bar{s}}\frac{\omega_t(s)}{\Omega_t}\frac{\int f_s(\mathbf{w}_t|\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{y}_t,\hat{\theta}_{t-1})r_w[i]d\mathbf{w}_t}{\hat{\ddot{g}}_{s_t}}\right)\\ &\quad /\left(\sum_{t=0}^{n}\sum_{\bar{s}}\frac{\omega_t(s)}{\Omega_t}\right)\\ &= \hat{\ddot{r}}_{\ddot{s}}[i]_{n-1}+\frac{1}{\Xi_n(\ddot{s})}\sum_{\bar{s}}\frac{\omega_n(s)}{\Omega_n}\cdot\\ &\quad \left(\frac{\int f_s(\mathbf{w}_n|\hat{\ddot{g}}_{s_n},\hat{\bar{g}}_{s_n},\mathbf{y}_n,\hat{\theta}_{n-1})r_w[i]d\mathbf{w}_n}{\hat{\ddot{g}}_{s_n}}-\hat{\ddot{r}}_{\ddot{s}}[i]_{n-1}\right), && (18)\end{aligned}$$

where

$$\Xi_n(\ddot{s}) \quad = \quad \sum_{t=0}^{n}\sum_{\bar{s}} \frac{\omega_t(s)}{\Omega_t} = \Xi_{n-1}(\ddot{s}) + \sum_{\bar{s}} \frac{\omega_n(s)}{\Omega_n}, \tag{19}$$

is a normalization factor. The expected value $\int f_s(\mathbf{w}_n|\hat{\ddot{g}}_{s_n}, \hat{\bar{g}}_{s_n}, \mathbf{y}_n, \hat{\theta}_{n-1}) r_w[i] d\mathbf{w}_n$ can be obtained in the frequency domain by applying the inverse Fourier transform of the expected noise sample spectrum [9,15]. After obtaining the autocorrelation sequence, the AR parameters are then calculated from the estimated autocorrelation sequence using the Levinson-Durbin recursive algorithm.

The optimal state transition probability $\ddot{a}_{\ddot{s}'\ddot{s}}$ with respect to the auxiliary function (15) can be solved under the constraint $\sum_{\ddot{s}} \ddot{a}_{\ddot{s}'\ddot{s}} = 1$. Let $\tau_t(\ddot{s}', \ddot{s}) = \sum_{\bar{s}} \sum_{\bar{s}'} \frac{\omega'_t(s',s)}{\Omega_t}$ denote the weighting factor for the transition from state $\ddot{s}'$ to $\ddot{s}$. The solution can be formulated recursively:

$$\hat{\ddot{a}}_{\ddot{s}'\ddot{s},n} = \hat{\ddot{a}}_{\ddot{s}'\ddot{s},n-1} + \frac{\sum_{\ddot{s}} \tau_n(\ddot{s}', \ddot{s})}{\Xi'_n(\ddot{s}')} \cdot \left(\frac{\tau_n(\ddot{s}', \ddot{s})}{\sum_{\ddot{s}} \tau_n(\ddot{s}', \ddot{s})} - \hat{\ddot{a}}_{\ddot{s}'\ddot{s},n-1} \right), \tag{20}$$

where

$$\Xi'_n(\ddot{s}') \quad = \quad \Xi'_{n-1}(\ddot{s}') + \sum_{\ddot{s}} \tau_n(\ddot{s}', \ddot{s}). \tag{21}$$

The remainder of the noise model parameters are also estimated using recursive estimation algorithms. The derivation is similar to the derivation in [12]. The update equations for the gain model parameters are

$$\hat{\ddot{\mu}}_{\ddot{s},n} \quad = \quad \hat{\ddot{\mu}}_{\ddot{s},n-1} + \frac{1}{\Xi_n(\ddot{s})} \sum_{\bar{s}} \frac{\omega_n(s)}{\Omega_n} \left(\hat{\ddot{g}}'_{s_n} - \hat{\ddot{\mu}}_{\ddot{s},n-1} \right) \tag{22}$$

$$\hat{\ddot{\sigma}}^2_{\ddot{s},n} \quad = \quad \hat{\ddot{\sigma}}^2_{\ddot{s},n-1} + \frac{1}{\Xi_n(\ddot{s})} \sum_{\bar{s}} \frac{\omega_n(s)}{\Omega_n} \cdot \left(\left(\hat{\ddot{g}}'_{s_n} - \hat{\ddot{\mu}}_{\ddot{s},n-1} \right)^2 - \hat{\ddot{\sigma}}^2_{\ddot{s},n-1} \right), \tag{23}$$

where the most probable log-gain $\hat{\ddot{g}}'_{s_n}$ is used.

The practical implementation of the proposed estimator requires two additional adjustments:

1) Forgetting factor: To estimate time-varying parameters of the noise model, exponential forgetting factors are introduced in the update equations to restrict the impact of the past observations [12]. Hence, the modified normalization terms of (19) and (21) are evaluated by recursive summation of the past values

$$\Xi_n(\ddot{s}) \quad = \quad \rho\Xi_{n-1}(\ddot{s}) + \sum_{\bar{s}} \frac{\omega_n(s)}{\Omega_n}. \tag{24}$$

$$\Xi'_n(\ddot{s}') \quad = \quad \rho\Xi'_{n-1}(\ddot{s}') + \sum_{\ddot{s}} \tau_n(\ddot{s}', \ddot{s}), \tag{25}$$

where $0 \leq \rho \leq 1$ is an exponential forgetting factor and $\rho = 1$ corresponds to no forgetting.

2) Gain variance compensation: The noise gain variance estimate (23) is biased towards a lower value due to the approximation of the integration in (42), which essentially ignores the support of the $f_{s_t}(\ddot{g}_t, \bar{g}_t, \mathbf{y}_t)$ function. In some cases, e.g., $\hat{\ddot{g}}'_{s_n} \approx \hat{\ddot{\mu}}_{\ddot{s},n-1}$ for a few consecutive frames, (23) may lead to a noise gain variance that is close to zero. This leads to slow convergence when the noise gain varies thereafter. The underestimation of noise gain variance can be compensated by adding a small constant ε to the estimated variance. A similar approach was used in [16]. The constant was set such that the resulting $\ddot{\sigma}^2_{\ddot{s}}$ was lower-bounded by the gain variance of white Gaussian noise [5].

C. *Safety-net state strategy*

The parameter estimation algorithm is based on the recursive EM using forgetting factors. It is expected that the algorithm adapts to slowly-varying model parameters[1]. However, slow convergence and local optima are potential problems when the noise characteristics changes abruptly, e.g., the noise environment switches from one noise type to another. In extreme cases, the algorithm may converge to local optima far off from the globally optimal solution, and not be able to further improve the model parameters.

With the aforementioned problem in mind, we introduce a strategy based on an additional *safety-net state* in the noise model. The probability of a noise state $\ddot{s}$ for the past N frames is given by

$$\Psi_n(\ddot{s}) \quad = \quad \sum_{t=n-N}^{n} \sum_{\bar{s}} \frac{\omega_t(s)}{\Omega_t}. \tag{26}$$

We select the state with the lowest $\Psi_n(\ddot{s})$, corresponding to the least likely state over this period, as the safety-net state. The safety-net state is selected once every few seconds.

The on-line estimation procedure (section II-B) is not applied to the parameters of the safety-net state. Instead, they are constructed from the noise PSD estimated by a reference noise estimation algorithm that does not suffer from local convergence problems. In this work, we use the method based on minimum statistics [3]. For a safety-net state $\ddot{s}$, $\ddot{\mu}_{\ddot{s}}$ and $\ddot{\alpha}_{\ddot{s}}$ parameters are obtained from the estimated noise spectrum using the Levinson-Durbin recursive equations, and $\ddot{\sigma}^2_{\ddot{s}}$ is set to a small constant.

The behavior of the safety-net state can be summarized as follows. If the on-line estimation procedure performs better than the reference method, the safety-net state will have lower weights $\omega_n(s)$ in average, $\Psi_n(\ddot{s})$ will remain small, and no safety-net state relocation will occur. The safety-net state then has little contribution to the overall performance. Safety-net state relocation only occurs when the reference method produces a better model than the on-line estimation procedure. The previous safety-net state becomes again adaptive. The switching procedure corresponds to re-initialization of the previous safety-net state using the reference noise estimation method. In this case, re-initialization of the least probable noise state allows for a new starting point for the recursive EM algorithm. An experiment for demonstrating the behavior of the safety-net state strategy is presented in section IV-F.

[1] As for the state dependent gain models, the means and variances are considered slowly-varying.

D. Discussion

The proposed scheme is based on prior knowledge of speech using a hidden Markov model. Other speech enhancement methods based on prior knowledge of speech are, e.g., [5–10, 14, 15, 17–20].

The methods of [9, 19–21] require knowledge of the noise statistics, which has to be estimated separately using an additional algorithm. The methods based on prior knowledge of both speech and noise include [5–8, 14]. In these methods, the noise models are obtained through training using off-line recorded noise, and the on-line adaptation is restricted to the noise gain only.

In [10,15,18,22], more elaborate noise models are estimated from the noisy observations. The proposed method differs in how the parameters are estimated and the underlying noise model structure. The algorithm in [10, 15] only applies to a *batch* of noisy data, e.g., one sentence, and is not directly applicable for on-line estimation. Also, the noise model in [10] is limited to stationary Gaussian noise (white or colored). In [15, 18, 22], the noise HMM is considered adaptive and the parameters are estimated on-line from the observed noisy speech. To a certain extent, the methods deal with non-stationary noise. In many real-world noise situations, however, the noise energy varies rapidly with the movement of the noise source. In such cases, many components are required to model the changes with sufficient accuracy. Moreover, the methods of [15,22] apply the on-line learning during noise-only frames, and a VAD is required. In non-stationary noisy environments, a VAD is difficult to design and misclassification may be catastrophic.

The novelty of our proposed noise estimation algorithm is in the effective modeling of the noise gain and shape model using SG-HMM, and the continuous estimation of the model parameters without requiring a VAD. As the model is parameterized per state, it is capable of dealing with non-stationary noise with rapidly changing spectral contents. The noise gain models the time-varying noise energy level due to, e.g., movement of the noise source. The separation of the noise gain and shape modeling allows for improved modeling efficiency over [15, 18, 22], i.e. the noise model would require fewer mixture components and we may assume that model parameters change less quickly with time.

III. Noise power spectrum estimation

In this section, we discuss a noise PSD estimation algorithm using the SG-HMM based speech and noise models. By producing a noise PSD estimate, the proposed scheme can be combined with any short-time spectral attenuation (STSA) based speech estimation algorithms for generating the enhanced speech. Let $\mathcal{Y}_n[k]$ denote the k'th spectrum band of the noisy signal frame in the discrete Fourier domain. Then the STSA approach processes the noisy speech by applying a multiplicative attenuation factor $H_n[k]$ to $\mathcal{Y}_n[k]$.

The proposed noise PSD estimate is based on a recursive averaging of the instantaneous estimate of the noise sample spectrum (periodogram). We denote the speech and noise power spectra associated to each composite state by $\bar{\lambda}_{s_n}[k]$ and $\ddot{\lambda}_{s_n}[k]$, where

$$\bar{\lambda}_{s_n}[k] = \frac{\hat{\bar{g}}_{s_n}}{\left|\sum_{j=0}^{\bar{p}} \bar{\alpha}_{\bar{s}}[j] e^{-2\pi jk/K}\right|^2} \quad (27)$$

$$\ddot{\lambda}_{s_n}[k] = \frac{\hat{\ddot{g}}_{s_n}}{\left|\sum_{j=0}^{\ddot{p}} \hat{\ddot{\alpha}}_{\ddot{s}_n}[j] e^{-2\pi jk/K}\right|^2}. \quad (28)$$

The minimum mean square error (MMSE) estimate of the noise sample spectrum is given by [9]

$$\mathrm{E}\left[|W_n[k]|^2|\mathbf{Y}_0^n = \mathbf{y}_0^n\right] = \frac{1}{\Omega_n}\sum_s \omega_n(s) \cdot \left(|(1 - H_{s_n}[k])\mathcal{Y}_n[k]|^2 + H_{s_n}[k]\ddot{\lambda}_{s_n}[k]\right), \quad (29)$$

where $H_{s_n}[k]$ is the attenuation factor of the Wiener filter for state s_n,

$$H_{s_n}[k] = \frac{\bar{\lambda}_{s_n}[k]}{\bar{\lambda}_{s_n}[k] + \ddot{\lambda}_{s_n}[k]}. \quad (30)$$

We apply a recursive averaging of the noise sample spectrum estimate to obtain an estimate of the noise PSD, denoted by $\hat{\ddot{\lambda}}_n[k]$,

$$\hat{\ddot{\lambda}}_n[k] = \hat{\ddot{\lambda}}_{n-1}[k] + \xi \cdot \left(\mathrm{E}\left[|W_n[k]|^2|\mathbf{Y}_0^n = \mathbf{y}_0^n\right] - \hat{\ddot{\lambda}}_{n-1}[k]\right), \quad (31)$$

where ξ is a forgetting factor.

IV. Experiments and results

In this section, we discuss the implementational details of the proposed scheme. We then present the experimental setup and the results.

A. System Implementation

We implemented the system for 8 kHz sampled speech. The noisy signal was processed on a frame-by-frame basis with frames of 32 ms and an overlap of 16 ms. The signal frame was first windowed using the Hann window, and then converted to the frequency domain using the discrete Fourier transform (DFT). Both noise estimation and enhancement were performed in the frequency domain. The enhanced frame was converted to the time domain using the inverse DFT, and the speech was synthesized using the overlap-and-add technique.

The speech HMM had eight states and 16 mixture components per state, and the AR model was of order ten. Training of the speech HMM was performed using 640 utterances from the training set of the TIMIT database according to [5]. We set the number of states in the noise model to five, based on an objective experiment (section IV-C.2). We assumed broadband noise and the noise AR model was experimentally selected to be of order six. For a tonal noise, such as the siren noise, or a noise with many spectral peaks and/or deep valleys, a higher model order would be needed (e.g., as prior information). The forgetting factor ρ (section II-B.1) was experimentally set to .97. The constant for gain variance compensation ε (section II-B.2) was set to 0.001. To construct the noise model of the safety-net state, we used the minimum statistics method [3]. The AR model parameters were obtained from the estimated noise PSD by the inverse DFT followed by the Levinson-Durbin recursion, and the mean $\ddot{\mu}_{\ddot{s}}$ was set to the excitation variance. The variance $\ddot{\sigma}^2_{\ddot{s}}$ was set to a constant 0.0107, which corresponds to the gain variance of white Gaussian noise [5]. The safety-net state was reselected every N=300 frames (every 4.8 seconds). The forgetting factor ξ for the noise PSD estimate was experimentally set to 0.15.

Initialization of the noise model was performed using the first five noisy signal frames that were considered to be noise-only. Each frame was used for initializing the noise model of one particular state. The states and their transitions were assumed to be initially uniformly distributed.

B. Experimental setup

The experiments were performed using utterances from the core test set of the TIMIT database (resampled to 8kHz). The experiments were performed using additive noise of the following types: 1) computer generated white Gaussian noise; 2) traffic noise, recorded on the side of a busy freeway; 3) babble noise, from the Noisex-92 database; 4) white-2 noise, an amplitude modulated white Gaussian noise using a sinusoid function. The sinusoid had a period of two seconds, and the maximum amplitude was four times higher than the minimum amplitude. The amplitude modulation simulates the change of noise energy level, which characterizes many real-world non-stationary noise types. Both the traffic and white-2 noises are highly non-stationary noises with rapidly time-varying energy. The speech utterances were concatenated and added to the noise signals to generate the noisy speech signals. We consider low SNR scenarios ranging from zero to ten dB where non-stationary noise is more annoying. For all experiments, the noisy utterances were processed concatenated.

The main focus of this work is the on-line estimation of a SG-HMM-based noise model. The estimated noise model can be used for estimation of the speech waveform [5], or estimation of the noise PSD (section III). Therefore, objective experiments were conducted to evaluate: 1) modeling accuracy of the estimated noise model, 2) noise PSD estimation performance, 3) speech enhancement using the noise PSD estimate. In addition, we performed an experiment to demonstrate the behavior of the safety-net state strategy. The detailed experimental setup and the results are presented in sections IV-C to IV-F.

C. Evaluation of noise model accuracy

In this experiment, we measure the accuracy of an estimated noise model. From a statistical estimation perspective, accuracy of the assumed statistical model is crucial for the estimation performance. The model accuracy was measured using the log-likelihood (LL) score of the model evaluated on the *true* noise signals. For the n'th frame, the score was calculated using [5]

$$LL(\mathbf{w}_n) \quad = \quad \log\left(\frac{1}{\Omega_n}\sum_{s}\omega_n(s)f_{\ddot{s}}(\mathbf{w}_n|\hat{\ddot{g}}_n)\right), \qquad (32)$$

where $f_{\ddot{s}}(\mathbf{w}_n|\hat{\ddot{g}}_n)$ is the density function (5) evaluated using the estimated noise gain $\hat{\ddot{g}}_n$. Since the model parameters were estimated using the noisy signals only, the LL score measures how well the estimated model fits to the true noise signals. In this experiment, we used 16 utterances from the test set, one male and one female speaker from each of the eight dialects. The total length of the evaluation utterances was about one minute. The score was evaluated for every signal frame, and averaged over all utterances to obtain the mean value for the overall performance.

1) Reference systems: For evaluation of the noise model accuracy, we used the following reference methods: A) the SG-HMM based method with prior noise information [5], B) a simple noise AR model converted from a noise PSD estimate, C) an AR-HMM noise model estimated on noise-only frames [15], D) an AR-HMM noise model estimated using the recursive EM algorithm [18].

Reference method A is based on the same SG-HMM modeling of speech and noise as the proposed system. Following [5], nine states were used in the noise model. Reference method A additionally assumes that the type of the noise environment is known and the corresponding (off-line trained) noise model can be used. Also, its AR model order was optimized for each noise environment. For instance, the babble noise model in reference method A was set to ten, due to its spectral similarity to speech. Therefore, reference method A can be considered as the ideal solution, and it is expected that reference method A would have a better performance than the proposed method. Reference method B is based on an AR model without stochastic gain modeling, in order to create a reference scale for interpretation of the numerical LL scores. The AR parameters are obtained from the noise PSD estimate using minimum statistics [3] by applying the inverse DFT followed by Levinson-Durbin recursion. Both reference method C and D are based on AR-HMM modeling of speech and noise, with on-line estimation of the noise model. Reference method C estimates the noise model using noise-only frames, determined by a VAD. In the experiments, we used an ideal VAD obtained using the clean speech. For reference method D, the noise model is updated in all frames using the recursive EM algorithm.

2) Optimization of noise model size: First, we determine the number of states to be used in the noise model. For a fixed SNR level of 5 dB, the LL scores were evaluated for a different number of noise model states. The experimental results of the proposed system and reference systems A and B are shown in Fig. 3. The proposed system achieved better results with an increased number of noise model states. For white noise, the performance improved only insignificantly with more than one state. For other noise types, the LL improvement with two states was significant. For a complex noise such as the babble noise, it was beneficial to have more than two states. Based on the experimental results, we concluded that five noise model states were sufficient to achieve a good performance for all tested noise types.

3) Noise model accuracy: Using a noise model with five states, we evaluated the noise model accuracy of the proposed system for SNRs ranging from 0 to 25 dB. Reference methods A to D were evaluated for comparison. Again, we expected that reference method A would outperform the proposed

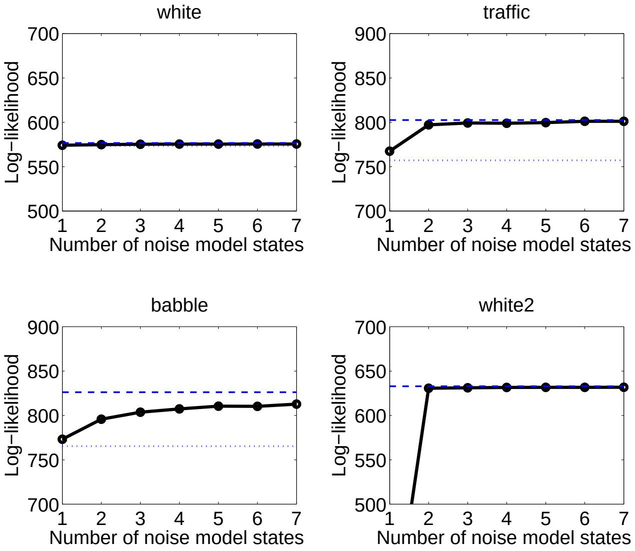

Fig. 3. LL of the estimated noise models versus number of noise model states for the proposed system (solid lines). The input SNR is 5 dB. Reference method A (with nine states) and reference method B are shown in dashed and dotted lines, respectively. For the white noise, all three methods perform similarly and the lines in the figure overlap. For the white-2 noise, reference method B has a likelihood score of 300, and is not shown in the figure for clarity.

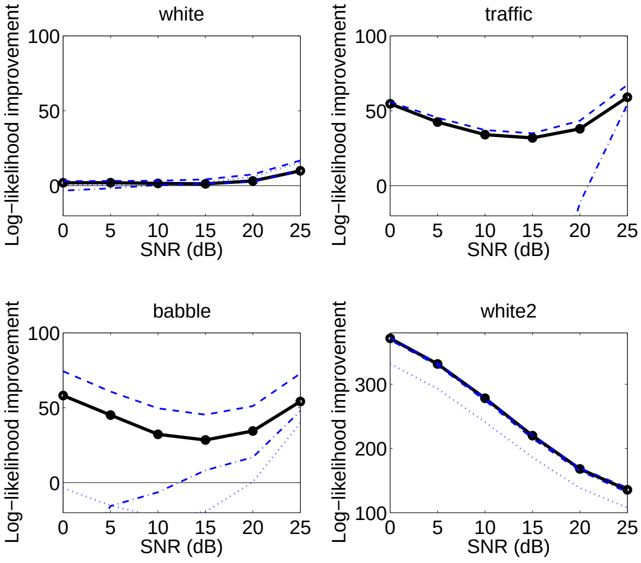

Fig. 4. LL improvement (over reference method B) of the estimated noise models for different SNR conditions. The solid lines with dots are from the proposed method, dashed lines from reference method A, dotted lines from reference method C and dash-dotted lines from reference method D. For the white noise, the results are close to each other and some lines overlap. For the traffic noise, the results of reference C are worse than reference B, and the dotted line falls outside the plot. For the white-2 noise, the results of the proposed system, reference A and reference D are close to each other and the lines overlap.

method, since additional a-priori information was used. Compared to reference method B, the proposed method consistently produced better LL scores. For clarity of presentation, the LL improvements over reference method B as functions of input SNR are shown in Fig. 4. A score of zero indicates identical performance as reference B, and a positive score indicates an improved performance.

As expected, the proposed method performed slightly worse than reference method A. This is particularly true for babble noise. For the remaining noise types, the proposed method performed closely to reference method A. Compared to reference C and D, the proposed method performed significantly better for the traffic noise and babble noise cases. We believe that the improvement was due to the SG-HMM modeling of both speech and noise, as the time-varying energy was explicitly modeled. The improvement over reference C was additionally due to the continuous learning (reference D was performing better than reference C). For the artificially generated white and white-2 noises, both reference C and D performed well or reasonably well, likely due to the less complex spectral shapes. We conclude that the proposed method achieves more accurate noise models than reference C and D in complex noise types such as babble and traffic noises.

D. *Evaluation of noise PSD estimate*

In this experiment, we evaluate the noise PSD estimation performance. The estimation performance was measured as the log spectral distance (LSD) between the ideal noise PSD estimate, denoted by $\lambda_n[k]$, and the noise PSD estimated by the test methods. We used the smoothed periodogram of the noise signal as the ideal noise PSD estimate. The smoothing was obtained using a rectangular window of nine frames (160 ms in total). The LSD measure of the n'th frame is given by

$$LSD_n = \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(10\log_{10}\hat{\hat{\lambda}}_n[k] - 10\log_{10}\lambda_n[k]\right)^2}. \quad (33)$$

We used the same test set of 16 utterances as in Sec. IV-C. For each frame, the LSD score was evaluated, and the overall score was obtained by averaging over all frames. The 95% confidence interval was also evaluated.

Two reference method were used in this experiment: A) SG-HMM based method with prior noise information [5] (as reference A in the previous experiment) adapted to use the noise PSD estimator (Sec. III), and E) the minimum statistics method [3]. Again, reference A makes use of additional prior information and is expected to perform better than the proposed method.

The LSD scores and 95% confidence intervals for various SNR levels are given in TABLE I. The results show that the proposed method is slightly worse than reference A for white, traffic and babble noises. For white-2 noise, the proposed method has a minor advantage, likely to due to the more flexible gain model. Compared to reference E, the proposed method achieves a consistently lower LSD level.

Fig. 5 demonstrates the estimated noise PSD for a fixed frequency bin using the ideal estimator, the proposed estimator

	white	traffic	babble	white-2
		0 dB SNR		
Prop.	1.36±0.01	2.38±0.02	3.35±0.02	1.95±0.02
Ref.A	1.35±0.01	2.28±0.02	3.02±0.02	2.09±0.02
Ref.E	1.86±0.01	3.55±0.05	4.26±0.03	6.47±0.09
		5 dB SNR		
Prop.	1.45±0.01	2.72±0.03	3.55±0.03	2.09±0.02
Ref.A	1.40±0.01	2.51±0.03	3.17±0.02	2.28±0.03
Ref.E	1.96±0.01	3.57±0.04	4.20±0.03	6.37±0.09
		10 dB SNR		
Prop.	1.63±0.02	3.37±0.05	3.73±0.03	2.38±0.04
Ref.A	1.45±0.01	2.74±0.03	3.33±0.02	2.67±0.04
Ref.E	2.04±0.01	3.69±0.04	4.25±0.03	6.23±0.08

TABLE I

COMPARISON OF LOG SPECTRAL DISTANCE (LSD) MEASURE WITH THE 95% CONFIDENCE INTERVAL FOR THE PROPOSED SYSTEM (PROP.), NOISE PSD ESTIMATE BASED ON SG-HMM WITH PRIOR NOISE INFORMATION (REF.A), AND THE MINIMUM STATISTICS METHOD (REF.E). THE LSD IS MEASURED USING AN IDEAL NOISE PSD ESTIMATE.

and the minimum statistics estimator. Clearly, the proposed method obtains a closer match to the ideal solution than the minimum statistics method. The improvement is significant when a rapid increase in noise energy level occurs, e.g., for the traffic noise (at 1-2 seconds) and white-2 noise. In such cases, the proposed method yields a more accurate noise PSD estimate and is expected to produce a lower level of residual noise when integrated in a speech enhancement system.

E. Evaluation of speech enhancement performance

To demonstrate the advantage of our noise estimation algorithm for speech enhancement, the proposed noise PSD estimate was integrated into a speech enhancement system based on the Ephraim-Malah MMSE short-time spectral amplitude estimator [23]. The reference systems for evaluation were based on the same speech estimator with different noise estimation algorithms. The two reference noise estimation methods from Sec. IV-D were used. The speech enhancement system was not specifically tuned for any of the noise estimation algorithms, and the tuning parameters were set according to [23].

Fig. 6 illustrates an example utterance processed by the proposed system and the reference system 2 (using minimum statistics). In the example, the enhanced speech using the proposed system has a lower level of residual noise. The difference is largest during the end of the utterance, when a rapid increase in noise energy level occurs. We note that both systems are based on the Ephraim-Malah speech estimator with the same tuning parameters, and the reduced residual noise level was only due to the more accurate noise PSD estimation.

We evaluated the speech enhancement performance through objective evaluations and informal listenings. According to recent studies, e.g., [24], most commonly used objective evaluation methods for the perceptual quality of a speech enhancement system perform poorly. Therefore, we used the segmental signal-to-noise ratio (SSNR) [25] that only measures the waveform similarity. We believe that using an improved noise estimation algorithm, the waveform of the enhanced speech will be more similar to the original speech. Since the tested speech enhancement system is not jointly tuned with the noise estimation algorithms, all tested methods will benefit from additional perceptual tuning of the enhancement system.

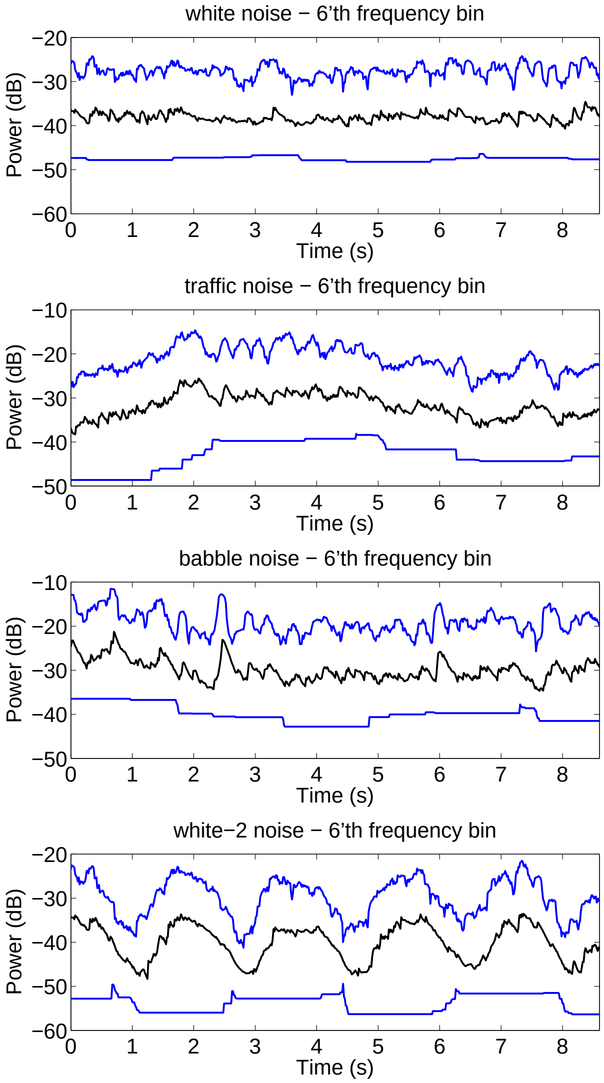

Fig. 5. Examples of noise PSD estimation for 0 dB SNR in the 6th frequency bin. For each sub-plot, the line on the top is from the ideal noise estimate, the line in the middle is from the proposed noise estimate (with 10 dB offset), and the line on the bottom is from the minimum statistics method (with 20 dB offset). The ideal noise estimate was obtained by convolving the true noise periodograms with a rectangular window of nine frames.

The complete core test set (192 utterances) of the TIMIT database was used in this experiment. The SSNR measure was evaluated for each utterance on frames of 16 ms. Frames with energy 40 dB below the average energy of the utterance were excluded from the computation. The final score was obtained by averaging the scores from the utterances. The first utterance

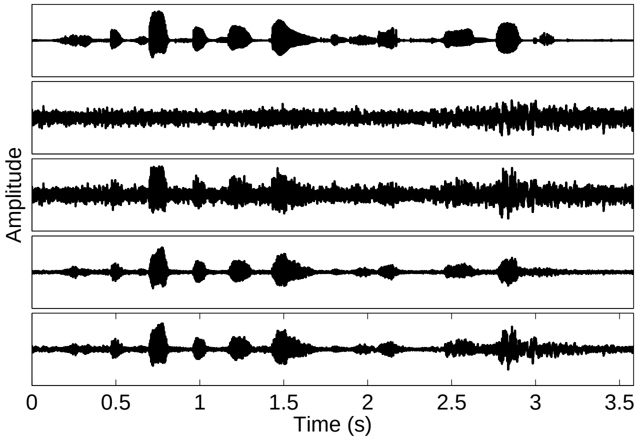

Fig. 6. An example of speech enhancement using the proposed noise PSD estimator and the minimum statistics estimator. The utterance is ”She said sharks have no bones and shrimp swam backward” from the TIMIT database corrupted by traffic noise in 0-dB SNR. The five sub-plots demonstrate: 1) clean speech 2) traffic noise 3) noisy speech 4) enhanced speech using the proposed noise PSD estimate 5) enhanced speech using the minimum statistics noise PSD estimate.

was removed from the averaging to avoid biased results due to initializations.

TABLE II summarizes the experimental results. The system using the proposed noise PSD estimate achieved a comparable performance as reference A for white, traffic and white-2 noises. For the babble noise, reference A has a clear advantage due to a more accurate noise model (as seen in Sec. IV-C). Compared to reference E, the proposed system achieved consistently better SSNR results under all test conditions. The improvement is more significant for non-stationary noise types. From informal listenings, we conclude that the system using the proposed noise estimate achieves a lower level of residual noise than reference E without introducing additional artifacts.

F. *Evaluation of safety-net state strategy*

In this experiment, we demonstrate the need for and the behavior of the safety-net state strategy. The experiment was conducted for two test scenarios. Both scenarios consisted of two artificial noises generated using white Gaussian noise filtered by FIR filters, one low-pass filter with coefficients [.5 .5] and one high-pass filter with coefficients [.5 -.5]. The two noise sources are alternated every 10 seconds (scenario one) and 1 second (scenario two). The first scenario simulates a change of noise environment, e.g., entering a cafe from a street. Such a change does not occur often, but the noise characteristics change abruptly. The second scenario simulates a non-stationary noise environment that contains multiple noise sources with different noise characteristics.

The proposed noise estimation was tested with and without the safety-net state. Three states were used in the noise models. Reference method B (section IV-C.3) was also evaluated for comparison. The LL score (32) was used as the performance measure. Instead of evaluating the averaged score, we compare the LL scores on a frame-by-frame basis. Fig. 7 and 8 demonstrate the experimental results.

Type	white	traffic	babble	white-2
		0 dB SNR		
Noisy	-9.51±0.26	-8.27±0.43	-8.76±0.29	-7.89±0.28
Prop.	0.03±0.20	0.62±0.30	-1.96±0.25	0.80±0.20
Ref.A	0.05±0.20	0.42±0.34	-1.34±0.24	0.71±0.20
Ref.E	-0.30±0.20	-1.40±0.41	-3.36±0.28	-4.11±0.27
		5 dB SNR		
Noisy	-4.51±0.26	-3.27±0.43	-3.76±0.29	-2.89±0.28
Prop.	2.99±0.20	3.63±0.29	1.50±0.23	3.75±0.21
Ref.A	3.04±0.19	3.61±0.32	2.00±0.23	3.67±0.20
Ref.E	2.72±0.20	2.35±0.38	0.65±0.26	0.07±0.26
		10 dB SNR		
Noisy	0.49±0.26	1.73±0.43	1.25±0.29	2.11±0.28
Prop.	5.87±0.21	6.69±0.30	4.92±0.23	6.62±0.22
Ref.A	6.00±0.20	6.89±0.32	5.50±0.23	6.59±0.21
Ref.E	5.78±0.20	6.07±0.36	4.60±0.25	4.19±0.25

TABLE II
EXPERIMENTAL RESULTS OF THE SEGMENTAL SNR DISTORTION MEASURE USING THE PROPOSED METHOD, NOISE PSD ESTIMATE BASED ON SG-HMM WITH PRIOR NOISE INFORMATION (REF.A), AND THE MINIMUM STATISTICS METHOD (REF.E).

For test scenario one (Fig. 7), all three methods achieved good performance during the first ten seconds. When the noise characteristics changed at the 10’th second, all methods performed poorly. Without using the safety-net state strategy, the algorithm was caught in a local optimum, and had poor performance during the following ten seconds. Using the safety-net state strategy, the safety-net state (the third state) was recovered after a short delay. The first state becomes the safety net state just before the 15’th second. The third state was therefore re-initialized and useful again for continuous learning. The behavior repeated near 20’th second, when the first noise state had also been re-initialized.

The test scenario two (Fig. 8) simulates a non-stationary noise environment with two alternating noise sources. The safety-net state was initially set to the third state, and switched to the first state near the 5’th second. The re-initialization allowed the third state to successfully converge to the correct model of the second noise source. After the 10’th second, the two noise characteristics were properly learned in two noise states, as demonstrated in the opposing behavior in LL scores of these states. Therefore, the overall model performed well on both noise sources. After the initial re-location, the safety-net state remained in the first state and had little effect on the overall performance. We conclude that the proposed scheme is inherently capable of learning such a dynamic noise environment through multiple noise states and stochastic gain models, and that the safety-net state strategy facilitates robust model re-initialization and helps prevent local convergence towards a non-optimal noise model.

A real-world noisy environment is often more dynamic and

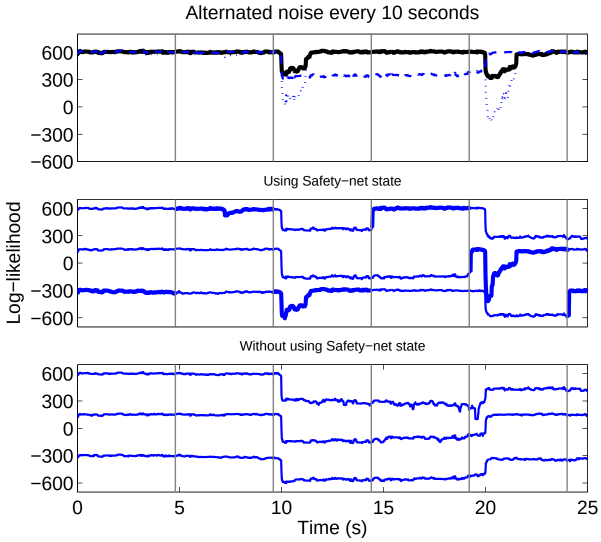

Fig. 7. LL scores of the estimated noise models on a frame-by-frame basis. The sub-plot on the top shows the LL scores using the safety-net state (solid line), without safety-net state (dashed line) and reference method B (dotted line). The sub-plots in the middle and on the bottom show the individual LL scores of each noise state with 450 offset. In the middle sub-plot, the LL results of the safety-net state is highlighted using bold lines. The vertical lines indicate the frames when reselection of the safety-net state occurs. For clarity of presentation, the scores were filtered by a 15 sample median filter.

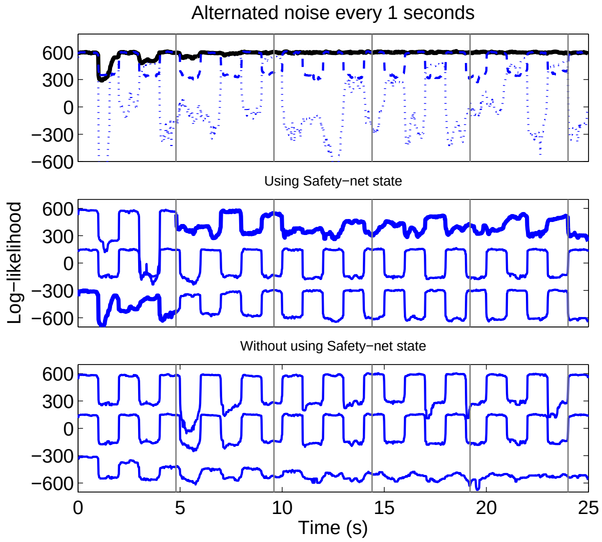

Fig. 8. LL scores of the estimated noise models on a frame-by-frame basis. The sub-plot on the top shows the LL scores using the safety-net state (solid line), without safety-net state (dashed line) and reference method B (dotted line). The sub-plots in the middle and on the bottom show the individual LL scores of each noise state with 450 offset. In the middle sub-plot, the LL results of the safety-net state is highlighted using bold lines. The vertical lines indicate the frames when reselection of the safety-net state occurs. For clarity of presentation, the scores were filtered by a 15 sample median filter.

complex than the test scenarios considered in this experiment. However, we believe that the safety-net state strategy helps improving the robustness of the EM based learning algorithm in situations where noise of a specific character re-occurs. Good examples of such situations are mobile-telephone and hearing-aid environments, with the user moving between environments. While the method can not improve over existing systems when the noise signal character changes continuously (without re-occurrences), such situations occur rarely in reality.

V. Conclusions

We have proposed an on-line noise estimation algorithm using the stochastic-gain hidden Markov modeling (SG-HMM) of speech and noise. The model parameters of the noise model are estimated on-line using the recursive EM algorithm. The strength of the proposed algorithm in a non-stationary noise environment is in 1) continuous adaptation to change of spectral shapes; and 2) continuous adaptation to the change of noise energy level. We have also proposed a noise PSD estimate using the SG-HMM framework. We showed through objective evaluations that the proposed scheme achieves a more accurate noise model and PSD estimate, particularly for non-stationary noise sources with rapidly-changing energy level. Integrated to a speech enhancement system, the proposed scheme facilitates a lower level of residual noise.

Appendix I
Derivation of (15-16)

In this appendix, detailed derivations of (15-16) are given. We start by expansion and simplification of the auxiliary $\mathcal{Q}_n(\cdot)$ function (10). The derivation follows the derivations in, e.g., [26]. The logarithmic term of (10) can be written as

$$\begin{aligned}\log f(\mathbf{z}_0^n, \mathbf{y}_0^n|\theta) &= \log f(\mathbf{y}_0^n|\mathbf{z}_0^n, \theta) + \log f(\mathbf{z}_0^n|\theta)\\ &= \log f(\mathbf{y}_0^n|s_0^n, \ddot{g}_0^n, \bar{g}_0^n, \mathbf{x}_0^n, \theta) + \log f(s_0^n|\theta)\\ &+ \log f(\ddot{g}_0^n|\ddot{s}_0^n, \theta) + \log f(\bar{g}_0^n|\bar{s}_0^n) + \log f(\mathbf{x}_0^n|\bar{s}_0^n, \bar{g}_0^n) \quad (34)\end{aligned}$$

The last two terms are independent of θ, and are neglected in the following derivations. Using the conditional independency property of HMM[2], the remaining terms can be expanded, and we get

$$\log f(\mathbf{y}_0^n|s_0^n, \ddot{g}_0^n, \bar{g}_0^n, \mathbf{x}_0^n, \theta) = \sum_{t=0}^{n} \log f_{s_t}(\mathbf{y}_t|\ddot{g}_t, \bar{g}_t, \mathbf{x}_t, \theta), \quad (35)$$

$$\log f(s_0^n|\theta) = \sum_{t=0}^{n} \log\left(\bar{a}_{\bar{s}_{t-1}\bar{s}_t} \ddot{a}_{\ddot{s}_{t-1}\ddot{s}_t}\right), \quad (36)$$

$$\log f(\ddot{g}_0^n|\ddot{s}_0^n, \theta) = \sum_{t=0}^{n} \log f_{\ddot{s}_t}(\ddot{g}_t|\theta). \quad (37)$$

The integral of (10) over the hidden variables can then be simplified by separating each time index t and integrate over

[2] The conditional independency property of HMM refers to [26]: 1) state variable s_n given s_{n-1} is independent of previous state and observation variables; 2) the n'th observation, $\mathbf{w}_n$, given state s_n, is independent of other state and observation variables.

each hidden variable. The $\mathcal{Q}_n(\cdot)$ function can be rewritten as

$$\begin{aligned}\mathcal{Q}_n(\theta|\hat{\theta}_0^{n-1}) \;=\; \sum_{t=0}^{n}\Bigg[&\sum_{s_t}\iiint f(s_t,\ddot{g}_t,\bar{g}_t,\mathbf{x}_t|\mathbf{y}_0^n,\hat{\theta}_0^{n-1})\\ &\Big(\log f_{s_t}(\mathbf{y}_t|\ddot{g}_t,\bar{g}_t,\mathbf{x}_t,\theta)+\log f_{\ddot{s}_t}(\ddot{g}_t|\theta)\Big)\,d\ddot{g}_t d\bar{g}_t d\mathbf{x}_t+\sum_{s_{t-1}}\\ &\sum_{s_t}\iint f(s_{t-1},s_t,\ddot{g}_t,\bar{g}_t|\mathbf{y}_0^n,\hat{\theta}_0^{n-1})\log\ddot{a}_{\ddot{s}_{t-1}\ddot{s}_t}d\ddot{g}_t d\bar{g}_t\Bigg].\end{aligned}\tag{38}$$

Implementation of $f(s_t,\ddot{g}_t,\bar{g}_t,\mathbf{x}_t|\mathbf{y}_0^n,\hat{\theta}_0^{n-1})$ requires backward probability calculations from frame $n-1$ to t, for each t. This leads to additional computational complexity and storage, which makes it impractical in the on-line learning algorithm. To facilitate low complexity and low memory implementation, we neglect contributions from future observations with respect to each time index t, and therefore avoid the backward probability calculations. Although the approximation was shown to have a small negative impact on the performance [12], it significantly simplifies the practical implementation. Thus, we approximate $f(s_t,\ddot{g}_t,\bar{g}_t,\mathbf{x}_t|\mathbf{y}_0^n,\hat{\theta}_0^{n-1})$ of (38) by

$$\begin{aligned}f(s_t,\ddot{g}_t,\bar{g}_t,\mathbf{x}_t|\mathbf{y}_0^n,\hat{\theta}_0^{n-1}) \;&\approx\; f(s_t,\ddot{g}_t,\bar{g}_t,\mathbf{x}_t|\mathbf{y}_0^t,\hat{\theta}_0^{t-1})\\ &= \frac{\gamma_t(s_t)f_{s_t}(\ddot{g}_t,\bar{g}_t,\mathbf{y}_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})f_{s_t}(\mathbf{x}_t|\ddot{g}_t,\bar{g}_t,\mathbf{y}_0^t,\hat{\theta}_0^{t-1})}{f(\mathbf{y}_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})}\\ &= \frac{\gamma_t(s_t)f_{s_t}(\ddot{g}_t,\bar{g}_t,\mathbf{y}_t|\hat{\theta}_{t-1})f_{s_t}(\mathbf{x}_t|\ddot{g}_t,\bar{g}_t,\mathbf{y}_t,\hat{\theta}_{t-1})}{f(\mathbf{y}_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})},\end{aligned}\tag{39}$$

where $\gamma_t(s_t)$ is the probability of being in the composite state s_t given all past noisy observations up to frame $t-1$,

$$\begin{aligned}\gamma_t(s_t) \;&=\; f(s_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})\\ &=\; \sum_{s_{t-1}} f(s_{t-1}|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})f(s_t|s_{t-1},\hat{\theta}_{t-1}),\end{aligned}\tag{40}$$

in which $f(s_{t-1}|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})$ is the forward probability at frame $t-1$, which can be obtained using the forward algorithm.

Applying the same approximation (39) to $f(s_{t-1},s_t,\ddot{g}_t,\bar{g}_t|\mathbf{y}_0^n,\hat{\theta}_0^{n-1})$ of (38), we get

$$\begin{aligned}f(s_{t-1},s_t,\ddot{g}_t,\bar{g}_t|\mathbf{y}_0^n,\hat{\theta}_0^{n-1}) \;&\approx\; f(s_{t-1},s_t,\ddot{g}_t,\bar{g}_t|\mathbf{y}_0^t,\hat{\theta}_0^{t-1})\\ =&\frac{f(s_{t-1}|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})f(s_t|s_{t-1},\hat{\theta}_{t-1})f_{s_t}(\ddot{g}_t,\bar{g}_t,\mathbf{y}_t|\hat{\theta}_{t-1})}{f(\mathbf{y}_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})}.\end{aligned}\tag{41}$$

Now we apply an additional approximation to simplify the integral terms over the gain variables. We assume [7]:

$$f_{s_t}(\ddot{g}_t,\bar{g}_t,\mathbf{y}_t) \;\approx\; f_{s_t}(\ddot{g}_t,\bar{g}_t,\mathbf{y}_t)\delta(\bar{g}_t-\hat{\bar{g}}_{s_t})\delta(\ddot{g}_t-\hat{\ddot{g}}_{s_t}),\tag{42}$$

where $\delta(\cdot)$ denotes the Dirac delta function. The approximation can be motivated by the observation that $f_{s_t}(\ddot{g}_t,\bar{g}_t,\mathbf{y}_t)$ typically decades rapidly from its peak. Using the approximation, the integral terms of (10) become simple to evaluate. A practical solution for obtaining $\{\hat{\bar{g}}_{s_t},\hat{\ddot{g}}_{s_t}\}$ is given in [5, appendix I]. The approximation applies both to the nominators and denominators of (39) and (41), to ensure that (39) and (41) integrate to one over the remaining variables. The denominators of (39) and (41) are approximated by

$$\begin{aligned}f(\mathbf{y}_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1}) \;&\approx\; \sum_{s_t} f(s_t,\hat{\ddot{g}}_{s_t},\hat{\bar{g}}_{s_t},\mathbf{y}_t|\mathbf{y}_0^{t-1},\hat{\theta}_0^{t-1})\\ &=\; \sum_{s}\omega_t(s)=\Omega_t\end{aligned}\tag{43}$$

Combining (39, 41, 42, 43) with the auxiliary $\mathcal{Q}_n(\cdot)$ function (10), (15-16) can be obtained.

Acknowledgement

The authors wish to thank the anonymous reviewers for their valuable comments that helped in significantly improving the paper.

References

[1] "Enhanced Variable Rate Codec, speech service option 3 for wideband spread spectrum digital systems," TIA/EIA/IS-127, Jul. 1996.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 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 and Signal Processing*, vol. 3, June 2000, pp. 1875–1878.

[5] D. Y. Zhao and W. B. Kleijn, "HMM-based gain-modeling for enhancement of speech in noise," in *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, Mar. 2007, pp. 882–892.

[6] 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 and Signal Processing*, vol. 1, May 2001, pp. 669–672.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 1077–1080.

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

[9] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[10] Y. Zhao, "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 255–266, May 2000.

[11] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Statist. Soc. B*, vol. 46, no. 2, pp. 257–267, 1984.

[12] V. Krishnamurthy and J. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Processing*, vol. 41, no. 8, pp. 2557–2573, Aug. 1993.

[13] B.-H. Juang, "On hidden Markov model and dynamic time warping for speech recognition - a unified view," *Bell System Technical Journal*, vol. 63, no. 7, pp. 1213–1244, Sep. 1984.

[14] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, pp. 1–14, Jan. 2006.

[15] B. Logan and T. Robinson, "Adaptive model-based speech enhancement," *Speech Communication*, vol. 34, no. 4, pp. 351–368, Jul. 2001.

[16] D. Y. Zhao and W. B. Kleijn, "On noise gain estimation for HMM-based speech enhancement," in *Proc. Interspeech*, Sep. 2005, pp. 2113–2116.

[17] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[18] Y. Zhao, S. Wang, and K.-C. Yen, "Recursive estimation of time-varying environments for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, May 2001, pp. 225–228.

[19] J. Hansen and M. Clements, “Constrained iterative speech enhancement with application to speech recognition,” *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 795–805, April 1991.

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

[21] Y. Ephraim, “Gain-adapted hidden Markov models for recognition of clean and noisy speech,” *IEEE Trans. Signal Processing*, vol. 40, no. 6, pp. 1303–1316, Jun. 1992.

[22] B. L. McKinley and G. H. Whipple, “Noise model adaptation in model based speech enhancement,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, 1996, pp. 633 – 636.

[23] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[24] Y. Hu and P. Loizou, “Evaluation of objective measures for speech enhancement,” in *Proceedings of INTERSPEECH-2006*, 2006, pp. 1447–1450.

[25] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Prentice Hall, 1984.

[26] J. Bilmes, “A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” Technical Report, University of Berkeley, ICSI-TR-97-021, 1997.