# ADAPTIVE SUPPRESSION OF NON-STATIONARY NOISE BY USING THE VARIATIONAL BAYESIAN METHOD

*Takuya Yoshioka and Masato Miyoshi*

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

## ABSTRACT

This paper proposes an adaptive noise suppression method for non-stationary noise based on the Bayesian estimation method. The following conditions are assumed: (1) Speech and noise samples are statistically independent, and they follow auto-regressive (AR) processes. (2) The prior distribution of the parameters of the noise AR model of a current frame is identical to the posterior distribution of those parameters calculated in the previous frame. Under these conditions, the proposed method approximates the joint posterior distribution of the AR model parameters and the speech samples by using the variational Bayesian method. Furthermore, we describe an efficient implementation by assuming that all involved covariance matrices have the Toeplitz structure. The proposed method was tested on real speech and noise signals and compared with other noise suppression methods.

***Index Terms***— Noise suppression, Bayesian estimation, variational Bayesian method, auto-regressive process

## 1. INTRODUCTION

The task of suppressing noise in degraded speech signals observed at microphones has been of great interest for decades. Noise suppressors working in realistic environments must be able to cope with non-stationary noise. Traditional noise suppression methods that are based on voice activity detectors (VADs) are unable to suppress non-stationary noise very much because they update noise estimates only during periods when speech is absent. As a result, adaptive noise suppression methods that can update noise estimates even during speech activity have been studied in the last decade.

One approach to adaptive noise suppression is based on statistical models of speech and noise [1, 2, 3]. The statistical model based approach estimates the model parameters or the speech signals from the observed noisy speech signals by using the maximum likelihood (ML) estimation method, the Bayesian estimation method, or the like. Auto-regressive (AR) models are usually employed for both the speech and noise models. The statistical model based approach seems to produce less distorted speech signals than the minimum statistics based approach [4]. This may be partly because the minimum statistics based approach estimates noise power spectral components independently for each frequency bin, and partly because this approach involves spectral smoothing, which makes speech signals reverberant.

The statistical model based methods are further classified into two categories. One requires the speech and noise models to be trained in advance [1, 2]. The methods in this category fail to recover the speech signals contaminated by noise signals out of the training data. The other estimates both the model parameters and the speech signals online without the need for prior training of the models [3]. However, the methods in this category are, as far as we know, based on the ML estimation method, which is in principle inferior to the Bayesian estimation method.

In this paper, we propose an adaptive noise suppression method that performs the Bayesian estimation of the model parameters and the speech signals. That is, the proposed method calculates the joint posterior distribution of the AR parameters and the speech signals on a frame-by-frame basis. The proposed method assumes that the speech and noise signal samples follow auto-regressive (AR) processes. Moreover, the prior distribution of the parameters of the noise AR model of a current frame is assumed to be identical to the posterior distribution of those parameters calculated in the previous frame. Since it is difficult to derive the exact posterior distribution, in reality, the proposed method approximates the true posterior distribution by using the variational Bayesian method [5]. Furthermore, we describe an efficient implementation of the proposed method, resulting in only a slightly higher computational cost than the ML estimation method. Note that the proposed method can be distinguished from the method described in [6], which is also based on the variational Bayesian method, in that the proposed method does not assume the noise to be stationary.

## 2. BAYESIAN APPROACH TO ADAPTIVE NOISE SUPPRESSION

### 2.1. Task formulation of adaptive noise suppression

Let $s_t(n)$, $v_t(n)$, and $x_t(n)$ denote a speech sample, a noise sample, and an observed noisy speech sample, respectively, in the $t$-th short time frame of length $N$. The vector of the noisy speech samples contained in the $t$-th frame, $\boldsymbol{x}_t = [x_t(N), \cdots, x_t(1)]^T$, is represented by

$$\boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{v}_t, \tag{1}$$

where $\boldsymbol{s}_t = [s_t(N), \cdots, s_t(1)]^T$, $\boldsymbol{v}_t = [v_t(N), \cdots, v_t(1)]^T$, and superscript $T$ is the transpose operator. The noise suppression task addressed in this paper is to estimate the speech samples of the $t$-th frame, $\boldsymbol{s}_t$, from the observed noisy speech samples up to the $t$-th frame, $\boldsymbol{x}_1^t = \{\boldsymbol{x}_u\}_{1 \le u \le t}$.

If we are to solve the noise suppression task, we must introduce some constraints on the speech and noise. In this paper, we assume the following conditions.

(a1) The speech sample $s_t(n)$ follows an auto-regressive (AR) process of order $P$. Thus, the probability density function (PDF) of $s_t(n)$ conditioned on the past $P$ samples $\boldsymbol{s}_t^P(n-1) = [s_t(n-1), \cdots, s_t(n-P)]^T$, the regression coefficient vector $\boldsymbol{a}_t = [a_{t,1}, \cdots, a_{t,P}]^T$, and the innovation variance

$\sigma_t^2$ is given by

$$p(s_t(n)|\boldsymbol{s}_t^P(n-1), \boldsymbol{a}_t, \sigma_t^2) = \mathcal{N}\{s_t(n); \boldsymbol{s}_t^P(n-1)^T \boldsymbol{a}_t, \sigma_t^2\}, \quad (2)$$

where $\mathcal{N}\{\boldsymbol{x}; \boldsymbol{\mu}, \Sigma\}$ denotes the PDF of a random variable $\boldsymbol{x}$ following the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$. We hereafter represent the parameters of this speech model by $\Phi_t = \{\boldsymbol{a}_t, \sigma_t^2\}$.

(a2) The noise sample $v_t(n)$ follows an AR process of order $Q$. Thus, the PDF of $v_t(n)$ conditioned on the past $Q$ samples $\boldsymbol{v}_t^Q(n-1) = [v_t(n-1), \cdots, v_t(n-Q)]^T$, the regression coefficient vector $\boldsymbol{b}_t = [b_{t,1}, \cdots, b_{t,Q}]^T$, and the innovation variance $\gamma_t^2$ is given by

$$p(v_t(n)|\boldsymbol{v}_t^Q(n-1), \boldsymbol{b}_t, \gamma_t^2) = \mathcal{N}\{v_t(n); \boldsymbol{v}_t^Q(n-1)^T \boldsymbol{b}_t, \gamma_t^2\}. \quad (3)$$

We hereafter represent the parameters of this noise model by $\Psi_t = \{\boldsymbol{b}_t, \gamma_t^2\}$.

(a3) $s_t(m)$ and $v_u(n)$ are mutually independent random variables for any $t$, $u$, $m$, and $n$. $\Phi_t$ and $\Psi_u$ are also random variables that are independent of each other.

Under the above assumptions, the conditional distribution of $\boldsymbol{x}_t$ and the conditional distribution of $\boldsymbol{s}_t$ are characterized respectively by

$$p(\boldsymbol{x}_t|\boldsymbol{s}_t, \Psi_t) = \prod_{n=1}^{N} \mathcal{N}\{x_t(n); s_t(n) + (\boldsymbol{x}_t^Q(n-1) - \boldsymbol{s}_t^Q(n-1))^T \boldsymbol{b}_t, \gamma_t^2\} \quad (4)$$

$$p(\boldsymbol{s}_t|\Phi_t) = \prod_{n=1}^{N} \mathcal{N}\{s_t(n); \boldsymbol{s}_t^P(n-1)^T \boldsymbol{a}_t, \sigma_t^2\}. \quad (5)$$

In addition to the above, we further assume the following conditions as regards the dynamics of the parameters of the speech and noise models.

(a4) The parameters of the noise model vary so slowly that the prior distribution of $\Psi_t$ given $\boldsymbol{x}_1^{t-1}$ is identical to the posterior distribution of $\Psi_{t-1}$ given $\boldsymbol{x}_1^{t-1}$ , *i.e.*

$$p(\Psi_t|\boldsymbol{x}_1^{t-1}) = p(\Psi_{t-1}|\boldsymbol{x}_1^{t-1}). \quad (6)$$

Furthermore, this distribution is assumed to be the conjugate distribution as follows:

$$p(\Psi_t|\boldsymbol{x}_1^{t-1}) = \mathcal{N}\Big\{\boldsymbol{b}_t; \boldsymbol{\nu}_{v,t-1}, \Big(\frac{\zeta_{v,t-1}\Xi_{v,t-1}}{\gamma_t^2}\Big)^{-1}\Big\} \times \chi^{-2}\{\gamma_t^2; \rho_{v,t-1}, \lambda_{v,t-1}\}, \quad (7)$$

where $\chi^{-2}\{\xi^2; \rho, \lambda\}$ denotes the PDF of a random variable $\xi^2$ following the scaled inverse chi square distribution with degree of freedom $\rho$ and scale parameter $\lambda$. The choice of the conjugate prior is for mathematical tractability.

(a5) The prior distribution of the speech model parameters $\Phi_t$ given $\boldsymbol{x}_1^{t-1}$ is free from $\boldsymbol{x}_1^{t-1}$ and is independent of the frame index $t$. Specifically, the prior distribution is given by

$$p(\Phi_t) = \mathcal{N}\Big\{\boldsymbol{a}_t; \boldsymbol{\nu}_s{}', \Big(\frac{\zeta_s'\Xi_s'}{\sigma_s^2}\Big)^{-1}\Big\}\chi^{-2}\{\sigma_t^2; \rho_s', \lambda_s'\}, \quad (8)$$

where we let $\zeta'$ and $\rho'$ be a very small value to make the prior distribution non-informative.

Under the above conditions, we shall estimate $\boldsymbol{s}_t$ from $\boldsymbol{x}_1^t$.

### 2.2. Bayesian approach

In this paper, we approach the above task based on the Bayesian estimation. The goal of the Bayesian estimation is to estimate the joint posterior distribution of the speech samples $\boldsymbol{s}_t$ and the parameters $\Theta_t = \{\Phi_t, \Psi_t\}$ of the $t$-th frame, $p(\boldsymbol{s}_t, \Theta_t|\boldsymbol{x}_1^t)$. The posterior distribution can be factorized as

$$p(\boldsymbol{s}_t, \Theta_t|\boldsymbol{x}_1^t) \propto p(\boldsymbol{x}_t|\boldsymbol{s}_t, \Psi_t)p(\boldsymbol{s}_t|\Phi_t)p(\Phi_t)p(\Psi_t|\boldsymbol{x}_1^{t-1}). \quad (9)$$

The terms of the right hand side of (9) have been already defined in (4), (5), (8), and (7). Moreover, by integrating out $\boldsymbol{s}_t$ and $\Phi_t$ from (9), we obtain the posterior distribution for the $t$-th frame, $p(\Psi_t|\boldsymbol{x}_1^t)$, which is used as the prior distribution for the $(t+1)$-th frame, $p(\Psi_{t+1}|\boldsymbol{x}_1^t)$. Importantly, with the variational approximation described in the next section, $p(\Psi_t|\boldsymbol{x}_1^t)$ has the same form as $p(\Psi_t|\boldsymbol{x}_1^{t-1})$ by virtue of $p(\Psi_t|\boldsymbol{x}_1^{t-1})$ being conjugate. Thus, in principle, every time noisy speech samples of a new frame are observed, we can obtain the posterior distribution of the speech samples and the parameters of that frame adaptively.

## 3. ALGORITHM BASED ON VARIATIONAL BAYESIAN METHOD

### 3.1. Variational Bayesian method

We employ the variational Bayesian method, which approximates the posterior distribution $p(\boldsymbol{s}_t, \Theta_t|\boldsymbol{x}_1^t)$ because it is difficult to derive the analytic form of $p(\boldsymbol{s}_t, \Theta_t|\boldsymbol{x}_1^t)$. The difficulty arises from the fact that $\boldsymbol{s}_t$ is a hidden variable in the model that generates observed variable $\boldsymbol{x}_t$. The variational Bayesian method approximates the true posterior distribution $p(\boldsymbol{s}_t, \Theta_t|\boldsymbol{x}_1^t)$ by using the hypothetical posterior distribution of the following factorized form:

$$q(\boldsymbol{s}_t, \Theta_t) = q(\boldsymbol{s}_t)q(\Theta_t). \quad (10)$$

$q(\boldsymbol{s}_t)$ and $q(\Theta_t)$ are calculated so that the Kullback-Leibler (KL) divergence between $p(\boldsymbol{s}_t, \Theta_t|\boldsymbol{x}_1^t)$ and $q(\boldsymbol{s}_t, \Theta_t)$ is minimized.

Such $q(\boldsymbol{s}_t)$ and $q(\Theta_t)$, which are called optimal variational posterior distributions, can be proven to satisfy

$$q(\boldsymbol{s}_t) \propto \exp\{\langle \log p(\boldsymbol{x}_t, \boldsymbol{s}_t|\Theta_t)\rangle_{q(\Theta_t)}\} \quad (11)$$

$$q(\Theta_t) \propto p(\Theta_t|\boldsymbol{x}_1^{t-1}) \exp\{\langle \log p(\boldsymbol{x}_t, \boldsymbol{s}_t|\Theta_t)\rangle_{q(\boldsymbol{s}_t)}\}, \quad (12)$$

where the complete data likelihood $p(\boldsymbol{x}_t, \boldsymbol{s}_t|\Theta_t)$ is obtained by multiplying (4) and (5). It is obvious that these two conditions are dependent on each other. Therefore, the variational Bayesian method first calculates $q(\boldsymbol{s}_t)$ according to (11) for a fixed $q(\Theta_t)$, and then calculates $q(\Theta_t)$ for a fixed $q(\boldsymbol{s}_t)$ according to (12). $q(\boldsymbol{s}_t)$ and $q(\Theta_t)$ are obtained by repeating these two steps until convergence. Below we derive specific forms of the optimal variational posterior distributions $q(\Theta_t)$ and $q(\boldsymbol{s}_t)$.

### 3.2. Optimal variational posterior distribution of parameters

It is shown that substituting (4) and (5) into (12) leads to

$$q(\Theta_t) = \mathcal{N}\Big\{\boldsymbol{a}_t; \boldsymbol{\nu}_{s,t}, \Big(\zeta_{s,t}\frac{\Xi_{s,t}}{\sigma_t^2}\Big)^{-1}\Big\}\chi^{-2}\{\sigma_t^2; \rho_{s,t}, \lambda_{s,t}\} \times \mathcal{N}\Big\{\boldsymbol{b}_t; \boldsymbol{\nu}_{v,t}, \Big(\zeta_{v,t}\frac{\Xi_{v,t}}{\gamma_t^2}\Big)^{-1}\Big\}\chi^{-2}\{\gamma_t^2; \rho_{v,t}, \lambda_{v,t}\}. \quad (13)$$

In (13), the hyperparameters of the speech model (the first line of (13)) are defined as follows.

$$\zeta_{s,t} = N + \zeta'_s \tag{14}$$

$$\Xi_{s,t} = \frac{N R_{s,t} + \zeta'_s \Xi'_s}{N + \zeta'_s} \tag{15}$$

$$\boldsymbol{\nu}_{s,t} = \Xi_{s,t}^{-1} \boldsymbol{\xi}_{s,t} \tag{16}$$

$$\rho_{s,t} = N + \rho'_s \tag{17}$$

$$\lambda_{s,t} = \lambda'_s + N r_{s,t} + \zeta'_s \boldsymbol{\nu}_s'^T \Xi'_s \boldsymbol{\nu}'_s - (N + \zeta'_s) \boldsymbol{\xi}_{s,t}^T \Xi_{s,t}^{-1} \boldsymbol{\xi}_{s,t}, \tag{18}$$

where

$$\boldsymbol{\xi}_{s,t} = \frac{N \boldsymbol{r}_{s,t} + \zeta'_s \Xi'_s \boldsymbol{\nu}'_s}{N + \zeta'_s}. \tag{19}$$

$R_{s,t}$, $\boldsymbol{r}_{s,t}$, and $r_{s,t}$ that appear in (14) to (19) are respectively the $P$-th order autocovariance matrix, the one-sample delayed autocovariance vector of order $P$, and the variance of the speech samples $\{s_t(n)\}_{1 \le n \le N}$ expected on $q(\boldsymbol{s}_t)$:

$$R_{s,t} = \frac{1}{N} \sum_{n=1}^{N} \langle \boldsymbol{s}_t^P(n-1) \boldsymbol{s}_t^P(n-1)^T \rangle_{q(\boldsymbol{s}_t)} \tag{20}$$

$$\boldsymbol{r}_{s,t} = \frac{1}{N} \sum_{n=1}^{N} \langle s_t(n) \boldsymbol{s}_t^P(n-1) \rangle_{q(\boldsymbol{s}_t)} \tag{21}$$

$$r_{s,t} = \frac{1}{N} \sum_{n=1}^{N} \langle s_t(n)^2 \rangle_{q(\boldsymbol{s}_t)}. \tag{22}$$

As shown later, $q(\boldsymbol{s}_t)$ is a normal distribution with mean $\boldsymbol{\mu}_t$ and covariance matrix $\Upsilon_t$. Hence, the expected product of $s_t(n)$ and $s_t(m)$ that constitute $R_{s,t}$, $\boldsymbol{r}_{s,t}$ and $r_{s,t}$ can be calculated as

$$\langle s_t(m) s_t(n) \rangle_{q(\boldsymbol{s}_t)} = \mu_t(m) \mu_t(n) + \upsilon_{m,n}, \tag{23}$$

where $\mu_t(n)$ is the $(N-n+1)$-th component of $\boldsymbol{\mu}_t$ and $\upsilon_{m,n}$ is the $(N-m+1, N-n+1)$-th component of $\Upsilon_t$. Recall and note that the elements of $\boldsymbol{s}_t$ are arranged in time-descending order.

The hyperparameters of the noise model are defined in the same way as those of the speech model given by (14) to (19). To take the nonstationarity of noise into account, however, we propose multiplying the hyperparameters of the prior distribution of the noise model parameters by forgetting factor $\alpha$. Thus, the update equations for the hyperparameters of the noise model are as follows.

$$\zeta_{v,t} = N + \alpha \zeta_{v,t-1} \tag{24}$$

$$\Xi_{v,t} = \frac{N R_{v,t} + \alpha \zeta_{v,t-1} \Xi_{v,t-1}}{N + \alpha \zeta_{v,t-1}} \tag{25}$$

$$\boldsymbol{\nu}_{v,t} = \Xi_{v,t}^{-1} \boldsymbol{\xi}_{v,t} \tag{26}$$

$$\rho_{v,t} = N + \alpha \rho_{v,t-1} \tag{27}$$

$$\lambda_{v,t} = \alpha \lambda_{v,t-1} + N r_{v,t} + \alpha \zeta_{v,t-1} \boldsymbol{\nu}_{v,t-1}^T \Xi_{v,t-1} \boldsymbol{\nu}_{v,t-1} - (N + \alpha \zeta_{v,t-1}) \boldsymbol{\xi}_{v,t}^T \Xi_{v,t} \boldsymbol{\xi}_{v,t}, \tag{28}$$

where

$$\boldsymbol{\xi}_{v,t} = \frac{N \boldsymbol{r}_{v,t} + \alpha \zeta_{v,t-1} \Xi_{v,t-1} \boldsymbol{\nu}_{v,t-1}}{N + \alpha \zeta_{v,t-1}} \tag{29}$$

$R_{v,t}$, $\boldsymbol{r}_{v,t}$, and $r_{v,t}$ in (24) to (29) are respectively the $Q$-th order autocovariance matrix, the one-sample delayed autocovariance vector of order $Q$, and the variance of the noise samples $\{x_t(n) - s_t(n)\}_{1 \le n \le N}$ expected on $q(\boldsymbol{s}_t)$. These are given in the same way as (20) to (22).

To summarize this subsection, once the posterior distribution of the speech samples, $q(\boldsymbol{s}_t)$, is given, the posterior distribution of the parameters is updated by updating its hyperparameters according to (14) to (18) and (24) to (28).

### 3.3. Optimal variational posterior distribution of speech samples

Next, let us derive the update procedure for $q(\boldsymbol{s}_t)$ given $q(\Theta_t)$. By substituting (4) and (5) into (11), we finally have

$$q(\boldsymbol{s}_t) = \mathcal{N}\{\boldsymbol{s}_t; \boldsymbol{\mu}_t, \Upsilon_t\}. \tag{30}$$

The covariance matrix $\Upsilon_t$ and the mean $\boldsymbol{\mu}_t$ are defined respectively as follows:

$$\Upsilon_t = \Big( \frac{\rho_{s,t}}{\lambda_{s,t}} A_t^T A_t + \frac{1}{\zeta_{s,t}} \Omega_{s,t} + \frac{\rho_{v,t}}{\lambda_{v,t}} B_t^T B_t + \frac{1}{\zeta_{v,t}} \Omega_{v,t} \Big)^{-1} \tag{31}$$

$$\boldsymbol{\mu}_t = \Big( \frac{\rho_{v,t}}{\lambda_{v,t}} B_t^T B_t + \frac{1}{\zeta_{v,t}} \Omega_{v,t} \Big) \Upsilon_t \boldsymbol{x}_t, \tag{32}$$

where $A_t$ and $B_t$ are $N$-th order upper triangular Toeplitz matrices whose first rows are respectively

$$\overbrace{[1, -\boldsymbol{\nu}_{s,t}^T}^{P+1}, \overbrace{0, \cdots, 0]}^{N-P-1}, \text{ and } \overbrace{[1, -\boldsymbol{\nu}_{v,t}^T}^{Q+1}, \overbrace{0, \cdots, 0]}^{N-Q-1}. \tag{33}$$

$\Omega_{s,t}$ is defined as follows. Let $\Omega_{s,t}(n)$ denote the $N$-th order matrix whose $P$-th order submatrix beginning from the $(N-n+1, N-n+1)$-th component is equal to $\Xi_{s,t}^{-1}$ and whose components outside the submatrix are all zero. Now, $\Omega_{s,t}$ is defined as

$$\Omega_{s,t} = \sum_{n=1}^{N} \Omega_{s,t}(n-1). \tag{34}$$

The definition of $\Omega_{v,t}$ is similar to that of $\Omega_{s,t}$. Therefore, the posterior distribution of the speech samples, $q(\boldsymbol{s}_t)$, is updated by updating its mean $\boldsymbol{\mu}_t$ and covariance matrix $\Upsilon_t$ according to (32) and (31), respectively, when the posterior distribution of the parameters $q(\Theta_t)$ is given.

It is noteworthy that if we ignore $\Omega_{s,t}$ and $\Omega_{v,t}$, the mean $\boldsymbol{\mu}_t$ is the minimum mean square error (MMSE) estimate of $\boldsymbol{s}_t$ under the condition that the parameters $\Theta_t$ are deterministic as $\boldsymbol{a}_t = \boldsymbol{\nu}_{s,t}$, $1/\sigma_t^2 = \rho_{s,t}/\lambda_{s,t}$, $\boldsymbol{b}_t = \boldsymbol{\nu}_{v,t}$, and $1/\gamma_t^2 = \rho_{v,t}/\lambda_{v,t}$. Furthermore, the covariance matrix $\Upsilon_t$ corresponds to the associated error covariance matrix. The terms $\Omega_{s,t}/\zeta_{s,t}$ and $\Omega_{v,t}/\zeta_{v,t}$ reflect the degree of uncertainty of the parameters $\Theta_t$.

### 3.4. Efficient Implementation

Based on the above derivation, the proposed method is summarized as follows.

(s1) Initialize the hyperparameters of the posterior distribution of the speech and noise models, $\zeta_{s,t}$, $\Xi_{s,t}$, $\boldsymbol{\nu}_{s,t}$, $\rho_{s,t}$, $\lambda_{s,t}$, $\zeta_{v,t}$, $\Xi_{v,t}$, $\boldsymbol{\nu}_{v,t}$, $\rho_{v,t}$, and $\lambda_{v,t}$.

(s2) Update the mean $\boldsymbol{\mu}_t$ and covariance matrix $\Upsilon_t$ of the posterior distribution of the speech samples according to (32) and (31), respectively.

(s3) Update the hyperparameters according to (14) to (18) and (24) to (28).

(s4) Unless convergence is reached, return to (s2).

The above algorithm can be implemented efficiently in the following two respects.

- We suppose that both

$$\begin{bmatrix} * & \boldsymbol{\xi}_{s,t}^T \\ \boldsymbol{\xi}_{s,t} & \Xi_{s,t} \end{bmatrix} \text{ and } \begin{bmatrix} * & \boldsymbol{\xi}_{v,t}^T \\ \boldsymbol{\xi}_{v,t} & \Xi_{v,t} \end{bmatrix}, \tag{35}$$

  where $*$ denotes a certain value, are symmetric Toeplitz matrices. Let us represent the first column of the former matrix by $[\bar{r}_{s,t}(0), \cdots, \bar{r}_{s,t}(P)]^T$. Then, $\boldsymbol{\nu}_{s,t}$ and $\lambda_{s,t}$ of (16) and (18) are calculated by applying the Levinson-Durbin algorithm [7] of order $P$ to $\{\bar{r}_{s,t}(0), \cdots, \bar{r}_{s,t}(P)\}$. $\boldsymbol{\nu}_{v,t}$ and $\lambda_{v,t}$ of (26) and (28) can also be calculated by using the $Q$-th order Levinson-Durbin algorithm.

- We suppose that $A_t^T A_t$, $\Omega_{s,t}$, $B_t^T B_t$, and $\Omega_{v,t}$ are all Toeplitz matrices. Then, $\boldsymbol{\mu}_t$ and $\Upsilon_t$ of (32) and (31) can be calculated in the frequency domain. Moreover, based on the result described in [8], $\Omega_{s,t}$ is found to be calculated from the $(P-1)$-th order regression coefficients for $\{\bar{r}_{s,t}(0), \cdots, \bar{r}_{s,t}(P-1)\}$. Surprisingly and fortunately, the $(P-1)$-th order regression coefficients have already been obtained in the calculation of $\boldsymbol{\nu}_{s,t}$. The same holds for $\Omega_{v,t}$.

## 4. EXPERIMENTAL RESULTS

We conducted experiments to evaluate the performance of the proposed method. Japanese utterances of five males and five females were taken from the JNAS database. The sampling frequency was 8 kHz. The speech signals of these utterances were contaminated by eight types of additive noise with a signal to noise ratio (SNR) of 10 dB, which were taken from the AURORA-2 database. The system parameters were set as follows: the frame size $N$ was 256 samples, the frame shift was 128 samples, the window function was a Hanning window, the order of the speech AR model, $P$, was 12, the order of the noise AR model, $Q$, was 6, and the forgetting factor $\alpha$ was 0.9.

The proposed method was compared with the minimum statistics based method of [4] and the ML estimation based method. The ML estimation based method is derived by forcing $\Omega_{s,t}$ and $\Omega_{v,t}$ of (31) and (32) to be zero matrices. The noise suppression performance was evaluated by using the segmental SNR (SSNR) and the segmental Itakura-Saito distortion (SISD) measure. The SSNR and SISD were calculated by using median SNRs and ISDs for each frame as in [3] to mitigate the influence of outliers.

Figs. 1 and 2 show the experimental results. It can be seen that the proposed method achieved the best performance in terms of both measures. The reason for the high SISDs of the minimum statistics based method may be that the method is likely to introduce spectral distortion owing to spectral smoothing. Hence, it can be concluded that the proposed method is able to suppress non-stationary noise while reducing spectral distortion.

## 5. CONCLUSION

In this paper, we have described a Bayesian method for estimating the parameters of speech and noise AR models and speech signals. The variational Bayesian method was used to obtain the approximate posterior distribution of the parameters and the speech signals. An efficient implementation was also developed by assuming the Toeplitz structure for all involved covariance matrices. Experimental results showed that the proposed method performed better than alternative methods in terms of SSNR and SISD.

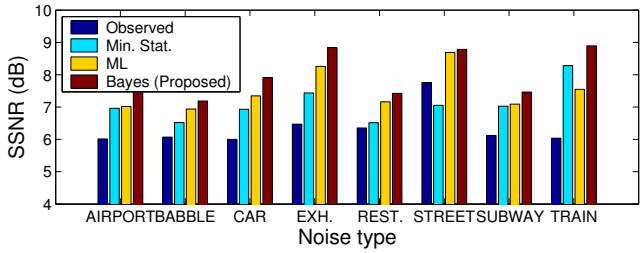


**Fig. 1**. Average SSNRs for each noise type. Greater SSNRs indicate better performance.

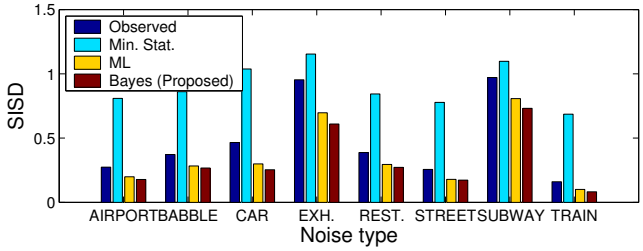


**Fig. 2**. Average SISDs for each noise type. Smaller SISDs indicate better performance.

## 6. REFERENCES


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

[2] S. Srinivasan, J. Samuelsson, and B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech, Audio Process.*, vol. 6, no. 4, pp. 373–385, 1998.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[5] S. R. Waterhouse, D. MacKay, and A. J. Robinson, "Bayesian methods for mixture of experts," in *Adv. Neural Inform. Process. Syst. 8*, 1995, pp. 351–357.

[6] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Adv. Neural Inform. Process. Syst. 13*, 2000, pp. 758–764.

[7] J. Durbin, "The fitting of time-series models," *Rev. Inst. Int. Stat.*, vol. 28, no. 3, pp. 233–243, 1960.

[8] W. F. Trench, "An algorithm for the inversion of finite Toeplitz matrices," *J. Soc. Indust. Appl. Math.*, vol. 12, no. 3, pp. 515–522, 1964.