# HMM-BASED SPEECH ENHANCEMENT USING EXPLICIT GAIN MODELING

*David Y. Zhao and W. Bastiaan Kleijn*

KTH (Royal Institute of Technology)
School of Electrical Engineering
10044 Stockholm, Sweden
{ david.zhao, bastiaan.kleijn}@ee.kth.se

**ABSTRACT**

We propose a hidden Markov model (HMM) based speech enhancement method using explicit modeling of speech and noise gains. The gains are considered to be stochastic variables in an HMM framework. The speech gain models the energy variations of 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 time-varying parameters of the gain models are estimated on-line using the recursive expectation maximization (EM) algorithm. The performance of the proposed enhancement system is evaluated through both objective and subjective tests. The experimental results confirm the advantage of explicit gain modeling, particularly for non-stationary noise sources.

## 1. INTRODUCTION

The enhancement of speech from corrupted noisy observations is often based on probabilistic models of speech and noise, e.g., [1]. Accurate modeling and estimation of the speech and noise statistics is, therefore, of great importance. While methods based on the traditional noise estimation algorithms perform reasonably well for stationary noise, their performance under non-stationary noise conditions is still unsatisfactory.

The hidden Markov model (HMM) has been applied successfully to model the statistics of speech [1,2] and noise [3] for speech enhancement. 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. Applied to speech, an AR-HMM models the change of spectral characteristics, assuming a finite number of AR processes, each with a fixed excitation variance. While it is reasonable to assume limited variations of AR coefficients due to the physical constraints of the human vocal tract, the standard AR-HMM does not explicitly model the variations in speech energy levels of a phone, typically due to differences in pronunciation and/or different vocalizations of individual speakers. 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 target the aforementioned problems and propose explicit parameterization and modeling of speech and noise gains, incorporated in the HMM framework. The speech and noise gains are defined as the parameters modeling the energy levels of speech and noise, respectively, and are considered as stochastic variables. The state-dependent probabilistic density function (PDF) of speech/noise signal is then a function depending on the gain. For the speech gain model, we assume that different states have different gain distributions. Thus, the model facilitates that a voiced sound typically has a larger gain than an unvoiced sound. The time-varying parameters of the gain models are estimated on-line using the recursive expectation maximization (EM) algorithm. The proposed HMMs with explicit gain models are applied to a Bayesian speech estimator, as shown in Fig. 1.

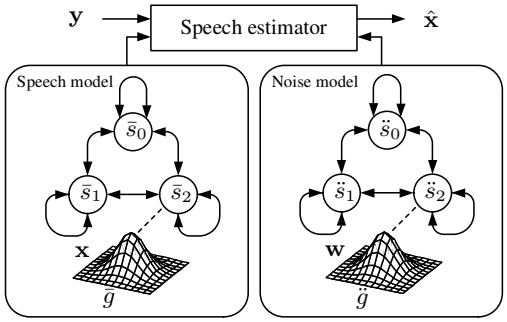


**Fig. 1**. Schematic diagram of the proposed HMM-based speech enhancement method. $\mathbf{x}$, $\mathbf{w}$ and $\mathbf{y}$ denote speech, noise and noisy signals, respectively. $s$ denotes an HMM state and $g$ denotes a gain variable. The overbar $\bar{}$ is used for the variables in the speech model,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

The proposed speech HMM generalizes the AR HMM based method [1], and the gain-adaptive HMM based method [1, 2]. In the gain-adaptive HMM [1, 2], the speech gain (referred to as the gain contour), is estimated on-line using the noisy observation in the maximum likelihood (ML) sense. Hence, the method implicitly assumed a uniform prior of the gain in a Bayesian framework. The performance of the gain-adaptive HMM method was shown to be inferior to the AR-HMM method [1], partly due to the weak gain modeling. In our work, stronger prior gain knowledge is introduced to the HMM framework using state-dependent gain distributions. For the noise HMM, a heuristic noise gain adaptation using a voice activity detector (VAD) was proposed in [3], where the adaptation is performed in speech pauses longer than 100 ms. In our recent work [4], 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.

## 2. 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 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 variables of the noisy signal, clean speech and noise, respectively.

### 2.1. Speech model

We describe the statistics of the speech using an ergodic 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 PDF of $\mathbf{x}_0^{N-1}$ is modeled as

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

where the summation is over the set of all possible state sequences $\bar{\mathbf{S}}$. 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 for a given state $\bar{s}$, $f_{\bar{s}}(\mathbf{x}_n)$, is defined as the integral over all possible speech gains, modeling the speech energy, in the logarithmic domain,

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

where $\bar{g}'_n = \log \bar{g}_n$ and $\bar{g}_n$ denotes the speech gain in the linear domain. The logarithmic domain formulation facilitates 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 $\bar{g}_n$. 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 change of recording conditions.

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. The density function is given by

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

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 consist of the AR coefficients including the leading one, $[1, \bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_{\bar{p}}]^T$.

### 2.2. Noise model

Elaborate noise models [3, 5] are useful to capture the high diversity and variability of acoustical noise. In this work, similar HMMs are used for speech (cf. 2.1) and noise. The model parameters for noise are denoted using double dots $\ddot{}$ (instead of bar $\bar{}$ 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 noise states,

$$f(\ddot{g}'_n) = \frac{1}{\sqrt{2\pi\ddot{\psi}^2}} \exp\!\left(-\frac{1}{2\ddot{\psi}^2}\left(\ddot{g}'_n - \ddot{\phi}_n\right)^2\right), \tag{6}$$

i.e., with mean $\ddot{\phi}_{\ddot{s}_n} = \ddot{\phi}_n$ and variance $\ddot{\psi}_{\ddot{s}}^2 = \ddot{\psi}^2$. The mean $\ddot{\phi}_n$ is considered to be a time-varying parameter to model the unknown noise energy.

### 2.3. Noisy signal model

The PDF of the noisy signal can be derived based on the 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, each state $s$ consists of the composition of the state $\bar{s}$ of the speech component and the state $\ddot{s}$ of the noise component. The noisy PDF corresponding to state $s$ is

$$f_s(\mathbf{y}_n) \quad = \quad \int\int f_s(\mathbf{y}_n, \bar{g}'_n, \ddot{g}'_n) d\bar{g}'_n d\ddot{g}'_n \tag{7}$$

$$\quad = \quad \int\int f_{\bar{s}}(\bar{g}'_n) f(\ddot{g}'_n) f_s(\mathbf{y}_n|\bar{g}'_n, \ddot{g}'_n) d\bar{g}'_n d\ddot{g}'_n, \tag{8}$$

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

$$\mathbf{D}_s \quad = \quad \bar{g}_n \bar{\mathbf{D}}_{\bar{s}} + \ddot{g}_n \ddot{\mathbf{D}}_{\ddot{s}}. \tag{9}$$

The integral of (7) can be evaluated numerically, e.g., by stochastic integration. To facilitate real-time implementation, we approximate the integral using the point estimates of the gains,

$$f_s(\mathbf{y}_n) \quad \approx \quad f_s(\mathbf{y}_n, \hat{\bar{g}}'_n, \hat{\ddot{g}}'_n), \tag{10}$$

$$\{\hat{\bar{g}}'_n, \hat{\ddot{g}}'_n\} \quad = \quad \arg\max_{\bar{g}'_n, \ddot{g}'_n} \log f_s(\mathbf{y}_n, \bar{g}'_n, \ddot{g}'_n), \tag{11}$$

where $\{\hat{\bar{g}}'_n, \hat{\ddot{g}}'_n\}$ is obtained numerically. The approximation is valid if the only significant peak of the integrand in (7) is at $\{\hat{\bar{g}}'_n, \hat{\ddot{g}}'_n\}$. The integrand can then be considered as a scaled Dirac delta function centered at $\{\hat{\bar{g}}'_n, \hat{\ddot{g}}'_n\}$. A more rigorous analysis of a similar approximation is provided in [5].

## 3. BAYESIAN SPEECH ESTIMATION

We consider the estimation of the clean speech from the observed noisy signal. Motivated by our previous work [4], we consider the Bayesian speech estimator based on a criterion that results in an adjustable level of residual noise in the enhanced speech,

$$\hat{\mathbf{x}}_n \quad = \quad \arg\min_{\tilde{\mathbf{x}}_n} \mathrm{E}[C(\mathbf{X}_n, \mathbf{W}_n, \tilde{\mathbf{x}}_n)|\mathbf{Y}_0^n = \mathbf{y}_0^n], \tag{12}$$

minimizing the Bayes risk for the cost function

$$C(\mathbf{x}_n, \mathbf{w}_n, \tilde{\mathbf{x}}_n) = ||(\mathbf{x}_n + \epsilon\mathbf{w}_n) - \tilde{\mathbf{x}}_n||^2, \tag{13}$$

where $||\cdot||$ denotes the vector norm and $0 \leq \epsilon \ll 1$ defines the adjustable residual noise level. By explicitly leaving some level of residual noise, the criterion facilitates reduction of processing artifacts, i.e., speech distortions. It converges to the minimum mean squared error (MMSE) waveform estimator, when $\epsilon$ is set to zero.

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

$$f(\mathbf{x}_n|\mathbf{y}_0^n) \;=\; \frac{f(\mathbf{x}_n,\mathbf{y}_n|\mathbf{y}_0^{n-1})}{f(\mathbf{y}_n|\mathbf{y}_0^{n-1})} = \frac{\sum_s \gamma_n(s) f_s(\mathbf{x}_n,\mathbf{y}_n)}{f(\mathbf{y}_n|\mathbf{y}_0^{n-1})}, \quad (14)$$

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

$$\gamma_n(s) \;=\; f(s_n|\mathbf{y}_0^{n-1}) = \sum_{s_{n-1}} f(s_{n-1}|\mathbf{y}_0^{n-1}) a_{s_{n-1}s_n}. \quad (15)$$

Using the approximation (10), the posterior PDF is approximately

$$\begin{aligned} f(\mathbf{x}_n|\mathbf{y}_0^n) \;&=\; \frac{1}{\Omega_n}\sum_s \gamma_n(s) \int\!\!\int f_s(\mathbf{y}_n,\bar{g}_n',\ddot{g}_n') \\ &\qquad f_s(\mathbf{x}_n|\mathbf{y}_n,\bar{g}_n',\ddot{g}_n') d\bar{g}_n' d\ddot{g}_n' \\ &\approx\; \frac{1}{\Omega_n}\sum_s \omega_n(s) f_s(\mathbf{x}_n|\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n'), \quad (16) \\ \omega_n(s) \;&=\; \gamma_n(s)\, f_s(\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n'), \quad (17) \\ \Omega_n \;&=\; f(\mathbf{y}_n|\mathbf{y}_0^{n-1}) = \int f(\mathbf{x}_n,\mathbf{y}_n|\mathbf{y}_0^{n-1}) d\mathbf{x}_n \\ &\approx\; \sum_s \gamma_n(s) f_s(\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n') = \sum_s \omega_n(s). \quad (18) \end{aligned}$$

The conditional PDF $f_s(\mathbf{x}_n|\mathbf{y}_n,\hat{\bar{g}}_n',\hat{\ddot{g}}_n')$ can be shown to be a Gaussian distribution, e.g., [1]. The speech estimator (12) can then be obtained as

$$\hat{\mathbf{x}}_n = \frac{1}{\Omega_n}\sum_s \omega_n(s)(\hat{\bar{g}}_n \bar{\mathbf{D}}_{\bar{s}} + \epsilon \hat{\bar{g}}_n \ddot{\mathbf{D}}_{\bar{s}})(\hat{\bar{g}}_n \bar{\mathbf{D}}_{\bar{s}} + \hat{\ddot{g}}_n \ddot{\mathbf{D}}_{\bar{s}})^{-1}\mathbf{y}_n \quad (19)$$

The estimator (19) can be implemented efficiently in the frequency domain, e.g., [1], assuming that the covariance matrix of each state is circulant. The assumption is asymptotically valid, e.g., when the signal block length $K$ is large compared to the AR model order $\bar{p}$.

## 4. PARAMETER ESTIMATION

The estimation of the speech and noise HMM parameters is considered in this section. The proposed estimation algorithm consists of two parts: an iterative algorithm, generalizing the standard Baum-Welch algorithm, for off-line estimation of the time-invariant parameters, and an on-line algorithm to estimate the time-varying parameters.

### 4.1. Off-line parameter estimation

The parameters of the speech HMM, $\bar{\theta} = \{\bar{a}, \bar{\phi}, \bar{\psi}^2, \bar{\alpha}\}$, are estimated using recordings of clean speech utterances. Similarly to the Baum-Welch algorithm, we propose an iterative algorithm based on the expectation-maximization (EM) technique. We consider the missing data 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 auxiliary function $\mathcal{Q}(\theta|\theta^{(j-1)})$ of iteration $j$ is [6]

$$\mathcal{Q}(\theta|\hat{\theta}^{(j-1)}) = \mathrm{E}\left[\log(f(\mathbf{Z}_0^{N-1},\mathbf{x}_0^{N-1}|\theta))|\mathbf{x}_0^{N-1},\hat{\theta}^{(j-1)}\right] \quad (20)$$

Taking the first derivative with respect to the variables of interests and setting the resulting expression to zero, we obtain the update equations as:

$$\bar{\phi}_{\bar{s}}^{(j)} = \frac{1}{\bar{\Omega}}\sum_n \bar{\omega}_n(\bar{s}) \int \bar{g}_n' f_{\bar{s}}(\bar{g}_n'|\mathbf{x}_n,\hat{\theta}^{(j-1)}) d\bar{g}_n', \quad (21)$$

$$\bar{\psi}_{\bar{s}}^{2(j)} = \frac{1}{\bar{\Omega}}\sum_n \bar{\omega}_n(\bar{s}) \int (\bar{g}_n' - \bar{\phi}_{\bar{s}}^{(j)})^2 f_{\bar{s}}(\bar{g}_n'|\mathbf{x}_n,\hat{\theta}^{(j-1)}) d\bar{g}_n', \quad (22)$$

where $\bar{\Omega} = \sum_n \bar{\omega}_n(\bar{s})$, and $\bar{\omega}_n(\bar{s})$ is the state probability from the forward/backward calculation. The AR coefficients, $\bar{\alpha}^{(j)}$, are obtained from the estimated autocorrelation sequence, $\bar{r}_{\alpha_{\bar{s}}}^{(j)}$,

$$\bar{r}_{\alpha_{\bar{s}}}^{(j)}[i] = \frac{1}{\bar{\Omega}}\sum_n \bar{\omega}_n(\bar{s}) r_{x_n}[i] \int (\bar{g}_n)^{-1} f_{\bar{s}}(\bar{g}_n'|\mathbf{x}_n,\hat{\theta}^{(j-1)}) d\bar{g}_n', \quad (23)$$

$$r_{x_n}[i] = \sum_{j=0}^{K-i-1} x_n[j] x_n[j+i], \quad (24)$$

using the Levinson-Durbin recursion algorithm. The integrals in the update equations are difficult to solve analytically. Applying the 2nd order Taylor expansion of $f_{\bar{s}}(\bar{g}_n'|\mathbf{x}_n,\hat{\bar{\theta}}^{(j-1)})$ around the maximizing location, approximate solutions of the update equations can be obtained.

The training of the noise model is simplified by the state-independent gain model. The noise model is obtained using the standard Baum-Welch algorithm 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.

### 4.2. On-line parameter estimation

The time-varying parameters $\{\bar{q}_n, \ddot{\phi}_n\}$ are to be estimated on-line using the observed noisy data. Under the assumption that the parameters vary slowly, we apply the recursive EM algorithm [7] 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.

Following the derivations of [7], and applying the approximation (10), the update equations can be shown to be

$$\hat{\bar{\phi}}_n \;=\; \hat{\bar{\phi}}_{n-1} + \frac{1}{\Xi_n}\sum_s \frac{\omega_n(s)}{\Omega_n}\left(\hat{\bar{g}}_n' - \hat{\bar{\phi}}_{n-1}\right), \quad (25)$$

$$\hat{\bar{q}}_n \;=\; \hat{\bar{q}}_{n-1} + \frac{1}{\Xi_n'}\sum_s \frac{\omega_n(s)}{\Omega_n \bar{\psi}_{\bar{s}}^2}\left(\hat{\bar{g}}_n' - \bar{\phi}_{\bar{s}} - \hat{\bar{q}}_{n-1}\right), \quad (26)$$

where $\Xi_n = \sum_t \rho_{\ddot{\phi}}^{n-t}$ and $\Xi' = \sum_t \rho_{\bar{q}}^{n-t} \sum_s \omega_t(s)/(\Omega_t \bar{\psi}_{\bar{s}}^2)$, for $t = 0,\ldots,n$, and $\rho_{\ddot{\phi}}$ and $\rho_{\bar{q}}$ are two exponential forgetting factors.

## 5. EXPERIMENTS AND RESULTS

In this section, we describe the experimental setup and results from the objective and subjective evaluations. The evaluation is performed using 16 utterances, resampled to 8 kHz, from the core test set of the TIMIT database, one male and one female speaker from each of the eight dialects. The noise environments considered are: traffic noise, white Gaussian noise, babble noise (Noisex-92), and white-2, generated from the white noise, amplitude-modulated by

a sinusoid function (period of 2 s). The noisy signals are generated by adding the speech and noise for an input SNR of 10 dB. The utterances are processed concatenated.

The analysis is in blocks of 32 ms windowed using the Hann window. The synthesis is performed using 50% overlap-and-add. The HMMs are implemented using Gaussian mixture models (GMM) in each state. The speech HMM consists of eight states and 16 mixture components per states. The noise HMMs are pre-trained for each noise environment. Each noise HMM consists of three states and three mixture components per state. We assume prior knowledge of the type of the noise environment, such that the correct noise model is used in the enhancement. The forgetting factors for adapting the time-varying gain model parameters are experimentally set to $\rho_{\ddot{\phi}} = 0.9$ and $\rho_{\bar{q}} = 0.99$.

The scores from the evaluation of signal-to-noise ratio (SNR), segmental SNR (SSNR), and the Perceptual Evaluation of Speech of Speech Quality (PESQ) [8], are shown in Table 1. The residual noise level, $\epsilon$, is set to zero, corresponding to the MMSE waveform estimator. The reference methods are the AR-HMM based method (ref. A) [1], the gain-adaptive HMM based method (ref. B) [1, 2] and the method using HMM based noise adaptation (ref. C) [3]. The ref. A and B methods are implemented using the noise estimation algorithm based on minimum statistics [9]. The ref. C method uses the ideal VAD, estimated from the clean signal, for noise classification and gain adaptation. The measures are evaluated for each utterance separately and averaged over the utterances to get the final scores. The first utterance is removed from the averaging to avoid biased results due to initializations. The results from the objective evaluation show consistent improvements over the reference methods in all evaluated measures. The improvement is significant for non-stationary noise types, such as the traffic and white-2 noises.

| Type | Noisy | Sys. | Ref.A | Ref.B | Ref.C |
|---|---|---|---|---|---|
| | | | SNR (dB) | | |
| white | 10.00 | 15.17 | 14.94 | 14.26 | 15.00 |
| traffic | 10.10 | 14.59 | 12.87 | 13.31 | 13.04 |
| babble | 10.23 | 13.54 | 12.61 | 12.52 | 11.92 |
| white-2 | 10.03 | 14.99 | 11.65 | 11.43 | 13.31 |
| | | | SSNR (dB) | | |
| white | 0.55 | 8.07 | 7.47 | 5.27 | 7.87 |
| traffic | 1.67 | 8.17 | 6.00 | 6.21 | 6.26 |
| babble | 1.27 | 6.61 | 5.18 | 4.43 | 5.33 |
| white-2 | 2.19 | 8.42 | 4.84 | 4.36 | 6.53 |
| | | | PESQ (MOS) | | |
| white | 2.13 | 2.82 | 2.73 | 2.56 | 2.79 |
| traffic | 2.48 | 2.96 | 2.77 | 2.77 | 2.69 |
| babble | 2.49 | 2.74 | 2.63 | 2.65 | 2.43 |
| white-2 | 2.22 | 2.72 | 2.43 | 2.40 | 2.44 |

**Table 1**. Results from the objective evaluation (10 dB input SNR).

The perceptual quality of the proposed method is evaluated in a listening test similar to the Comparison Category Rating (CCR) [10] test. Ten listeners participated the test. The residual noise level, $\epsilon$, is experimentally set to 0.15. The method of [11] is applied as a post-processor to suppress the residual noise between spectral harmonics due to the AR modeling of speech. No other perceptual tuning is performed. The listening test is performed in comparison to the noise suppression module of the Enhanced Variable Rate Codec (EVRC) [12]. The results are shown in Table 2. Again, the results show a clear preference to the proposed method, particularly for the non-stationary noise types. We believe that the results can be further improved by additional perceptual tuning.

| white | traffic | babble | white-2 |
|---|---|---|---|
| 0.95±0.10 | 1.22±0.13 | 0.39±0.14 | 1.43±0.13 |

**Table 2**. Scores from the CCR listening test with 95% confidence intervals (10 dB input SNR). The scores are rates from -3 to 3 in step of one, corresponding to from *much worse* to *much better* [10]. Positive scores indicate a preference for the proposed method.

## 6. CONCLUSIONS

In this paper, a new HMM-based speech enhancement method using speech and noise gain modeling is presented. Through the introduction of HMM-based gain distributions, energy variation in speech and noise is explicitly modeled. The time-varying parameters of the gain models are estimated on-line using the recursive expectation maximization (EM) algorithm. The advantage of explicit gain modeling for speech enhancement is shown through both objective and subjective evaluations. The performance of the proposed method is consistently better than the reference methods, with significant improvement for non-stationary noise sources.

## 7. REFERENCES


[1] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, pp. 725–735, Apr. 1992.

[2] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Signal Processing*, vol. 40, pp. 1303–1316, Jun. 1992.

[3] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in non-stationary noise," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 445–455, Sep. 1998.

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

[5] 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, pp. 1077–1080, March 2005.

[6] 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.

[7] 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, pp. 2557–2573, Aug. 1993.

[8] "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.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 504–512, Jul. 2001.

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

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

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