NON-STATIONARY NOISE ESTIMATION METHOD BASED ON BIAS-RESIDUAL COMPONENT DECOMPOSITION FOR ROBUST SPEECH RECOGNITION

Masakiyo Fujimoto Shinji Watanabe Tomohiro Nakatani

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

ABSTRACT
This paper addresses a noise suppression problem, namely the estimation of non-stationary noise sequences. In this problem, we assume that non-stationary noise can be decomposed into stationary and non-stationary components. These components are described respectively as the bias factor and the residual signal between the bias component and noise at each frame. This decomposition clarifies the role of each component, thus enabling us to apply a suitable parameter estimation technique to each component. In this paper, the bias component is estimated by the EM algorithm with the entire observed signal sequence. On the other hand, the residual component is sequentially estimated by multiplying the extended Kalman filter with the EM algorithm. In the evaluation results, we confirmed that the proposed method improved speech recognition accuracy compared with the noise estimation methods without component decomposition.

Index Terms—speech recognition, noise suppression, non-stationary noise, component decomposition

1. INTRODUCTION
Noise robustness is one of the most crucial problems for automatic speech recognition (ASR). Noise suppression plays an important role as regards this problem, because it reduces the mismatch between corrupted speech signals and acoustic models.

The traditional noise suppression techniques are spectral subtraction [1] and the use of a Wiener filter [2]. As an alternative to these traditional methods, various noise suppression techniques have been proposed. In particular, statistical model-based approaches are widely used and various improved methods have been proposed including the minimum mean squared error (MMSE)-based approach [3], the Gaussian mixture model (GMM)-based approach [4], the vector Taylor series (VTS)-based approach [5], and the switching linear dynamical system (SLDM) [6, 7]. We have also proposed a model-based approach, which we call the model-based Wiener filter (MBWF) technique [8].

The common problem with noise suppression methods is noise estimation. If a noise has stationary characteristics, we can easily estimate the noise parameter in a non-speech period of the observed signal, e.g., the beginning of the observed signal. However, in most cases, a noise observed in the real world has non-stationary characteristics and is usually unknown in advance. Thus, robustness in the presence of non-stationary noise and the absence of a need for an a priori knowledge of noise are the most important factors for ASR in the real world. To solve such problems, various methods have been proposed for estimating non-stationary noise sequence including the Kalman filter [8], the particle filter [9, 10], and the sequential EM algorithm [11]. These methods realize sequential noise tracking, and their effectiveness is proved through ASR evaluation in non-stationary noise environments. However, since the reliable estimation of non-stationary noise is a very difficult task, conventional methods sometimes degrade the ASR accuracies due to serious noise estimation errors.

A serious estimation error may be caused by insufficient modeling of the non-stationary noise. To capture the non-stationary characteristics of noise, we introduce a noise model based on bias-residual component decomposition as shown in Fig. 1. Fig. 1 shows an example of the 2-dimensional feature space of noise $N_t$ in the logarithm output energy of the Mel-filter bank (LMFB) domain at the $t$-th frame. By using this representation, the noise $N_t$ can be decomposed into the bias component $\mu_N$ and the residual component $\tilde{N}_t$. Namely, the bias component $\mu_N$ is equivalent to the centroid of the noise feature space, and is estimated with the stationary assumption of the model. On the other hand, the non-stationary characteristics of noise are represented as a residual component $\tilde{N}_t$ from the centroid.

However, even with this component decomposition, the estimation of non-stationary model parameters tends to be unreliable, and the behavior may often be unstable. To mitigate this problem, this paper adopts an auto-regressive (AR) process for modeling the residual component $\tilde{N}_t$ in each Mel-filter bank, which can capture the dynamical characteristics of the noise. As the first step of this approach, we limit the parameters of the AR process so that they are time invariant in this paper. Even with this limitation, the noise component decomposition and the use of the dynamical noise model allow us to greatly reduce the modeling errors in the non-stationary noise characteristics. In addition, with this configuration, we can employ reliable and stable parameter estimation techniques for each noise component. These advantages are particularly important for non-stationary noise estimation.

In this paper, the bias component $\mu_N$ is estimated using the EM algorithm with the entire observed signal sequence. On the other hand, the residual component $\tilde{N}_t$ is sequentially estimated by an AR process-based multiplied extended Kalman filter (MEKF). The parameters of the AR process are also optimized using the EM algorithm with the assumption that the state transition process is stationary for $\tilde{N}_t$. In each component estimation, a statistical model
of clean speech is used for parameter estimation. The statistical model of clean speech has two internal states, namely states of silence and clean speech. Therefore, the proposed method estimates each component by considering the speech activity or the speech absence at each frame. In other words, the proposed method implicitly involves voice activity detection (VAD), and simultaneously executes the VAD and estimation of each component.

The estimated noise is utilized for our noise suppression method, MBWF. This method is evaluated for ASR in highly non-stationary noise environments, and proved that the proposed method improves ASR accuracy in results obtained for non-stationary noise environments using a noise estimation method without decomposition.

2. NON-STATIONARY NOISE ESTIMATION BASED ON NOISE COMPONENT DECOMPOSITION

2.1. Noise component decomposition

In noise estimation, \( \tilde{N}_t \) is decomposed into bias and residual components as follows:

\[
N_t = \mu_N + \tilde{N}_t
\]

In Eq. (1), the bias component \( \mu_N \) can be defined as the mean vector of the noise component included in sequence of the observed signal \( O_t \), and a stationary parameter of \( N_t \). On the other hand, the residual component \( \tilde{N}_t \) is defined as the residual signal between \( N_t \) and \( \mu_N \), and is the non-stationary parameter of \( N_t \) at each frame. Here, we assume that the non-stationary characteristics of \( N_t \) are represented by the AR process as follows:

\[
\tilde{N}_{t+1} = F \tilde{N}_t + U_t \quad U_t \sim \mathcal{N}(0, \Sigma_U)
\]

where \( F \) is a time invariant AR coefficient matrix. \( U_t \) and \( \Sigma_U \) denote the error signal of the AR process and the diagonal variance matrix of \( U_t \), respectively. By substituting Eq. (2) into Eq. (1), \( N_t \) can be represented by a biased AR process as follows:

\[
N_t = \mu_N + F \tilde{N}_{t-1} + U_{t-1}
\]

Thus, the proposed method is equivalent to the application of the biased AR process to the signal model of \( N_t \).

In Eq. (3), \( \tilde{N}_t \) is modeled by the stationary AR process with the AR matrix \( F \). Although the non-stationary AR process with the time varying parameter \( F_t \) is more desirable for the model of \( N_t \) in this paper we focus on the effectiveness of the noise component decomposition method with time invariant \( F \).

In the model, \( \mu_N \) is estimated by using the EM algorithm described in Sec. 2.2. \( \tilde{N}_t \) is sequentially estimated by using the AR-based MEKF with the EM algorithm described in Sec. 2.3.

2.2. Bias component estimation

In the EM algorithm-based estimation, the bias component \( \mu_N \) is estimated as the vector that maximizes the likelihood between a sequence of \( O_t \) and the observed signal model. When the parameters of speech model are given, the parameters of the observed signal model at the \( t \)-th frame can be derived by VTS [5]. In the proposed method, the speech model has two internal states, i.e., states of silence and clean speech, and each state is modeled by a GMM in the LMFH domain in advance.

[Initialization]

The initial value of \( \mu_N \) is estimated as \( \mu_N^{(0)} = \frac{1}{T} \sum_{t=0}^{T-1} O_t \) by using first \( M \) frames of \( \tilde{N}_t \). The superscript \( i \) denotes the iteration index. Then, the diagonal variance matrix is also derived as \( \Sigma_N = \text{diag} \left\{ \sum_{t=0}^{T-1} (O_t - \mu_N^{(0)}) (O_t - \mu_N^{(0)})^T \right\} \).

[E-step]

When \( O_{0:T-1} = O_0, \cdots, O_{T-1} \) is given, the expectation of an auxiliary function related to \( \mu_N \) is derived as

\[
Q \left( O_{0:T-1}, \mu_N^{(i)} \right) = \sum_{t,j,k} P_{t,j,k}^{(i)} \left[ \log w_{j,k} + \log \mathcal{N}(O_t; \mu_{O,j,k}^{(i)}, \Sigma_{O,j,k}^{(i)}) \right]
\]

where

\[
\mu_{O,j,k}^{(i)} = \mu_{S,j,k} + \log \left( 1 + \exp \left( \mu_{S,j,k}^{(i)} - \mu_{S,j,k} \right) \right)
\]

\[
\Sigma_{O,j,k}^{(i)} = \left( I - H_{j,k}^{(i)} \right) \Sigma_{S,j,k} \left( I - H_{j,k}^{(i)} \right)^T + H_{j,k}^{(i)} \Sigma_{N,t} H_{j,k}^{(i)T}
\]

\[
H_{j,k}^{(i)} = \text{diag} \left\{ \partial h \left( \mu_{S,j,k}, \mu_{O,j,k}^{(i)} \right) / \partial \mu_{O,j,k}^{(i)} \right\}
\]

\[
P_{t,j,k}^{(i)} = \frac{\sum_{j',k'} w_{j',k'} \mathcal{N}(O_t; \mu_{O,j',k'}^{(i)}, \Sigma_{O,j',k'}^{(i)})}{\sum_{j',k'} \sum_{j,k'} w_{j',k'} \mathcal{N}(O_t; \mu_{O,j',k'}^{(i)}, \Sigma_{O,j',k'}^{(i)})}
\]

[M-step]

Since the auxiliary function defined by Eq. (5) is given by a highly non-linear function, the closed-form solution related to \( \mu_N^{(i)} \) is impossible. Therefore, we employ Newton’s method to maximize Eq. (5) as follows:

\[
\mu_{N}^{(i+1)} \leftarrow \mu_{N}^{(i)} - \left( \nabla^2 Q^{(i)} \right)^{-1} \nabla Q^{(i)},
\]

where \( \nabla Q^{(i)} \) and \( \nabla^2 Q^{(i)} \) denote the gradient vector and Hessian matrix of the auxiliary function, respectively.

By iterating the E-step and the M-step until convergence, the optimum estimation is obtained as \( \mu_N = \mu_N^{(i)} \). The convergence criterion is \( \left| Q \left( O_{0:T-1}, \mu_N^{(i)} \right) - Q \left( O_{0:T-1}, \mu_N^{(i-1)} \right) \right| / T \leq \eta \).

2.3. Residual component estimation

2.3.1. State-space model

The Kalman filter-based techniques require a definition of the signal model called a state-space model, which consists of a state transition equation and an observation equation. For the state transition equation, we employ the signal model of Eq. (2). On the other hand, the observation equation is given by the following non-linear function,

\[
O_t = h \left( \mu_{S,j,k}, \mu_N + \tilde{N}_t \right) + V_{t,j,k}
\]

\[
V_{t,j,k} \sim \mathcal{N}(0, \Sigma_{S,j,k})
\]

4817
where \( V_{t,j,k} \) denotes an error signal between the LMFB of clean speech \( S_t \) and \( S_{t,j,k} \).

Since the proposed method utilizes GMMs of silence and clean speech and each GMM consists of \( K \) Gaussians, \( 2K \) types of observation equations are derived from Eq. (12). Using these observation equations, the Kalman filter is multiplied into \( 2K \) types. This method is called the MEKF technique.

### 2.3.2. Optimization of AR matrix

The stationary AR matrix \( F \) is also optimized by the EM algorithm.

#### Initialization

The initial value of \( F \) is set at \( \hat{F}^{(i=0)} = I \).

#### E-step

By using the state space model described in Sec. 2.3.1, the likelihood function of the MEKF is derived as

\[
Q \left( O_{0:T-1} | F^{(i)} \right) = \sum_t \log \mathcal{N} \left( \tilde{N}_t | \hat{F}^{(i)} \tilde{N}_{t-1} , \Sigma_U \right) + \sum_{t,j,k} \tilde{P}_{t,j,k}^{(i)} \tilde{P}_{t,j,k}^{(i)} \left( \log w_{j,k} + \log \mathcal{N} \left( O_t | \hat{\mu}^{(i)}_{O,t,j,k} , \Sigma^{(i)}_{O,t,j,k} \right) \right),
\]

where \( b \left( O_t \right) \) is the likelihood of the MEKF at the \( t \)-th frame. Therefore, the expectation of an auxiliary function related to \( \hat{F} \) is derived as

\[
\mathbb{E} \left[ Q \left( O_{0:T-1} | F \right) \right] = \log \mathcal{N} \left( \tilde{N}_t | \hat{F}^{(i)} \tilde{N}_{t-1} , \Sigma_U \right) + \sum_{t,j,k} \tilde{P}_{t,j,k}^{(i)} \tilde{P}_{t,j,k}^{(i)} \left( \log w_{j,k} + \log \mathcal{N} \left( O_t | \hat{\mu}^{(i)}_{O,t,j,k} , \Sigma^{(i)}_{O,t,j,k} \right) \right),
\]

where

\[
\hat{\mu}^{(i)}_{O,t,j,k} = h \left( \mu_{S,t,j,k} , \mu_N + \tilde{N}_{t,j,k}^{(i)} \right)
\]

\[
\Sigma^{(i)}_{O,t,j,k} = g \left( \Sigma_{S,t,j,k} , \Sigma_{N,t,j,k} , \tilde{H}^{(i)}_{t,j,k} \right)
\]

\[
\tilde{H}^{(i)}_{t,j,k} = \text{diag} \left\{ \partial h \left( \mu_{S,t,j,k} , \mu_N + \tilde{N}_{t,j,k}^{(i)} \right) / \partial \tilde{N}_{t,j,k}^{(i)} \right\},
\]

\[
\tilde{P}_{t,j,k}^{(i)} = \tilde{P}_{t,j,k}^{(i)} \frac{\tilde{P}_{t,j,k}^{(i)}}{\tilde{P}_{t,j,k}^{(i)} + \Sigma_{N,t,j,k}^{(i)}}
\]

#### Unification step

Each estimated candidate of the MEKF is unified by weighted averaging with a posteriori probabilities \( \tilde{P}_{t,j,k}^{(i)} \) as follows:

\[
\hat{N}_t^{(i)} = \sum_{j,k} \tilde{P}_{t,j,k}^{(i)} \tilde{N}_{t,j,k}^{(i)}
\]

\[
\Sigma_{N,t}^{(i)} = \sum_{j,k} \tilde{P}_{t,j,k}^{(i)} \tilde{N}_{t,j,k}^{(i)}
\]

### 3. Noise Suppression

#### 3.1. Model-based Wiener filter

The noise suppression is carried out using an MBWF in accordance with our previous work [8]. In this method, the optimum Wiener filter \( W_{t,t}^{\text{opt}} \) is given as an MMSE estimate by using the parameters of the speech model, estimated noise, and a posteriori probabilities \( \tilde{P}_{t,j,k} \) as follows:

\[
W_{t,t}^{\text{opt}} = \frac{\sum_{j,k} \tilde{P}_{t,j,k} \tilde{P}_{t,j,k} \Sigma_{N,t,j,k} \Sigma_{N,t,j,k}}{h \left( \mu_{S,t,j,k} , \mu_N + \tilde{N}_{t,j,k} \right)}
\]

where \( l \) denotes the index of the vector element.

By applying a third order spline interpolation, \( W_{t,t}^{\text{opt}} \) can be transformed into a linear-scaled filter \( W_{t,t}^{\text{opt}} \). The noise suppressed signal \( s_t \) is obtained by applying \( W_{t,t}^{\text{opt}} \) and an inverse fast Fourier transform to a complex spectrum of the observed signal \( o_{t,n} \).

#### 3.2. Processing flow

The following algorithm summarizes the overall proposed method.

**Algorithm 1**: Noise estimation and suppression

1. Initialize \( \mu_N \)
2. repeat
3. Update \( \mu_N \) (See Sec. 2.2.)
4. until convergence is achieved
5. Initialize \( \hat{N}_t \) and \( F \)
6. repeat
7. Update \( \hat{N}_t \) for all \( t \) (See Sec. 2.3.3.)
8. Update \( F \) (See Sec. 2.3.2.)
9. until convergence is achieved
10. Apply the MBWF with \( \mu_N \) and \( \hat{N}_t \) (See Sec. 3.1.)
Table 1. ASR results in WER (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>Noise model</th>
<th>Airport lobby noise</th>
<th>Platform noise</th>
<th>Street noise</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10 dB</td>
<td>5 dB</td>
<td>0 dB</td>
<td>10 dB</td>
</tr>
<tr>
<td>w/o noise suppression</td>
<td>$N = 0$</td>
<td>30.2</td>
<td>65.5</td>
<td>57.5</td>
<td>35.6</td>
</tr>
<tr>
<td>MBWF</td>
<td>$N = \mu N$</td>
<td>15.7</td>
<td>34.7</td>
<td>66.9</td>
<td>21.2</td>
</tr>
<tr>
<td>Bias estimation</td>
<td>$N = \mu N$</td>
<td>13.4</td>
<td>32.8</td>
<td>62.9</td>
<td>20.8</td>
</tr>
<tr>
<td>Residual estimation</td>
<td>$N = N_t$</td>
<td>14.6</td>
<td>32.5</td>
<td>64.8</td>
<td>18.4</td>
</tr>
<tr>
<td>Proposed</td>
<td>$N = \mu N + N_t$</td>
<td>12.9</td>
<td>30.9</td>
<td>61.2</td>
<td>17.6</td>
</tr>
</tbody>
</table>

4. EXPERIMENTS

4.1. Experimental setup

The experimental materials were 100 utterances spoken by 23 Japanese males that were taken from the IPA (information-technology promotion agency, Japan)-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 signal by changing the SNR at 3 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 clean speech had 256 Gaussians. The parameters $M$, $\Sigma_D$, and $n$ were set at 10, diag(0.001), and 0.001, respectively. In the biased AR process, we used a first order AR matrix $F$.

The proposed method was evaluated by using ASR. The ASR was carried out by using a weighted finite state transducer-based decoder [13]. We used speaker independent triphone HMMs trained by clean speech. The HMM topology was a three state left-to-right HMM and the amount of HMM states was 2,000. Each state had 16 Gaussians. The feature parameters for the ASR consisted of 12th order MFCCs and log energy with their first and second order derivatives. Cepstral mean normalization is not applied to each utterance. The language model was a back-off tri-gram with Witten-Bell discounting. 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 3.8 %.

4.2. Experimental results

Table 1 shows the ASR results for each method. In the table, “MBWF,” “Bias estimation,” “Residual estimation,” and “Proposed” represent results obtained with the MBWF described in Sec.3.1 with noise parameter $\mu N_{0}$ estimated by using the first $M$ frames of $O_{t}$, the “MBWF” with bias estimation alone ($N_{t}$ is set at 0), the “MBWF” with residual estimation alone ($\mu N$ is set at 0), and the “MBWF” with both bias and residual estimation, respectively.

As seen in the table, the proposed method “Proposed” improves the WER compared with the other methods. The WERs of “Bias estimation” and “Residual estimation” are also improved from those obtained with “MBWF”. The improvement with “Bias estimation” is obtained by using an optimally estimated bias component. However, since the bias component is a stationary parameter, the non-stationary characteristics of noise are not considered. On the other hand, the improvement with “Residual estimation” is obtained by considering the non-stationary characteristics of noise. However, since the bias component is not utilized, a serious estimation error may occur in the estimated noise sequence. These facts show that the proposed method can complementarily realize “Bias estimation” and “Residual estimation,” and we can confirm the effectiveness of the noise component decomposition and the estimation technique for each component.

5. CONCLUSIONS

This paper presented a non-stationary noise estimation method based on the biased AR model. The proposed method is designed to achieve the effective suppression of non-stationary noise by decomposing noise into a bias component and a residual component. The evaluation results show that our proposed method improves the accuracy of ASR in non-stationary noise environments compared with bias or residual estimation alone. In the future, we will investigate a non-stationary AR process for the MEKF.

6. REFERENCES