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 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 estimated [1], described using a hidden Markov model (HMM), or modeled using a Gaussian mixture model (GMM). For 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 signal 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) coefficients, 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

\[
y_n = x_n + w_n,
\]

where \( y_n = [y_n[0], \ldots, y_n[L-1]]^T \), \( x_n = [x_n[0], \ldots, x_n[L-1]]^T \) and \( w_n = [w_n[0], \ldots, w_n[L-1]]^T \) are the complex spectra of noisy, clean speech and noise, respectively, for frequency channels \( 0 \leq l < L \). Furthermore, we assume that the noise \( w_n \) can be decomposed as \( w_n = \sqrt{\text{var}}(w) \), where \( \text{var}(w) \) denotes the noise gain variable, and \( \sqrt{\text{var}}(w) \) 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, \( \pi \) denotes the initial state probabilities, \( \tilde{\alpha} = [\tilde{\alpha}_t] \) denotes the state transition probability matrix from state \( s \) to \( t \) and \( \tilde{\rho} = \{\tilde{\rho}_{is}\} \) 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

\[
f_{s,i}(x_n) = \prod_{k=0}^{K-1} \frac{1}{\sqrt{2\pi c_{i,k}^2}} \exp\left( -\frac{1}{2} \frac{E_{x,n}^2[k]}{c_{i,k}^2} \right),
\]

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

As in [1], 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.

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.
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 \( y^0_n = \{ y_n, \tau = 0, \ldots, n \} \), and \( f(x) \) instead of \( fx(x) \) 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 estimated spectrum components model multiple spectra of various classes of sounds. The mixture component index on the state is also dropped, e.g., we write \( x_{s_n} \) instead of \( x_{s_n|s} \).

\[
\beta_{n} = \arg \min_{x_{s_n}} E[C(\beta_{n}, w_{n}, x_{s_n})|y^{0}_n = y^{0}_n], \tag{3}
\]

minimizing the Bayes risk for the cost function

\[
C(\beta_{n}, w_{n}, x_{s_n}) = |x_{s_n} + \epsilon w_{n} - \beta_{n}|^2, \tag{4}
\]

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 \(\epsilon\) can be extended to be time- and frequency dependent, to introduce perceptual shaping of the noise.

To solve the speech estimation (3), we first assume that the noise gain \(g_{w_{n}}\) is given. The PDF of the noisy signal \(f(y_{n}|g_{w_{n}})\) 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 \(s_{n}\) and a noise model state \(s_{n}\). The covariance matrix of the \(ij\)th mixture component of the composite state \(s_{n}\) has \(\sigma_{n}^{2}[k] + g_{w_{n}}\sigma_{n}^{2}[k]\) on the diagonal.

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

\[
f(x_{n}|y^{0}_n, g_{w_{n}}) = \sum_{s_{n}} \gamma_{n} s_{n} \bar{p}_{s_{n}} f_{x_{n}}(y_{n}|g_{w_{n}}) f(y^{0}_n|y^{0}_{n-1}, g_{w_{n}}) \tag{5}
\]

where \(\gamma_{n}\) is the probability of being in the composite state \(s_{n}\) given all past noisy observations up to block \(n-1\),

\[
\gamma_{n} = p(s_{n}|y^{0}_{n-1}) = \sum_{s_{n-1}} p(s_{n-1}|y^{0}_{n-1}) a_{s_{n-1}, s_{n}}, \tag{6}
\]

where \(p(s_{n-1}|y^{0}_{n-1})\) is the scaled forward probability. The posterior noise PDF \(f(w_{n}|g_{w_{n}}, y_{n})\) has the same structure as (5), with \(x_{n}\) replaced by \(w_{n}\).

The proposed speech estimator (3) becomes

\[
x_{n} = \sum_{s_{n}} \gamma_{n} s_{n} \bar{p}_{s_{n}} f_{x_{n}}(y_{n}|g_{w_{n}}) \mu_{ij}(g_{w_{n}}) f(y^{0}_n|y^{0}_{n-1}, g_{w_{n}}), \tag{7}
\]

where for the \(l\)th frequency bin,

\[
\mu_{ij}(g_{w_{n}})[l] = \frac{\epsilon_{ij}^{2}[k] + g_{w_{n}} \epsilon_{ij}^{2}[k]}{\sigma_{n}^{2}[k] + g_{w_{n}} \sigma_{n}^{2}[k]} y_{n}[l], \tag{8}
\]

for the subband \(k\) fulfilling low\((k) \leq l \leq \text{high}(k)\). 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 \(\epsilon\). 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 \(\epsilon\). 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, \(f(g_{w_{n}}|y^{0}_{n-1}) \approx N(\phi_{w_{n}}, \psi_{w_{n}})\). To model the time-varying noise energy level, we assume that \(g_{w_{n}}\) is a first-order Gauss-Markov process

\[
g_{w_{n}} = g_{w_{n-1}} + u_{n}, \tag{9}
\]

where \(u_{n}\) is a white Gaussian process with zero mean and variance \(\sigma_{w_{n}}^{2}, \sigma_{w_{n}}^{2}\) models how fast the noise gain changes. For simplicity, we set \(\sigma_{w_{n}}^{2}\) to be a constant for all noise types.

The posterior speech PDF can be reformulated as an integration over all possible realizations of \(g_{w_{n}}\).

\[
f(x_{n}|y_{n}) = \int f(x_{n}|y_{n}, g_{w_{n}}) f(g_{w_{n}}|y^{0}_{n}) dg_{w_{n}} \tag{10}
\]

\[
= \frac{1}{B} \sum_{s_{n}, i, j} \gamma_{n} s_{n} \bar{p}_{s_{n}} f_{x_{n}}(y_{n}|g_{w_{n}}) f(y^{0}_n|y^{0}_{n-1}, g_{w_{n}}) f(g_{w_{n}}|y^{0}_{n-1}) \tag{11}
\]

for \(\xi_{ij}(g_{w_{n}}) = f_{x_{n}}(y_{n}|g_{w_{n}}) f(g_{w_{n}}|y^{0}_{n-1})\) and \(B\) ensures that the PDF integrates to one. The speech estimator (7), assuming stochastic noise gain becomes

\[
x_{n}^{A} = \frac{1}{B} \sum_{s_{n}, i, j} \gamma_{n} s_{n} \bar{p}_{s_{n}} \int \xi_{ij}(g_{w_{n}}) \mu_{ij}(g_{w_{n}}) dg_{w_{n}}. \tag{11}
\]
The integral (11) can be evaluated using numerical integration algorithms. It has been shown in [4] that the component likelihood function $f_{ij}(y_{n}|g_{w_n})$ 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 $\hat{g}_{w_n,ij} = \arg\max_{g_{w_n}} \log \xi_{ij}(g_{w_n})$,

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

where $A_{ij}^2 = - \left( \frac{\partial^2 \log \xi_{ij}(g_{w_n})}{\partial g_{w_n}^2} \right)_{\hat{g}_{w_n,ij}}^{-1}$. To obtain the mode $\hat{g}_{w_n,ij}$, we use the Newton-Raphson algorithm, initialized using the expected value $\hat{g}_{w_n,ij}$. 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 $\mu_{ij}(g_{w_n}) \approx \mu_{ij}(\hat{g}_{w_n,ij})$, and integrate only $\xi_{ij}(g_{w_n})$:

$$\hat{x}_n \approx \frac{1}{B} \sum_{s_n,t_n} \gamma_n \hat{p}_n \hat{p}_j A_{ij} \xi_{ij}(\hat{g}_{w_n,ij}), \quad (13)$$

We now discuss how to obtain the parameters of $f(\hat{g}_{w_n+1}|y_0^s)$. Using Bayes’ rule, it can be shown that

$$f(\hat{g}_{w_n}|y_0^s) = \frac{1}{B} \sum_{s_n,t_n} \gamma_n \hat{p}_n \hat{p}_j A_{ij} \xi_{ij}(\hat{g}_{w_n,ij}), \quad (14)$$

and $f(g_{w,n+1}|y_0^s)$ 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 $f(g_{w,n+1}|y_0^s)$ are obtained by

$$\hat{\phi}_{n+1} \approx \frac{1}{B} \sum_{s_n,t_n} \gamma_n \hat{p}_n \hat{p}_j A_{ij} \xi_{ij}(\hat{g}_{w_n,ij}) \hat{g}_{w,n,ij} \quad (15)$$

$$\hat{\psi}_{n+1} \approx \sigma_n^2 + \frac{1}{B} \sum_{s_n,t_n} \gamma_n \hat{p}_n \hat{p}_j A_{ij} \xi_{ij}(\hat{g}_{w_n,ij}) \left( A_{ij}^2 + (\hat{g}_{w,n,ij} - \hat{\phi}_{n+1})^2 \right). \quad (16)$$

To summarize, the method approximates the noise gain PDF using the log-normal distribution. The 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 signals for the noise gain estimation. The ML noise gain estimator is then defined as

$$\hat{g}_{w_n} = \arg\max_{g_{w_n}} \sum_{m=n-M}^{n+M} \log f(y_m|y_0^{m-1}, g_{w_n}). \quad (17)$$

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

$$\log f(y_n|y_0^{n-1}, g_{w_n}) = \log \frac{1}{B} \sum_{s_n,t_n} \gamma_n \hat{p}_n \hat{p}_j f_{ij}(y_n|g_{w_n}) \approx \log \left( \max_{s_n,t_n} \frac{\gamma_n \hat{p}_n \hat{p}_j}{B} f_{ij}(y_n|g_{w_n}) \right). \quad (18)$$

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

$$\hat{g}_{w_n} = \frac{\hat{g}_{w_{n-1}}}{\exp \frac{f_{ij}[\log f_{ij}(y_n|g_{w_n})]}{\partial g_{w_n}}}, \quad (19)$$

where $i_{\max}$ in (19) is the index of the most likely mixture component, evaluated using the previous estimate $\hat{g}_{w_n}$. The step-size $\Delta[n]$ controls the rate of the noise gain adaptation, and is set to a constant $\Delta$. 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 $\Delta$ is set to 0.015 and the noise variance $\sigma_n^2$ 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 $\epsilon$ 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 8 kHz, 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 than 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
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


