Frame-wise model re-estimation method based on Gaussian pruning with weight normalization for noise robust voice activity detection

Masakiyo Fujimoto*, Shinji Watanabe, Tomohiro Nakatani

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

Received 10 August 2010; received in revised form 24 August 2011; accepted 25 August 2011
Available online 16 September 2011

Abstract

This paper proposes a robust voice activity detection (VAD) method that operates in the presence of noise. For noise robust VAD, we have already proposed statistical models and a switching Kalman filter (SKF)-based technique. In this paper, we focus on a model re-estimation method using Gaussian pruning with weight normalization. The statistical model for SKF-based VAD is constructed using Gaussian mixture models (GMMs), and consists of pre-trained silence and clean speech GMMs and a sequentially estimated noise GMM. However, the composed model is not optimal in that it does not fully reflect the characteristics of the observed signal. Thus, to ensure the optimality of the composed model, we investigate a method for its re-estimation that reflects the characteristics of the observed signal sequence. Since our VAD method works through the use of frame-wise sequential processing, processing with the smallest latency is very important. In this case, there are insufficient re-training data for a re-estimation of all the Gaussian parameters. To solve this problem, we propose a model re-estimation method that involves the extraction of reliable characteristics using Gaussian pruning with weight normalization. Namely, the proposed method re-estimates the model by pruning non-dominant Gaussian distributions that express the local characteristics of each frame and by normalizing the Gaussian weights of the remaining distributions. In an experiment using a speech corpus for VAD evaluation, CENSREC-1-C, the proposed method significantly improved the VAD performance with compared that of the original SKF-based VAD. This result confirmed that the proposed Gaussian pruning contributes to an improvement in VAD accuracy.

© 2011 Elsevier B.V. All rights reserved.

Keywords: Voice activity detection; Switching Kalman filter; Gaussian pruning; Posterior probability; Gaussian weight normalization

1. Introduction

Voice activity detection (VAD) that automatically detects a period containing a target speech signal from a continuously observed signal is playing a crucial role in the development of speech processing technology. VAD can be employed in various speech-oriented technology fields, e.g., speech enhancement, speech coding for cellular or IP phones, and the front-end processing of automatic speech recognition. Thus, a lot of research on VAD has been proposed and its performance discussed.

VAD usually consists of two parts: a part for extracting a distinctive feature parameter and a part for detecting the absence of speech or speech activity. The feature extraction part extracts acoustic features that distinguish between speech absence and speech activity. Traditionally, the zero-crossing rate and the energy difference between nonspeech and speech (Rabiner and Sambur, 1975; ETSI EN 301 708 v.7.1.1, 1999) have been widely used as distinctive feature parameters. However, these traditional parameters are not robust in the presence of interference noises, and so various noise robust features and their combinations have been proposed as follows:

* Corresponding author.
E-mail address: fujimoto.masakiyo@lab.ntt.co.jp (M. Fujimoto).
• All-band spectra, sub-band spectra of output signal of Wiener filter, and variance of spectra (ETSI ES 202 050 v.1.1.4, 2006).
• Higher order statistics (Nemer et al., 2001; Li et al., 2005; Cournapeau and Kawahara, 2007).
• Long-term spectral divergence (Ramírez et al., 2004).
• Speech periodicity (Kristjansson et al., 2005) and speech periodic component to aperiodic component ratio (Ishizuka et al., 2010).
• Volatility (covariance variance) obtained from generalized autoregressive conditional heteroskedasticity (GARCH) filtering (Tahmasbi and Razaei, 2007; Kato et al., 2006).

These parameters can improve VAD accuracy, however, employing robust feature parameters alone is insufficient for noise robust VAD. In particular, in a low signal to noise ratio (SNR) environment, the discriminative characteristics of the feature parameter inevitably degrade due to the strong noise energy, even if noise robust feature parameters are used. This problem suggests the importance of a decision mechanism for noise robust VAD. If a robust decision mechanism is introduced into VAD, the VAD accuracy will improve, even if the discriminative characteristics of the feature parameter become unreliable owing to crucial interference noises.

Some statistical model-based VAD techniques have been proposed as one of the most robust decision mechanisms (Sohn et al., 1999; Ramírez et al., 2007; Fujimoto and Ishizuka, 2007; Weiss and Kristjansson, 2005; Ramirez et al., 2005; Chang et al., 2006; Fujimoto et al., 2008). Most of these methods define a non-speech/speech state transition model, and calculate the speech state to non-speech state likelihood ratio. Whether speech is absent or active is indicated by a likelihood ratio test (LRT) with a threshold.

The method proposed by Sohn et al. (1999) is a fundamental technique for statistical model-based VAD. However, assumptions regarding stationary noise environments and a priori knowledge are indispensable to Sohn’s method, and its applicable noise environments are restricted to specific ones. In most cases, a noise observed in the real world has non-stationary characteristics and is unknown in advance. Thus, robustness in the presence of non-stationary noise without a priori knowledge of the noise is the most important factor for robust and useful VAD in the real world. For VAD based on such assumptions, we have already proposed a technique that employs a switching Kalman filter (SKF), which is robust in non-stationary noise environments (Fujimoto and Ishizuka, 2007).

SKF-based VAD consists of the following three steps:

1. **Noise estimation step**: the parameters of a noise Gaussian mixture model (GMM) are estimated based on SKF by using the mean and the variance parameters of silence and clean speech GMMs, which are estimated in advance by using clean speech corpora.
2. **Composition step**: two internal states of non-speech (silence + noise) and speech (clean speech + noise) in the noisy speech (observed signal) model are constructed by composing silence and clean speech GMMs with the estimated noise GMM as shown in Fig. 1.
3. **Discrimination step**: speech absence or activity in the observed signal is discriminated by using the likelihood ratio of the observed signal between non-speech and speech GMMs.

In this framework, the characteristics of the observed signal are not directly reflected in the composition step, unlike the noise estimation. Namely, the composition step lacks optimality in terms of the consideration of the observed signal. Thus, one of the most important factors as regards our proposed approach is finding a way to ensure the optimality of the noisy speech model by incorporating the characteristics of the observed signal. A standard way of achieving this is to employ a Gaussian parameter re-estimation method, e.g., a maximum likelihood (ML) estimation or a maximum a posterior (MAP) estimation, in the noisy speech model by using the observed signal at the composition step. However, since VADs strongly require the smallest possible latency in many applications, a re-estimation method with one frame sample is desired in the composition step. Here, in this paper, the frame length and the frame shift length are set at 20 and 10 ms, respectively. Thus, the frame interval (algorithm latency) is 10 ms. In this situation, it is almost impossible to re-estimate all of the Gaussian parameters by ML or MAP estimation due to the well-known over-fitting problem, and frame-by-frame processing-based VAD using statistical approaches suffers inherently from this problem.

To satisfy the requirement and to avoid this over-fitting problem, we proposed another model re-estimation method from a different standpoint. The method proposed here is a Gaussian pruning method that selects dominant Gaussian distributions of non-speech and speech GMMs depending on the observed signal of the current frame. Thus, this method extracts beneficial Gaussian distributions that optimally represent the observed signal in the current frame. Fig. 2 shows the processing flow of our VAD method, and the proposed Gaussian pruning method is employed after the composition estimation step.

![Fig. 1. Non-speech/speech state transition model with noise dynamics.](image-url)

The symbols $H_0$, $H_1$, and $N_t$ denote a silence state, a clean speech state, and a noise state sequence, respectively.

---

230 M. Fujimoto et al. / Speech Communication 54 (2012) 229–244
The composed non-speech and speech GMMs include several Gaussian distributions, and each Gaussian distribution represents various characteristics of a noisy speech signal. Since the observed signal may have various characteristics, various Gaussian distributions that can satisfactorily represent the characteristics of the observed signal are usually required to achieve an exact likelihood calculation. However, this demand is restricted for a likelihood calculation involving the entire frame signal of an observed signal. In sequential processing, only partial aspects of whole speech characteristics are assumed to appear in each frame. Therefore, contributions made by a small number of Gaussian distributions would be more dominant and superior to the others in expressing the local characteristics of the observed signal in each frame. On the other hand, non-dominant Gaussian distributions may be both unimportant and harmful due to the serious gap between the observed signal and the distributions.

Under these assumptions, we investigate a way of improving the optimality of the noisy speech model by using Gaussian pruning-based model re-estimation. Then, we also apply Gaussian weight normalization to the remaining dominant Gaussian distributions in each frame. This normalization method can enhance the likelihood given by each remaining Gaussian distribution, and improves the effectiveness of the proposed Gaussian pruning.

Usually, the characteristics of the ambient noise and speech do not suddenly change in neighboring frames if these sound sources are in the stable state. However, if the sound sources are in a transient state, i.e., speech onset, speech offset, or the appearance of burst noise, the acoustical characteristics of an observed signal will vary greatly in a short time range. In addition, this sudden change is very difficult to predict. Thus, continual frame-wise optimization is a reasonable way to cope with unpredictable sudden changes in the sound sources included in the observed signal.

We employ the posterior probability as the Gaussian pruning criterion. Namely, the Gaussian distributions with high posterior probabilities are extracted as dominant Gaussian distributions for likelihood calculation. We investigate and evaluate three techniques as the posterior probability-based pruning method, i.e., $N$-best selection, threshold processing for each posterior probability, and threshold processing for posterior probability accumulation. Furthermore, we investigate a method for determining the pruning parameters of each Gaussian pruning technique, e.g., the sequential determination of the pruning threshold or the number of selected distributions, and prove that our pruning parameter determination method satisfies the aim of the Gaussian pruning in terms of sequential model re-estimation.

The Gaussian pruning technique is usually used for reducing computational complexity without degrading speech recognition accuracy (Ogawa and Takahashi, 2008; Shinoda and Iso, 2002; Fischer and Rob, 1999). Most Gaussian pruning techniques reduce the number of Gaussian distributions by merging similar distributions using certain clustering methods. These methods involve the prior processing of speech recognition, and the characteristics of the observed signal are not usually reflected in the pruning result. On the other hand, our proposed method involves the posterior processing of VAD, and improves the optimality of the noisy speech model by using the characteristics of the observed signal. The proposed method does not aim to reduce computational complexity, which makes it different from conventional pruning techniques.

The proposed method was evaluated with the CENSREC-I-C database (corpora and environments for noisy speech recognition-1 concatenated) (Kitaoka et al., 2007), which is Japanese noisy speech data for VAD evaluation. The evaluation results revealed that the proposed Gaussian pruning method significantly improves VAD accuracy compared with the conventional methods. In particular, we confirmed that Gaussian pruning contributes greatly to the improvement of VAD accuracy.

In this paper, Section 2 reviews statistical model-based VAD and SKF-based VAD, Section 3 describes Gaussian pruning in detail, Section 4 describes a frame-wise method for determining the pruning parameter, Section 5 reports evaluation results obtained with the CENSREC-I-C framework, and Section 6 summarizes the paper and describes future directions.

2. Reviews of statistical model-based VAD and SKF-based VAD

This section reviews statistical model-based VAD (Sohn et al., 1999) and our previous work, SKF-based VAD (Fujimoto and Ishizuka, 2007).
2.1. Fundamentals of statistical model-based VAD

Statistical model-based VAD discriminates between speech absence and speech activity by using an LRT with a statistical model. In the LRT, a typical statistical model has a state transition process with non-speech (speech absence) and speech (speech activity) states as shown in Fig. 3.

When the state transition model is given, the easiest way to perform the LRT is derived from the following equation

\[ q_t = \begin{cases} H_0 & \text{if } R_t < \text{Threshold} \\ H_1 & \text{otherwise} \end{cases} \quad (1) \]

\[ R_t = \frac{b_1(O_t)}{b_0(O_t)} \quad (2) \]

where \( q_t, b_j(O_t), \) and \( O_t \) denote the estimated state at the \( t \)th frame, the likelihood at the state \( j (j = 0: \text{non-speech state, } j = 1: \text{speech state}) \), and the observed signal at the \( t \)th frame, respectively.

Eq. (2) is a simple and fundamental LRT technique. However, it is not a robust technique, because the state transition process of non-speech and speech states is not considered in Eq. (2). As a robust LRT technique with a state transition process, Sohn et al. proposed the HMM-based hang-over scheme (Sohn et al., 1999). In the HMM-based hang-over scheme, when the sequence of observed signals \( O_{t:j} = \{ O_t, \ldots, O_j \} \) is given, the state \( q_t \) is decided with respect to the conditional probability \( p(q_t|O_{0:j}) \).

The conditional probability \( p(q_t|O_{0:j}) \) has the following relationship with the joint probability \( p(O_{0:j}, q_t) \)

\[ p(q_t|O_{0:j}) = \frac{p(O_{0:j}, q_t)}{p(O_{0:j})} \propto p(O_{0:j}, q_t) \quad (3) \]

Therefore, we focus on the joint probability \( p(O_{0:j}, q_t) \) instead of the conditional probability \( p(q_t|O_{0:j}) \). The joint probability \( p(O_{0:j}, q_t) \) can be represented by the recursive formula of Eq. (4) by assuming a first order Markov chain

\[ p(O_{0:j}, q_t) = \sum_{q_{t-1}} p(q_t|q_{t-1}) p(O_{t:j-1}, q_{t-1}) \quad (4) \]

The right hand side of Eq. (4) is equivalent to the forward probability \( a_{j:t} \). Thus, Eq. (4) can be efficiently computed by the following recursive formula

\[ p(O_{0:j}, q_t = H_j) = a_{j:t} = \sum_{t=0}^{\infty} a_{i:j} \cdot b_j(O_i) \cdot q_{i-1} \quad (5) \]

where \( a_{i:j} = p(q_i = H_j|q_{i-1} = H_j) \) (the state transition probability) and \( b_j(O_i) = p(O_i|q_i = H_j) \) (the likelihood given by state \( H_j \) at the \( t \)th frame). When \( t = 0 \), the beginning frame is assumed to be a non-speech frame. Thus, the initial values \( a_{0,0} = 1 \) and \( a_{1,0} = 0 \) are given.

Finally, the likelihood ratio, which is given as

\[ R_t = \frac{a_{1:t}}{a_{0:t}} \quad (6) \]

is used for the LRT of Eq. (1) instead of Eq. (2).

2.2. Review of SKF-based VAD

This section briefly reviews each step of SKF-based VAD. The processing of each step is detailed in Appendix A.

2.2.1. State transition model with noise transition process

The LRT proposed by Sohn et al. employs a single Gaussian distribution whose feature parameters are complex spectra for the statistical model, and the likelihood ratio is calculated by using \textit{a priori} and \textit{a posteriori} SNRs (Sohn et al., 1999). However, this method assumes that the noise environment is known, and the state transition process of the noise is not considered. Thus, this method is not robust for unknown and non-stationary noise environments.

To cope with unknown non-stationary noise environments, we must construct environmentally matched model sets where the noise has a state transition process and without \textit{a priori} knowledge of the noise. To deal with this problem, we first define the non-speech and speech states as follows:

\[ q_t = H_0 : \text{Non-speech state} = \text{Silence + Noise} \]

\[ q_t = H_1 : \text{Speech state} = \text{Speech + Noise} \]

With this definition, we can construct a statistical model that distinguishes between the state transition processes of speech and noise. The state transition process of speech, i.e., speech absence or speech activity, is given as the “clean speech model” in Fig. 1. This model is the same as the ergodic model used in Sohn’s method as shown in Fig. 3. However, the model used in the proposed method has clean speech and silence states, and each state is given by a GMM.

Next, we assume that noise has non-stationary characteristics, thus, the noise sequence is modeled by using the sequential state transition model shown as the “noise model” in Fig. 1. Finally, by composing clean speech and noise models, we can construct the non-speech/speech state transition model with noise dynamics shown as the “noisy speech model” in Fig. 1. Namely, this model has state transition processes for both speech and noise. Speech has a discrete state transition process and noise has a sequential state transition process.

With this approach, the silence and clean speech GMMs can be modeled in advance by using a clean speech corpus. On the other hand, the noise statistics are unknown. Thus, we estimate the noise statistics sequentially by using a Kalman filter.
2.2.2. Noise estimation step

The mean and variance parameters of a noise GMM are sequentially estimated and updated at each frame with the SKF in the log Mel-spectral domain. In this step, parts of the parameter sets of the SKF are obtained by the pre-trained silence and clean speech GMMs. Furthermore, the observed signal \( O_t \) is an L-dimensional vector of the log of the output energy of the Mel-filter bank (LMFB) at the current th frame, utilized in the SKF, thus the SKF can reflect the characteristics of the observed signal in the estimation results. The noise estimation step is detailed in Appendix A.1.

2.2.3. Composition step

In the LMFB domain, the composition of the noisy speech model is derived by the following equations:

\[
P_{O_t,j,k} = \mu_{S,j,k} + \log(1 + \exp(\mu_{N,t,j,k} - \mu_{S,j,k}))
\]

\[
\Sigma_{O_t,j,k} = F_{t,j,k} \Sigma_{N,t,j,k} F_{t,j,k}^T + \Sigma_{S,j,k}
\]

\[
F_{t,j,k} = \text{diag}\left\{ \frac{\partial \Sigma_{O_t,j,k}}{\partial \mu_{N,t,j,k}} \right\}
\]

where \( \mu_{S,j,k}, \Sigma_{S,j,k}, \mu_{N,t,j,k}, \Sigma_{N,t,j,k}, \mu_{O_t,j,k}, \) and \( \Sigma_{O_t,j,k} \) denote the mean vectors and the diagonal variance matrices of the \( \text{th} \) Gaussian distribution in a silence (\( j = 0 \)) or clean speech (\( j = 1 \)) GMM, a noise GMM, and a noisy speech GMM, respectively. The operations \( \log(\cdot) \) and \( \exp(\cdot) \) are independently applied to each vector element, and the vector \( \mathbf{1} = [1, \ldots, 1]^T \).

This step merely composes the silence and the clean speech GMMs with the estimated noise GMM, thus the characteristics of the observed signal are not fully reflected in the composed noisy speech model.

2.2.4. Discrimination step

When the noisy speech model is given, speech absence or activity is discriminated by using the LRT with the HMM-based hang-over scheme described in Section 2.1.

In the LRT, the likelihood of state \( H_j \) at the th frame is the following probability density function (PDF) of GMM

\[
b_j(O_t) = \sum_{k=1}^{K} w_{j,k} \cdot N(O_t; \mu_{O_t,j,k}, \Sigma_{O_t,j,k})
\]

where \( w_{j,k} \) and \( N(\cdot) \) denote the Gaussian weight of a silence or clean speech GMM and the PDF of a Gaussian distribution, respectively.

3. Model re-estimation based on Gaussian pruning with weight normalization

3.1. Problem of frame-wise model re-estimation

SKF-based VAD discriminates between the speech absence and speech activity of an observed signal by using the likelihood ratio of non-speech and speech GMMs. The non-speech and speech GMMs, namely, noisy speech GMMs, consist of the parameters derived using Eqs. (7) and (8). With this method, the noise parameters are optimally estimated by the SKF, because the estimation scheme of the Kalman filter ensures an optimal estimation result by reflecting the information of the observed signal. However, the parts of the parameter sets of the SKF that are provided by the silence and clean speech GMMs are not optimal for the observed signal, because the Gaussian parameters are trained in advance by using speech corpora, which are usually different from the observed signal. Therefore, the Gaussian parameters derived by using Eqs. (7) and (8) are not fully optimal for the observed signal.

The optimization of noisy speech GMMs is one of the most important factors for all the statistical model-based VAD techniques. Thus, to ensure the optimality of the noisy speech model, the Gaussian parameters are usually re-estimated reflecting the characteristics of the observed signal. Here, VAD techniques require the smallest latency, i.e., real time frame by frame processing without any processing delay. Therefore, the data of some frames succeeding the current frame are unavailable as re-training data, because the use of the succeeding frames directly causes a processing delay. On the other hand, if the observed signal has temporally stable characteristics, we can use the data of preceding frames as the re-training data. However, the acoustical characteristics of an observed signal can vary greatly in the presence of speech onset, speech offset, or the appearance of burst noise, and these variations are very difficult to predict. This sudden change will be detrimental to parameter optimization at the current frame.

Thus, the data of preceding frames may be unsuitable as re-training data. In this case, we should re-estimate all the Gaussian parameters by using only the current frame (one frame sample), however, this is almost impossible to achieve. Thus, we investigate a model re-estimation method based on Gaussian pruning with weight normalization.

3.2. Aim of Gaussian pruning

Each noisy speech GMM includes \( K \) Gaussian distributions. Each Gaussian distribution contained in each GMM represents various noisy speech signal characteristics. Furthermore, the observed signal also has various characteristics. For an exact likelihood calculation, we usually need various Gaussian distributions to cope with the various characteristics included in all (or several) frame sequences of the observed signal. For example, Fig. 4(a) shows the correspondence between the characteristics of several frames and Gaussian distributions. As shown in Fig. 4(a), the GMM of the figure consists of four Gaussian distributions, and we can see that all the distributions are needed to represent the characteristics of several frames. Thus, we must utilize various Gaussian distributions that can satisfactorily represent all the characteristics of the observed signal for an exact likelihood calculation.

However, the above approach can only be applied to a likelihood calculation that involves the whole frame
sequence of the observed signal. If the likelihood calculation is restricted to a local (specific) characteristic of a current frame, Gaussian distributions that are not dominant in expressing the local characteristic may be unimportant. For example, Fig. 4(b) shows the correspondence between the characteristics of the current frame and Gaussian distributions. As shown in Fig. 4(b), the likelihood calculation may be performed satisfactorily by using only two Gaussian distributions, \( k = 2 \) and \( k = 3 \), which are dominant distributions for expressing the local characteristic of the current frame. Other Gaussian distributions, \( k = 1 \) and \( k = 4 \), are not dominant distributions for expressing the local characteristic, and so may be unimportant for the likelihood calculation. Therefore, since only some of the speech characteristics appear in each frame in the sequential processing, the contributions of a small number of Gaussian distributions are more dominant than the others. Under these assumptions, we investigate the exact likelihood calculation by pruning non-dominant Gaussian distributions.

The effect of proposed Gaussian pruning technique in a GMM may be similar effect to that of making an ergodic HMM that has several states with same total number of Gaussian components. This method selects the most dominant state at each frame, namely, it employs a state transition scheme as a substitute for the Gaussian pruning.

In the proposed method, the optimum number of Gaussian components and optimum combination of Gaussian components change at each frame. However, the number of Gaussian components assigned in each state is fixed in the ergodic HMM-based approach. In this case, the ergodic HMM-based approach cannot select the optimum number of Gaussian components. Thus, the utilization of an ergodic HMM is not a reasonable approach to the frame-wise model re-estimation.

3.3. Gaussian pruning techniques

As the criterion for Gaussian pruning, we employ the posterior probability \( P(k|O_{i,j}) \) derived using the following equation

\[
P(k|O_{i,j}) = \frac{w_{j,k} \cdot N(O_{i}; \mu_{O_{i},j,k}, \Sigma_{O_{i},j,k})}{\sum_{k'=1}^{K} w_{j,k'} \cdot N(O_{i}; \mu_{O_{i},j,k'}, \Sigma_{O_{i},j,k'})}
\]

(11)

With this method, the posterior probability is used as a confidence measure for each Gaussian distribution. Namely, the Gaussian distributions with a high posterior probability are given precedence in a likelihood calculation. We investigate three techniques for a posterior probability-based pruning approach, i.e., \( N \)-best selection (Section 3.3.1), threshold processing for each posterior probability (Section 3.3.2), and threshold processing for posterior probability accumulation (Section 3.3.3). Each method is described in detail in the following sections.

3.3.1. Gaussian pruning based on \( N \)-best approach

By sorting \( P(k|O_{i,j}) \) in descending order, we obtain the sorted posterior probability \( P_{\text{sorted}}(k|O_{i,j}) \) and corresponding original Gaussian index \( Idx_{i,j,k} \). Fig. 5 shows an example of the descending order sorting of \( P(k|O_{i,j}) \).

In the results of the descending order sorting, \( P_{\text{sorted}}(k|O_{i,j}) \) with a smaller \( k \) has a higher value. Thus, we can obtain the important Gaussian distributions simply by utilizing those with high \( P_{\text{sorted}}(k|O_{i,j}) \) values. After the descending order sorting, \( N \) Gaussian distributions, i.e., the Gaussian distributions labeled from \( Idx_{i,j,1} \) to \( Idx_{i,j,N} \), are selected from those with the highest posterior probability. Then the set of selected Gaussian distributions \( \hat{\mathcal{M}}_{i,j} \) is defined as \( \{ Idx_{i,j,1}, \ldots, Idx_{i,j,N} \} \subset \hat{\mathcal{M}}_{i,j} \).

With this method, since the remaining distributions depend on \( O_{i} \), different Gaussian distributions remain in each frame. However, the number of selected Gaussian distributions is fixed at \( N \).

![Diagram](https://example.com/diagram.png)

Fig. 5. Example of descending order sorting of \( P(k|O_{i,j}) \).
3.3.2. Gaussian pruning based on posterior probability thresholding

This method selects Gaussian distributions based on posterior probability thresholding as follows:

\[ \text{Idx}_{t,j,k} \Rightarrow \tilde{M}_{t,j} \text{ if } P^s(k|\mathbf{O}_t,j) \geq \beta \]  

(12)

If the sorted posterior probability \( P^s(k|\mathbf{O}_t,j) \) is less than the pruning threshold \( \beta \), \( \text{Idx}_{t,j,k} \) is not added to \( \tilde{M}_{t,j} \). Thus, Eq. (12) directly selects the Gaussian distributions that have a sufficiently large \( P^s(k|\mathbf{O}_t,j) \). With this method, different Gaussian distributions are selected in each frame, and the number of remaining Gaussian distributions varies in each frame.

3.3.3. Gaussian pruning based on posterior probability accumulation

This method is almost same as the N-best approach described in Section 3.3.1. This method also first applies descending order sorting to \( P(k|\mathbf{O}_t,j) \). Then the number of remaining distributions, \( \tilde{M}_j(\mathbf{O}_t) \), is decided as the minimum \( M \) that makes the accumulation of \( P^s(k|\mathbf{O}_t,j) \) greater than the pruning threshold \( Z \) as follows:

\[ \tilde{M}_j(\mathbf{O}_t) = \arg\min_M \left\{ \sum_{k=1}^{M} P^s(k|\mathbf{O}_t,j) \geq Z \right\} \quad (0 < Z \leq 1) \]  

(13)

Then the set of selected Gaussian distributions \( \tilde{M}_{t,j} \) is defined as \( \{ \text{Idx}_{t,j,1}, \ldots, \text{Idx}_{t,j,M(\mathbf{O}_t)} \} \in \tilde{M}_{t,j} \).

Fig. 6 shows an example of an \( \tilde{M}_j(\mathbf{O}_t) \) decision. In the figure, \( Z = 0.7 \) gives \( \tilde{M}_j(\mathbf{O}_t) = 5 \) and \( Z = 0.9 \) gives \( \tilde{M}_j(\mathbf{O}_t) = 9 \). With this method, different Gaussian distributions are selected in each frame, and the number of remaining Gaussian distributions varies in each frame.

3.4. Gaussian weight normalization

After the Gaussian pruning described in Section 3.3, the Gaussian weights of the remaining distributions are normalized, because the sum of the Gaussian weights is equal to 1 in accordance with the following equation:

\[ \hat{w}_{t,j,m} = \frac{W_{t,m}}{\sum_{m' \in \tilde{M}_{t,j}} W_{t,m'}} \]  

(14)

where \( \hat{w}_{t,j,m} \) denotes the normalized Gaussian weight of the remaining distributions, and \( \text{Idx}_{t,j,k} \) is substituted with \( m \).

This normalization is an important factor in Gaussian pruning, because it enhances the likelihood of the remaining distributions. For example, when the Gaussian weights prior to pruning are given as \( w_{j,k} = \{0.25, 0.35, 0.25, 0.15\} \) and the two Gaussian distributions, \( k = 2 \) and \( k = 4 \), remain after the pruning, the Gaussian weights after the pruning are normalized as \( \hat{w}_{j,2} = 0.7 \) and \( \hat{w}_{j,4} = 0.3 \). By using these values, the likelihoods of Gaussian distributions \( k = 2 \) and \( k = 4 \) can be effectively enhanced.

By using \( \hat{w}_{t,j,m} \), the likelihood is given by

\[ \hat{b}_j(\mathbf{O}_t) = \sum_{m \in \tilde{M}_{t,j}} \hat{w}_{t,j,m} \cdot N(\mathbf{O}_t; \mu_{O,t,
 j,m}, \Sigma_{O,t,
 j,m}) \]  

(15)

The likelihood \( \hat{b}_j(\mathbf{O}_t) \) is used for the LRT derived with Eqs. (1), (5) and (6), is fed back for SKF-based noise updating (see Eqs. (A.20) and (A.21) in Appendix A.2.2). Moreover, the posterior probability is also recalculated as

\[ \tilde{P}(k|\mathbf{O}_t,j) = \begin{cases} \frac{\hat{w}_{t,j,m} \cdot N(\mathbf{O}_t; \mu_{O,t,
 j,m}, \Sigma_{O,t,
 j,m})}{\sum_{m' \in \tilde{M}_{t,j}} \hat{w}_{t,j,m'} \cdot N(\mathbf{O}_t; \mu_{O,t,
 j,m'}, \Sigma_{O,t,
 j,m'})} & \text{if } k \in \tilde{M}_{t,j} \\ 0 & \text{otherwise} \end{cases} \]  

(16)

The posterior probability \( \tilde{P}(k|\mathbf{O}_t,j) \) is fed back for SKF-based noise updating (see Eqs. (A.17) and (A.18) in Appendix A.2.2).

4. Frame-wise pruning parameter determination based on maximum likelihood criterion

With the Gaussian pruning methods described in Section 3.3, we have to set appropriate pruning parameters depending on variations in the speakers or in the noise environments. In addition, each pruning parameter is independent of the frame index \( t \) and the model index \( j \). Thus, these methods represent a stationary determination property for pruning parameters \( N \), \( \beta \), and \( Z \), because each pruning parameter is manually adjusted to a constant value. This stationary determination is contrary to the aim of the Gaussian pruning described in Section 3.2 in terms of sequential model re-estimation. To avoid these problems, this section investigates a way to determine the pruning parameter of each technique, i.e., \( N \), \( \beta \), and \( Z \). With this approach, the pruning parameter of each method is determined in each frame and each model based on the ML criterion.

With the \( N \)-best approach described in Section 3.3.1, the number of selected Gaussian distributions \( N(\mathbf{O}_t) \) that depend on the observed signal \( \mathbf{O}_t \) at the current frame and the model index \( j \) is determined as the number of Gaussian distributions that maximize likelihood \( \hat{b}_j(\mathbf{O}_t) \) with Gaussian weight normalization derived with Eq.
(14). \( N_j(O_j) \) is determined using a round-robin strategy. Since both non-speech and speech GMMs have \( K \) Gaussian distributions, \( K \) types of pruned GMMs are obtained by using a set of selected Gaussian distributions \( \{ \tilde{M}_{i,j} \} \). \( \{ \tilde{M}_{i,j} \} \) is determined by the pruning threshold \( Z \). Essentially, the accumulation approach determines \( \tilde{M}_{i,j} \) with a variable pruning threshold \( Z_j(O_j) \), which maximizes \( \tilde{b}_j(O_j) \) with Gaussian weight normalization.

The three pruning parameter determination methods described above represent the property of frame-wise (local) determination for pruning parameters \( N_j(O_j) \), \( \beta_j(O_j) \), and \( Z_j(O_j) \), because each pruning parameter is locally determined as a variable that maximizes the likelihood \( \tilde{b}_j(O_j) \) depending on the observed signal \( O_j \) at the current frame and the model index \( j \). Thus, these pruning parameter determination methods further satisfy the aim of achieving Gaussian pruning in terms of sequential model re-estimation.

5. Experiments

5.1. Evaluation database

The proposed method was evaluated by using CENSREC-1-C (Kitaoka et al., 2007). CENSREC-1-C was designed as an evaluation framework for VAD in noisy environments and has two types of evaluation data sets, i.e., artificial data and real data. In this paper, we chose the real data set for the evaluation.

The real data were recorded in two real noisy environments (a restaurant and a street) with two different sound pressure levels (high SNR and low SNR). The low SNR recordings were made in a crowded restaurant and near a main highway at sound pressure levels of 69.2 dBA and 69.7 dBA (approximately −5 to 5 dB SNR), respectively. On the other hand, the high SNR recordings were made in an uncrowded restaurant and near a subsidiary highway at sound pressure levels of 53.4 dBA and 58.4 dBA (approximately 5–15 dB SNR), respectively. The microphone was about 50 cm from the speaker’s mouth. The data was originally recorded at a sampling rate of 48 kHz (with 16 bit quantization), and was down-sampled to 8 kHz.

There were 10 speakers (five males and five females) and the recorded speech consisted of four sound files per speaker in each recording situation. A single sound file included 8–10 utterances of continuous 1–12 digit numbers with two-second intervals. Thus, there were a total of 160 sound files (40 sound files per recording situation) and the recorded data included 1532 utterances (383 utterances per recording situation).

For the evaluation, reference VAD labels were generated by employing hand-labeled transcription, which included temporal information about the speech onset, speech offset, and pause.

5.2. Experimental setup

The feature parameter for the SKF-based VAD were 12th order LMFBs that were extracted by using a Hamming window with a 20 ms frame length and a 10 ms frame shift length. The state transition probabilities were set at \( a_{ij} = \{0.90, 0.10, 0.45, 0.55\} \). The silence and clean speech GMMs with 32 Gaussian distributions were trained by using speech data for the clean HMM training of CENSREC-1, also known as AURORA-2J (Nakamura et al., 2005). The digit contexts of AURORA-2J are exactly the same as those of AURORA-2 (Hirsch and Pearce, 2000),
however, English digits are translated into Japanese digits. The AURORA-2J data were recorded by using the translated digit contexts. The training data consisted of 8440 utterances spoken by 110 speakers. These training data are different from the real data of CENSREC-1-C.

The VAD experiments were performed with two evaluation schemes, i.e., frame-level evaluation and utterance-level evaluation. Frame-level evaluation is the main focus of this section. This evaluation was carried out to prove the robustness of the proposed pruning methods against various noise environments. On the other hand, utterance-level evaluation is proved robust in regard to threshold.

5.3. Experimental results of frame-level evaluation

Frame-level evaluation is the most reasonable approach to VAD evaluation. The criteria for frame-level evaluation are the false rejection rate (FRR) and the false acceptance rate (FAR) as shown by Eqs. (17) and (18), respectively.

\[
\text{FRR} = \frac{N_{FR}}{N_s} \times 100\% \\
\text{FAR} = \frac{N_{FA}}{N_{ns}} \times 100\%
\]

where \(N_s\), \(N_{ns}\), \(N_{FR}\), and \(N_{FA}\) are the total number of speech frames, the total number of non-speech frames, the number of speech frames detected as non-speech frames, and the number of non-speech frames detected as speech frames, respectively.

FAR and FRR are controlled by a threshold or certain parameters in the VAD technique, and have a trade-off relationship. Thus, we draw the receiver operating characteristics (ROC) curves by using several FARs and FRRs, which we obtained by changing the threshold for LRT.

Figs. 8–10 shows the results of frame-level evaluations. In each figure, “Baseline,” “Sohn,” “AFE,” “G.729B,” “SKF,” “N-best,” “Threshold,” and “Accumulation” represent results obtained with the baseline VAD technique of CENSREC-1-C (energy-based VAD with adaptive threshold) (Kitaoka et al., 2007), the statistical model-based VAD method proposed by Sohn et al. (1999), EISI ES 202 050 (advanced front-end) (ETSI ES 202 050 v.1.1.4, 2006), ITU-T G.729 Annex B (ITU-T Recommendation G.729 Annex B., 1996), the original SKF-based VAD, the N-best approach, posterior probability thresholding, and posterior probability accumulation, respectively. In each figure, the ROC curve closest to the origin shows the best performance. Only one result each for “AFE” and “G.729B” are shown in the figure, because these methods have no adjustable parameters.

In addition, Table 1 summarizes the results of the frame-level evaluation. In the table, all the results with the exception of “AFE” and “G.729B” were obtained by adjusting the threshold for LRT so that the FRR and FAR were approximately equal. These results are called the equal error rate.

![ROC curves obtained by Gaussian pruning based on N-best approach with automatic pruning parameter determination.](image)
Fig. 9. ROC curves obtained by Gaussian pruning based on posterior probability thresholding with automatic pruning parameter determination.

Fig. 10. ROC curves obtained by Gaussian pruning based on posterior probability accumulation with automatic pruning parameter determination.
As seen in the figures and the table, each Gaussian pruning method outperforms “SKF” and the other conventional methods as regards both the ROC curves and the equal error rate. In particular, the results in the presence of restaurant noise show significant improvement. These results prove that the proposed Gaussian pruning methods are effective for statistical model-based VAD. The proposed methods re-estimate the statistical model using only the information of the observed signal at the current frame, namely, the model re-estimation is realized with the smallest amount of re-training data and processing latency. The proposed method may be easily employed with other statistical model-based techniques, e.g., speech recognition.

Table 2 shows the average number of remaining Gaussian distributions. As seen in the table, we can confirm that the proposed method is improved by adaptively selecting a small number of Gaussian distributions. This method also justifies the aim of the proposed Gaussian pruning techniques.

From the results in Table 2, the number of Gaussian distributions is reduced almost to 1. Thus, we can assume that the Gaussian pruning achieved by selecting the most dominant Gaussian distribution provides the results same as the proposed methods. To prove the effectiveness of the proposed method, we carried out an evaluation with the selection of the 1-best Gaussian distribution, i.e., “N-best” with N = 1.

Table 3 compares the 1-best selection (“1-best”) and the proposed methods. As seen in the table, the results for “1-best” are better than for “SKF”, however, they are worse than those of the proposed methods. These results prove that the number of indispensable dominant Gaussian distributions changes for each frame and only the use of the most dominant Gaussian distribution is unsuitable for frame-wise Gaussian pruning. In particular, if the sound sources are in a transient state, i.e., speech onset, speech offset, or the appearance of burst noise, the acoustical characteristics of the observed signal vary greatly in a short time range, and become ambiguous. In such a time period, several Gaussian distributions are required to represent the ambiguous characteristics of the observed signal. Since the proposed method adaptively controls the optimum number of Gaussian distributions at each frame, it outperforms “1-best”.

The three proposed Gaussian pruning techniques provide almost the same results. Of these methods, “Threshold” and “Accumulation” require the parameter range and resolution of β and Z to be adjusted. On the other hand, the parameter range and resolution of “N-best” are fixed. Namely, the maximum number of N is given as K.

### Table 1
Results of frame-level evaluation by using automatic pruning parameter determination with FRR and FAR criteria.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Restaurant</th>
<th>Street</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>FRR criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>16.90</td>
<td>34.40</td>
<td>37.10</td>
</tr>
<tr>
<td>Sohn</td>
<td>29.70</td>
<td>31.00</td>
<td>30.40</td>
</tr>
<tr>
<td>AFE</td>
<td>7.30</td>
<td>2.50</td>
<td>7.00</td>
</tr>
<tr>
<td>G.729B</td>
<td>15.90</td>
<td>23.40</td>
<td>38.30</td>
</tr>
<tr>
<td>SKF</td>
<td>14.10</td>
<td>5.00</td>
<td>6.70</td>
</tr>
<tr>
<td>N-best</td>
<td>10.20</td>
<td>5.80</td>
<td>6.30</td>
</tr>
<tr>
<td>Threshold</td>
<td>10.10</td>
<td>6.10</td>
<td>5.90</td>
</tr>
<tr>
<td>Accumulation</td>
<td>10.20</td>
<td>5.80</td>
<td>6.30</td>
</tr>
</tbody>
</table>

| FAR criteria (%) |            |        |      |     |      |
| Baseline         | 18.40      | 33.00  | 35.20| 27.90|      |
| Sohn             | 28.90      | 31.70  | 31.20| 33.00|      |
| AFE              | 75.90      | 46.90  | 29.50| 56.93|      |
| G.729B           | 47.10      | 24.90  | 39.45|      |      |
| SKF              | 14.00      | 9.00   | 14.78|      |      |
| N-best           | 10.30      | 5.80   | 5.90 | 11.25|      |
| Threshold        | 10.10      | 6.20   | 5.70 | 11.13|      |
| Accumulation     | 10.30      | 6.20   | 4.90 | 11.13|      |

### Table 2
The average number of the remaining Gaussian distributions. The values in parentheses are the standard deviations.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Restaurant</th>
<th>Street</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>N-best method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Gaussians of non-speech model</td>
<td>1.34</td>
<td>1.53</td>
<td>2.14</td>
</tr>
<tr>
<td>(#)</td>
<td>(1.08)</td>
<td>(1.03)</td>
<td>(1.03)</td>
</tr>
<tr>
<td># of Gaussians of speech model</td>
<td>1.03</td>
<td>1.30</td>
<td>2.14</td>
</tr>
<tr>
<td>(#)</td>
<td>(0.30)</td>
<td>(0.20)</td>
<td>(0.23)</td>
</tr>
</tbody>
</table>

| Threshold method |
| # of Gaussians of non-speech model | 1.52 | 1.64 | 2.32 | 1.91 |
| (#) | (1.52) | (1.24) | (1.15) | (1.34) |
| # of Gaussians of speech model | 1.02 | 1.38 | 2.23 | 1.71 |
| (#) | (0.59) | (0.47) | (0.40) | (0.53) |

| Accumulation method |
| # of Gaussians of non-speech model | 1.34 | 1.53 | 2.13 | 1.72 |
| (#) | (1.08) | (1.03) | (1.03) | (1.08) |
| # of Gaussians of speech model | 1.03 | 1.29 | 2.12 | 1.61 |
| (#) | (0.30) | (0.20) | (0.23) | (0.34) |
and the resolution of $N$ is fixed at 1. Therefore, “N-best” is the most reasonable approach to the Gaussian selection, because it has no tuning parameters.

5.4. Experimental results of utterance-level evaluation

This section describes the results obtained with utterance-level evaluations.

The criteria for utterance-level evaluation are the utterance correct rate and the utterance accuracy rate as shown by Eqs. (19) and (20)

$$\text{Corr} = \frac{N_c}{N_u} \times 100\% \quad (19)$$

$$\text{Acc} = \frac{N_c - N_f}{N_u} \times 100\% \quad (20)$$

where $N_u$, $N_c$, and $N_f$ denote the total number of speech utterances, the number of correctly detected utterances, and the number of incorrectly detected utterances, respectively. Acc and Corr allow errors, i.e., a short time lag at utterance start and utterance end between the reference time label and the detected utterance. In this evaluation, the allowable length is set at 100 ms. With the LRT-based approach, “Sohn,” “SKF,” “1-best,” “N-best,” “Threshold,” and “Accumulation”, the threshold is set at a common value, 1.0E+1 (=10.0) for all the noise environments. This threshold is determined as the value that maximizes the Corrs and Accs of each environment. Here, the aim of these evaluations is to indicate the upper performance limit of the proposed method; automatic threshold optimization is not the aim of this work.

As seen in the table, the use of environmentally dependent threshold provides significant improvement compared with the results obtained with the common threshold. The

<table>
<thead>
<tr>
<th>Noise</th>
<th>Restaurant</th>
<th>Street</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>FRR criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKF</td>
<td>14.10</td>
<td>25.30</td>
<td>5.00</td>
</tr>
<tr>
<td>1-Best</td>
<td>11.10</td>
<td>23.20</td>
<td>5.40</td>
</tr>
<tr>
<td>N-best</td>
<td>10.20</td>
<td>23.40</td>
<td>5.80</td>
</tr>
<tr>
<td>Threshold</td>
<td>10.10</td>
<td>23.40</td>
<td>6.10</td>
</tr>
<tr>
<td>Accumulation</td>
<td>10.20</td>
<td>23.40</td>
<td>5.80</td>
</tr>
<tr>
<td>FAR criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKF</td>
<td>14.00</td>
<td>24.10</td>
<td>12.00</td>
</tr>
<tr>
<td>1-Best</td>
<td>11.40</td>
<td>23.20</td>
<td>7.60</td>
</tr>
<tr>
<td>N-best</td>
<td>10.30</td>
<td>23.10</td>
<td>6.20</td>
</tr>
<tr>
<td>Threshold</td>
<td>10.10</td>
<td>23.10</td>
<td>6.10</td>
</tr>
<tr>
<td>Accumulation</td>
<td>10.30</td>
<td>23.10</td>
<td>6.20</td>
</tr>
</tbody>
</table>

means that the proposed methods correctly detect speech activity periods and also discriminate correctly between the characteristics of non-speech and speech.

Tables 5 and 6 show the results obtained with a threshold depending on the noise and SNR and the values of the threshold for the LRT. The values in the table give the results shown in Table 4. These thresholds are determined as the values that maximize the Corrs and Accs of each environment. Here, the aim of these evaluations is to indicate the upper performance limit of the proposed method; automatic threshold optimization is not the aim of this work.

As seen in the table, the use of environmentally dependent threshold provides significant improvement compared with the results obtained with the common threshold. The

<table>
<thead>
<tr>
<th>Noise</th>
<th>Restaurant</th>
<th>Street</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>FRR criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKF</td>
<td>14.10</td>
<td>25.30</td>
<td>5.00</td>
</tr>
<tr>
<td>1-Best</td>
<td>11.10</td>
<td>23.20</td>
<td>5.40</td>
</tr>
<tr>
<td>N-best</td>
<td>10.20</td>
<td>23.40</td>
<td>5.80</td>
</tr>
<tr>
<td>Threshold</td>
<td>10.10</td>
<td>23.40</td>
<td>6.10</td>
</tr>
<tr>
<td>Accumulation</td>
<td>10.20</td>
<td>23.40</td>
<td>5.80</td>
</tr>
<tr>
<td>FAR criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKF</td>
<td>14.00</td>
<td>24.10</td>
<td>12.00</td>
</tr>
<tr>
<td>1-Best</td>
<td>11.40</td>
<td>23.20</td>
<td>7.60</td>
</tr>
<tr>
<td>N-best</td>
<td>10.30</td>
<td>23.10</td>
<td>6.20</td>
</tr>
<tr>
<td>Threshold</td>
<td>10.10</td>
<td>23.10</td>
<td>6.10</td>
</tr>
<tr>
<td>Accumulation</td>
<td>10.30</td>
<td>23.10</td>
<td>6.20</td>
</tr>
</tbody>
</table>

Table 4 shows the results with utterance-level evaluations. As seen in the table, the proposed methods outperform the conventional methods even if a common threshold is used. Since a huge number of incorrectly detected utterances are inserted, the Accs of “AFE” and “G.729B” are very low. The other conventional methods also include serious detection errors. In spite of these errors, the proposed methods perform well as regards Acc. This

<table>
<thead>
<tr>
<th>Noise</th>
<th>Restaurant</th>
<th>Street</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>FRR criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKF</td>
<td>14.10</td>
<td>25.30</td>
<td>5.00</td>
</tr>
<tr>
<td>1-Best</td>
<td>11.10</td>
<td>23.20</td>
<td>5.40</td>
</tr>
<tr>
<td>N-best</td>
<td>10.20</td>
<td>23.40</td>
<td>5.80</td>
</tr>
<tr>
<td>Threshold</td>
<td>10.10</td>
<td>23.40</td>
<td>6.10</td>
</tr>
<tr>
<td>Accumulation</td>
<td>10.20</td>
<td>23.40</td>
<td>5.80</td>
</tr>
<tr>
<td>FAR criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKF</td>
<td>14.00</td>
<td>24.10</td>
<td>12.00</td>
</tr>
<tr>
<td>1-Best</td>
<td>11.40</td>
<td>23.20</td>
<td>7.60</td>
</tr>
<tr>
<td>N-best</td>
<td>10.30</td>
<td>23.10</td>
<td>6.20</td>
</tr>
<tr>
<td>Threshold</td>
<td>10.10</td>
<td>23.10</td>
<td>6.10</td>
</tr>
<tr>
<td>Accumulation</td>
<td>10.30</td>
<td>23.10</td>
<td>6.20</td>
</tr>
</tbody>
</table>

Table 5

<table>
<thead>
<tr>
<th>Noise</th>
<th>Restaurant</th>
<th>Street</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Corr criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sohn</td>
<td>74.20</td>
<td>56.52</td>
<td>39.42</td>
</tr>
<tr>
<td>AFE</td>
<td>70.72</td>
<td>33.33</td>
<td>93.86</td>
</tr>
<tr>
<td>G.729B</td>
<td>43.77</td>
<td>46.67</td>
<td>79.13</td>
</tr>
<tr>
<td>SKF</td>
<td>51.88</td>
<td>43.48</td>
<td>43.86</td>
</tr>
<tr>
<td>1-Best</td>
<td>89.28</td>
<td>59.42</td>
<td>97.39</td>
</tr>
<tr>
<td>N-best</td>
<td>88.99</td>
<td>56.32</td>
<td>97.46</td>
</tr>
<tr>
<td>Threshold</td>
<td>90.72</td>
<td>61.16</td>
<td>97.97</td>
</tr>
<tr>
<td>Accumulation</td>
<td>89.28</td>
<td>62.03</td>
<td>97.68</td>
</tr>
<tr>
<td>Acc criteria (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sohn</td>
<td>21.45</td>
<td>43.48</td>
<td>15.65</td>
</tr>
<tr>
<td>AFE</td>
<td>24.35</td>
<td>21.16</td>
<td>79.13</td>
</tr>
<tr>
<td>G.729B</td>
<td>73.62</td>
<td>94.20</td>
<td>245.51</td>
</tr>
<tr>
<td>SKF</td>
<td>120.35</td>
<td>199.42</td>
<td>204.35</td>
</tr>
<tr>
<td>1-Best</td>
<td>68.12</td>
<td>5.22</td>
<td>93.33</td>
</tr>
<tr>
<td>N-best</td>
<td>68.99</td>
<td>12.17</td>
<td>95.07</td>
</tr>
<tr>
<td>Threshold</td>
<td>71.30</td>
<td>11.59</td>
<td>95.07</td>
</tr>
<tr>
<td>Accumulation</td>
<td>69.86</td>
<td>10.43</td>
<td>94.49</td>
</tr>
</tbody>
</table>
tendency of the optimum threshold value is similar to that of all the LRT-based approaches. In the restaurant noise environment, the interference is mainly babble noise, and its acoustical characteristics are similar to speech signals. In this environment, the discrimination of non-speech/speech periods is seriously confused. Therefore, to extract a period reliably, a high value threshold should be chosen. On the other hand, the street noise interference is car noise, and this has different acoustical characteristics from those of speech signals. Therefore, confusion as regards discrimination decreases and accurate discrimination can be achieved by using a threshold with small value. These facts show that the automatic determination of the optimum threshold for LRT is an important factor for VAD, and we will investigate this problem to realize a more robust VAD technique.

6. Conclusion

This paper presented Gaussian pruning techniques for statistical model-based VAD. The aim of Gaussian pruning is to realize an exact likelihood calculation that reflects the local characteristics of the observed signal. We investigated three different Gaussian pruning techniques based on posterior probability and corresponding pruning parameter determination techniques. The evaluation results show that our proposed method significantly improves VAD accuracy compared with our previous work, i.e., SKF-based VAD, and the conventional methods. In the future, we plan to determine the optimum threshold for LRT.

In addition, we also plan to investigate the integration of VAD and other speech processing techniques. In this investigation, we have already proposed a technique for integrating VAD and noise suppression (Fujimoto et al., 2009). As the next step, we will attempt to integrate the proposed VAD and a weighted finite state transducer (WFST)-based speech recognizer (Mohri et al., 2000; Hori et al., 2007). Both the proposed VAD and the WFST-based speech recognizer utilize a statistical approach and graphical models. Thus, the integration of these techniques offers good prospects for achieving an effective speech recognition technique.

Acknowledgements

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

Appendix A. Details of SKF-based VAD

This appendix describes SKF-based VAD in detail.

A.1. Noise estimation step

The sequential updating of the noise Gaussian parameter is carried out by Kalman filtering. In the formulation of a Kalman filter-based noise estimation, \( \mathbf{O}_t \) and \( \mathbf{N}_t \) denote the \( L \)-dimensional vector of the LMFB of the observed signal and noise at the \( t \)th frame, respectively. This step consists of the following three sub-steps.

(a) Definition of state-space model.
(b) Prediction step.
(c) Estimation step.

A.1.1. Definition of state-space model

The Kalman filter requires a definition of the signal model called a state-space model (dynamical system). Typically, a state-space model is defined by two equations: a state transition equation that represents the dynamics of the target signal, and an observation equation that represents the output system of the observed signal.

For the state transition process of noise, a random walk process is applied to the state equation as follows:

\[
N_{t+1} = N_{t} + W_{t} \tag{A.1}
\]

\[
W_{t} \sim \mathcal{N}(0, \sigma^2_{W, t}) \tag{A.2}
\]

where \( N_{t} \), \( W_{t} \) and \( \sigma^2_{W, t} \) denote the \( t \)th element of \( \mathbf{N}_t \), the driving noise for the state transition process and the variance of \( W_{t} \), respectively.

On the other hand, the observation equation is given by the following non-linear function,

\[
O_{t} = S_{t} + \log(1 + \exp(N_{t} - S_{t})) = f(S_{t}, N_{t}) \tag{A.3}
\]

where \( S_{t} \) denotes the LMFB of silence or clean speech given by the \( l \)th Mel frequency filter at the \( t \)th frame. Here, the Kalman filter, which is derived from a non-linear state-space model, is called an extended Kalman filter (EKF) or a non-linear Kalman filter.

In Eq. (A.3), parameter \( S_{t} \) is usually unknown, therefore, the Gaussian parameters of silence or clean speech GMMs are substituted for parameter \( S_{t} \) as follows:

\[
O_{t,j} = f(\mu_{S,j,k,l}, N_{t,j}) + V_{t,j,k,l} \quad \text{(A.4)}
\]

\[
V_{t,j,k,l} \sim \mathcal{N}(0, \sigma^2_{S,j,k,l}) \quad \text{(A.5)}
\]

where \( \mu_{S,j,k,l} \) and \( \sigma^2_{S,j,k,l} \) denote the mean and variance of the \( k \)th Gaussian distribution in silence (\( j = 0 \)) or speech (\( j = 1 \)) GMM, respectively. \( V_{t,j,k,l} \) denotes an error signal between \( S_{t,j} \) and \( \mu_{S,j,k,l} \).

Since a GMM consists of \( K \) Gaussian distributions, \( K \) types of observation equations are derived from Eq. (A.4). Using these observation equations, the EKF is multiplied into \( K \) types and we can obtain \( K \) types of

Table 6
The values of the environmentally dependent threshold for the LRT.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Restaurant</th>
<th>Street</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Sohn</td>
<td>3.0E+1</td>
<td>4.0E+1</td>
</tr>
<tr>
<td>SKF</td>
<td>4.0E+1</td>
<td>3.0E+1</td>
</tr>
<tr>
<td>1-Best</td>
<td>4.0E+1</td>
<td>1.0E+1</td>
</tr>
<tr>
<td>N-best</td>
<td>5.0E+1</td>
<td>2.0E+1</td>
</tr>
<tr>
<td>Threshold</td>
<td>5.0E+1</td>
<td>3.5E+1</td>
</tr>
<tr>
<td>Accumulation</td>
<td>5.0E+1</td>
<td>4.0E+1</td>
</tr>
</tbody>
</table>

This appendix describes SKF-based VAD in detail.
estimation candidates for GMMs of silence and clean speech. The multiplied Kalman filter is called a switching Kalman filter (SKF).\(^1\) Usually, SKF estimates a target signal by switching (selecting) the state-space models with a hard decision scheme in each time slice. However, if the decision criterion for the state-space model selection is not certain, undesirable errors will occur in the hard decision scheme. To avoid this problem, we employ soft decision-based SKF, which estimates the noise candidates by using all the state-space models. After estimation with the multiplied state-space model, the noise estimate candidates are unified by employing the two stage weighted averaging described in Appendix A.2.2. The unified estimate is utilized for estimation at the next time slice.

\(\text{A.1.2. Prediction step}\)

The formula for estimating each EKF is given by iterating the prediction step and the estimation step. The prediction step predicts the parameters of the \(t\)th frame from the estimated parameters of the \(t-1\)th frame. With the proposed method, the input parameters of the prediction step are \(\hat{N}_{t-1,l}\) and \(\hat{\sigma}^2_{N,t-1,l}\) which denote the mean and variance of the noise GMM estimated by the soft decision-based SKF at the previous frame \(t-1\), respectively. The output parameters are \(N_{t-1,j,k,l}\) and \(\sigma^2_{N,t-1,j,k,l}\), which denote the predicted mean and variance of the noise GMM, respectively.

The prediction step is derived as follows:

\[
\begin{align*}
N_{t-1,j,k,l} & = \hat{N}_{t-1,l} \\
\sigma^2_{N,t-1,j,k,l} & = \hat{\sigma}^2_{N,t-1,l} + \sigma^2_{W,l}
\end{align*}
\]  
(A.6)

where subscripts \(t\)\(t-1\) denotes the predicted parameter from the \(t-1\)th frame.

As the intermediate parameters, this step also outputs \(\mu_{O,t-1,j,k,l}\) and \(\sigma^2_{O,t-1,j,k,l}\), which denote the predicted mean and variance of the observed signal \(O_t\) given by the previous frame \(t-1\), respectively. \(\mu_{O,t-1,j,k,l}\) and \(\sigma^2_{O,t-1,j,k,l}\) are derived by the following equations:

\[
\begin{align*}
\mu_{O,t-1,j,k,l} & = f(\mu_{S,j,k,l}, N_{t-1,j,k}) \\
\sigma^2_{O,t-1,j,k,l} & = F_{t-1} \cdot \sigma^2_{N,t-1,j,k,l} \cdot F_{t-1} + \sigma^2_{S,j,k,l}
\end{align*}
\]  
(A.8)

where

\[
F_{t-1,j,k,l} = \frac{\partial f(\mu_{S,j,k,l}, N_{t-1,j,k,l})}{\partial N_{t-1,j,k,l}} = \frac{\exp(N_{t-1,j,k,l} - \mu_{S,j,k,l})}{1 + \exp(N_{t-1,j,k,l} - \mu_{S,j,k,l})}
\]  
(A.10)

This method requires initial values of \(\hat{N}_{0,l}\) and \(\hat{\sigma}^2_{N,0,l}\). In this paper, we assumed the first (0th) frame of the observed signal to be a noise frame. Thus, we used initial values of \(\hat{N}_{0,l} = O_{0,l}\) and \(\hat{\sigma}^2_{N,0,l} = 0\).

\(\text{A.1.3. Estimation step}\)

The estimation step estimates the parameters of the \(t\)th frame based on the predicted parameters and the observed signal. In other words, the estimation step corrects the predicted parameters by using information about the observed signal.

With the proposed method, the input parameters of the estimation step are \(N_{t-1,j,k,l}\) and \(\sigma^2_{N,t-1,j,k,l}\). The output parameters are \(N_{t,j,k,l}\) and \(\sigma^2_{N,t,j,k,l}\), which denote an \(N_{t,j}\) candidate estimated using the parameters of the \(t\)th Gaussian distribution contained in model \(f\) (silence or speech GMM) and the squared error variance of each candidate, respectively.

The estimation step is derived as follows:

\[
N_{t,j,k,l} = N_{t-1,j,k,l} + G_{t,j,k,l}(O_t - \mu_{O,t-1,j,k,l}) \\
\sigma^2_{N,t,j,k,l} = (1 - G_{t,j,k,l} \cdot F_{t-1,j,k,l}) \sigma^2_{N,t-1,j,k,l}
\]  
(A.11)

where

\[
G_{t,j,k,l} = \frac{\sigma^2_{S,t-1,j,k,l} \cdot F_{t-1,j,k,l}}{\sigma^2_{O,t-1,j,k,l}}
\]  
(A.13)

\(\text{A.2. Composition step}\)

This section describes the composition of the noisy speech model by using the pre-trained clean speech model and the estimated noise model.

\(\text{A.2.1. Model composition}\)

In the LMFB domain, the noisy speech model, which is modeled by a GMM is composed using the following equations:

\[
\begin{align*}
\mu_{O,t,j,k,l} & = f(\mu_{S,j,k,l}, N_{t,j,k,l}) \\
\sigma^2_{O,t,j,k,l} & = F_{t,j,k,l} \cdot \sigma^2_{N,t,j,k,l} + \sigma^2_{S,j,k,l}
\end{align*}
\]  
(A.14)

where

\[
F_{t,j,k,l} = \frac{\exp(N_{t,j,k,l} - \mu_{S,j,k,l})}{1 + \exp(N_{t,j,k,l} - \mu_{S,j,k,l})}
\]  
(A.16)

\(\mu_{O,t,j,k,l}\) and \(\sigma^2_{O,t,j,k,l}\) denote the composed mean and variance of \(O_{t,k}\) respectively.

\(\text{A.2.2. Unification of multiplied noise estimation}\)

Each estimated candidate for the soft decision-based SKF is unified by two stage weighted averaging. The first weighted averaging is derived as follows:

\[
N_{t,j,l} = \sum_{k=1}^{K} P(k|O_t) \cdot N_{t,j,k,l}
\]  
(A.17)

\[
\sigma^2_{N,t,j,l} = \sum_{k=1}^{K} P(k|O_t) \cdot \sigma^2_{N,t,j,k,l}
\]  
(A.18)

where

\[
P(k|O_t) = \frac{w_{j,k} \cdot N(O_t; \mu_{O,j,k,l}, \sigma^2_{O,j,k,l})}{\sum_{k'=1}^{K} w_{j,k'} \cdot N(O_t; \mu_{O,j,k,l}, \sigma^2_{O,j,k,l})}
\]  
(A.19)

\(N_{t,j,l}\) and \(\sigma^2_{N,t,j,l}\) denote averaged results. \(P(k|O_t)\) and \(w_{j,k}\) denote the posterior probability of each Gaussian
distribution and the Gaussian (mixture) weight of the speech or silence GMM, respectively. In this method, \( P(k|O_{ij}) \) is used for the weight of each estimated candidate. \( \mu_{O_{ij},k} \) is a vector that has \( \mu_{O_{ij},k,l} \) in each element and \( \Sigma_{O_{ij},k} \) is a matrix that has \( \sigma^2_{O_{ij},k,l} \) in each diagonal element.

Here, GMMs of non-speech and speech, i.e., the noisy speech GMMs at the \( t \)th frame, are constructed by using \( w_{j,k} \mu_{O_{ij},k} \) and \( \Sigma_{O_{ij},k} \). Thus, the denominator of Eq. (A.19) is equivalent to the likelihood \( b_j(O_j) \) of the non-speech state (\( j = 0 \), silence + noise) or speech state (\( j = 1 \), speech + noise).

After the first weighted averaging of Eqs. (A.17) and (A.18), the final estimation result of the noise used for estimation at the next frame is given by the weighted average (second stage) with normalized likelihood \( b_j(O_j) \) as follows:

\[
\tilde{N}_{ij} = \frac{b_0(O_j)}{b_0(O_j) + b_1(O_j)} N_{i,0,j} + \frac{b_1(O_j)}{b_0(O_j) + b_1(O_j)} N_{i,1,j} \quad \text{(A.20)}
\]

\[
\tilde{\sigma}^2_{N_{ij}} = \frac{b_0(O_j)}{b_0(O_j) + b_1(O_j)} \tilde{\sigma}^2_{N_{i,0,j}} + \frac{b_1(O_j)}{b_0(O_j) + b_1(O_j)} \tilde{\sigma}^2_{N_{i,1,j}} \quad \text{(A.21)}
\]

### A.3. Discrimination step

In the formulation of the LRT with a transition process of noise, when \( O_{0,j} = \{O_0, \ldots, O_j\} \) and \( N_{0,j} = \{N_0, \ldots, N_j\} \) are given, the speech state \( q_t \) is decided with respect to the conditional probability \( p(q_t|O_{0,t}, N_{0,t}) \). The conditional probability \( p(\hat{q}_t|O_{0,t}, N_{0,t}) \) has the following relationship with the joint probability \( p(q_t|O_{0,t}, N_{0,t}) \):

\[
p(\hat{q}_t|O_{0,t}, N_{0,t}) = \frac{p(O_{0,t}, q_t, N_{0,t})}{p(O_{0,t}, N_{0,t})} \propto p(O_{0,t}, q_t, N_{0,t}) \quad \text{(A.22)}
\]

Here, we assume that speech state \( q_t \) and noise state \( N_t \) are mutually independent, therefore, the joint probability \( p(q_t|O_{0,t}, N_{0,t}) \) can be represented by the recursive formula of Eq. (A.22) by assuming a first order Markov chain:

\[
p(O_{0,t}, q_t, N_{0,t}) = \sum_{q_{t-1}} p(q_t|q_{t-1}) p(O_{t}, q_{t}, N_{t}) p(N_{t}|N_{t-1}) \times p(O_{0,t-1}, q_{t-1}, N_{0,t-1}) \quad \text{(A.23)}
\]

The right hand side of Eq. (A.23) is equivalent to the forward probability \( \alpha_{j,t} \). Thus, Eq. (A.23) can be efficiently computed by the following recursive formula:

\[
p(O_{0,t}, q_t, N_{0,t}) = \alpha_{j,t} = \sum_{i=0}^{t} a_{j,i} \cdot b_j(O_i) \cdot \epsilon_{i,t-1} \cdot \alpha_{i,t-1} \quad \text{(A.24)}
\]

where \( a_{j,i} = p(q_t = H_i|q_{t-1} = H_i) \) (the state transition probability of speech), \( b_j(O_i) = p(O_t|q_t = H_j, N_t) \) (the likelihood given by state \( H_j \) at the \( t \)th frame), and \( \epsilon_{i,t-1} = p(N_{t-1}|N_{t-1}) \) (the state transition probability of noise). When \( t = 0 \), i.e., the first (beginning) frame is assumed to be a non-speech frame. Thus, the initial values \( \alpha_{0,0} = 1 \) and \( \alpha_{1,0} = 0 \) are given.

In Eq. (A.24), since the noise has a sequential state transition process, we assume \( \epsilon_{i,t-1} = 1 \). Thus, Eq. (A.24) is simplified with the same formulation as Eq. (5). The LRT is also given by the same formulations as Eqs. (1) and (6). The difference with regard to Sohn’s method is the likelihood calculation that considers the noise sequence.

### References


ETSI EN 301 708 v.7.1.1, 1999. Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels.

ETSI ES 202 050 v.1.1.4, 2006. Speech processing, transmission and quality aspects (STQ), advanced distributed speech recognition; front-end feature extraction algorithm; compression algorithms.


