A Robust Estimation Method of Noise Mixture Model for Noise Suppression

Masakiyo Fujimoto, Shinji Watanabe, and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Japan

{fujimoto.masakiyo, watanabe.shinji, nakatani.tomohiro}@lab.ntt.co.jp

Abstract

Vector Taylor series (VTS)-based noise suppression usually employs a single Gaussian distribution for the noise model. However, it is insufficient for non-stationary noise which has a multi-peak distribution. It is very complex to estimate multi-peak distribution of the noise, when we deal with the noise as random variables or hidden variables. To solve these problems, we investigate a way of estimating the noise mixture model by using a minimum mean squared error (MMSE) estimate of the noise. By iterating the MMSE estimation of noise and noise model estimation, the proposed method realizes the simultaneous optimization of both the observed signal model and the noise model. The proposed method significantly outperformed the VTS-based approach, and the maximum improvement in the word error rate was about 12%.

Index Terms: noise suppression, noise model estimation, MMSE estimation.

1. Introduction

Ensuring noise robustness is an important factor as regards the practical use of automatic speech recognition (ASR). Noise robust ASR techniques are typically classified into two approaches. One is front-end processing including robust feature extraction [1] and noise suppression [2]-[6]. The other is back-end processing including model compensation [7, 8] and model adaptation [9, 10] techniques. Among them, the noise suppression is the most general technique, because it is applied to both ASR and hearing aids for human communication.

In recent research, statistical model-based approaches have attracted attention as powerful tools for noise robust ASR. Widely used statistical model-based techniques are the minimum mean squared error (MMSE)-based approach [3], the vector Taylor series (VTS)-based approach [4], and the switching linear dynamical system (SLDM) [5]. We have also proposed a model-based approach, which we call the model-based Wiener filter (MBWF) technique [6].

Of the various statistical model-based approaches, the VTS-based approach is one of the strongest techniques for noise robust ASR. The VTS-based approach estimates the statistics of the non-linear mismatch function between a clean speech signal and an observed signal by utilizing the parameters of a clean speech model and a noise model. To cope with fluctuations in the noise environment, the parameter update scheme of the non-linear mismatch function is simplified based on a Taylor series-based linear approximation.

An accurate update of the mismatch function, i.e., an estimation of the noise model parameters is a crucial factor in the VTS-based approach. Usually, a VTS-based approach (for front-end processing) employs a single Gaussian distribution for the noise model. In this case, if the noise has a unimodal distribution e.g., stationary noise characteristics, the simple probability density function (PDF), namely, a single Gaussian distribution, may be sufficient for the noise model. However, most of the noises observed in the real world have non-stationary characteristics. If the noise has highly non-stationary characteristics, the PDF of the noise will have a multi-peak distribution such as a Gaussian mixture model (GMM) or a hidden Markov model (HMM) based on the long-term statistical analysis. In such a case, a single Gaussian distribution is insufficient for the noise model. In addition, the noise parameters are estimated as hidden variables which maximize the likelihood function respected to the observed signal. However, it is very complex to estimate multi-peak distribution of the noise, when we deal with the noise as hidden variables.

We considered the problem of the VTS-based approach, and propose a robust unsupervised estimation method for a noise mixture model. Usually, the observable parameter of the front-end processing is restricted to the observed noisy speech signal. In this paper, to estimate the noise distribution characteristics by the noise mixture model in a computational tractable way, we use the noise estimate obtained from observed signal based on the MMSE estimation. The proposed method involves the voice activity detection (VAD) scheme based on statistical models [6], and the MMSE estimates of noise are obtained by utilizing the information of speech absent or speech activity. After the MMSE estimation of the noise, the noise model is estimated with a suitable model topology that matches to the distribution characteristics of the noise. The MMSE estimation of the noise and noise model estimation is performed iteratively based on the EM algorithm.

The proposed method was evaluated for ASR in highly non-stationary noise environments, and it was proved that the proposed method improves ASR accuracy in results obtained for non-stationary noise environments using a VTS-based approach.

2. Review of vector Taylor series approach

This section reviews the VTS-based approach [4]. The VTS-based approach applies noise compensation to a clean speech model by using the parameters of a noise model and a non-linear mismatch function. In this paper, we utilize a clean speech model with two internal states, i.e., states of silence (speech absent) and speech (speech activity). Each state is modeled by a GMM with K Gaussians in the M-dimensional logarithm output energy of the Mel-filter bank (LMFB) domain in advance.

In the VTS-based approach, the zero-th order VTS (VTS-0) and the first order VTS (VTS-1) are widely used for noise compensation.

2.1. VTS-0: Zero-th order VTS

With VTS-0, model compensation is applied only to the mean vector. When the M-dimensional observation (noisy speech) vector $\boldsymbol{O}_t$ in the LMFB domain at the t-th frame is given, the initial mean vector of the noise model is derived as $\boldsymbol{\mu}_N^{Ini} =$

$\frac{1}{U}\sum_{t=0}^{U-1}\boldsymbol{O}_t$. With this parameter, VTS-0 is derived as follows:

$$w_{O,j,k} = w_{S,j,k} \tag{1}$$

$$\begin{aligned}\boldsymbol{\mu}_{O,j,k} &= \boldsymbol{\mu}_{S,j,k} + \log\left(\mathbf{1} + \exp\left(\boldsymbol{\mu}_N^{Ini} - \boldsymbol{\mu}_{S,j,k}\right)\right) \\ &= h\left(\boldsymbol{\mu}_{S,j,k}, \boldsymbol{\mu}_N^{Ini}\right)\end{aligned} \tag{2}$$

$$\boldsymbol{\Sigma}_{O,j,k} \simeq \boldsymbol{\Sigma}_{S,j,k}\ , \tag{3}$$

where $w_{O,j,k}$, $\boldsymbol{\mu}_{O,j,k}$, $\boldsymbol{\Sigma}_{O,j,k}$, $w_{S,j,k}$, $\boldsymbol{\mu}_{S,j,k}$, and $\boldsymbol{\Sigma}_{S,j,k}$ denote the Gaussian weights, mean vectors, and diagonal variance matrices of the compensated model and clean speech model, respectively. j and k denote indices of the state and Gaussian, respectively. The operations $\log(\cdot)$ and $\exp(\cdot)$ are independently applied to each vector element, and $\mathbf{1} = \{1, \cdots, 1\}^{\mathrm{T}}$.

2.2. VTS-1: First order VTS

VTS-1 compensates for the parameter difference between the initial noise model and the target noise model by using the first order Taylor series-based linear approximation as follows[1]:

$$\boldsymbol{\mu}_{O,j,k} \simeq h\left(\boldsymbol{\mu}_{S,j,k}, \boldsymbol{\mu}_N^{Ini}\right) + \boldsymbol{H}_{j,k}\left(\boldsymbol{\mu}_N - \boldsymbol{\mu}_N^{Ini}\right) \tag{4}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{O,j,k} &\simeq \left(\boldsymbol{I} - \boldsymbol{H}_{j,k}\right)\boldsymbol{\Sigma}_{S,j,k}\left(\boldsymbol{I} - \boldsymbol{H}_{j,k}\right)^{\mathrm{T}} + \boldsymbol{H}_{j,k}\boldsymbol{\Sigma}_N\boldsymbol{H}_{j,k}^{\mathrm{T}} \\ &= g\left(\boldsymbol{\Sigma}_{S,j,k}, \boldsymbol{\Sigma}_N, \boldsymbol{H}_{j,k}\right)\ ,\end{aligned} \tag{5}$$

with the Jacobian matrix $\boldsymbol{H}_{j,k} = diag\left\{\frac{\partial h(\boldsymbol{\mu}_{S,j,k}, \boldsymbol{\mu}_N^{Ini})}{\partial \boldsymbol{\mu}_N^{Ini}}\right\}$, where $\boldsymbol{\mu}_N$, $\boldsymbol{\Sigma}_N$, and $\boldsymbol{I}$ denote the mean vector and variance matrix of the target noise model and the identity matrix, respectively. The initial variance matrix of the noise model is derived as $\boldsymbol{\Sigma}_N^{Ini} = diag\left\{\frac{1}{U}\sum_{t=0}^{U-1}\left(\boldsymbol{O}_t - \boldsymbol{\mu}_N^{Ini}\right)\left(\boldsymbol{O}_t - \boldsymbol{\mu}_N^{Ini}\right)^{\mathrm{T}}\right\}$.

Under the single Gaussian assumption, the parameter set of the noise model $\boldsymbol{\lambda}_N = \{\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\}$ is optimized as the parameters that maximize the following likelihood function $P_O(\cdot)$:

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_N &= \arg\max_{\boldsymbol{\lambda}_N}\sum_t \log P_O\left(\boldsymbol{O}_t\,|\boldsymbol{\lambda}_O\right) \\ &= \arg\max_{\boldsymbol{\lambda}_N}\sum_t \log P_O\left(\boldsymbol{O}_t\,|\mathrm{VTS}\left(\boldsymbol{\lambda}_S, \boldsymbol{\lambda}_N\right)\right)\ ,\end{aligned} \tag{6}$$

with the parameter sets $\boldsymbol{\lambda}_S = \left\{w_{S,j,k}, \boldsymbol{\mu}_{S,j,k}, \boldsymbol{\Sigma}_{S,j,k}\right\}$ and $\boldsymbol{\lambda}_O = \left\{w_{O,j,k}, \boldsymbol{\mu}_{O,j,k}, \boldsymbol{\Sigma}_{O,j,k}\right\}$, where the operation $\mathrm{VTS}(\cdot)$ is the VTS-1 transformation given by Eqs. (4) and (5).

2.3. Problem with VTS-based approach

Fig. 1 shows an example of the noise distribution and the noise model estimated by VTS-1. In the figure, the noise histogram has a multi-peak characteristic, thus, a single Gaussian PDF is unsuitable for the noise model. In addition, the noise model estimated by VTS-1 has a fatal error compared with the true noise histogram or the initial model parameterized by $\boldsymbol{\mu}_N^{Ini}$ and $\boldsymbol{\Sigma}_N^{Ini}$. In VTS-1, the optimization criterion of the parameters set $\boldsymbol{\lambda}_N$ is given by Eq. (6). However, this criterion is primarily the optimization scheme for the observed signal model, and the optimality of the noise parameter estimation is not directly ensured. Therefore, the performance of VTS-1 degrades in terms of noise with the multi-peak distribution due to the unsuitable model topology and optimization criterion.

[1] Although VTS-1 can compensate for the fluctuation of the speech parameter, this paper focus on only noise compensation.

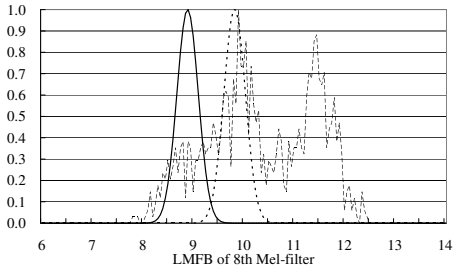


Figure 1: Examples of noise models estimated by the VTS. This figure shows the LMFB distribution of the 8-th Mel filter (center frequency of 1022.4 Hz) in platform noise with 0 dB SNR. The broken, dotted, and solid lines show the true noise histograms, the initial noise model used for VTS-0 and VTS-1, and the noise model estimated by VTS-1 respectively.

3. Estimation of noise mixture model

To overcome the problem with the VTS-based approach, the proposed method investigates an unsupervised technique for estimating for the noise mixture model by using the LMFB vector of noise $\boldsymbol{N}_t$ estimated by the MMSE estimator. In the proposed method, the MMSE estimation of $\boldsymbol{N}_t$ and the noise mixture model estimation are performed iteratively based on the EM algorithm. In this paper, the topology of the noise model is defined by a GMM with L Gaussians, thus, the parameter set $\boldsymbol{\lambda}_N$ is expanded as $\boldsymbol{\lambda}_N = \left\{w_{N,l}, \boldsymbol{\mu}_{N,l}, \boldsymbol{\Sigma}_{N,l}\right\}$, where l and $w_{N,l}$ denote the Gaussian index and Gaussian weight, respectively.

The MMSE estimate of $\boldsymbol{N}_t$ is derived as:

$$\hat{\boldsymbol{N}}_t = \mathcal{E}\left\{\boldsymbol{N}_t\,|\boldsymbol{O}_t, \boldsymbol{\lambda}_S, \boldsymbol{\lambda}_N\right\}\ . \tag{7}$$

Then, with the estimated noise $\hat{\boldsymbol{N}}_t$, the parameter set $\boldsymbol{\lambda}_N$ is directly estimated based on following criterion instead of Eq. (6),

$$\hat{\boldsymbol{\lambda}}_N = \arg\max_{\boldsymbol{\lambda}_N}\sum_t \log P_N\left(\hat{\boldsymbol{N}}_t\,|\boldsymbol{\lambda}_N\right)\ , \tag{8}$$

where $P_N(\cdot)$ denotes the likelihood function of the noise model. By iterating these processes, the noise model is successfully optimized.

3.1. Initialization

The initial parameter set of the noise model is given as $\boldsymbol{\lambda}_N^{(i=0)} = \left\{w_{N,l}^{(i=0)} = \frac{1}{L}, \boldsymbol{\mu}_{N,l}^{(i=0)} = \boldsymbol{\mu}_N^{Ini}, \boldsymbol{\Sigma}_{N,l}^{(i=0)} = \boldsymbol{\Sigma}_N^{Ini}\right\}$.

3.2. E-step

3.2.1. Model composition

The first stage of **E-step** is the composition of an observed signal model with parameter set of previous iteration, $\boldsymbol{\lambda}_N^{(i-1)}$. During the model composition, each state of the clean speech model has K Gaussians and the noise model has L Gaussians. Thus, the number of Gaussians contained in each state of the composed model is expanded to $K \times L$. At the i-th iteration, the model parameter set $\boldsymbol{\lambda}_O^{(i)}$ is derived as

$$\boldsymbol{\lambda}_O^{(i)} = \left\{\begin{array}{l} w_{O,j,k,l}^{(i)} = w_{S,j,k} \cdot w_{N,l}^{(i-1)}\ , \\ \boldsymbol{\mu}_{O,j,k,l}^{(i)} = h\left(\boldsymbol{\mu}_{S,j,k}, \boldsymbol{\mu}_{N,l}^{(i-1)}\right)\ , \\ \boldsymbol{\Sigma}_{O,j,k,l}^{(i)} = g\left(\boldsymbol{\Sigma}_{S,j,k}, \boldsymbol{\Sigma}_{N,l}^{(i-1)}, \boldsymbol{H}_{j,k,l}^{(i)}\right) \end{array}\right\}\ . \tag{9}$$

3.2.2. Expectation of cost function

When $\boldsymbol{O}_{0:T-1} = \boldsymbol{O}_0, \cdots, \boldsymbol{O}_{T-1}$ is given, the expectation of the cost function related to the parameter set $\boldsymbol{\lambda}_O^{(i)}$ is derived as follows:

$$Q\left(\boldsymbol{O}_{0:T-1}, \boldsymbol{\lambda}_O^{(i)}\right) = \sum_{t,j,k,l} P_{t,j}^{(i)} P_{t,j,k,l}^{(i)} \times \left(\log w_{O,j,k,l}^{(i)} + \log \mathcal{N}\left(\boldsymbol{O}_t; \boldsymbol{\mu}_{O,j,k,l}^{(i)}, \boldsymbol{\Sigma}_{O,j,k,l}^{(i)}\right)\right) , \quad (10)$$

where $\mathcal{N}(\cdot)$, $P_{t,j}^{(i)}$, and $P_{t,j,k,l}^{(i)}$ denote the PDF of a Gaussian distribution and *a posteriori* probabilities with respect to the state index j and the Gaussian indices k and l, respectively.

3.3. M-step

3.3.1. MMSE estimation of noise

The MMSE estimate of $\boldsymbol{N}_t^{(i)}$ defined by Eq. (7) is derived as

$$\begin{aligned}\hat{\boldsymbol{N}}_t^{(i)} &= P_{t,j=1}^{(i)} \boldsymbol{O}_t + P_{t,j=2}^{(i)} \left(\boldsymbol{O}_t - \mathcal{E}\left\{\boldsymbol{G}_t^{(i)}\right\}\right) \\ &= P_{t,j=1}^{(i)} \boldsymbol{O}_t + P_{t,j=2}^{(i)} \left(\boldsymbol{O}_t - \sum_{k,l} P_{t,j=2,k,l}^{(i)} \boldsymbol{G}_{k,l}^{(i)}\right) ,\end{aligned} \quad (11)$$

with the statistics of the mismatch function,

$$\begin{aligned}\boldsymbol{G}_{k,l}^{(i)} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P\left(\boldsymbol{S}\,|j=2,k,\boldsymbol{O}_t\right) P\left(\boldsymbol{N}^{(i-1)}\,|l,\boldsymbol{O}_t\right) \\ &\quad \times \left(h\left(\boldsymbol{S},\boldsymbol{N}^{(i-1)}\right) - \boldsymbol{N}^{(i-1)}\right) d\boldsymbol{S} d\boldsymbol{N}^{(i-1)} \\ &\simeq \boldsymbol{\mu}_{O,j=2,k,l} - \boldsymbol{\mu}_{N,l}^{(i-1)} ,\end{aligned} \quad (12)$$

where $\boldsymbol{G}_t^{(i)}$ denotes the MMSE estimate of the mismatch function. In Eq. (11), the MMSE estimate of $\boldsymbol{N}_t^{(i)}$ is given as the weighted average of two $\boldsymbol{N}_t^{(i)}$ candidates. Since the state $j=1$ is the silence (speech absence) state, the candidate for $\boldsymbol{N}_t^{(i)}$ is obtained as observed signal $\boldsymbol{O}_t$. With the speech activity state, $j=2$, the candidate for $\boldsymbol{N}_t^{(i)}$ is obtained by using the MMSE estimation with $P_{t,j,k,l}^{(i)}$. Namely, the proposed method implicitly involves the VAD to the noise model estimation [6].

3.3.2. Noise mixture model estimation with MMSE estimates

After the MMSE estimation of $\boldsymbol{N}_t^{(i)}$, $\boldsymbol{\lambda}_N^{(i)}$ is estimated by using a standard maximum likelihood estimation based on Eq. (8).

4. Noise suppression

4.1. Model-based Wiener filter

The noise is suppressed using an MBWF in accordance with our previous work [6]. With this method, the optimum Wiener filter $W_{t,m}^{Mel}$ is given as an MMSE estimate by using the parameters of the speech model, the estimated noise model, and *a posteriori* probabilities $P_{t,j}$ and $P_{t,j,k,l}$ as follows:

$$W_{t,m}^{Mel} = \sum_{j,k,l} P_{t,j} P_{t,j,k,l} \frac{\exp\left(\mu_{S,j,k,m}\right)}{\exp\left(h\left(\mu_{S,j,k,m}, \mu_{N,l,m}\right)\right)} , \quad (13)$$

where m denotes the index of the vector element.

By applying a third order spline interpolation, $W_{t,m}^{Mel}$ can be transformed into a linear-scaled filter $W_{t,n}^{Lin}$. The noise suppressed signal $\hat{s}_\tau$ is obtained by applying $W_{t,n}^{Lin}$ and an inverse fast Fourier transform to the complex spectrum of the observed signal $O_{t,n}$.

4.2. Processing flow

The following algorithm summarizes the proposed method, and is applied to each utterance.

Algorithm 1 Noise model estimation and noise suppression

1: Initialize $\boldsymbol{\lambda}_N^{(i=0)}$ (See Sec. 3.1.)
2: **repeat**
3: Model composition (See Sec. 3.2.1.)
4: Compute expectation of cost function (See Sec. 3.2.2.)
5: Estimate $\boldsymbol{N}_t^{(i)}$ for all t (See Sec. 3.3.1.)
6: Update $\boldsymbol{\lambda}_N^{(i)}$ with ML estimation (See Sec. 3.3.2.)
7: **until** convergence is achieved
8: Apply the MBWF (See Sec. 4.1.)

5. Experiments

5.1. Experimental setup

The experimental materials were 100 utterances spoken by 23 Japanese males that were taken from the Information-technology Promotion Agency (IPA)-98-TestSet. The speaking style of the speech data is read speech. Three types of highly non-stationary noises, i.e., airport lobby noise, platform noise, and street noise, were artificially added to clean speech signals by changing the SNR at three levels; 10, 5, and 0 dB. These noises include various sound sources such as babble, trains, vehicles, footsteps, and chimes. Thus, these noises have highly non-stationary characteristics. The sampling frequency of the speech data and noises was 16 kHz. The feature parameters for the noise suppression were 24th-order LMFBs that were extracted by using a Hamming window with a 20 msec frame length and a 10 msec frame shift length. The GMMs of silence and speech had $K = 128$ Gaussians. The number of Gaussians of the noise model was set at $L = 1, 2, 3, 4$. The parameter U was set at 10.

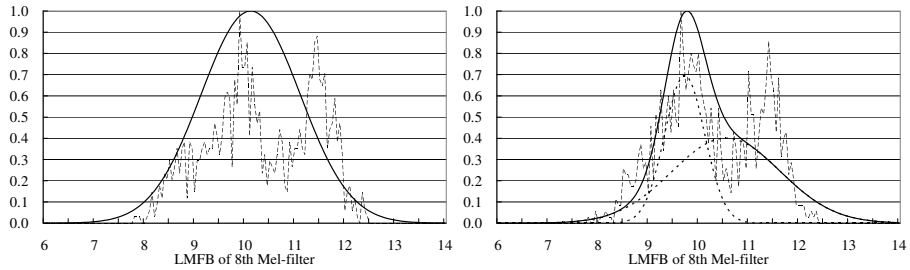
The proposed method was evaluated by using ASR. The ASR was carried out by employing a weighted finite state transducer-based decoder [11]. We used speaker independent triphone HMMs trained by clean speech. The HMM training was performed using a variational Bayesian approach [12]. The HMM topology was a three state left-to-right HMM and there were 2,364 HMM states. Each state had 16 Gaussians. The feature parameters for the ASR consisted of 12th-order MFCCs and the log energy with their first and second order derivatives. Cepstral mean normalization was applied to each utterance. The training materials for the GMMs and the HMMs were 33,820 phonetically balanced sentences spoken by 180 Japanese males.

The language model was a back-off tri-gram with Witten-Bell discounting. It was trained using 75 months' worth of Japanese newspaper articles. The vocabulary size was 20k words. The evaluation criterion for ASR was the word error rate (WER). The WER of a clean speech signal was 3.9 %.

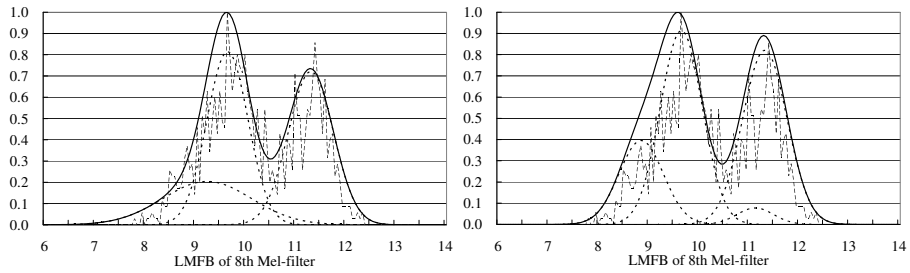
5.2. Experimental results

Fig. 2 shows noise estimation results. As seen in the figure, the noise model with three Gaussians (Fig. 2c) is well-matched to the multi-peak characteristic of the true noise histogram. On the other hand, the estimation results in Fig. 2a and 2b are not sufficient to represent the true noise histogram and over-fitting is observed on the left side of Fig. 2d. In the comparisons with the VTS-0 and VTS-1 shown in Fig. 1, the proposed method shows significant improvement of the noise model estimation.

Table 1 shows the ASR results for each method. In the table, the WERs of spectral subtraction [2] are much worse than with the other methods. As mentioned in Sec. 2.3, the WERs of VTS-1 are worse than those of VTS-0 due to the characteristics of highly non-stationary noise. With "Proposed (1 Gaussian)," the WER improvements are insignificant due to the poor



(a) Estimated noise model (1 Gaussian) (b) Estimated noise model (2 Gaussians)



(c) Estimated noise model (3 Gaussians) (d) Estimated noise model (4 Gaussians)

Figure 2: Examples of noise models estimated with the proposed method. Each panel shows the LMFB distribution of the 8-th Mel filter (center frequency of 1022.4 Hz) in platform noise with 0 dB SNR. The broken, solid, and dotted lines show the true noise histograms, the estimated noise model, and each Gaussian component in the estimated noise model, respectively.

Table 1: ASR results in WER (%)

Method	Airport lobby noise			Platform noise			Street noise			Avg.
	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	
w/o noise suppression	26.10	59.10	87.10	27.20	55.10	79.00	11.50	28.70	61.00	48.31
Spectral subtraction	25.80	46.10	78.50	30.00	50.50	75.90	12.20	21.30	36.90	41.91
VTS-0	13.90	32.40	65.80	20.80	42.50	67.30	7.50	14.30	29.00	32.61
VTS-1	17.00	39.50	72.00	24.00	43.90	70.70	7.30	14.50	29.90	35.42
Proposed (1 Gaussian)	17.10	34.20	63.10	23.10	42.40	64.50	8.40	15.60	28.80	33.02
Proposed (2 Gaussians)	13.10	**28.80**	**59.90**	20.40	37.50	60.20	**6.90**	13.20	26.60	29.62
Proposed (3 Gaussians)	13.10	29.80	60.10	17.80	37.10	**59.10**	7.00	**12.40**	**26.10**	29.17
Proposed (4 Gaussians)	**12.90**	29.70	60.00	**17.10**	**35.40**	59.30	7.60	13.30	26.90	**29.13**

topology of the noise model. However, in the other cases, the proposed method shows significant WER improvements with the mixture topology of the noise model. These results also prove the effectiveness of the proposed method.

The next problem is the selection of an adaptive model topology. This problem may be solved by using the variational Bayesian approach [12] or a Gaussian pruning technique [13].

6. Conclusions

This paper presented a method of estimating a noise mixture model with the MMSE estimates of the noise. The proposed method makes it possible to estimate the accurate model parameters of non-stationary noise with a multi-peak distribution. The evaluation results show that the proposed method significantly improves the accuracy of ASR in highly non-stationary noise environments. In the future, we will investigate the selection of an adaptive model topology.

7. References

[1] K. Ishizuka and T. Nakatani, "A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1447–1457, Nov. 2006.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, vol. 32, pp. 1109–1121, Dec. 1984.

[4] P. J. Moreno, *et al.,* "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP '96*, vol. II, pp. 733–736, May 1996.

[5] S. Windmann and R. Haeb-Umbach, "Modeling the dynamics of speech and noise for speech feature enhancement in ASR," in *Proc. of ICASSP '08*, pp. 4409–4412, Apr 2008.

[6] M. Fujimoto, *et al.,* "A study of mutual front-end processing method based on statistical model for noise robust speech recognition," in *Proc. of Interspeech '09*, pp. 1235–1238, Sept. 2009.

[7] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on SAP*, vol. 4, no. 5, pp. 352–359, May 1996.

[8] R. C. van Dalen and M. J. F. Gales, "Extended VTS for noise-robust speech recognition," *IEEE Trans. on SAP*, vol. 19, no. 4, pp. 733–743, May 2011.

[9] C. L. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[10] C. H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1241–1269, Aug. 2000.

[11] T. Hori, *et al.,* "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Trans. on ASLP*, vol. 15, no. 4, pp. 1352–1365, May 2007.

[12] S. Watanabe, *et al.,* "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. on SAP*, vol. 12, no. 4, pp. 365–381, July 2004.

[13] M. Fujimoto, *et al.,* "Voice activity detection using Gaussian pruning-based frame-wise model re-estimation," in *Proc. of Interspeech '10*, pp. 3102–3105, Sept. 2010.