# NOISE ESTIMATION USING A CONSTRAINED SEQUENTIAL HMM IN LOG-SPECTRAL DOMAIN

*Dongwen Ying*[1], *Xugang Lu*[2], *Junfeng Li*[1], *Yonghong Yan*[1], *Jianwu Dang*[3], *and Frank Soong*[4]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences
[2]National Institute of Information and Communications Technology, Japan
[3]School of Information Science, Japan Advanced Institute of Science and Technology
[4]Speech Group, Microsoft Research Asia.

## ABSTRACT

How to utilize the time correlation of speech/nonspeech presence is a crucial problem faced by noise estimators. The popular technique of exploiting such correlation is to smooth noisy spectra by using a temporal recursive filter with a time-varying smoothing factor. But this technique cannot warrant the statistical optimality. In theory, hidden Markov model (HMM) is more desirable than this technique. It can give an elaborate description of speech/nonspeech transition. Moreover, some theoretical frameworks, such as maximum likelihood (ML), are available for optimal estimation. This paper presents a constrained sequential HMM to model the time correlation of speech/nonspeech presence of an individual log-power sequence. Its parameter set is on-line adapted to varying signals based on a ML framework. We compared its performance with that of well-established algorithms by speech enhancement experiments. The results confirmed its promising performance.

***Index Terms***— Noise estimation, time correlation, sequential hidden Markov model, constraints.

## 1. INTRODUCTION

Time correlation of speech/nonspeech presence is a widely used clue for noise estimation. The most popular technique of exploiting the time correlation is to use a temporal recursive averaging filter, where the forgetting factor is controlled by speech presence probability (SPP) [1] - [4]. But this techniques is not so perfect because it cannot warrant the optimality of noise estimate. In addition, its estimation process cannot be unified into a theoretical framework, and thus it seems to be heuristic somewhat.

In theory, the hidden Markov model (HMM) is more desirable to model the time correlation than the popular technique. On one hand, some theoretical frameworks, such as maximum likelihood (ML) re-estimation, are available to estimate its parameter set in a statistically optimal sense. On the other hand, it can give an elaborate description of speech/nonspeech transition by using transition probabilities. The binary-state HMM consisting of a speech and nonspeech state has been proposed to model the correlation of speech/nonspeech presence [5], [6]. However, the transition probability of those algorithms is fixed and does not adapt to varying signals, which is problematic since the time correlation varies with times and differs from band to band. Hence this HMM-based method should be further perfected.

In this paper, we propose a constrained sequential HMM (CSHMM) to model the time correlation of a log-power sequence. A sequential scheme is derived from a ML framework of HMM, where the HMM parameter set including the noise estimate is sequentially regulated at varying signals based upon the ML criterion. Another advantage of this algorithm over conventional algorithms concerns with its initialization methods. Conventional noise estimators usually initialize their models based on an assumption that the first several frames of an utterance to be nonspeech. If an utterance begins with speech, the noise model of conventional ones will be incorrectly initialized. Fortunately, the CSHMM initialization process is conducted by expectation-maximization (EM), and thus does not require this assumption. This advantage is very valuable to practical applications. The proposed algorithm is conducted at each band in parallel, and the detail in a single band is given in the following.

## 2. MODELING A LOG-POWER SEQUENCE USING HMM

We firstly consider a high-SNR band in which both speech and nonspeech signals are supposed to be present. In addition, the speech/nonspeech logarithmic power is assumed to satisfy the Gaussian distribution. The transition dynamics of the power sequence between speech and nonspeech states is modeled by using a Markov chain, each state of which consists of a unique Gaussian component. For the interests of brevity, the bin index $k$ is omitted since an individual band is concerned in this algorithm. Let $\lambda_\ell$ denote the parameter set of HMM estimated from a log-power sequence, $\mathbf{x}_\ell \triangleq \{x_1, x_2, \ldots, x_\ell\}$. Let $\mathbf{s}_\ell \triangleq \{s_1, s_2, \ldots, s_\ell\}$, $s_\ell \in \{0, 1\}$ be a state sequence corresponding to $\mathbf{x}_\ell$, where 1 denotes the speech state and 0 for the nonspeech state. The HMM probability density function (PDF) is given by

$$p(\mathbf{x}_\ell|\lambda_\ell) = \sum_{\mathbf{s}_\ell} p(\mathbf{s}_\ell|\lambda_\ell) p(\mathbf{x}_\ell|\mathbf{s}_\ell, \lambda_\ell), \tag{1}$$

where $p(\mathbf{s}_\ell|\lambda_\ell)$ is the probability of the state sequence $\mathbf{s}_\ell$,

$$p(\mathbf{s}_\ell|\lambda_\ell) = \prod_{t=1}^{\ell} a_{s_{t-1}, s_t}, \tag{2}$$

where $a_{s_{t-1}, s_t}$ denotes the transition probability from state $s_{t-1}$ at time $t-1$ to state $s_t$ at time $t$, and $a_{0,s_1} \triangleq \pi_{s_1}$ denotes the probability of the initial state $s_1$. $p(\mathbf{x}_\ell|\mathbf{s}_\ell, \lambda_\ell)$ is the PDF of given the sequence of states $\mathbf{s}_\ell$, described as

$$p(\mathbf{x}_\ell|\mathbf{s}_\ell, \lambda_\ell) = \prod_{t=1}^{\ell} p(x_t|s_t, \lambda_\ell) \triangleq \prod_{t=1}^{\ell} b(x_t|s_t, \lambda_\ell), \tag{3}$$

where $b(x_t|s_t, \lambda_\ell)$ is the PDF of the observed data $x_t$ given the state $s_t$ and the parameter set $\lambda_\ell$, denoted as

$$b(x_t|s_t = i, \lambda_\ell) = \frac{1}{\sqrt{2\pi\kappa_{i,\ell}}} \exp\Big\{ -\frac{1}{2}(x_t - \mu_{i,\ell})^2/\kappa_{i,\ell}\Big\}, \quad (4)$$

where $\mu_{i,\ell}$ and $\kappa_{i,\ell}$ are respectively the mean and variance of the Gaussian function for the given state $s_t = i$.

How to estimate the parameter set $\lambda_\ell = \{\boldsymbol{\pi}, \mathbf{a}_\ell, \boldsymbol{\mu}_\ell, \boldsymbol{\kappa}_\ell\}$ is the key problem, where $\boldsymbol{\mu}_\ell \triangleq \{\mu_{0,\ell}, \mu_{1,\ell}\}$, $\boldsymbol{\kappa}_\ell \triangleq \{\kappa_{0,\ell}, \kappa_{1,\ell}\}$, and $\mathbf{a}_\ell$ is a $2 \times 2$ transition matrix. $\boldsymbol{\pi} \triangleq \{\pi_0, \pi_1\}$ denotes the initial probability respectively in nonspeech and speech states. Given a training sequence $\mathbf{x}_\ell$ from the observed noisy speech spectra, a maximum-likelihood estimate of the parameter set $\lambda_\ell$ is obtained from

$$\lambda_\ell = \arg\max_{\lambda} \ln \sum_{\mathbf{s}_\ell} p(\mathbf{x}_\ell, \mathbf{s}_\ell|\lambda). \quad (5)$$

where $\mu_{0,\ell}$ is the optimal estimate of the noise power.

## 3. SEQUENTIAL ESTIMATION OF HMM PARAMETERS

The classical method of estimating HMM parameters is the expectation-maximization (EM) algorithm. It is a batch algorithm that requires processing the received data as a whole. However, the speech and nonspeech signals are assumed to be piecewise stationary. Their statistical characteristics vary with time; hence a sequential scheme can be adaptive in nature to track the varying parameters. The typical sequential schemes for HMM will induce a heavy computational load since the whole Markov chain is updated in each time. For this reason, we present a simplified scheme, where the current model depends on the last model and the current observation. It is actually a first-order recursive process described as $\lambda_{\ell+1} = \Phi(x_{\ell+1}, \lambda_\ell)$. The initial Markov chain is constructed from the first $M$ samples by the EM algorithm in an off-line manner, and then sequentially updated frame by frame. After initialization, the parameter $\boldsymbol{\pi}$ does not vary with time going. In the following, we give only the sequential scheme and the EM-based initialization can refer to textbook.

The criterion of the sequential estimate is based on the principle of maximum likelihood,

$$\lambda_{\ell+1} = \max_{\lambda} Q_{\ell+1|\lambda_\ell}(\lambda), \quad (6)$$

where the likelihood function is defined as

$$\begin{aligned} Q_{\ell+1|\lambda_\ell}(\lambda) &\triangleq E\{\log p(\mathbf{x}_{\ell+\mathbf{1}}, \mathbf{s}_{\ell+\mathbf{1}}|\lambda)\Big|\mathbf{x}_{\ell+\mathbf{1}}, \lambda_\ell\} \\ &= \sum_{t=1}^{\ell+1} \psi_{t|\lambda_\ell}(\lambda) + \sum_i \gamma_{0|\lambda_\ell}(i) \log \pi_i, \end{aligned} \quad (7)$$

with

$$\begin{aligned} \psi_{t|\lambda_\ell}(\lambda) = &\sum_i \sum_j \xi_{t|\lambda_\ell}(i,j) \log a_{ij} \\ &+ \sum_i \gamma_{t|\lambda_\ell}(i) \log \frac{1}{\sqrt{2\pi\kappa_i}} * \exp\Big\{ -\frac{(x_t - \mu_i)^2}{2\kappa_i}\Big\}. \end{aligned} \quad (8)$$

Here, $\pi_i, a_{ij}, \mu_i, \kappa_i$ are the unknown parameters in $\lambda$, $\xi_{t|\lambda_\ell}(i,j) = p(s_{t-1} = i, s_t = j|x_t, \lambda_\ell)$ is the conditional transition probability, and $\gamma_{t|\lambda_\ell}(i) = p(s_t = i|x_t, \lambda_\ell)$ is the speech/nonspeech presence probability. Because of the limited space, the sequential scheme is directly given, and please refer to [7] for its derivation in details.

The recursive processes for mean and variance are represented respectively as

$$\mu_{i,\ell+1} = \tilde{\alpha}_{\ell+1}(i)\mu_{i,\ell} + \big[1 - \tilde{\alpha}_{\ell+1}(i)\big]x_{\ell+1}, \quad (9)$$

$$\kappa_{i,\ell+1} = \tilde{\alpha}_{\ell+1}(i)\kappa_{i,\ell} + \big[1 - \tilde{\alpha}_{\ell+1}(i)\big](x_{\ell+1} - \mu_{i,\ell})^2, \quad (10)$$

where $\tilde{\alpha}_{\ell+1}(i)$ is a time-varying and frequency-dependent smoothing factor, represented as

$$\tilde{\alpha}_{\ell+1}(i) = \alpha\bar{\gamma}_\ell(i)/\bar{\gamma}_{\ell+1}(i), \quad (11)$$

with

$$\bar{\gamma}_{\ell+1}(i) = \alpha\bar{\gamma}_\ell(i) + (1-\alpha)\gamma_{\ell+1|\lambda_\ell}(i), \quad (12)$$

where $\alpha$ is a constant forgetting factor.

The transition probability is described as a non-linear recursive equation,

$$\begin{aligned} a_{ij,\ell+1} = a_{ij,\ell} + \\ \frac{\frac{\xi_{\ell+1|\lambda_\ell}(i,j)}{a_{ij,\ell}} - \frac{\xi_{\ell+1|\lambda_\ell}(i,1-j)}{1-a_{ij,\ell}}}{\frac{K}{a_{ij,\ell}^2}\bar{\xi}_{\ell+1}(i,j) + \frac{K}{(1-a_{ij,\ell})^2}\bar{\xi}_{\ell+1}(i,1-j)}, \end{aligned} \quad (13)$$

where $K = \lfloor \alpha/(1-\alpha) \rfloor$ makes the transition probability be forgotten with the same speed as that of the means and variances. $\bar{\xi}_{\ell+1}(i,j)$ is the smoothed transition probability, described as

$$\bar{\xi}_{\ell+1}(i,j) = \alpha\bar{\xi}_\ell(i,j) + (1-\alpha)\xi_{\ell+1|\lambda_\ell}(i,j). \quad (14)$$

In the sequential process of Eqs. 9 - 14, $\lambda_{\ell+1}$ is determined by four variables, namely $x_{\ell+1}, \lambda_\ell, \xi_{\ell+1|\lambda_\ell}(i,j)$, and $\gamma_{\ell+1|\lambda_\ell}(i)$. If the conditional probabilities $\gamma_{\ell+1|\lambda_\ell}(i)$ and $\xi_{\ell+1|\lambda_\ell}(i,j)$ can be represented as a function of $x_{\ell+1}$ and $\lambda_\ell$, the sequential scheme will be a first-order recursive process.

These conditional probabilities are usually calculated by using the forward and backward factors. Therefore, this algorithm makes an approximation to those factors in order to construct a first-order sequential scheme. Assuming the model $\lambda_\ell$ varies with time slowly, the forward factor is defined as a sequential variable,

$$F_{\ell+1|\lambda_\ell}(i) = \sum_j F_{\ell|\lambda_{\ell-1}}(j) a_{ji,\ell} b(x_{\ell+1}|s_{\ell+1} = i, \lambda_\ell). \quad (15)$$

As the future information is unavailable in the on-line estimation, the backward factor $B_{t|\lambda_\ell}(i)$ is set to be 1. Hence, the following conditional probabilities are obtained,

$$\gamma_{\ell+1|\lambda_\ell}(i) = \frac{F_{\ell+1|\lambda_\ell}(i)}{\sum_i F_{\ell+1|\lambda_\ell}(i)}, \quad (16)$$

$$\xi_{\ell+1|\lambda_\ell}(i,j) = \frac{F_{\ell+1|\lambda_\ell}(i) a_{ij,\ell} b(x_{\ell+1}|s_{\ell+1} = j, \lambda_\ell)}{\sum_{ij} F_{\ell+1|\lambda_\ell}(i) a_{ij,\ell} b(x_{\ell+1}|s_{\ell+1} = j, \lambda_\ell)}. \quad (17)$$

From Eqs. 15-17, one can see $\gamma_{\ell+1|\lambda_\ell}(i)$ and $\xi_{\ell+1|\lambda_\ell}(i,j)$ are the functions of $x_{\ell+1}$ and $\lambda_\ell$. Eventually, the sequential scheme is efficiently realized based on a first-order time-recursive process.

## 4. CONSTRAINTS TO THE HMM

The binary-state HMM introduced in section 3 has the advantage of well treating the binary-mode data set consisting of speech and nonspeech samples. Due to sparsity of speech distribution in frequency domain, however, speech signal may be absent or very weak in some

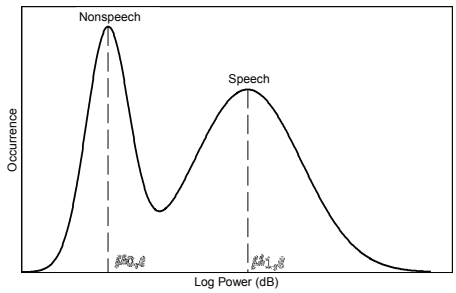


**Fig. 1**. Schematic illustration of the log-power distribution.

low-SNR bands, where the data set consists of only nonspeech samples, and thus the speech state is difficult to be modeled. For this reason, some constraints to HMM are utilized to adapt the binary-state model to the nonspeech-only-mode data set.

These constraints are mainly induced from the relationships between speech and nonspeech distributions. Fig. 1 schematically illustrates the binary-mode histogram of logarithmic power at a noisy band. Since "nonspeech" denotes noise while "speech" represents the superposition of noise and clean speech signals in this paper, the nonspeech mean is smaller than the speech mean. So, one mode with smaller mean is identified as nonspeech mode, and the other as speech mode. The mean difference $\mu_{1,\ell} - \mu_{0,\ell}$ represents the posterior SNR of this band. Assuming the noise signal is more stationary than the speech signal, the nonspeech variance is smaller than the speech variance.

These distribution relationships are reflected in the binary-state HMM by some constraints. First, the relationship between speech and nonspeech means is described by the constraint,

$$\mu_{1,\ell} = \max\{\mu_{1,\ell}, \mu_{0,\ell} + \delta\}, \tag{18}$$

where $\delta > 0$. Second, according to the variance relationship, the following constraint is introduced.

$$\kappa_{1,\ell} = \max\{\kappa_{0,\ell}, \kappa_{1,\ell}\}. \tag{19}$$

The process that the constraints adapt the binary-state model to the nonspeech-only-mode data set is demonstrated by EM iterations. When speech signal is absent or very weak, the actual posterior SNR is less than $\delta$ and the speech mean will be set as $\mu_{0,M} + \delta$ by the constraint of Eq.18. As a result, the speech likelihood of most samples $b(x_\ell|s_\ell = 1, \lambda_M)$ is decreased, and vice versa for the nonspeech likelihood, and then the nonspeech presence probability of most samples will increase. Accordingly, this adaptation attempts to identify more samples as nonspeech. After several iterations, the nonspeech presence probability of all samples will approach to 1. Therefore, the nonspeech mean and variance is calculated by using all samples while the speech mean and variance are respectively set as $\mu_{1,M} = \mu_{0,M} + \delta$ and $\kappa_{1,M} = \kappa_{0,M}$ by the constraints.

Here, $\delta$ is an important parameter determining the first constraint to be activated or not. In general conditions, the bands with the posterior SNR less than $\delta$ are approximated to be speech-absent, where the weak speech components are taken as nonspeech, which is referred to as speech leakage. If $\delta$ is too large, some strong nonspeech components are likely to be taken as speech, which is named as nonspeech leakage. Therefore, $\delta$ regulates the tradeoff between speech and nonspeech leakage.

The third constraint is introduced due to the first constraint. When the first constraint is activated in initialization, the SPP of all samples will be close to zero. The denominators of iteration equations in EM algorithms will be zero when the speech mean and variance are re-estimated, which results in the EM algorithm failure. Therefore, we introduce the third constraint,

$$\frac{1}{M}\sum_{\ell=1}^{M} p(s_\ell = 1|x_\ell, \lambda_M) > \epsilon, \tag{20}$$

where $\epsilon$ is close and greater than zero. When the condition in Eq.20 is not satisfied, the EM iteration will terminate.

The last constraint concerns the state transition in the sequential process at low-SNR bands. If the inter-state transition probability $a_{ij}$ $(i \neq j)$ is very small, one state is difficult to transit to the other, which results in over-smoothing. Particularly when speech samples are unavailable for initialization at speech-absent bands, the inter-state transition probability will be zero. Accordingly, the inter-state transition is impossible to occur even if strong speech signal comes in the following sequential process. Hence the transition probabilities must be regulated in a certain range,

$$\begin{aligned} a_{0,1,\ell} &= \max(a_{0,1,\ell}, \zeta) \quad a_{0,0,\ell} = 1 - a_{0,1,\ell} \\ a_{1,0,\ell} &= \max(a_{1,0,\ell}, \zeta) \quad a_{1,1,\ell} = 1 - a_{1,0,\ell}. \end{aligned} \tag{21}$$

The CSHMM estimator can run in low-SNR bands by utilizing these constraints. Even when speech signal is absent, these constraints fabricate a virtual speech state to guarantee the binary-state model works well. This virtual one can convert into a real one when the CSHMM is updated with speech samples. In addition, these constraints are unlikely to being activated in high-SNR bands, and they hardly affect the HMM parameters.

## 5. IMPLEMENTATION AND EVALUATIONS

The above section shows noise estimation in one frequency band. The noise power of each band is estimated in parallel. The constraints are applied after each parameter is updated. For example, the constraint of Eq.18 follows Eq.9, Eq.19 to Eq.10, and Eq.21 to Eq.13. Before estimation, the noisy logarithmic spectrum is smoothed with a three-point median filter in order to take into account the strong correlation of speech presence in neighboring frequency bins. The parameters for signal with sampling rate 16 KHz are set as $\alpha = 0.98$, $\delta = 0.55$, $\epsilon = 0.03$, $\zeta = 0.05$, $M = 30$.

The noise and speech signals in our evaluation are respectively taken from the NOISEX-92 database and TIMIT database. They include white Gaussian, F16 cockpit, and babble noises. In addition, the non-stationary white Gaussian noise (NONWGN) is simulated by increasing the level of the stationary white at a rate of 3 dB/s for a period of three seconds, and sometime afterwards decreasing it back to the original level at the same rate. All noise signals are artificially added to clean speech signals at SNR of 0, 5, 10 dB. Each clean speech signal mixed with the NONWGN is constructed from every four short utterances while the speech signals for other noises are constructed from every two short utterances. There are ten long clean utterances and totally thirty noisy utterances for each noise. The sampling rate of all signals is 16 kHz.

Two well-established noise estimators, namely MS [8] and IMCRA [2], are utilized as the competing ones. The IMCRA is a typical time-recursive noise estimator. The crucial distinction between IMCRA and CSHMM is the method of employing the time correlation. The CSHMM incorporate the state transition into the recursive process while IMCRA does not.

**Table 1**. Segmental SNR of enhanced speech under various conditions (dB).

| SNR | *White noise* | | | *NONWGN* | | | *F16 cockpit noise* | | | *Babble noise* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MS | IMCRA | CSHMM | MS | IMCRA | CSHMM | MS | IMCRA | CSHMM | MS | IMCRA | CSHMM |
| 0dB | 3.95 | 4.23 | *4.97* | 4.04 | 4.23 | *4.93* | 3.66 | 3.45 | *4.11* | 4.10 | 4.30 | *4.41* |
| 5dB | 6.61 | 6.74 | *7.52* | 6.60 | 6.68 | *7.49* | 6.38 | 6.05 | *6.85* | 6.97 | 7.06 | *7.36* |
| 10dB | 9.68 | 9.66 | *10.50* | 9.63 | 9.59 | *10.46* | 9.62 | 9.33 | *10.12* | 10.22 | 10.27 | *10.54* |

**Table 2**. PESQ score of enhanced speech under various conditions (dB).

| SNR | *White noise* | | | *NONWGN* | | | *F16 cockpit noise* | | | *Babble noise* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MS | IMCRA | CSHMM | MS | IMCRA | CSHMM | MS | IMCRA | CSHMM | MS | IMCRA | CSHMM |
| 0dB | 1.96 | 2.00 | *2.11* | 1.96 | 2.02 | *2.14* | 1.85 | 1.81 | *1.95* | 1.79 | 1.77 | *1.80* |
| 5dB | 2.31 | 2.32 | *2.43* | 2.35 | 2.37 | *2.48* | 2.20 | 2.15 | *2.26* | 2.15 | 2.15 | *2.18* |
| 10dB | 2.61 | 2.63 | *2.73* | 2.67 | 2.68 | *2.78* | 2.54 | 2.50 | *2.60* | 2.50 | 2.49 | *2.53* |

Two objective evaluations based on speech enhancement are used to evaluate the noise estimators. A decision-directed Wiener filter for speech enhancement [9] is combined with these noise estimators. They are utilized to track the priori and posteriori SNR of the Wiener filter. Then, the noise estimators are indirectly assessed via objective evaluation of the enhanced speech quality. One evaluation is segmental SNR (SegSNR), but not the improved SegSNR. The other is perceptual evaluation of speech quality (PESQ), which is a mean opinion score (MOS)-like objective evaluation that facilitates the objective evaluation of audio signal quality based upon perceptual criteria. Tables 1 and 2 respectively show the objective evaluations of the enhanced speech quality under various conditions. Incorporating SegSNR and PESQ for all conditions, it is readily seen that the CSHMM estimator consistently achieves the best results under all conditions.

## 6. CONCLUSION

In this paper, we propose a binary-state CSHMM to model the log-power sequence. CSHMM is actually a temporal recursive filter derived from the ML framework of HMM. It has three distinctions from the popular time-recursive technique [1] - [4]. Firstly, CSHMM incorporates the state transition between speech and nonspeech into the time-recursive process while the popular techniques do not. Secondly, the noise logarithmic power is estimated in the ML sense while the popular ones can not warrant the estimation optimality. Lastly, the initialization condition of CSHMM is relaxed since it does not require the first several frames of an utterance to be nonspeech. Even if speech signal is absent, the proposed model can be correctly initialized by using the constraints. It is therefore more practical than the popular techniques. This advantage is confirmed by our preliminary experiments. The proposed estimator is an extension of the unsupervised learning framework in [10]. Since the time correlation is fully considered within HMM, the proposed model is more desirable than that in [10].

## 7. REFERENCES


[1] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12 - 15, 2002.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466 - 475, 2003.

[3] S. Rangachari, C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, pp. 220-231, 2006.

[4] J. Erkelens, R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. on Speech, Audio, and Language Processing*, vol. 16, no. 6, pp. 1112 - 1123, Aug. 2008.

[5] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1 - 3, Jan. 1999.

[6] S. Gazor, W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498 - 505, Sep. 2003.

[7] V. Krishnamurthy, J. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback - Leibler information measure," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2557-2573, 1993.

[8] 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.

[9] P. Scalart, J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, USA, 1996, pp. 629 - 632.

[10] Dongwen Ying, Yonghong Yan, Jianwu Dang, Frank Soong, "Voice activity detection based on an unsupervised learning framework," *To appear in IEEE Trans. on Audio, Speech, and Language Processing*.