HMM-based Gain-Modeling for Enhancement of Speech in Noise

David Y. Zhao, *Student Member, IEEE,* and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*— Accurate modeling and estimation of speech and noise gains facilitate good performance of speech enhancement methods using data-driven prior models. In this work, we propose a hidden Markov model (HMM) based speech enhancement method using explicit gain modeling. Through the introduction of stochastic gain variables, energy variation in both speech and noise is explicitly modeled in a unified framework. The speech gain models the energy variations of the speech phones, typically due to differences in pronunciation and/or different vocalizations of individual speakers. The noise gain helps to improve the tracking of the time-varying energy of non-stationary noise. The expectation-maximization (EM) algorithm is used to perform off-line estimation of the time-invariant model parameters. The time-varying model parameters are estimated on-line using the recursive EM algorithm. The proposed gain modeling techniques are applied to a novel Bayesian speech estimator, and the performance of the proposed enhancement method is evaluated through objective and subjective tests. The experimental results confirm the advantage of explicit gain modeling, particularly for non-stationary noise sources.**

***Index Terms*— noise suppression, speech enhancement, HMM, gain modeling**

I. Introduction

The estimation of clean speech from a noisy observation is of interest for a wide range of applications, such as speech communication, speech recognition, speaker identification and hearing aids. Motivated by the stochastic nature of acoustic background noises, the separation of speech from noise is often based on probabilistic models, e.g., [1]. Accurate modeling of the speech and noise statistics is, therefore, of great importance. Traditional approaches such as spectral subtraction, e.g., [2], [3], rely on a separate noise estimation algorithm to obtain the noise statistics. The statistics of speech is typically assumed to have Gaussian, or supergaussian [4] distributions. State-of-the-art noise estimation algorithms use, e.g., a voice activity detector (VAD) [3], recursive averaging with soft VAD [5] and order statistics [6], [7]. While the methods perform reasonably well for stationary noise, their performance under non-stationary noise conditions is unsatisfactory. In particular, they share the common limitation that they react slowly to changes in the noise energy, which are typical in non-stationary noise environments.

Methods based on data-driven prior modeling of speech [1], [8]–[12] and noise [13]–[17] have provided the basis of a framework that can significantly improve the performance in non-stationary noise environments. These prior models are based on generic models trained using recorded signals. Again, the performance depends on the accuracy of the obtained prior models.

The hidden Markov model (HMM) has been applied successfully to speech and noise modeling for applications such as robust speech recognition [8] and enhancement [1], [13]. For an auto-regressive (AR) HMM, e.g., [1], the signal is modeled as an AR process for a given state. The states are connected through transition probabilities of a Markov chain. While such HMMs can model the change of spectral characteristics in speech, they do not explicitly model the different speech energy levels of a phone, typically due to differences in pronunciation and/or different vocalizations of individual speakers. Phones of similar characteristics, e.g., those corresponding to the same phoneme, typically vary significantly in energy in different utterances, although the long-term average of energy is kept constant [18]. A similar problem appears in noise modeling, as a result of changes in the noise environment, movements of the noise source, etc.

In this paper, we propose a unified solution to the aforementioned problems using an explicit parameterization and modeling of speech and noise gains that is incorporated in the HMM framework. The speech and noise gains are defined as stochastic variables modeling the energy levels of speech and noise, respectively. The separation of speech and noise gains facilitates incorporation of prior knowledge of these entities. For instance, the speech gain is assumed to have distributions that depend on the HMM states. Thus, the model facilitates that a voiced sound typically has a larger gain than an unvoiced sound. The dependency of gain and spectral shape (parameterized in the AR coefficients) is then implicitly modeled, as they are tied to the same state.

Time-invariant parameters of the speech and noise gain models are obtained off-line using training data, together with the remainder of the HMM parameters. The time-varying parameters are estimated in an on-line fashion using the observed noisy speech signal. That is, the parameters are updated recursively for each observed block of the noisy speech signal. We propose solutions to the parameter estimation problems, based on the regular and recursive expectation maximization (EM) frameworks [19], [20]. The proposed HMMs with explicit gain models are applied to a Bayesian speech estimator, and the basic system structure is shown in Fig. 1.

The proposed speech HMM generalizes the AR HMM based method [1], and the gain-adaptive HMM based method [1], [8]. In [1], the speech gain is implicitly modeled as a constant of the state-dependent AR models. Thus, the variation of the speech gain within a state is not considered. In [8], the speech

The authors 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).

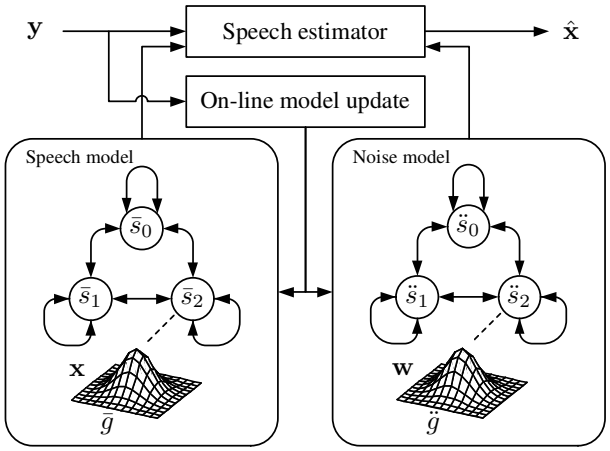

Fig. 1. Schematic diagram of the proposed speech enhancement system using the HMM-based gain modeling technique. $\mathbf{x}$, $\mathbf{w}$ and $\mathbf{y}$ denote speech, noise and noisy speech signals, respectively. s denotes an HMM state and g denotes a gain variable. The overbar $\bar{}$ is used for the variables in the speech model, and double dots $\ddot{}$ for the noise model.

gain (referred to as the gain contour), is estimated on-line using the noisy observation optimizing the maximum likelihood (ML) criterion. Hence, the method implicitly assumed a uniform prior of the gain in a Bayesian framework. The subjective quality of the gain-adaptive HMM method was shown to be inferior to the AR-HMM method [1], partly due to the uniform gain modeling. In our work, stronger prior gain knowledge is introduced to the HMM framework using state-dependent gain distributions.

As for the HMM-based noise gain modeling and estimation, little work has been reported. A heuristic noise gain adaptation using VAD has been proposed in [13], where the adaptation is performed only in speech pauses longer than 100 ms. In our recent work [21], continuous noise gain model estimation techniques were proposed. Herein, the framework is extended to modeling of both speech and noise gains in a unified framework.

The main contribution of this work is the new HMM based gain-modeling technique which is shown to improve the modeling of the non-stationarity of speech and noise. An off-line training algorithm is proposed based on the EM technique. For time-varying parameters, an on-line estimation algorithm is proposed based on the recursive EM technique. Moreover, the superior performance of the explicit gain modeling is demonstrated in the speech enhancement application, where the proposed speech and noise models are applied to a novel Bayesian speech estimator.

II. Signal model

We consider the estimation of the clean speech signal from speech contaminated by independent additive noise. The signal is processed in blocks of K samples, within which we can assume the stationarity of the speech and noise. The n'th noisy speech signal block is modeled as

$$\mathbf{Y}_n = \mathbf{X}_n + \mathbf{W}_n, \tag{1}$$

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.

The statistical modeling of speech $\mathbf{X}$ and noise $\mathbf{W}$ with explicit speech and noise gain models is discussed in section II-A and II-B. The modeling of the noisy speech signal $\mathbf{Y}$ is discussed in section II-C.

A. Speech model

We describe the statistics of the speech using an HMM with state-dependent gain models. We use overbar $\bar{}$ to denote the parameters of the speech HMM. Let $\mathbf{x}_0^{N-1} = \{\mathbf{x}_0, \ldots, \mathbf{x}_{N-1}\}$ denote the sequence of the speech block realizations from 0 to $N-1$, the probability density function (PDF) of $\mathbf{x}_0^{N-1}$ is then modeled as

$$f(\mathbf{x}_0^{N-1}) = \sum_{\bar{\mathbf{s}} \in \bar{\mathbf{S}}} \prod_{n=0}^{N-1} \bar{a}_{\bar{s}_{n-1}\bar{s}_n} f_{\bar{s}_n}(\mathbf{x}_n), \tag{2}$$

where the summation is over the set of all possible state sequences $\bar{\mathbf{S}}$, and for each realization of the state sequence $\bar{\mathbf{s}} = [\bar{s}_0, \bar{s}_1, \ldots, \bar{s}_{N-1}]$, $\bar{s}_n$ denotes the state of block n, $\bar{a}_{\bar{s}_{n-1}\bar{s}_n}$ denotes the transition probability from state $\bar{s}_{n-1}$ to $\bar{s}_n$ with $\bar{a}_{\bar{s}_{-1}\bar{s}_0}$ being the initial state probability. The probability density function of $\mathbf{x}_n$ for a given state $\bar{s}$ is the integral over all possible speech gains[1]. Modeling the speech energy in the logarithmic domain, we then have

$$f_{\bar{s}}(\mathbf{x}_n) = \int_{-\infty}^{\infty} f_{\bar{s}}(\bar{g}'_n) f_{\bar{s}}(\mathbf{x}_n|\bar{g}'_n) d\bar{g}'_n, \tag{3}$$

where $\bar{g}'_n = \log \bar{g}_n$ and $\bar{g}_n$ denotes the speech gain in the linear domain. The integral is formulated in the logarithmic domain for the convenient modeling of the non-negative gain. Since the mapping between $\bar{g}_n$ and $\bar{g}'_n$ is one-to-one, we use an appropriate notation based on the context below.

The extension over the traditional AR-HMM is the stochastic modeling of the speech gain $\bar{g}_n$, where $\bar{g}_n$ is considered as a stochastic process. The PDF of $\bar{g}_n$ is modeled using a state-dependent log-normal distribution, motivated by the simplicity of the Gaussian PDF and the appropriateness of the logarithmic scale for sound pressure level. In the logarithmic domain, we have

$$f_{\bar{s}}(\bar{g}'_n) = \frac{1}{\sqrt{2\pi\bar{\psi}_{\bar{s}}^2}} \exp\!\left(-\frac{1}{2\bar{\psi}_{\bar{s}}^2}\left(\bar{g}'_n - \bar{\phi}_{\bar{s}} - \bar{q}_n\right)^2\right), \tag{4}$$

with mean $\bar{\phi}_{\bar{s}} + \bar{q}_n$ and variance $\bar{\psi}_{\bar{s}}^2$. The time-varying parameter $\bar{q}_n$ denotes the *speech-gain bias*, which is a global parameter compensating for the overall energy level of an utterance, e.g., due to a change of physical location of the recording device. The parameters $\{\bar{\phi}_{\bar{s}}, \bar{\psi}_{\bar{s}}^2\}$ are modeled to be

[1]It is common to use mixture PDFs, such as Gaussian mixture models (GMM), as the state-dependent PDF. For clarity of the derivations, we assume only one component per state. The extension to mixture models is straightforward by considering the mixture components as sub-states of the HMM, and is applied in the experiments.

time-invariant, and can be obtained off-line using training data, together with the other speech HMM parameters.

For a given speech gain $\bar{g}_n$, the PDF $f_{\bar{s}}(\mathbf{x}_n|\bar{g}'_n)$ is considered to be a $\bar{p}$-th order zero-mean Gaussian AR density function, equivalent to white Gaussian noise filtered by the all-pole AR model filter. The density function is given by

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

where $|\cdot|$ denotes the determinant, $^{\sharp}$ denotes the Hermitian transpose and the covariance matrix $\bar{\mathbf{D}}_{\bar{s}} = (\mathbf{A}_{\bar{s}}^{\sharp}\mathbf{A}_{\bar{s}})^{-1}$, where $\mathbf{A}_{\bar{s}}$ is a $K \times K$ lower triangular Toeplitz matrix with the first $\bar{p}+1$ elements of the first column consisting of the AR coefficients including the leading one, $[1, \bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_{\bar{p}}]^T$.

B. Noise model

Elaborate noise models [13]–[15] are useful to capture the high diversity and variability of acoustical noise. In this work, similar HMMs are used for speech (cf. II-A) and noise. The model parameters for noise are denoted using double dots ¨ (instead of overbar ¯ for speech).

For simplicity, we assume further that a single noise gain model, $f_{\ddot{s}}(\ddot{g}'_n) = f(\ddot{g}'_n)$, is shared by all HMM noise states. The noise PDF for a given state $\ddot{s}$ is

$$f_{\ddot{s}}(\mathbf{w}_n) = \int_{-\infty}^{\infty} f(\ddot{g}'_n) f_{\ddot{s}}(\mathbf{w}_n|\ddot{g}'_n) d\ddot{g}'_n, \quad (6)$$

with the noise gain model given by

$$f(\ddot{g}'_n) = \frac{1}{\sqrt{2\pi\ddot{\psi}^2}}\exp\!\left(-\frac{1}{2\ddot{\psi}^2}\left(\ddot{g}'_n - \ddot{\phi}_n\right)^2\right). \quad (7)$$

i.e., with mean $\ddot{\phi}_n$ and variance $\ddot{\psi}^2$ being fixed for all noise states. The mean $\ddot{\phi}_n$ is considered to be a time-varying parameter to model the unknown noise energy, and is to be estimated on-line using the noisy observations. The variance $\ddot{\psi}^2$ and the remaining noise HMM parameters are considered to be time-invariant variables, which can be estimated off-line using recorded signals of the noise environment.

The simplified model implies that the noise gain and the noise shape, defined as the gain normalized noise spectrum, are considered independent. This assumption is valid mainly for continuous noise, where the energy variation can be generally modeled well by a global noise gain variable with time-varying statistics. The change of the noise gain is typically due to movement of the noise source or the recording device, which is assumed independent of the acoustics of the noise source itself. For intermittent or impulsive noise, the independent assumption is not valid. State-dependent gain models can then be applied to model the energy differences in different states of the sound. This scenario is not considered in this paper.

C. Noisy signal model

The PDF of the noisy speech signal can be derived based on the assumed models of speech and noise. Let us assume that the speech HMM contains $|\bar{S}|$ states and the noise HMM $|\ddot{S}|$ states. Then, the noisy model is an HMM with $|\bar{S}||\ddot{S}|$ states, where each composite state s consists of combination of the state $\bar{s}$ of the speech component and the state $\ddot{s}$ of the noise component. The transition probabilities of the composite states are obtained using the transition probabilities in the speech and noise HMMs.

The noisy PDF corresponding to state s is

$$f_s(\mathbf{y}_n) \;=\; \int\!\!\int f_s(\mathbf{y}_n, \bar{g}'_n, \ddot{g}'_n) d\bar{g}'_n d\ddot{g}'_n \quad (8)$$

$$\;=\; \int\!\!\int f_{\bar{s}}(\bar{g}'_n) f(\ddot{g}'_n) f_s(\mathbf{y}_n|\bar{g}'_n, \ddot{g}'_n) d\bar{g}'_n d\ddot{g}'_n, \quad (9)$$

where $f_s(\mathbf{y}_n|\bar{g}'_n, \ddot{g}'_n)$ is a Gaussian PDF with zero-mean and covariance matrix $\mathbf{D}_s$,

$$\mathbf{D}_s \;=\; \bar{g}_n\bar{\mathbf{D}}_{\bar{s}} + \ddot{g}_n\ddot{\mathbf{D}}_{\ddot{s}}. \quad (10)$$

The integral of (9) can be evaluated numerically, e.g., by stochastic integration. To facilitate real-time implementation, we approximate $f_s(\mathbf{y}_n, \bar{g}'_n, \ddot{g}'_n)$ by a scaled Dirac delta function [16]

$$f_s(\mathbf{y}_n, \bar{g}'_n, \ddot{g}'_n) \approx f_s(\mathbf{y}_n, \bar{g}'_n, \ddot{g}'_n)\delta(\bar{g}'_n - \hat{\bar{g}}'_n)\delta(\ddot{g}'_n - \hat{\ddot{g}}'_n), \quad (11)$$

where $\delta(\cdot)$ denotes the Dirac delta function and

$$\{\hat{\bar{g}}'_n, \hat{\ddot{g}}'_n\} \;=\; \arg\max_{\bar{g}'_n, \ddot{g}'_n} \log f_s(\mathbf{y}_n, \bar{g}'_n, \ddot{g}'_n). \quad (12)$$

The noisy PDF of state s, $f_s(\mathbf{y}_n)$, is then approximated to

$$f_s(\mathbf{y}_n) \approx f_s(\mathbf{y}_n, \hat{\bar{g}}'_n, \hat{\ddot{g}}'_n). \quad (13)$$

The approximation is valid if the only significant peak of the integrand in (8) is at $\{\hat{\bar{g}}'_n, \hat{\ddot{g}}'_n\}$, and the function decays rapidly from the peak. The behavior was confirmed through simulations. A more rigorous analysis of a similar approximation is provided in [16]. The procedure to obtain $\{\hat{\bar{g}}'_n, \hat{\ddot{g}}'_n\}$ is discussed in appendix I.

III. Speech estimation

We consider the enhancement of speech in noise by estimating speech from the observed noisy speech signal. Motivated by our previous work [21], we consider the Bayesian speech estimator based on a criterion that results in an adjustable level of residual noise in the enhanced speech. The speech is estimated as

$$\hat{\mathbf{x}}_n \;=\; \arg\min_{\tilde{\mathbf{x}}_n} \mathrm{E}[C(\mathbf{X}_n, \mathbf{W}_n, \tilde{\mathbf{x}}_n)|\mathbf{Y}_0^n = \mathbf{y}_0^n], \quad (14)$$

where $\mathrm{E}[\ ]$ denotes the expectation and the Bayes risk is defined for the cost function

$$C(\mathbf{x}_n, \mathbf{w}_n, \tilde{\mathbf{x}}_n) = ||(\mathbf{x}_n + \epsilon\mathbf{w}_n) - \tilde{\mathbf{x}}_n||^2, \quad (15)$$

where $||\cdot||$ denotes the vector norm and $0 \le \epsilon \ll 1$ defines the adjustable residual noise level. The cost function is the squared error for the estimated speech compared to the clean speech plus some residual noise. By explicitly leaving some level of residual noise, the criterion reduces the processing artifacts, which are commonly associated with traditional speech enhancement systems. When ϵ is set to zero, the estimator converges to the MMSE speech waveform estimator.

Using the Markov assumption, the posterior speech PDF given the noisy observations can be formulated as

$$f(\mathbf{x}_n|\mathbf{y}_0^n) \;=\; \frac{f(\mathbf{x}_n,\mathbf{y}_n|\mathbf{y}_0^{n-1})}{f(\mathbf{y}_n|\mathbf{y}_0^{n-1})} = \frac{\sum_s \gamma_n(s) f_s(\mathbf{x}_n,\mathbf{y}_n)}{f(\mathbf{y}_n|\mathbf{y}_0^{n-1})}, \quad (16)$$

where $\gamma_n(s)$ is the probability of being in the composite state s_n given all past noisy observations up to block $n-1$,

$$\gamma_n(s) \;=\; f(s_n|\mathbf{y}_0^{n-1}) = \sum_{s_{n-1}} f(s_{n-1}|\mathbf{y}_0^{n-1}) a_{s_{n-1}s_n}, \quad (17)$$

in which $f(s_{n-1}|\mathbf{y}_0^{n-1})$ is the forward probability at block $n-1$, obtained using the forward algorithm.

Applying the approximation (11), the posterior PDF can be rewritten as

$$\begin{aligned} f(\mathbf{x}_n|\mathbf{y}_0^n) \;&=\; \frac{1}{\Omega_n}\sum_s \gamma_n(s) \iint f_s(\mathbf{y}_n,\bar{g}_n',\ddot{g}_n') \\ &\qquad f_s(\mathbf{x}_n|\mathbf{y}_n,\bar{g}_n',\ddot{g}_n') d\bar{g}_n' d\ddot{g}_n' \\ &\approx\; \frac{1}{\Omega_n}\sum_s \omega_n(s) f_s(\mathbf{x}_n|\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n'), \quad (18) \end{aligned}$$

where

$$\begin{aligned} \omega_n(s) \;&=\; \gamma_n(s)\; f_s(\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n'), \quad (19) \\ \Omega_n \;&=\; f(\mathbf{y}_n|\mathbf{y}_0^{n-1}) = \int f(\mathbf{x}_n,\mathbf{y}_n|\mathbf{y}_0^{n-1}) d\mathbf{x}_n \\ &\approx\; \sum_s \gamma_n(s) f_s(\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n') = \sum_s \omega_n(s). \quad (20) \end{aligned}$$

Using the AR-HMM signal model, the conditional PDF $f_s(\mathbf{x}_n|\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n')$ for state s can be shown to be a Gaussian distribution, e.g., [1], with mean given by

$$\mathrm{E}_s[\mathbf{X}_n|\mathbf{Y}_n=\mathbf{y}_n,\bar{g}_n'=\hat{\bar{g}}_n',\ddot{g}_n'=\hat{\ddot{g}}_n'] \;= \hat{\bar{g}}_n\bar{\mathbf{D}}_{\bar{s}}(\hat{\bar{g}}_n\bar{\mathbf{D}}_{\bar{s}}+\hat{\ddot{g}}_n\ddot{\mathbf{D}}_{\ddot{s}})^{-1}\mathbf{y}_n, \quad (21)$$

which is the Wiener filtering of $\mathbf{y}_n$. The posterior noise PDF $f(\mathbf{w}_n|\mathbf{y}_0^n)$ has the same structure as (18), with $\mathbf{x}_n$ replaced by $\mathbf{w}_n$.

The Bayesian speech estimator (14) can then be obtained as

$$\begin{aligned} \hat{\mathbf{x}}_n \;&=\; \int \mathbf{x}_n f(\mathbf{x}_n|\mathbf{y}_0^n) d\mathbf{x}_n + \epsilon \int \mathbf{w}_n f(\mathbf{w}_n|\mathbf{y}_0^n) d\mathbf{w}_n \\ &=\; \mathbf{H}_n\mathbf{y}_n, \quad (22) \end{aligned}$$

where

$$\begin{aligned} \mathbf{H}_n \;&=\; \frac{1}{\Omega_n}\sum_s \omega_n(s)\mathbf{H}_s \quad (23) \\ \mathbf{H}_s \;&=\; (\hat{\bar{g}}_n\bar{\mathbf{D}}_{\bar{s}}+\epsilon\hat{\ddot{g}}_n\ddot{\mathbf{D}}_{\ddot{s}})(\hat{\bar{g}}_n\bar{\mathbf{D}}_{\bar{s}}+\hat{\ddot{g}}_n\ddot{\mathbf{D}}_{\ddot{s}})^{-1}. \quad (24) \end{aligned}$$

IV. Off-line parameter estimation

The training of the speech and noise HMM with gain models can be performed off-line using recordings of clean speech utterances and different noise environments. The training of the noise model is simplified by the assumption of independence between the noise gain and shape. The off-line training of the noise can be performed using the standard Baum-Welch algorithm [19], [22] using training data normalized by the long-term averaged noise gain. The noise gain variance $\ddot{\psi}^2$ is estimated as the sample variance of the logarithm of the excitation variances after the normalization.

The parameters of the speech HMM, $\bar{\theta}=\{\bar{a},\bar{\phi},\bar{\psi}^2,\bar{\alpha}\}$, are to be estimated using a training set that consists of R speech utterances. This training set is assumed to be sufficiently rich such that the general characteristics of speech are well represented. In addition, estimation of the speech gain bias $\bar{q}$ is necessary to calculate the likelihood score from the training data. For simplicity, we assume that the speech gain bias is constant for each training utterance. We use $\bar{q}_r$ to denote the speech gain bias of the r'th utterance. The block index n is now dependent on r, but is not explicitly shown in the notation for simplicity.

The parameters of interest, denoted $\theta=\{\bar{\theta},\bar{q}\}$, are optimized in the maximum likelihood sense. Similarly to the Baum-Welch algorithm, we propose an iterative algorithm based on the expectation-maximization (EM) framework [19]. The EM based algorithm is an iterative procedure that improves the log-likelihood score with each iteration. To avoid convergence to a local maximum, several random initializations are performed to select the best model parameters. The EM algorithm is particularly useful when the observation sequence is incomplete, i.e., the estimator is difficult to solve analytically without additional observations. In this case, the missing data is considered to be $\mathbf{z}_0^{N-1}=\{\bar{s}_0^{N-1},\bar{g}_0^{N-1}\}$, which are the sequence of the underlying states and speech gains.

The maximization step in the EM algorithm finds new model parameters that maximize the auxiliary function $\mathcal{Q}(\theta|\theta^{(j-1)})$ from the expectation step

$$\begin{aligned} \hat{\theta}^{(j)} \;&=\; \arg\max_\theta \mathcal{Q}(\theta|\hat{\theta}^{(j-1)}) \\ &=\; \arg\max_\theta \int_{\mathbf{z}_0^{N-1}} f(\mathbf{z}_0^{N-1}|\mathbf{x}_0^{N-1},\hat{\theta}^{(j-1)}) \\ &\qquad \log\left(f(\mathbf{z}_0^{N-1},\mathbf{x}_0^{N-1}|\theta)\right) d\mathbf{z}_0^{N-1}, \quad (25) \end{aligned}$$

where j denotes the iteration index.

Following the derivations in, e.g., [22], it can be shown that the auxiliary $\mathcal{Q}$ function can be written as

$$\begin{aligned} \mathcal{Q}(\theta|\hat{\theta}^{(j-1)}) = O(\theta|\hat{\theta}^{(j-1)}) + \sum_{r,n,\bar{s}} \bar{\omega}_n(\bar{s}) \int f_{\bar{s}}(\bar{g}_n'|\mathbf{x}_n,\hat{\theta}^{(j-1)}) \\ (\log f_{\bar{s}}(\bar{g}_n'|\theta) + \log f_{\bar{s}}(\mathbf{x}_n|\bar{g}_n',\theta))\, d\bar{g}_n', \quad (26) \end{aligned}$$

where the summations are over R utterances, N_r blocks of each utterance and $\bar{S}$ states. The posterior state probability

$$\bar{\omega}_n(\bar{s}) \;=\; f(\bar{s}_n|\mathbf{x}_0^{N-1},\hat{\theta}^{(j-1)}), \quad (27)$$

is evaluated using the forward-backward algorithm [23]. $O(\theta|\hat{\theta}^{(j-1)})$ contains all the terms associated with the parameters $\{\bar{a}\}$, which can be optimized following the standard Baum-Welch algorithm, e.g., [22].

Differentiating (26) with respect to the variables of interests and setting the resulting expression to zero, we can obtain the update equations for the j'th iteration. It turns out that the gradient terms with respect to $\{\bar{\phi},\bar{\psi}^2\}$ and $\bar{q}_r$ are not easily separable. Hence, an iterative estimation of $\bar{q}_r$ and $\bar{\theta}$ is performed. Assuming a fixed $\bar{q}_r$, the update equations for

$\{\bar{\phi}, \bar{\psi}^2\}$ are

$$\bar{\phi}_{\bar{s}}^{(j)} = \frac{1}{\bar{\Omega}} \sum_{r,n} \bar{\omega}_n(\bar{s}) \int \bar{g}'_n f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n, \hat{\theta}^{(j-1)}) d\bar{g}'_n - \bar{q}_r \quad (28)$$

$$\bar{\psi}_{\bar{s}}^{2(j)} = \frac{1}{\bar{\Omega}} \sum_{r,n} \bar{\omega}_n(\bar{s}) \int (\bar{g}'_n - \bar{\phi}_{\bar{s}}^{(j)} - \bar{q}_r)^2 f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n, \hat{\theta}^{(j-1)}) d\bar{g}'_n \quad (29)$$

where

$$\bar{\Omega} = \sum_{r,n} \bar{\omega}_n(\bar{s}). \quad (30)$$

The AR coefficients, $\bar{\alpha}$, can be obtained from the estimated autocorrelation sequence by applying the Levinson-Durbin recursion algorithm. Under the assumption of large K, e.g., [24], the autocorrelation sequence can be estimated as

$$\bar{r}_{\alpha_{\bar{s}}}^{(j)}[i] = \frac{1}{\bar{\Omega}} \sum_{r,n} \bar{\omega}_n(\bar{s}) r_{x_n}[i] \int (\bar{g}_n)^{-1} f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n, \hat{\theta}^{(j-1)}) d\bar{g}'_n, \quad (31)$$

where

$$r_{x_n}[i] = \sum_{j=0}^{K-i-1} x_n[j] x_n[j+i]. \quad (32)$$

Assume next that $\bar{\theta}$ is given, the update equation for $\bar{q}_r$ is

$$\bar{q}_r^{(j)} = \frac{1}{\bar{\Omega}'} \sum_{n,\bar{s}} \frac{\bar{\omega}_n(\bar{s})}{\bar{\psi}_{\bar{s}}^2} \left(\int \bar{g}'_n f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n, \hat{\theta}^{(j-1)}) d\bar{g}'_n - \bar{\phi}_{\bar{s}} \right), \quad (33)$$

for $\bar{\Omega}' = \sum_{n,\bar{s}} \bar{\omega}_n(\bar{s}) / \bar{\psi}_{\bar{s}}^2$.

By optimizing the EM criterion (25), the likelihood score of the parameters is non-decreasing in each iteration step [19]. Consequently, the iterative optimization will converge to model parameters that locally maximize the likelihood. The optimization is terminated when two consecutive likelihood scores are sufficiently close to each other.

The update equations contain several integrals that are difficult to solve analytically. One solution is to use the numerical techniques such as stochastic integration. In appendix II, we propose solutions by approximating the function $f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n, \hat{\theta}^{(j-1)})$ using the Taylor expansion.

V. On-line parameter estimation

The time-varying parameters $\theta = \{\bar{q}_n, \ddot{\bar{\phi}}_n\}$ as defined in (4) and (7) are to be estimated on-line using the observed noisy data. In addition, we restrict to the real-time constraint such that no additional delay is required by the estimation algorithm. Under the assumption that the model parameters vary slowly, we apply the recursive EM algorithm [20], [25] to perform the on-line parameter estimation. That is, the parameters are updated recursively for each observed noisy data block, such that the likelihood score is improved on average.

The recursive EM algorithm [20] is a technique based on the Robbins-Monro stochastic approximation principle, e.g., [26], for parameter re-estimation that involves incomplete or unobservable data. The recursive EM estimates of time-invariant parameters are shown [20] to be consistent and asymptotically Gaussianly distributed under conditions given in [20]. The technique is applicable to estimation of time-varying parameters by restricting the effect of the past observations, e.g. [25], by using forgetting factors. Applied to the estimation of HMM parameters, the Markov assumption makes the EM algorithm tractable [22]. The state probabilities are typically evaluated using the forward-backward algorithm. To facilitate low complexity and low memory implementation for the recursive estimation, we use the fixed-lag estimation approach of [25], neglecting the backward probabilities of the past states.

Let $\mathbf{z}_n = \{s_n, \bar{g}_n, \ddot{g}_n\}$ denote the hidden variables. The recursive EM algorithm optimizes for the auxiliary function defined as [25],

$$\mathcal{Q}_n(\theta|\hat{\theta}_0^{n-1}) = \int_{\mathbf{z}_0^n} f(\mathbf{z}_0^n|\mathbf{y}_0^n, \hat{\theta}_0^{n-1}) \log\left(f(\mathbf{z}_0^n, \mathbf{y}_0^n|\theta)\right) d\mathbf{z}_0^n, \quad (34)$$

where $\hat{\theta}_0^{n-1} = \{\hat{\theta}_j\}_{j=0..n-1}$ denotes the estimated parameters from the first block to the $n-1$'th block. Following lemma 3.2 of [25], the $\mathcal{Q}$ function (34) can be approximated as

$$\mathcal{Q}_n(\theta|\hat{\theta}_0^{n-1}) \approx \sum_{t=0}^{n} \mathcal{L}_t(\theta|\hat{\theta}_0^{t-1}) \quad (35)$$

$$\mathcal{L}_t(\theta|\hat{\theta}_0^{t-1}) \approx \sum_s \frac{\gamma_t(s)}{\Omega_t} \int\int f_s(\mathbf{y}_t, \bar{g}'_t, \ddot{g}'_t|\hat{\theta}_{t-1}) \left(\log f_{\bar{s}}(\bar{g}'_t|\theta) + \log f(\ddot{g}'_t|\theta)\right) d\bar{g}'_t d\ddot{g}'_t, \quad (36)$$

where the irrelevant terms with respect to the parameters of interest have been neglected. Applying the approximation (11), we get

$$\mathcal{L}_t(\theta|\hat{\theta}_0^{t-1}) \approx \sum_s \frac{\gamma_t(s)}{\Omega_t} f_s(\mathbf{y}_t, \hat{\bar{g}}'_t, \hat{\ddot{g}}'_t|\hat{\theta}_{t-1}) \left(\log f_{\bar{s}}(\hat{\bar{g}}'_t|\theta) + \log f(\hat{\ddot{g}}'_t|\theta)\right). \quad (37)$$

The recursive estimation algorithm optimizing the $\mathcal{Q}$ function can be implemented using the stochastic approximation technique, e.g., [26]. The update equations for the parameters have the form [25]

$$\hat{\theta}_n = \theta + \left(-\frac{\partial^2 \mathcal{Q}_n(\theta|\hat{\theta}_0^{n-1})}{\partial \theta^2} \right)^{-1} \frac{\partial \mathcal{L}_n(\theta|\hat{\theta}_0^{n-1})}{\partial \theta} \Bigg|_{\theta=\hat{\theta}_{n-1}}. \quad (38)$$

Taking the first and second derivatives of the auxiliary functions, the update equations can be solved analytically to

$$\hat{\ddot{\bar{\phi}}}_n = \hat{\ddot{\bar{\phi}}}_{n-1} + \frac{1}{\Xi_n} \sum_s \frac{\omega_n(s)}{\Omega_n} \left(\hat{\ddot{g}}'_n - \hat{\ddot{\bar{\phi}}}_{n-1} \right) \quad (39)$$

$$\hat{\bar{q}}_n = \hat{\bar{q}}_{n-1} + \frac{1}{\Xi'_n} \sum_s \frac{\omega_n(s)}{\Omega_n \bar{\psi}_{\bar{s}}^2} \left(\hat{\bar{g}}'_n - \bar{\phi}_{\bar{s}} - \hat{\bar{q}}_{n-1} \right), \quad (40)$$

where $\Xi_n = \sum_{t=0}^{n} \sum_s (\omega_t(s)/\Omega_t) = n+1$ and $\Xi'_n = \sum_{t=0}^{n} \sum_s \omega_t(s)/(\Omega_t \bar{\psi}_{\bar{s}}^2)$ are two non-decreasing normalization terms that control the impact of one new observation for increased number of past observations. As the parameters are considered time-varying, we apply exponential forgetting factors to the normalization term, to decrease the impact of the results from the past. Hence, the modified normalization

terms are evaluated by recursive summation of the past values

$$\Xi_n = \rho_{\ddot{\phi}}\, \Xi_{n-1} + 1 \quad (41)$$

$$\Xi'_n = \rho_{\bar{q}}\, \Xi'_{n-1} + \sum_s \frac{\omega_n(s)}{\Omega_n \bar{\psi}_{\bar{s}}^2}, \quad (42)$$

where $0 \leq \rho_{\ddot{\phi}}, \rho_{\bar{q}} \leq 1$ are two exponential forgetting factors and $\rho_{\ddot{\phi}}, \rho_{\bar{q}} = 1$ corresponds to no forgetting.

VI. Experiments and results

In this section, we first describe the implementation details of the proposed method. We then present the setup for the objective and subjective evaluations, followed by the experimental results and discussions.

A. System implementation

The proposed speech enhancement system was implemented for 8 kHz sampled speech. The HMMs were implemented using Gaussian mixture models (GMM) in each state. The speech HMM had eight states and 16 mixture components per state, with AR models of order ten. The training data for speech consisted of 640 clean utterances from the training set of the TIMIT database down-sampled to 8kHz. We used a set of pre-trained noise HMMs, each describing a particular noise environment. As discussed in previous works on prior noise models [13], [15], it is preferable to have a limited noise model that describes the current noise environment, than a general noise model that covers all possible noises. We trained a number of noise models, each describing one typical noise environment. Each noise model had three states and three mixture components per state. All noise models used AR models of order six, with the exception of the babble noise model, which was of order ten, motivated by the similarity of its spectra to speech. The noise signals used in the training were not used in the evaluation. During enhancement, the first 100 ms of the noisy signal was assumed to be noise only, and was used to select one active model from the inventory of noise models. The selection was based on the maximum likelihood criterion. The forgetting factors for adapting the time-varying gain model parameters were experimentally set to $\rho_{\ddot{\phi}} = 0.9$ and $\rho_{\bar{q}} = 0.99$. With these forgetting factors, as well as with other settings we tried, the on-line parameter estimation method (section V) was found to be numerically stable in our evaluations.

The noisy signal was processed in the frequency domain in blocks of 32 ms windowed using the Hann window. Using the approximation that the covariance matrix of each state was circulant, the estimator (22) was implemented in the frequency domain, e.g., [1]. The covariance matrices are then diagonalized by the Fourier transformation matrix, e.g., [27]. The estimator corresponds to applying an SNR dependent gain-factor to each of the frequency bands of the observed noisy spectrum. The gain-factors were obtained as in (23), with the matrices replaced by the frequency responses of the filters (24). The synthesis was performed using 50% overlap-and-add.

While the computational complexity has not been the focus of this work, it is one important constraint for applying the proposed method in practical environments. The computational complexity of the proposed method is proportional to the number of mixture components in the noisy model. Therefore, the key to reduce the complexity is pruning of mixture components that are unlikely to contribute to the estimators. In our implementation, we kept 16 speech mixture components in every block, and the selection was according to the likelihood scores calculated using the most likely noise component of the previous block. Our Matlab implementation (with the computation of (13) written in C/MEX) ran about 5-6 times real-time using a Pentium 4 computer at 2.8 GHz with one gigabyte of memory. We believe that more efficient pruning may further reduce the computational complexity.

B. Experimental setup

The evaluation was performed using the core test set of the TIMIT database (192 sentences) resampled to 8 kHz. The total length of the evaluation utterances was about ten minutes. The noise environments considered are: traffic noise, recorded on the side of a busy freeway, white Gaussian noise, babble noise (Noisex-92), and white-2, which is amplitude modulated white Gaussian noise using a sinusoid function. The amplitude modulation simulates the change of noise energy level, and the sinusoid function models that the noise source periodically passes by the microphone. The sinusoid we used had a period of two seconds, and the maximum amplitude of the modulation was four times higher than the minimum amplitude. The noisy signals were generated by adding the concatenated speech utterances to noise for various input SNRs. For all test methods, the utterances were processed concatenated.

Objective evaluations of the proposed method are described in sections VI-C to VI-E. The reference methods for the objective evaluations are the HMM based MMSE method (ref. A) [1], the gain-adaptive HMM based MAP method (ref. B) [8] and the HMM based MMSE method using HMM-based noise adaptation (ref. C) [13]. The reference methods were implemented using shared codes and similar parameter setups whenever possible to minimize irrelevant performance mismatch. The ref. A and B methods require a separate noise estimation algorithm, and we used the method based on minimum statistics [6]. The gain contour estimation of ref. B was performed according to [8]. The ref. C method requires a VAD for noise classification and gain adaptation, and we used the ideal VAD estimated from the clean signal. The global gain factor used in ref. A and C, which compensates for the speech model energy mismatch, is estimated according to [1].

The objective measures considered in the evaluations were signal-to-noise ratio (SNR), segmental SNR (SSNR) [28], log-spectral distortion (SD) [29] and the Perceptual Evaluation of Speech Quality (PESQ) [30]. For the SSNR and SD measures, the low energy blocks (40 dB lower than the long-term power level) were excluded from the evaluation [28]. The measures are evaluated for each utterance separately and averaged over the utterances to get the final scores. The first utterance was removed from the averaging to avoid biased results due to initializations. As the input SNR was evaluated over all

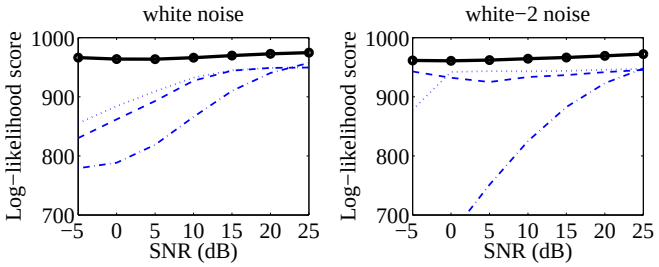

Fig. 2. The log likelihood (LL) scores of the speech models estimated from the noisy observations evaluated on true speech signals for different input SNRs. The solid lines with dots are from the proposed method. The dashed, dash-dotted and dotted lines are from the reference methods A, B, and C, respectively.

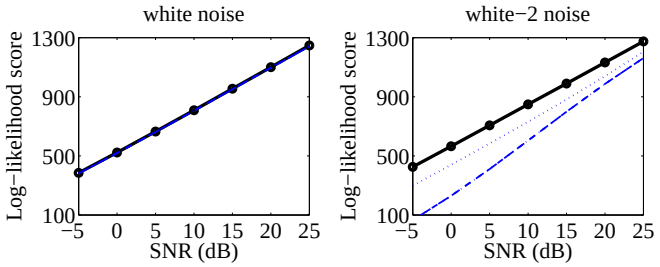

Fig. 3. The log likelihood (LL) scores of the noise models estimated from the noisy observations evaluated on true noise signals for different input SNRs. The solid lines with dots are from the proposed method. The dashed, dash-dotted and dotted lines are from the reference methods A, B, and C, respectively.

utterances concatenated, there is a small deviation in the evaluated SNR of the noisy signals in the presented results (TABLE I).

C. Evaluation of the modeling accuracy

One goal of this work is to improve the modeling accuracy for both speech and noise. The improved model is expected to result in improved speech enhancement performance. In this experiment, we evaluated the modeling accuracy of the methods by evaluating the log-likelihood (LL) score of the estimated speech and noise models using the true speech and noise signals.

The LL score of the estimated speech model for the n'th block is defined as

$$LL(\mathbf{x}_n) \quad = \quad \log\left(\frac{1}{\Omega_n}\sum_s \omega_n(s) f_{\bar{s}}(\mathbf{x}_n|\hat{\bar{g}}_n)\right), \qquad (43)$$

where the weight $\omega_n(s)$ is the state probability given the observations $\mathbf{y}_0^n$ (19), and $f_{\bar{s}}(\mathbf{x}_n|\hat{\bar{g}}_n)$ is the density function (5) evaluated using the estimated speech gain $\hat{\bar{g}}_n$. The likelihood score for noise is defined similarly. The values were averaged over all utterances to obtain the mean value. The low energy blocks (30 dB lower than the long-term power level) were excluded from the evaluation for the numerical stability.

The LL scores for the white and white-2 noises as functions of input SNRs are shown in Fig. 2 for the speech model and Fig. 3 for the noise model. The proposed method is shown in solid lines with dots, while the reference methods A, B and C are dashed, dash-dotted and dotted lines, respectively. The proposed method is shown to have higher scores than all reference methods for all input SNRs. Surprisingly, the ref. B. method performs poorly, particularly for low SNR cases. We speculate that this is due to the dependency on the noise estimation algorithm, which is sensitive to input SNR. As for the noise modeling, the performance of all the methods is similar for the white noise case. This is expected due to the stationarity of the noise. For the white-2 noise, the ref. C method performs better than the other reference methods, due to the HMM-based noise modeling. The proposed method has higher LL scores than all reference methods, as results from the explicit noise gain modeling.

D. Objective evaluation of MMSE waveform estimators

The improved modeling accuracy is expected to lead to increased performance of the speech estimator. In this experiment, we evaluated the MMSE waveform estimator by setting the residual noise level ϵ to zero. The MMSE waveform estimator optimizes the expected squared error between clean and reconstructed speech waveforms, which is measured in terms of SNR. Note that the ref. B method is a MAP estimator, optimizing for the hit-and-miss criterion, e.g., [31].

The SNR improvements of the methods as functions of input SNRs for different noise types are shown in Fig. 4. The estimated speech of the proposed method has consistently higher SNR improvement than the reference methods. The improvement is significant for non-stationary noise types, such as traffic and white-2 noises. The SNR improvement for the babble noise is smaller than the for other noise types, which is partly expected from the similarity of the speech and noise.

The objective quality measures for 10-dB input SNR are

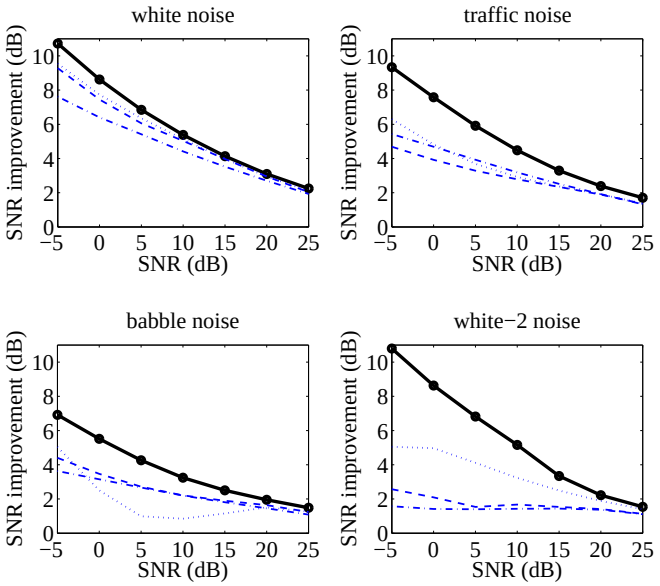

Fig. 4. SNR improvements in dB as functions of input SNRs. The solid lines with dots are from the proposed method. The dashed, dash-dotted and dotted lines are from the reference methods A, B, and C, respectively.

shown in TABLE I. The results for the SSNR measure are consistent with the SNR measure, where the improvement is significant for non-stationary noise types. While the MMSE estimator is not optimized for any perceptual measure, the results from PESQ show consistent improvement over the reference methods. The results for the SD measure are less consistent, where the Ref.A method seems to be better, particularly for the white noise. We believe the results are due to the fact the MMSE waveform estimator is optimized for the SNR criterion. The discussion follows in the next subsection.

Type	Noisy	Sys.	Ref.A	Ref.B	Ref.C
			SNR (dB)		
white	10.00	15.38	15.03	14.42	15.13
traffic	10.62	15.10	13.40	13.81	13.54
babble	10.21	13.45	12.42	12.41	11.06
white-2	10.04	15.20	11.71	11.46	13.27
			SSNR (dB)		
white	0.49	8.06	7.33	5.28	7.78
traffic	1.73	8.01	5.74	5.82	6.15
babble	1.25	6.13	4.57	4.16	4.04
white-2	2.11	8.21	4.66	4.19	6.24
			PESQ (MOS)		
white	2.16	2.86	2.72	2.61	2.78
traffic	2.50	2.97	2.75	2.76	2.70
babble	2.54	2.78	2.59	2.69	2.35
white-2	2.24	2.76	2.43	2.40	2.42
			SD (dB)		
white	6.52	5.64	5.23	6.47	5.44
traffic	4.64	4.21	4.27	4.88	4.38
babble	4.27	4.21	4.19	4.61	4.95
white-2	6.15	5.68	5.93	6.42	6.48

TABLE I
EXPERIMENTAL RESULTS FOR NOISY SPEECH SIGNALS OF 10-DB INPUT SNR USING MMSE WAVEFORM ESTIMATORS (REF. B IS A MAP ESTIMATOR).

E. Objective evaluation of sample spectrum estimators

As the MMSE waveform estimator is optimized for the SNR measure, the improved modeling may not necessarily result in improvement in measures such as SD. On the other hand, the MMSE sample spectrum estimator is the optimal preprocessor for AR model based vector quantization in the Itakura-Saito sense [32]. In the following test, the MMSE waveform estimator is replaced by the MMSE sample spectrum estimator, for the proposed method and the ref. A and C methods. The enhanced speech was reconstructed using the squared root of the estimated sample spectrum and the noisy phase. The ref. B method is unchanged.

The results using the SD measure for 10-dB input SNR are shown in TABLE II. The SD values for the proposed method reduce significantly compared to the corresponding MMSE waveform estimator, and the results are consistently better than the reference methods. For the ref. A and C methods, the difference between the estimators is relatively small. The reason is likely the inaccurate modeling of speech and noise statistics.

The objective results for different estimators suggest that accurate statistical modeling improves the objective performance, when a proper choice of the optimization criterion is used.

Type	Noisy	Sys.	Ref.A	Ref.B	Ref.C
			SD (dB)		
white	6.52	4.69	5.26	6.47	5.10
traffic	4.64	3.59	4.46	4.88	4.51
babble	4.27	3.61	4.26	4.61	4.89
white-2	6.15	4.77	5.75	6.42	5.46

TABLE II
SD RESULTS FOR NOISY SPEECH SIGNALS OF 10-DB INPUT SNR USING MMSE SAMPLE SPECTRUM ESTIMATORS (REF. B IS A MAP ESTIMATOR), CF. THE SD RESULTS IN TABLE I.

F. Perceptual quality evaluation

The objective evaluation in the previous subsections demonstrates the advantage of explicit gain modeling for HMM-based speech enhancement. Herein, we show how the proposed technique can be used in a practical speech enhancement system. The perceptual quality of the system was evaluated through listening tests. To make the tests relevant, the reference system must be perceptually well tuned (preferably a standard system). Hence, the noise suppression module of the Enhanced Variable Rate Codec (EVRC) [3] was selected as the reference system.

The proposed Bayesian speech estimator (15) facilitates adjustment of the residual noise level, ϵ. While the objective results (TABLE I) indicate good SNR/SSNR performance for $\epsilon = 0$, we found experimentally that $\epsilon = 0.15$ forms a good trade-off between the level of residual noise and audible speech distortion and this value was used in the listening tests.

The AR-based speech HMM does not model the spectral fine structure of voiced sounds in speech. Therefore, the estimated speech using (22) exhibits a low-level rumbling noise in some voiced segments, particularly high-pitched speakers. This problem is inherent for AR-HMM-based methods and is well documented, e.g., [1], [33]. Thus, the method of [34] was further applied to enhance the spectral fine-structure of voiced speech[2].

The subjective evaluation was performed under two test scenarios: 1) straight enhancement of noisy speech, and 2) enhancement in the context of a speech coding application.

[2]The method of [34] is constrained to make only small modifications to the signal, thus restricting it to strengthening the perceived periodicity of voiced speech.

Noisy speech signals of input SNR 10 dB were used in both tests. The evaluations were performed using 16 utterances from the TIMIT core test set, one male and one female speaker from each of the eight dialects. The tests were set up similarly to the Comparison Category Rating (CCR) test [35]. Ten listeners participated in the listening tests. Each listener was asked to score a test utterance in comparison to a reference utterance on an integer scale from -3 to +3, corresponding to *much worse* to *much better* [35]. Each pair of utterances was presented twice, with switched order. The utterance pairs were ordered randomly.

1) Evaluation of speech enhancement systems: The noisy speech signals were pre-processed by the 120 Hz high-pass filter from the EVRC system. The reference signals were processed by the EVRC noise suppression module. The encoding/decoding of the EVRC codec was not performed. The test signals were processed using the proposed speech estimator followed by the spectral fine-structure enhancer [34]. To demonstrate the perceptual importance of the spectral fine-structure enhancement, we also performed the test without this additional module. The mean CCR scores together with the 95% confidence intervals are presented in TABLE III.

white	traffic	babble	white-2
with [34]			
0.95±0.10	1.22±0.13	0.39±0.14	1.43±0.13
without [34]			
0.60±0.12	0.77±0.16	-0.22±0.14	0.96±0.14

TABLE III
TEST SCENARIO ONE: SCORES FROM THE CCR LISTENING TEST WITH 95% CONFIDENCE INTERVALS (10 DB INPUT SNR). THE SCORES ARE RATED ON AN INTEGER SCALE FROM -3 TO 3, CORRESPONDING TO *much worse* TO *much better* [35]. POSITIVE SCORES INDICATE A PREFERENCE FOR THE PROPOSED SYSTEM.

The CCR scores show a consistent preference to the proposed system when the fine-structure enhancement is performed. The scores are highest for the traffic and white-2 noises, which are non-stationary noises with rapidly time-varying energy. The proposed system has a minor preference for the babble noise, consistent with the results from the objective evaluations. As expected, the CCR scores are reduced without the fine-structure enhancement. In particular, the noise level between the spectral harmonics of voiced speech segments was relatively high and this noise was perceived as annoying by the listeners. Under this condition, the CCR scores still show a positive preference for the white, traffic and white-2 noise types.

2) Evaluation of enhancement in the context of speech coding: In the following test, the reference signals were processed by the EVRC speech codec with the noise suppression module enabled. The test signals were processed by the proposed speech estimator (without the fine-structure enhancement) as the preprocessor to the EVRC codec with its noise suppression module disabled. Thus, the same speech codec was used for both systems in comparison, and they differ only in the applied noise suppression system. The mean CCR scores together with the 95% confidence intervals are presented in TABLE IV. The test results show a positive preference for the white, traffic and white-2 noise types. Both systems perform similarly for the babble noise condition.

white	traffic	babble	white-2
0.62±0.12	0.92±0.15	0.02±0.13	0.98±0.14

TABLE IV
SCORES FROM THE CCR LISTENING TEST WITH 95% CONFIDENCE INTERVALS (10 DB INPUT SNR). THE NOISE SUPPRESSION SYSTEMS WERE APPLIED AS PRE-PROCESSORS TO THE EVRC SPEECH CODEC. THE SCORES ARE RATED ON AN INTEGER SCALE FROM -3 TO 3, CORRESPONDING TO *much worse* TO *much better* [35]. POSITIVE SCORES INDICATE A PREFERENCE FOR THE PROPOSED SYSTEM.

The results from the subjective evaluation demonstrate that the perceptual quality of the proposed speech enhancement system is better or equal to the reference system. The proposed system has a clear preference for noise sources with rapidly time-varying energy, such as traffic and white-2 noises, which is most likely due to the explicit gain modeling and estimation. The perceptual quality of the proposed system can likely be further improved by additional perceptual tuning.

VII. CONCLUSIONS

In this paper, a new HMM-based speech enhancement method using explicit speech and noise gain modeling is presented. Through the introduction of stochastic gain variables, energy variation in both speech and noise is explicitly modeled in a unified framework. The time-invariant model parameters are estimated off-line using the expectation-maximization (EM) algorithm, while the time-varying parameters are estimated on-line using the recursive EM algorithm. The experimental results demonstrate improvement in modeling accuracy of both speech and (non-stationary) noise statistics. The improved speech and noise models were applied to a novel Bayesian speech estimator. The combination of improved modeling and proper choice of optimization criterion was shown to result in consistent improvement over the reference methods. The improvement is significant for non-stationary noise types with fast time-varying energy. The performance in terms of perceptual quality was evaluated through listening tests. The subjective results confirm the advantage of the proposed scheme.

APPENDIX I
EM BASED SOLUTION TO (12)

Evaluation of the proposed speech estimator requires solving the maximization problem (12) for each state. Here, we provide a solution based on the EM algorithm. The problem corresponds to the maximum a-posteriori estimation of $\{\bar{g}_n, \ddot{g}_n\}$ for a given state s. We assume that the missing data of

interests are $\mathbf{x}_n$ and $\mathbf{w}_n$. We solve for $\{\hat{\bar{g}}_n, \hat{\ddot{g}}_n\}$ that maximize the $\mathcal{Q}$ function following the standard EM formulation. The optimization condition with respect to the speech gain $\bar{g}'_n$ of the j'th iteration is

$$\frac{1}{2}\frac{R_x^{(j-1)}}{\exp(\hat{\bar{g}}_n'^{(j)})} - \frac{\hat{\bar{g}}_n'^{(j)} - \bar{\phi}_{\bar{s}} - \bar{q}_n}{\bar{\psi}_{\bar{s}}^2} - \frac{K}{2} = 0 \qquad (44)$$

where

$$R_x^{(j-1)} \;=\; \int f(\mathbf{x}_n|\mathbf{y}_n, \hat{\theta}^{(j-1)})\mathbf{x}_n^T\bar{\mathbf{D}}_{\bar{s}}^{-1}\mathbf{x}_n d\mathbf{x}_n, \quad (45)$$

which is the expected residual variance of the speech filtered through the inverse filter. The condition equation of the noise gain $\ddot{g}_n$ has the similar structure as (44) with $\mathbf{x}$ replaced by $\mathbf{w}$. The equations can be solved using the Lambert W function [36]. Rearranging the terms in (44), we obtain

$$\begin{aligned}\hat{\bar{g}}_n'^{(j)} \;&=\; \bar{\phi}_{\bar{s}} + \bar{q}_n - \frac{K\bar{\psi}_{\bar{s}}^2}{2} + \\ &\quad W_0\left(\frac{\bar{\psi}_{\bar{s}}^2 R_x^{(j-1)}}{2}\exp\left(\frac{K\bar{\psi}_{\bar{s}}^2}{2} - \bar{\phi}_{\bar{s}} - \bar{q}_n\right)\right), \quad (46)\end{aligned}$$

where $W_0(\cdot)$ denotes the principle branch of the Lambert W function. Since the input term to $W_0(\cdot)$ is real and nonnegative, only the principle branch is needed and the function is real and nonnegative. Efficient implementation of $W_0(\cdot)$ is discussed in [37]. When the gain variance is large compared to the mean, taking the exponential function of (46) may result in values out of the numerical range of a computer. This can be prevented by ignoring the second term in (44) when the variance is too large. The approximation is equivalent to assuming uniform prior, which is reasonable for large variance.

Appendix II
Approximation of $f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n)$

We propose an approximation of $f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n)$ to simplify the integrals in (28,29,31,33). Let $f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n) = \frac{1}{C}f_{\bar{s}}(\bar{g}'_n, \mathbf{x}_n)$ for $C = f_{\bar{s}}(\mathbf{x}_n) = \int f_{\bar{s}}(\bar{g}'_n, \mathbf{x}_n)d\bar{g}'_n$, it can be shown that the second derivative of $\log f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n)$ with respect to $\bar{g}'_n$ is negative for all $\bar{g}'_n$, which suggests that $f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n)$ is a logconcave function and a unique global maximum exists. We approximate the function by applying the 2nd order Taylor expansion of $\log f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n)$ around its mode $\hat{\bar{g}}'_n$, and enforce proper normalization. The resulting PDF is Gaussian distributed

$$f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n) \approx \left(2\pi\bar{A}_n^2(\bar{s})\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\bar{A}_n^2(\bar{s})}(\bar{g}'_n - \hat{\bar{g}}'_n)^2\right), \quad (47)$$

for

$$\hat{\bar{g}}'_n \;=\; \arg\max_{\bar{g}'} \log f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n) \qquad (48)$$

$$\bar{A}_n^2(\bar{s}) \;=\; -\left(\frac{\partial^2 \log f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n)}{\partial \bar{g}_n'^2}\right)^{-1}\Bigg|_{\bar{g}'_n = \hat{\bar{g}}'_n}. \qquad (49)$$

Applying the approximated Gaussian PDF, the integrals in (3) and (28-33) can be solved analytically.

The maximizing $\hat{\bar{g}}'_n$ can be obtained by setting the first derivative of $\log f_{\bar{s}}(\bar{g}'_n|\mathbf{x}_n)$ to zero and solve for $\bar{g}'_n$. We get

$$\frac{1}{2}\frac{\mathbf{x}_n^T\bar{\mathbf{D}}_{\bar{s}}^{-1}\mathbf{x}_n}{\exp(\hat{\bar{g}}'_{\bar{s}})} - \frac{\hat{\bar{g}}'_{\bar{s}} - \bar{\phi}_{\bar{s}} - \bar{q}_n}{\bar{\psi}_{\bar{s}}^2} - \frac{K}{2} = 0, \qquad (50)$$

which again can be solved using the Lambert W function similarly as (44).

References

[1] 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.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, no. 2, pp. 113–120, Apr. 1979.

[3] "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," TIA/EIA/IS-127, Jul. 1996.

[4] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, 2002, pp. 253–256.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[12] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 568–580, Nov. 2003.

[13] 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.

[14] 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.

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

[16] ——, "Codebook-based Bayesian speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 1077–1080.

[17] H. Sameti and L. Deng, "Nonstationary-state hidden Markov model representation of speech signals for speech enhancement," *Elsevier Signal Processing Journal*, vol. 82, no. 2, pp. 205–227, Feb. 2002.

[18] H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter models: parameter selection and sensitivity to power normalization," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, pp. 80–89, Jan. 1994.

[19] A. P. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

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

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

[22] 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.

[23] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[24] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 6, pp. 1404–1413, Dec. 1985.

[25] 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.

[26] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. Springer Verlag, 2003.

[27] R. Gray, "Toeplitz and circulant matrices: a review," Technical Rept. No. 6504-1, Inform. Sys. Lab., Stanford Univ., (Revised), 2005.

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

[29] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, ch. 12, pp. 433–466.

[30] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, Feb. 2001.

[31] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

[32] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inform. Theory*, vol. 34, no. 4, pp. 826–834, Jul. 1988.

[33] M. Deisher and A. Spanias, "HMM-based speech enhancement using harmonic modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, 1997, pp. 1175–1178.

[34] W. B. Kleijn, "Enhancement of coded speech by constrained optimization," in *Proc. IEEE Workshop on Speech Coding*, Oct. 2002, pp. 163–165.

[35] "Methods for subjective determination of transmisson quality," ITU-T Recommendation P.800, Aug. 1996.

[36] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[37] D. A. Barry, P. J. Culligan-Hensley, and S. J. Barry, "Real values of the W-function," *ACM Transactions on Mathematical Software*, vol. 21, no. 2, pp. 161–171, Jun. 1995.