On Noise Gain Estimation for HMM-based Speech Enhancement

David Y. Zhao and W. Bastiaan Kleijn

KTH (Royal Institute of Technology)
Department of Signals, Sensors and Systems
10044 Stockholm, Sweden
{ david.zhao, bastiaan.kleijn }@s3.kth.se

Abstract

To address the variation of noise level in non-stationary noise signals, we study the noise gain estimation for speech enhancement using hidden Markov models (HMM). We consider the noise gain as a stochastic process and we approximate the probability density function (PDF) to be log-normal distributed. The PDF parameters are estimated for every signal block using the past noisy signal blocks. The approximated PDF is then used in a Bayesian speech estimator minimizing the Bayes risk for a novel cost function, that allows for an adjustable level of residual noise. As a more computationally efficient alternative, we also derive the maximum likelihood (ML) estimator, assuming the noise gain to be a deterministic parameter. The performance of the proposed gain-adaptive methods are evaluated and compared to two reference methods. The experimental results show significant improvement under noise conditions with time-varying noise energy.

1. Introduction

Noise robustness is a crucial requirement for a wide range of speech applications, such as speech communication and speech recognition. However, the performance of traditional single-channel noise suppression techniques under non-stationary noise conditions is still unsatisfactory. The main issue is the noise estimation problem, which is shown to be particularly difficult for non-stationary noises.

Methods based on prior knowledge of speech and noise [1–4] have provided significant improvement for non-stationary noise environments. The noise statistics for a specific noise environment can be, as in [1], described using a hidden Markov model (HMM), with multiple states and mixture components, or as in [4], using a codebook, with multiple codebook entries. By having more than one distinct spectral shapes in the noise model, the enhancement systems can handle rapid changes of the noise spectrum within a specific noise environment.

In the enhancement, the estimation of the noise gain is important, as the noise energy level in the noisy environment is inherently unknown, time-varying, and in most natural cases, different from the noise energy level in the training. In this paper, we refer to the noise gain as the variable that compensates for the energy mismatch between the noise signals used in the training and in the enhancement. The noise gain can be used to model the change of the noise energy level due to, e.g., movement of the noise source. In the codebook-based methods [2–4], only the noise spectral shapes, represented by linear prediction (LP) coeffients, are modeled in the noise model. The noise gain, or variance in the auto-regressive (AR) modeling, is estimated instantaneously for each signal block. For the HMM-based method, a heuristic noise gain adaptation has been proposed [1], in which the adaptation is performed in speech pauses longer than 100 ms. As the adaptation is only performed in longer speech pauses, the method is not capable of reacting to fast changes in the noise energy during speech activity.

In this work, we propose two extensions to the HMM-based methods [1, 5], that improve noise gain estimation. First, we consider the logarithm of the noise gain as a stochastic first-order Gauss-Markov process. That is, the noise gain is assumed to be log-normal distributed. The mean and variance are estimated for each signal block using the past noisy observations. The approximated PDF is then used in a novel Bayesian speech estimator, that allows for an adjustable level of residual noise. Later, we derive a computationally simpler alternative, based on the maximum likelihood (ML) criterion. This work differs from the method of [4] in that it uses different assumptions for the speech/noise models and the noise gain. For instance, the speech and noise PDFs in [4] are modeled using codebooks under the high-rate assumption, and the solution for the variance calculation assumes small modeling errors.

2. Signal model

We consider a noise suppression system for independent additive noise. The noisy signal is processed on a block-by-block basis in the frequency domain using the fast Fourier transform (FFT). The frequency domain representation of the noisy signal at block n is modeled as

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{w}_n, \tag{1}$$

where $\mathbf{y}_n = [y_n[0], \ldots, y_n[L-1]]^T$, $\mathbf{x}_n = [x_n[0], \ldots, x_n[L-1]]^T$ and $\mathbf{w}_n = [w_n[0], \ldots, w_n[L-1]]^T$ are the complex spectra of noisy, clean speech and noise, respectively, for frequency channels $0 \leq l < L$. Furthermore, we assume that the noise $\mathbf{w}_n$ can be decomposed as $\mathbf{w}_n = \sqrt{g_{w_n}}\ddot{\mathbf{w}}_n$, where g_{w_n} denotes the noise gain variable, and $\ddot{\mathbf{w}}$ is the gain-normalized noise signal block, whose statistics is modeled using an HMM.

As in [1], each output probability for a given state is modeled using a Gaussian mixture model (GMM). For the noise model, $\ddot{\boldsymbol{\pi}}$ denotes the initial state probabilities, $\ddot{\mathbf{a}} = [\ddot{a}_{st}]$ denotes the state transition probability matrix from state s to t and $\ddot{\boldsymbol{\rho}} = \{\ddot{\rho}_{i|s}\}$ denotes the mixture weights for a given state s. We define the component PDF for the i'th mixture component of the state s as

$$f_{i|s}(\mathbf{x}_n) = \prod_{k=0}^{K-1} \frac{1}{\sqrt{2\pi\ddot{c}^2_{i|s}[k]}} \exp\left(-\frac{1}{2}\frac{E^2_{x_n}[k]}{\ddot{c}^2_{i|s}[k]}\right), \tag{2}$$

where $E^2_{x_n}[k] = \sum_{l=\text{low}(k)}^{\text{high}(k)} |x_n[l]|^2$ is the speech energy in the subband $0 \leq k < K$, and $\text{low}(k)$ and $\text{high}(k)$ provide the

frequency boundaries of the subband. The corresponding parameters for the speech model are denoted using bar ¯ instead of double dots ¨.

The component model can be motivated by the filter-bank point-of-view, where the signal power spectrum is estimated in subbands by a filter-bank of band-pass filters. The subband spectrum of a particular sound is assumed to be a Gaussian (2) with zero-mean and diagonal covariance matrix. The mixture components model multiple spectra of various classes of sounds. Compared to the frequency domain model in [6], this approach has the advantage of a reduced parameter space, which leads to lower computational and memory requirements. The structure also allows for unequal frequency bands, such that a frequency resolution consistent with the human auditory system may be used. In our implementation, the frequency bands are defined as in the noise suppression system of the enhanced variable rate codec (EVRC-NS) [7].

The HMM parameters are obtained by training using the Baum-Welch algorithm and the expectation-maximization (EM) algorithm, from clean speech and noise signals. To simplify the notation, we write $\mathbf{y}_0^n = \{\mathbf{y}_\tau, \tau = 0, \ldots, n\}$, and $f(\mathbf{x})$ instead of $f_{\mathbf{X}}(\mathbf{x})$ in all PDFs. The dependency of the mixture component index on the state is also dropped, e.g., we write b_i instead of $b_{i|s}$.

3. Speech estimation

In this section, we derive a speech spectrum estimator based on a criterion that leaves an adjustable level of residual noise in the enhanced speech. We consider the Bayesian estimator

$$\hat{\mathbf{x}}_n = \arg\min_{\tilde{\mathbf{x}}_n} \mathrm{E}[C(\mathbf{X}_n, \mathbf{W}_n, \tilde{\mathbf{x}}_n)|\mathbf{Y}_0^n = \mathbf{y}_0^n], \quad (3)$$

minimizing the Bayes risk for the cost function

$$C(\mathbf{x}_n, \mathbf{w}_n, \tilde{\mathbf{x}}_n) = |(\mathbf{x}_n + \epsilon\mathbf{w}_n) - \tilde{\mathbf{x}}_n|^2, \quad (4)$$

where $|\cdot|$ denotes the complex vector norm and $0 \leq \epsilon \ll 1$ 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. Unlike the constrained optimization approach in [8], which is limited to linear estimators, the proposed Bayesian estimator can be nonlinear as well. The residual noise level ϵ can be extended to be time- and frequency dependent, to introduce perceptual shaping of the noise.

To solve the speech estimator (3), we first assume that the noise gain g_{w_n} is given. The PDF of the noisy signal $f(\mathbf{y}_n|g_{w_n})$ is an HMM [5] composed by combining of the speech and noise models. We use s_n to denote a composite state at the n'th block, which consists of the combination of a speech model state $\bar{s}_n$ and a noise model state $\ddot{s}_n$. The covariance matrix of the ij'th mixture component of the composite state s_n has $\bar{c}_i^2[k] + g_{w_n}\ddot{c}_j^2[k]$ on the diagonal.

Using the Markov assumption, the posterior speech PDF given the noisy observations and noise gain is

$$f(\mathbf{x}_n|\mathbf{y}_0^n, g_{w_n}) = \frac{\sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j f_{ij}(\mathbf{y}_n|g_{w_n}) f_{ij}(\mathbf{x}_n|\mathbf{y}_n, g_{w_n})}{f(\mathbf{y}_n|\mathbf{y}_0^{n-1}, g_{w_n})}, \quad (5)$$

where γ_n is the probability of being in the composite state s_n given all past noisy observations up to block $n-1$,

$$\begin{aligned} \gamma_n &= p(s_n|\mathbf{y}_0^{n-1}) \\ &= \sum_{s_{n-1}} p(s_{n-1}|\mathbf{y}_0^{n-1}) a_{s_{n-1}s_n}, \end{aligned} \quad (6)$$

where $p(s_{n-1}|\mathbf{y}_0^{n-1})$ is the scaled forward probability. The posterior noise PDF $f(\mathbf{w}_n|\mathbf{y}_0^n, g_{w_n})$ has the same structure as (5), with $\mathbf{x}_n$ replaced by $\mathbf{w}_n$.

The proposed speech estimator (3) becomes

$$\hat{\mathbf{x}}_n = \frac{\sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j f_{ij}(\mathbf{y}_n|g_{w_n}) \boldsymbol{\mu}_{ij}(g_{w_n})}{f(\mathbf{y}_n|\mathbf{y}_0^{n-1}, g_{w_n})}, \quad (7)$$

where for the l'th frequency bin,

$$\mu_{ij}(g_{w_n})[l] = \frac{\bar{c}_i^2[k] + \epsilon g_{w_n}\ddot{c}_j^2[k]}{\bar{c}_i^2[k] + g_{w_n}\ddot{c}_j^2[k]} y_n[l], \quad (8)$$

for the subband k fulfilling $\mathrm{low}(k) \leq l \leq \mathrm{high}(k)$. The proposed speech estimator (7) is a weighted sum of filters (8), and is nonlinear due to the signal dependent weights. The individual filter (8) differs from the Wiener filter by the additional noise term in the numerator. The amount of allowed residual noise is adjusted by ϵ. When $\epsilon = 0$, the filter converges to the Wiener filter. When $\epsilon = 1$, the filter is one, which does not perform any noise reduction. A particularly interesting difference between the filter (8) and the Wiener filter is that when there is no speech, the Wiener filter is zero while the filter (8) becomes ϵ. This lower bound on the noise attenuation is commonly used in speech enhancement to reduce the processing artifact [9], but was previously motivated as a correction term.

4. Noise gain estimation

This section discusses the algorithms for the noise gain estimation. We first derive a method based on the assumption that g_{w_n} is a stochastic process. Later, we propose a computationally simpler method using the maximum likelihood criterion.

4.1. The stochastic approach

In this section, we assume g_{w_n} to be a stochastic process and we assume that the PDF of $g'_{w_n} = \log g_{w_n}$ given the past noisy observations is a Gaussian, $f(g'_{w_n}|\mathbf{y}_0^{n-1}) \approx \mathcal{N}(\phi_n, \psi_n)$. To model the time-varying noise energy level, we assume that g'_{w_n} is a first-order Gauss-Markov process

$$g'_{w_n} = g'_{w_{n-1}} + u_n, \quad (9)$$

where u_n is a white Gaussian process with zero mean and variance σ_u^2. σ_u^2 models how fast the noise gain changes. For simplicity, we set σ_u^2 to be a constant for all noise types.

The posterior speech PDF can be reformulated as an integration over all possible realizations of g'_{w_n}

$$\begin{aligned} f(\mathbf{x}_n|\mathbf{y}_0^n) &= \int f(\mathbf{x}_n|\mathbf{y}_0^n, g'_{w_n}) f(g'_{w_n}|\mathbf{y}_0^n) dg'_{w_n} \\ &= \frac{1}{B} \sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j \int \xi_{ij}(g'_{w_n}) f_{ij}(\mathbf{x}_n|\mathbf{y}_n, g_{w_n}) dg'_{w_n}, \end{aligned} \quad (10)$$

for $\xi_{ij}(g'_{w_n}) = f_{ij}(\mathbf{y}_n|g'_{w_n}) f(g'_{w_n}|\mathbf{y}_0^{n-1})$ and B ensures that the PDF integrates to one. The speech estimator (7), assuming stochastic noise gain becomes

$$\hat{\mathbf{x}}_n^A = \frac{1}{B} \sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j \int \xi_{ij}(g'_{w_n}) \boldsymbol{\mu}_{ij}(g'_{w_n}) dg'_{w_n}. \quad (11)$$

The integral (11) can be evaluated using numerical integration algorithms. It has been shown in [4] that the component likelihood function $f_{ij}(\mathbf{y}_n|g_{w_n})$ decays rapidly from its mode. Thus, we make an approximation by applying the 2nd order Taylor expansion of $\log \xi_{ij}(g'_{w_n})$ around its mode $\hat{g}'_{w_{n,ij}} = \arg\max_{g'_{w_n}} \log \xi_{ij}(g'_{w_n})$,

$$\log \xi_{ij}(g'_{w_n}) \approx \log \xi_{ij}(\hat{g}'_{w_{n,ij}}) - \frac{1}{2A_{ij}^2}(g'_{w_n} - \hat{g}'_{w_{n,ij}})^2, \quad (12)$$

where $A_{ij}^2 = -\left(\frac{\partial^2 \log \xi_{ij}(g'_{w_n})}{\partial {g'_{w_n}}^2}\right)^{-1}$. To obtain the mode $\hat{g}'_{w_{n,ij}}$, we use the Newton-Raphson algorithm, initialized using the expected value ϕ_n. As the noise gain is typically slowly varying for two consecutive blocks, the method usually converges within a few iterations.

To further simplify the evaluation of (11), we approximate $\boldsymbol{\mu}_{ij}(g'_{w_n}) \approx \boldsymbol{\mu}_{ij}(\hat{g}'_{w_{n,ij}})$, and integrate only $\xi_{ij}(g'_{w_n})$:

$$\hat{\mathbf{x}}_n^A \approx \frac{1}{B} \sum_{sn,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j A_{ij} \xi_{ij}(\hat{g}'_{w_{n,ij}}) \boldsymbol{\mu}_{ij}(\hat{g}'_{w_{n,ij}}). \quad (13)$$

We now discuss how to obtain the parameters of $f(g'_{w_{n+1}}|\mathbf{y}_0^n)$. Using Bayes' rule, it can be shown that

$$f(g'_{w_n}|\mathbf{y}_0^n) = \frac{1}{B} \sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j \xi_{ij}(g'_{w_n}), \quad (14)$$

and $f(g'_{w_{n+1}}|\mathbf{y}_0^n)$ can be calculated using (9). To keep the problem tractable, we approximate (14) by a Gaussian, thus requiring only first and second order statistics. The parameters of $f(g'_{w_{n+1}}|\mathbf{y}_0^n) \approx \mathcal{N}(\phi_{n+1}, \psi_{n+1})$ are obtained by

$$\hat{\phi}_{n+1} \approx \frac{1}{B} \sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j A_{ij} \xi_{ij}(\hat{g}'_{w_{n,ij}}) \hat{g}'_{w_{n,ij}} \quad (15)$$

$$\hat{\psi}_{n+1} \approx \sigma_u^2 + \frac{1}{B} \sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j A_{ij} \xi_{ij}(\hat{g}'_{w_{n,ij}}) \cdot \left(A_{ij}^2 + (\hat{g}'_{w_{n,ij}} - \hat{\phi}_{n+1})^2\right). \quad (16)$$

To summarize, the method approximates the noise gain PDF using the log-normal distribution. The PDF parameters are estimated on a block-by-block basis using (15) and (16). Using the noise gain PDF, the Bayesian speech estimator (3) can be evaluated using (13). We refer to this method as system A in the experiments.

4.2. The maximum-likelihood approach

In this section, we present a computationally simpler noise gain estimator based on the ML estimation technique. To reduce the estimation variance, we make the assumption that the noise energy level is relatively constant over a longer period, such that we can utilize multiple noisy blocks for the noise gain estimation. The ML noise gain estimator is then defined as

$$\hat{g}_{w_n} = \arg\max_{g_{w_n}} \sum_{m=n-M}^{n+M} \log f(\mathbf{y}_m|\mathbf{y}_0^{m-1}, g_{w_n}), \quad (17)$$

where the optimization is over $2M+1$ blocks. The log-likelihood function of the n'th block is

$$\log f(\mathbf{y}_n|\mathbf{y}_0^{n-1}, g_{w_n}) = \log \frac{1}{B} \sum_{s_n,i,j} \gamma_n \bar{\rho}_i \ddot{\rho}_j f_{ij}(\mathbf{y}_n|g_{w_n}) \approx \log\left(\max_{s_n,i,j} \frac{\gamma_n \bar{\rho}_i \ddot{\rho}_j}{B} f_{ij}(\mathbf{y}_n|g_{w_n})\right), \quad (18)$$

where the log-of-a-sum is approximated using the logarithm of the largest term in the summation. The optimization problem can be solved numerically, and we propose a solution based on stochastic approximation [10]. The stochastic approximation approach can be implemented without any additional delay. Moreover, it has a reduced computational complexity, as the gradient function is evaluated only once for each block. To ensure $\hat{g}_{w_n}$ to be nonnegative, and to account for the human perception of loudness which is approximately logarithmic, the gradient steps are evaluated in the log domain. The noise gain estimate $\hat{g}_{w_n}$ is adapted once per block

$$\hat{g}'_{w_n} \approx \hat{g}'_{w_{n-1}} + \Delta[n] \frac{\partial \log f_{ij_{\max}}(\mathbf{y}_n|g_{w_n})}{\partial g'_{w_n}} \quad (19)$$

$$\hat{g}_{w_n} = \exp \hat{g}'_{w_n}, \quad (20)$$

where $ij_{\max}$ in (19) is the index of the most likely mixture component, evaluated using the previous estimate $\hat{g}_{w_{n-1}}$. The step-size $\Delta[n]$ controls the rate of the noise gain adaptation, and is set to a constant Δ. The speech spectrum estimator (7) can then be evaluated for $g_{w_n} = \hat{g}_{w_n}$. This method is referred to as system B in the experiments.

5. Experiments and results

Systems A and B are implemented for 8 kHz sampled speech. The FFT based analysis and synthesis follow the structure of the EVRC-NS system [7]. In the experiments, the step size Δ is set to 0.015 and the noise variance σ_u^2 in the stochastic gain model is set to 0.001. The parameters are set experimentally to allow a relatively large change of the noise gain, and at the same time to be reasonably stable when the noise gain is constant. As the gain adaptation is performed in the log domain, the parameters are not sensitive to the absolute noise energy level. The residual noise level ϵ is set to 0.1.

The training data of the speech model consists of 128 clean utterances from the training set of the TIMIT database down-sampled to 8kHz, with 50% female and 50% male speakers. The sentences are normalized on a per utterance basis. The speech HMM has 16 states and 8 mixture components in each state. We considered three different noisy environments in the evaluation: traffic noise, which was recorded on the side of a busy freeway, white Gaussian noise, and the babble noise from the Noisex-92 database. One minute of the recorded noise signal of each type was used in the training. Each noise model contains 3 states and 3 mixture components per state. The training data are energy normalized in blocks of 200 ms with 50% overlap to remove the long-term energy information. The noise signals used in the training were not used in the evaluation.

In the enhancement, we assume prior knowledge on the type of the noise environment, such that the correct noise model is used. We use one additional noise signal, white-2, which is created artificially by modulating the amplitude of a white noise signal 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. In the experiments, the sinusoid has a period of two seconds, and the maximum amplitude modulation is four times higher then the minimum one.

For comparison, we implemented two reference systems. Reference method C applies noise gain adaptation during detected speech pauses as described in [1]. Only speech pauses longer than 100 ms are used to avoid confusion with low energy speech. An ideal speech pause detector using the clean signal

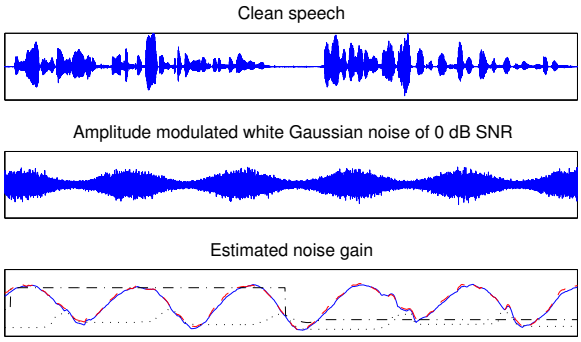

Figure 1: Comparison of different noise gain estimators. The solid line is the expected gain of sys. A. The dashed and dash-dotted lines are the estimated gain of sys. B and ref. C, respectively. The dotted line is the estimated noise energy of ref. D (minimum statistics method).

is used in the implementation of the reference method, which gives the reference method an advantage. To keep the comparison fair, the same speech and noise models as the proposed methods are used in reference C. Reference D is a spectral subtraction method [11] without using any prior speech or noise models. The noise power spectrum estimate is obtained using the minimum statistics algorithm from [12]. The residual noise levels of the reference systems are set to ϵ.

Figure 1 demonstrates one typical realization of different noise gain estimation strategies for the white-2 noise. Reference system C (dash-doted) updates the noise gain only during longer speech pauses, and is not capable of reacting to noise energy changes during speech activity. For reference system D, energy of the estimated noise is plotted (dotted). The minimum statistics method has an inherent delay of at least one buffer length, which is clearly visible from the figure. Both the proposed methods A (solid) and B (dashed) are capable of following the noise energy changes, which is a significant advantage over the reference systems.

We selected 16 utterances, eight male speakers and eight female speakers, all from the TIMIT database, for the object measure evaluation. The utterances are selected from the recommended core test set, such that neither the speakers nor the utterances were part of the training set for the speech HMM. The total length of the evaluation utterances is about one minute. The performance is compared in terms of the log spectral distortion (SD) and the segmental signal-to-noise ratio (SSNR). The first two sentences are removed from the objective measure evaluation to allow the necessary initialization for, e.g., the minimum statistics method.

In the comparison, noisy speech was synthesized with an input SNR of 10 dB. The objective measures are listed in figure 2. As expected, reference system D, which does not use a speech or a noise model, has a relatively poor performance, and is not comparable with other methods. Both proposed methods A and B are better than reference system C, for all the noise cases except the white Gaussian noise, where the performances are similar. The improvement is higher for the traffic and white-2 noises, which contain more rapid changes of the energy level. Thus, we conclude that the proposed extensions improve the performance of HMM-based enhancement under non-stationary noise conditions. System A has a slight advantage over system B, e.g., for the white-2 noise. The improvement can be explained by the more accurate noise gain modeling, and the soft-decision estimation approach. System B is, on the other hand, a simpler method to implement and has a lower computational complexity.

Dist.	Noise	Noisy	A	B	C	D
SD	traffic	3.47	2.51	2.53	2.74	2.98
(dB)	white	4.50	3.50	3.50	3.46	3.91
	white-2	4.13	3.38	3.43	3.62	3.90
	babble	3.15	2.54	2.54	2.59	2.98
SSNR	traffic	1.89	8.06	8.01	7.17	6.00
(dB)	white	0.58	6.64	6.64	6.59	5.22
	white-2	2.33	7.01	6.86	6.48	4.57
	babble	1.53	6.53	6.50	5.77	4.77

Figure 2: Experimental results for noisy signals of 10 dB input SNR.

6. Conclusions

We have presented two related methods to estimate the noise gain for HMM-based speech enhancement. The proposed methods allow faster adaptation to noise energy changes and are more suitable for suppression of non-stationary noises. The performance of the method A), based on a stochastic model, is better than the method B), based on the maximum likelihood criterion. However, method B requires less computations, and is more suitable for real-time implementations. Also, the gain estimation algorithms can be extended to adapt the speech model.

7. References

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

[2] 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.

[3] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Speech and Audio Processing*, accepted for publication.

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

[5] 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.

[6] 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.

[7] TIA/EIA/IS-127, *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, July 1996.

[8] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

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

[11] 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.

[12] 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.