A study of mutual front-end processing method based on statistical model for noise robust speech recognition

Masakiyo Fujimoto, Kentaro Ishizuka, and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Souraku-gun, Kyoto, 619-0237, Japan
{masakiyo, ishizuka, nak}@cslab.kecl.ntt.co.jp

Abstract

This paper addresses robust front-end processing for automatic speech recognition (ASR) in noise. Accurate recognition of corrupted speech requires noise robust front-end processing, e.g., voice activity detection (VAD) and noise suppression (NS). Typically, VAD and NS are combined as one-way processing, and are developed independently. However, VAD and NS should not be assumed to be independent techniques, because sharing each other’s information is important for the improvement of front-end processing. Thus, we investigate the mutual front-end processing by integrating VAD and NS, which can beneficially share each others’ information. In an evaluation of a concatenated speech corpus, CENSREC-1-C database, the proposed method improves the performance of both VAD and ASR compared with the conventional method.

Index Terms: voice activity detection, noise suppression, mutual front-end processing, speech recognition

1. Introduction

Speech signals observed in the real world are usually corrupted by ambient noise. To recognize corrupted speech accurately, it is necessary to employ robust methods, e.g., voice activity detection (VAD) and noise suppression (NS).

VAD, which automatically detects a target speech period from an observed signal, is one of the most important techniques for speech processing. We have proposed a noise robust VAD technique [1], which integrates a switching Kalman filter (SKF)-based robust speech / non-speech discriminator [2] and a robust feature parameter, the periodic to aperiodic component ratio of speech (PAR) [3].

Our proposed VAD works sufficiently well in various noise environments and improves the speech recognition accuracy by reducing the redundant non-speech period. However, this improvement is achieved by reducing the number of error words inserted in non-speech periods. For essential improvement, it is necessary to increase the number of correct words in speech periods. The NS plays an important role in this respect, because it reduces the mismatch between a corrupted speech signal and acoustic models.

VAD is important in terms of improving NS performance, i.e., in the accurate design of a NS filter. On the other hand, the signal to noise ratio (SNR) improvement resulting from NS leads to an improvement in VAD accuracy. These facts indicate that VAD and NS should not be considered separate techniques, but an integrated front-end processing scheme for ASR. Thus, we should investigate front-end processing with integrated VAD and NS, which can effectively share their input-output information as shown in Fig. 1–(b), instead of the conventional one-way combination shown in Fig. 1–(a).

To deal with the problem described above, we propose mutual front-end processing for ASR integrated with VAD and NS. Our proposed method adopts a statistical model-based technique for both VAD [1] and NS [4]. Thus, our aim is to be able to handle mutual input-output information for VAD and NS by sharing statistical models for each method. In this paper, we investigate the mutual front-end processing mentioned above. With the proposed method, VAD supplies NS with statistical models and NS supplies VAD with noise suppressed signals with low distortion.

2. Noise robust VAD

First, we describe our proposed noise robust VAD.

2.1. VAD based on switching Kalman filter

2.1.1. Speech / non-speech state transition model

The proposed method discriminates between speech and non-speech periods based on the likelihood ratio test (LRT) with a statistical model.

As shown by the clean speech model in Fig. 2, the proposed method trains Gaussian mixture models (GMMs) of clean speech and silence in advance. We assume that noise has non-stationary characteristics, thus, the noise sequence is modeled by using a sequential state transition model as shown by the noise model in Fig. 2. With this method, we assume that the noise model is not known in advance. Thus, we estimate the noise model sequentially by using a switching Kalman filter. Finally, by composing these models, we can construct the speech / non-speech transition model as shown by the noisy speech model in Fig. 2.

2.1.2. Formulation of likelihood ratio calculation

This section describes a speech / non-speech discrimination method based on the state transition model shown in Fig. 2.

In the method, \( O_t \), \( N_t \), and \( q_t \) denote the L-dimensional vector of the log Mel spectra of the observed signal and noise at the t-th frame, and the speech or the non-speech state at the t-th frame. When \( O_{0:t} = \{O_0, \ldots, O_t\} \) and \( N_{0:t} = \{N_0, \ldots, N_t\} \) are given, \( q_t \) is decided with respect to the con-
On the other hand, the observation model is processed by the following non-linear equation [2],
\[
    O_{t,l} = S_{t,l} + \log (1 + \exp (N_{t,l} - S_{t,l})) = f (S_{t,l}, N_{t,l})
\]
(7)
where \(O_{t,l}\) and \(S_{t,l}\) denote \(l\)-th element of \(O_t\) and log-Mel spectra of silence or clean speech.
In Eq. (7), the parameter \(S_{t,l}\) is usually unknown. Thus, the parameters of silence or clean speech GMMs are substituted for the parameter \(S_{t,l}\) as follows:
\[
    O_{t,l} = f (\mu_{S_{j,k,l}}, N_{t,l}) + V_{t,j,k,l}
\]
(8)

\[
    V_{t,j,k,l} \sim \mathcal{N} (0, \sigma_{S_{j,k,l}}^2)
\]
(9)
where \(\mu_{S_{j,k,l}}\) and \(\sigma_{S_{j,k,l}}^2\) denote the mean and variance of silence (\(j = 0\)) and speech (\(j = 1\)) GMMs, respectively. \(V_{t,j,k,l}\) denotes an error signal between \(S_{t,l}\) and \(\mu_{S_{j,k,l}}\).

Since a GMM consists of \(K\) Gaussian distributions, \(K\) types of observation processes are derived from Eq. (8). Using these observation processes, the (non-linear) Kalman filter is multiplied into \(K\) types and we can obtain \(K\) types of estimation results for each GMM. In addition, Eq. (8) switches the characteristic of the state-space representation according to the type of GMM (silence or clean speech), thus, the Kalman filter given by the state-space model of Eqs. (5) and (8) has the characteristic of a switching Kalman filter. The estimation formulas for each Kalman filter are described in [2].

After the noise updating, the mean and variance of the observation are given by the following equations.
\[
    \mu_{O_{t,j,k,l}} = f (\mu_{S_{j,k,l}}, N_{t,l})
\]
(10)
\[
    \sigma_{O_{t,j,k,l}}^2 = \frac{\partial \mu_{O_{t,j,k,l}}}{\partial \mu_{S_{j,k,l}}} \sigma_{S_{j,k,l}}^2 + \sigma_{S_{j,k,l}}^2
\]
(11)
where \(\mu_{S_{j,k,l}}\) and \(\sigma_{S_{j,k,l}}^2\) denote the mean and variance, respectively which are estimated by using the parameters of the \(k\)-th parameter included in GMM \(j\). \(\mu_{O_{t,j,k,l}}\) and \(\sigma_{O_{t,j,k,l}}^2\) denote the composed mean and variance of the observation in the \(l\)-th frame. The output probability of each state \(b_{j,l} (O_t)\) is given by a GMM that consists of \(\mu_{S_{j,k,l}}\) and \(\sigma_{S_{j,k,l}}^2\). Here, we use the mixture weight of a clean speech GMM and a silence GMM in place of those of the composed observation model.

### 2.2 Periodic to aperiodic component ratio

We provide a short explanation of the PAR calculation (see [3] for details). The dominant harmonic component in an observed signal is referred to as the periodic component, which is not always the target signal, and the other sound components are referred to as aperiodic components and include both ambient noise and the aperiodic components of the target speech. Although the estimated power of the periodic component is affected by changes in the aperiodic components, this effect can be mitigated in the PAR. Thus, the PAR is expected to be insensitive to dynamic changes in noise power and to be an effective feature for VAD. The decomposition of the periodic component and the aperiodic components is detailed in [3].

If the decomposition can ideally estimate the powers of periodic components, we can detect speech signals based solely on these estimates. However, the decomposition cannot completely avoid power estimation errors. By taking the estimation errors into account, our proposed VAD method statistically detects the existence of speech signals based on the likelihood derived from the error distributions estimated for the periodic
and aperiodic components. We assume that the error distribution follows a Gaussian distribution. Under this assumption, the output probabilities of PAR $b_{j,PAR} (\rho_i)$ are given by following equations [3].

$$b_{j=0,PAR} (\rho_i) = C_0 \exp \left( -\frac{1}{2} \frac{\rho_{p_{ij}}}{\rho_{p_{ij}}} \right)^2$$ (12)

$$b_{j=1,PAR} (\rho_i) = C_1 \exp \left( -\frac{1}{2} \frac{\rho_{n_{ij}}}{\rho_{n_{ij}}} \right)^2$$ (13)

where $\rho_{p_{ij}}$, $\rho_{n_{ij}}$, and $\rho_{n_{i}}$ denote the powers of observed signal, periodic components, and aperiodic components, respectively. $C_0$ and $C_1$ denote the constant values of each Gaussian distribution, respectively.

### 2.3. Integration of VAD methods

This section describes the combination of SKF and PAR mentioned in sections 2.1 and 2.2. The combination is carried out by employing joint likelihood $b_j (O_t, \rho_i)$ given by likelihoods SKF and PAR as follows:

$$b_j (O_t, \rho_i) = b_{j,NI} (O_t) \cdot b_{j,PAR} (\rho_i)$$ (14)

The forward probability given by Eq. (3) is calculated by using joint likelihood $b_j (O_t, \rho_i)$ instead of $b_{j,NI} (O_t)$.

### 3. Statistical model-based NS

This section describes a statistical model-based NS technique. This method is based on the Wiener filter and optimally estimates the Mel-scaled Wiener filter gain by using parameters of statistical models [5]. Usually, the Wiener filter gain $G_{\text{Ni}}$ is given by the following equation:

$$G_{\text{Ni}} = \exp (S_{\text{Ni}}) / \exp (O_{\text{Ni}})$$ (15)

In Eq. (15), $O_{\text{Ni}}$ is a given parameter, but $S_{\text{Ni}}$ is unknown parameter. Thus, $\mu_{S_{\text{Ni}},k,l}$ is substituted for $S_{\text{Ni}}, \mu_{O_{\text{Ni}},k,l}$ is also substituted for $O_{\text{Ni}}$. In addition, the statistical model is modeled by GMM with $K$-Gaussian distributions. Thus, we can design the plural filter gain $G_{\text{Ni},k,l}$ as follows:

$$G_{\text{Ni},k,l} = \exp (\mu_{S_{\text{Ni}},k,l}) / \exp (\mu_{O_{\text{Ni}},k,l})$$ (16)

All filter gains $G_{\text{Ni},k,l}$ are unified by weighted averaging with a posterior $p (k|O_{\text{Ni}})$ and the forward probability $\alpha_{\text{Ni}}$ derived by Eq. (3). Thus, Eq. (17) estimates the optimum filter gain $G_{\text{Ni}}$ by using the fluctuation of speech signal and the differences between the acoustic characteristics of silent periods and speech periods. This also means the utilization of information provided by VAD and NS [4].

$$G_{\text{Ni}} = \frac{1}{y=0} \sum_{y=0}^{K} \alpha_{\text{Ni},l} \cdot \sum_{k=1}^{K} p (k|O_{\text{Ni},l}) G_{\text{Ni},k,l}$$ (17)

The estimated filter gain is transformed into the impulse response with a Mel-warped inverse discrete cosine transform and the noise is suppressed by convoluting the impulse response to the observed signal.

### 4. Feedback of noise suppressed signal

The NS described in section 3 improves the SNR of the observed signal. Thus, VAD performance will be improved by feeding back the noise suppressed signal. However, a noise suppressed signal is usually distorted by several factors, e.g., noise estimation error, and the correlation between speech and noise.

When the distorted signal is used for VAD likelihood calculation, the likelihood will be degraded by the mismatch between the distorted feature parameters and the statistical models. This degradation is conspicuous in a low SNR environment. To solve the problem, we studied NS in several stages. This method is able to reduce speech distortion compared with conventional NS.

First, the Wiener filter given by Eq. (16) is redefined as

$$\hat{G}_{\text{Ni},k,l} = \frac{\exp (\mu_{S_{\text{Ni},k,l}}) + \epsilon \cdot \exp (N_{\text{Ni},l})}{\exp (\mu_{O_{\text{Ni},k,l}})}$$ (18)

where $\epsilon (0.0 \leq \epsilon \leq 1.0)$ denotes a control coefficient of NS gain. $\epsilon = 0.0$ and $\epsilon = 1.0$ indicate ordinal NS and no processing, respectively.

Under low SNR conditions, i.e., when expecting serious speech distortion after NS, $\epsilon = 0.5 \sim 0.9$ is used for NS. NS with this value is able to cope with both partial SNR improvement and speech distortion reduction. On the other hand, when the SNR is sufficiently high and speech distortion may be insignificant, $\epsilon = 0.0 \sim 0.5$ is used. This method aims to improve the performance of VAD and NS with recursive processing as shown in Fig. 3.

As a first investigation of this recursion, we evaluated the proposed method with one iteration. Thus, VAD and NS work twice in each frame. To avoid the residual noise, $\epsilon = 0.0$ is used for NS in second iteration. Above we discussed the effectiveness of the proposed method. Hereafter, we discuss the number of iterations and the optimization of $\epsilon$.

The ETSI Advanced Front-End (AFE) [6] also has iterative NS scheme which called two-stage Wiener filter approach. However, results of each Wiener filter are not reflected to VAD. Thus, the AFE has no mutual improvement scheme by VAD and NS. On the other hand, the proposed method has mutual improvement scheme by VAD and NS. Our aim of this proposal is integration of VAD and NS by handling mutual input-output information.

### 5. Experiments

#### 5.1. Experimental setup

The proposed method was evaluated by using the real recorded data of CENSREC-1-C [8]. The data were recorded in two real environments (a restaurant (Rest.) and a street (St.)) with two different sound pressure levels (avg. 60 dBA: high SNR (Hi.) and avg. 70 dBA: low SNR (Lo.)). The sampling rate was 8 kHz. There were ten speakers (five males and five females). The recorded speech consisted of four files per subject. A single file included 9–10 utterances of continuous 1–7 digit numbers with two-second intervals. The correct segment labels were tagged manually.

The feature parameters for the PAR-based VAD and SKF-based VAD were 1st order PAR and 12th order log-Mel spectra, respectively, which were extracted by using a Hamming window with a 20 msec frame length and a 10 msec frame shift length. We trained the silence and clean speech GMMs for SKF-based VAD by using the clean training data of CENSREC-1 [7]. Each GMM had 32 Gaussian distributions.
5.2. Experimental results of VAD
The evaluation criteria for VAD are the utterance correct rate and utterance accuracy rate as shown by Eqs. (19) and (20).

\[
Corr \equiv \frac{N_c}{N} \times 100 \% \tag{19}
\]

\[
Acc = \frac{(N_c - N_f)}{N} \times 100 \% \tag{20}
\]

where \( N \), \( N_c \), and \( N_f \) denote the total number of speech utterances, the number of correctly detected utterances, and the number of incorrectly detected utterances, respectively.

Table 1 shows the VAD results. In the table, “Baseline,” “Sohn,” “AFE,” “One-way,” “FB (\( \epsilon = 0.0 \)),” and “FB (\( \epsilon = 0.7 \))” represent results obtained by CENSREC-1-C baseline (energy-based VAD [8]), the statistical model-based VAD method proposed by Sohn [9], ETSI Advanced Front-End [6], conventional one-way processing of VAD and NS (our proposed VAD [1] alone), and the proposed integration method with \( \epsilon = 0.0 \) and \( \epsilon = 0.7 \), respectively. The value of \( \epsilon \) was decided experimentally.

In the table, since enhanced speech is seriously distorted, the results of “FB (\( \epsilon = 0.0 \))” are poorer than those of “One-way.” On the other hand, since the speech distortion is reduced, “FB (\( \epsilon = 0.7 \))” improves the VAD results more than “One-way” and “FB (\( \epsilon = 0.0 \)).” From these results, we confirmed that feeding back enhanced speech with a suitable \( \epsilon \) value is effective in improving VAD performance.

5.3. Experimental results of ASR
We used the HTK [10] for ASR and HMM training. The whole word HMMs (16 states, 20 Gaussian distributions per state) were trained by using the clean training data of CENSREC-1. The feature parameters used in this evaluation consisted of 39 MFCCs with 12 MFCCs, log-energy, and their first and second order derivatives.

In the table, “w/o FB” and “Ideal” represent results obtained by the integration of VAD and NS without feeding back enhanced speech [4] (only statistical model sharing, i.e., NS by \( \epsilon = 0.0 \) and no feedback) and NS with hand-labeled VAD, respectively. The table shows that the proposed method “FB (\( \epsilon = 0.7 \))” improves ASR accuracy more than “One-way,” “w/o FB,” and “FB (\( \epsilon = 0.0 \)).” These ASR improvements are obtained by the mutual improvement of VAD and NS. Thus, we can confirm the effectiveness of the proposed integration method with feeding back enhanced speech. On the other hand, the results of “AFE” are much worse than those of proposed method. This is caused by VAD error.

The difference between the proposed method and “Ideal” is VAD accuracy. This suggests that the improvement in VAD accuracy leads to an improvement in ASR with NS, and proves the suitability of feeding back enhanced speech for VAD with the proposed method. Thus, to improve the VAD performance, it is necessary to investigate the optimization of the \( \epsilon \) value by introducing a likelihood function.

Table 1: VAD results (%)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>74.20</td>
<td>56.52</td>
<td>39.42</td>
<td>41.45</td>
<td>52.90</td>
<td>21.45</td>
<td>-43.48</td>
<td>-15.65</td>
<td>-33.91</td>
<td>-17.90</td>
</tr>
<tr>
<td>Sohn</td>
<td>72.75</td>
<td>57.10</td>
<td>97.39</td>
<td>78.55</td>
<td>76.45</td>
<td>45.51</td>
<td>-6.38</td>
<td>94.49</td>
<td>57.39</td>
<td>47.75</td>
</tr>
<tr>
<td>AFE</td>
<td>44.35</td>
<td>80.58</td>
<td>47.25</td>
<td>72.17</td>
<td>61.09</td>
<td>-82.03</td>
<td>-245.51</td>
<td>-101.16</td>
<td>-168.70</td>
<td>-149.35</td>
</tr>
<tr>
<td>One-way</td>
<td>93.04</td>
<td>70.72</td>
<td>100.00</td>
<td>97.97</td>
<td>90.43</td>
<td>72.75</td>
<td>19.71</td>
<td>99.13</td>
<td>94.78</td>
<td>71.60</td>
</tr>
<tr>
<td>FB (( \epsilon = 0.0 ))</td>
<td>91.88</td>
<td>67.54</td>
<td>100.00</td>
<td>93.91</td>
<td>88.33</td>
<td>70.14</td>
<td>17.97</td>
<td>98.54</td>
<td>85.80</td>
<td>68.11</td>
</tr>
<tr>
<td>FB (( \epsilon = 0.7 ))</td>
<td>92.75</td>
<td>71.03</td>
<td>100.00</td>
<td>99.71</td>
<td>91.09</td>
<td>75.36</td>
<td>24.06</td>
<td>99.32</td>
<td>97.97</td>
<td>74.20</td>
</tr>
</tbody>
</table>

Table 2: Speech recognition results by word accuracy (%)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o VAD</td>
<td>45.17</td>
<td>1.28</td>
<td>34.43</td>
<td>25.23</td>
<td>26.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>44.16</td>
<td>18.12</td>
<td>29.96</td>
<td>21.62</td>
<td>28.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sohn</td>
<td>37.45</td>
<td>-3.81</td>
<td>33.41</td>
<td>29.58</td>
<td>24.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFE</td>
<td>42.71</td>
<td>-37.98</td>
<td>22.31</td>
<td>27.14</td>
<td>13.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-way</td>
<td>70.95</td>
<td>21.38</td>
<td>74.53</td>
<td>47.63</td>
<td>53.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o FB</td>
<td>71.31</td>
<td>21.68</td>
<td>76.96</td>
<td>56.85</td>
<td>56.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB (( \epsilon = 0.0 ))</td>
<td>70.12</td>
<td>17.07</td>
<td>74.86</td>
<td>53.46</td>
<td>54.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB (( \epsilon = 0.7 ))</td>
<td>73.86</td>
<td>28.87</td>
<td>83.97</td>
<td>62.39</td>
<td>62.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ideal</td>
<td>88.89</td>
<td>49.09</td>
<td>86.25</td>
<td>60.66</td>
<td>71.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusion
This paper presented a mutual front-end processing method for ASR in noise. The proposed method integrates VAD and NS, and utilizes information provided by each method. In the evaluation, the proposed method improved ASR accuracy compared with the conventional one-way combination. In the future, we are planning to investigate the design of likelihood function for optimization of coefficient \( \epsilon \).

7. Acknowledgements
The study was conducted using CENSREC-1 database and CENSREC-1-C database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

8. References