Voice Activity Detection Based on Augmented Statistical Noise Suppression

Yasunari Obuchi, Ryu Takeda, and Naoyuki Kanda
Central Research Laboratory, Hitachi Ltd., Tokyo, Japan
E-mail: yasunari.obuchi.jx@hitachi.com

***Abstract*—A new voice activity detection (VAD) algorithm using augmented statistical noise suppression is introduced. Statistical noise suppression is an effective tool for speech processing under noisy conditions. It achieves the best VAD performance when the noise suppression is augmented in various ways. The speech distortion, which is usually a severe side effect of strong noise suppression, does not affect the VAD performance, and the correctly estimated signal power provides accurate detection of speech. The performance of the proposed algorithm is evaluated using CENSREC-1-C public database, and it is confirmed that the proposed algorithm outperforms other algorithms such as the switching Kalman filter-based VAD.**

I. Introduction

Voice activity detection (VAD) is one of the key components of various speech applications. It can be used to reduce the bandwidth usage in communication systems. An accurate VAD preprocessor can also increase the robustness of speech recognition under noisy conditions. Moreover, an adaptive system such as an echo canceller or an adaptive beamformer needs to detect a noise-only period of the input audio stream to update its configuration.

Traditional VAD algorithm was based on thresholding of the instantaneous or short term power, which can be realized even by an analog circuit. As the digital technology developed, more sophisticated VAD algorithms have appeared. Some of them introduced new features, such as the cepstral feature [1], spectral entropy [2], periodic to aperiodic component ratio [3], and so on. Other methods try to trace the state sequence using statistical estimation [4], [5]. The latter approach, which has the origin at MMSE-based statistical noise suppression [6], has exhibited the better performance than the others, and it could be attributed either to the accurate noise estimation by the statistical model or to the precise state tracking.

In this paper, we start from the assumption that the accurate noise estimation is the main cause of the good performance of the state-of-the-art VAD algorithms, such as [4] and [5]. Accordingly, there are two points that would improve their algorithms: developing an even better noise estimation module, and optimizing the remaining part of the system after separating the noise estimation module. For the first point, we replace the noise estimation module with the optimally-modified log spectral amplitude (OM-LSA) speech estimator [7], which is known to be one of the best statistical noise suppression algorithms. Moreover, based on the empirical knowledge that over-subtraction improves the accuracy of the spectral subtraction-based VAD [8], we introduce some augmentation schemes in noise suppression. For the second point, we separate two main modules of the Sohn's and Fujimoto's algorithm, which are noise estimation and likelihood ratio test. We use the first module for noise suppression, and replace the second module with a more conservative power-based VAD. The effectiveness of combining these two approaches will be confirmed by the experiments using real data.

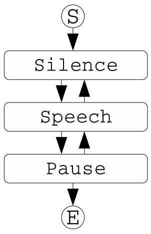

Fig. 1. State transition topology

TABLE I
State transition rules

Rule No.	From	condition $p > P$	$l_s > L_1$	$l_p > L_2$	To
1	**Silence**	yes			**Speech**
2	**Speech**	no	no		**Silence**
3	**Speech**	no	yes		**Pause**
4	**Pause**	yes			**Speech**
5	**Pause**	no		yes	**E**

The remainder of this paper is organized as follows. In the next section, we briefly review the traditional power-based VAD algorithm. The third section presents the outline of statistical noise suppression, particularly the OM-LSA speech estimator. More importantly, various augmentation techniques to improve the VAD accuracy will be provided in this section. The new VAD algorithm is evaluated using CENSREC-1-C [9], a public database for VAD evaluation, and the results are shown in Section 4 with reference to some published works. The last section is for the concluding remarks.

II. Review of Power-based VAD

The basic idea of the power-based VAD is simple. The input audio stream is divided into short term frames (typically 20 ms long and half overlapping), and the power is calculated for each of them. Finally, a frame is regarded as speech if its power is larger than the threshold, noise (or silence) otherwise. However, such a simple classification causes fragmentation of the spoken utterances, because even a connected utterance may include many short pauses. Such fragmentation can be avoided by introducing a simple state transition model with additional thresholds related to the duration.

Figure 1 shows a typical state transition topology, and Table I shows the corresponding transition rules. Transition always starts at **Silence**. If the power p, speech length l_s, and pause length l_p satisfy the condition of a rule whose "From" state matches the current state, then the current state changes to "To" state. If no rule matches, the current state is kept. For example, if the current state is **Speech**, p is not larger than P, and l_s is larger than L_1, then the current state changes to **Pause** (see Rule No. 3). In the above-mentioned process, the speech length is incremented at every frame belonging to the **Speech** or **Pause** state, and the pause length is incremented at every frame belonging to the **Pause** state.

Finally, an speech segment candidate is made when the transition reached **E**. The starting point of the segment corresponds to the last transition from **Silence** to **Speech**, and the ending point corresponds to the last transition from **Speech** to **Pause**. The candidate is discarded if it is too short or too long. Otherwise, small margins are added to the both sides of the segment if necessary, and the system tells that a speech segment has been detected.

III. Statistical Noise Suppression-based VAD

A. OM-LSA Speech Estimator for Power-based VAD

Statistical noise suppression takes place in the frequency domain, in which each time-frequency bin is identified by the frequency index k and frame index l. Once the noise-suppressed signal $\hat{X}(k,l)$ is obtained, the power of the l-th frame, which is to be used in the state transition judgment, is calculated as

$$p(l) = \sum_{k=0}^{K-1} w(k)|\hat{X}(k,l)|^2 \qquad (1)$$

where K is the number of frequency bins, and $\mathbf{w}$ is a weight vector. According to the preliminary experiment results, an A-weighting filter is used as $\mathbf{w}$.

Noise suppression by the OM-LSA speech estimator is defined as

$$|\hat{X}(k,l)|^2 = G(k,l)|Y(k,l)|^2. \qquad (2)$$

where $Y(k,l)$ denotes the observed noisy signal, and $G(k,l)$ is the gain function. First, the gain function of the LSA algorithm, $G_H(k,l)$, can be obtained by applying the MMSE criterion to the log-spectral amplitude error function. The solution is expressed as follows.

$$\begin{aligned} G_H(k,l) &= f(\xi(k,l),\gamma(k,l)) &(3)\\ f(\xi,\gamma) &= \frac{\xi}{1+\xi}\exp(\frac{1}{2}\int_{\gamma\xi/(1+\xi)}^{\infty}\frac{e^{-t}}{t}dt) &(4)\\ \gamma(k,l) &= \frac{|Y(k,l)|^2}{\alpha\sigma_m^2(k,l)} &(5)\\ \xi(k,l) &= c_1G_H^2(k,l-1)\gamma(k,l-1)\\ &\quad +(1-c_1)max\{\gamma(k,l)-1,0\} &(6)\end{aligned}$$

where ξ, γ, and α are called *a priori SNR*, *a posteriori SNR*, and *subtraction coefficient* respectively. The OM-LSA algorithm takes into account the speech presence probability $p(k,l)$, and modifies $G_H(k,l)$ to obtain the better form of $G(k,l)$.

$$\begin{aligned} G(k,l) &= [G_H(k,l)]^{p(k,l)}G_{min}^{1-p(k,l)} &(7)\\ p(k,l) &= [1+c_2(1+\xi(k,l))e^{-\nu(k,l)}]^{-1} &(8)\\ \nu(k,l) &= \gamma(k,l)\xi(k,l)/(1+\xi(k,l)) &(9)\end{aligned}$$

In the equations above, c_1, c_2, and G_{min} are adjustable parameters, and their values are fixed as 0.99, 0.25, and 0.01 throughout this paper.

B. Augmentation of Noise Suppression

When we use noise suppression for VAD, the top priority should be given to removing any unreliable components, regardless of the possibility to cause signal distortion. To further pursue the preference for noise removal instead of signal reconstruction, we apply some augmentation techniques for the OM-LSA speech estimator.

The first step of augmentation is realized by using a large α in eq.(5). It is known that the optimal value of α for speech recognition is smaller than 1.0 [10], because a large α causes distortion of the speech. However, distortion does not affect VAD severely, and it is known in the spectral subtraction case that over-subtraction is more favorable for VAD. It suggests that we should use much larger value of α.

The second step is the augmentation of the gain function. Since a smaller value of $G(k,l)$ means that the power of this frequency bin is less reliable, it is reasonable to lower the contribution of such bins by modifying eq.(2) as

$$|\hat{X}(k,l)|^2 = G^{\beta}(k,l)|Y(k,l)|^2 \qquad (10)$$

where β should be larger than 1.0. If we use large β, only the speech-dominant frequency bins contribute to the frame power.

The third augmentation aims at avoiding a false acceptance caused by very strong noises. Although it is difficult to remove such noises completely, the fact that a prominent frequency component exists indicates that the prominent component itself is made of noise. It is because some noises such as electrical beep and musical instrument noise have strong peaks at limited frequencies. Therefore, such prominent components should be removed, which modifies eq.(10) again as

$$|\hat{X}(k,l)|^2 = \begin{cases} G^{\beta}(k,l)|Y(k,l)|^2 & \text{rank}(k) \geq \eta K \\ 0 & \text{otherwise} \end{cases} \qquad (11)$$

where rank(k) is the number of frequency components in the same frame whose magnitude is larger than the k-th component.

After the augmentation by introducing large α, β, and η, the resulted audio signal is sometimes heavily distorted and not quite recognizable. However, one should use the output of eq.(11) only for VAD and prepare another noise-suppressed signal for speech recognition using a more conservative setting.

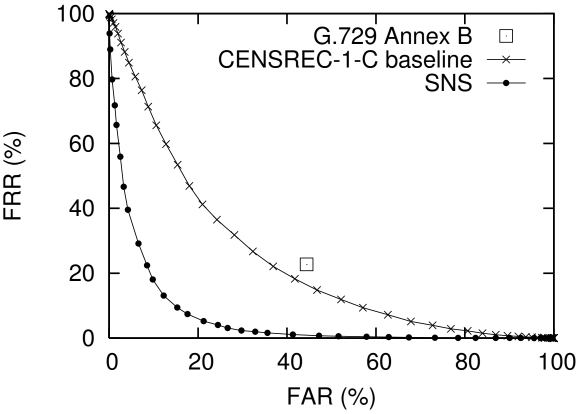

Fig. 2. ROC curve of baseline SNS algorithm.

IV. Evaluation Experiments Using CENSREC-1-C

A. Experimental setup

The proposed VAD algorithm was evaluated using CENSREC-1-C, a public database for evaluation of VAD algorithms. We used so-called *real* set of CENSREC-1-C, which was recorded in real noisy environments. The test data consist of four subsets, Restaurant/High-SNR, Restaurant/Low-SNR, Street/High-SNR, and Street/Low-SNR, each of which consists of 345 utterances in 36 files. The 8 kHz downsampled data were distributed with the correct endpoint labels. Some scripts to calculate the false acceptance rate (FAR) and false rejection rate (FRR) were also included in the distribution.

B. Baseline experiments

First, the baseline statistical noise suppression (SNS) algorithm was applied to CENSREC-1-C. Various values of the threshold P of TABLE I were tried, while $L_1 = 50\,\mathrm{ms}$ and $L_2 = 800\,\mathrm{ms}$ were fixed. Segment candidates shorter than 100 ms were discarded, and the other segments were accepted, however long they are. Noise suppression was executed using 64 ms half-overlapping frames, but the frame power p was re-calculated using 20 ms half-overlapping frames. It should also be noted that eq.(11) was applied after the second framing was performed in the experiments described later.

Figure 2 shows the receiver operating characteristic (ROC) curve of the baseline SNS algorithm, obtained by various thresholds. The output of VAD was scored by the frame-by-frame basis using the attached script. The results of CENSREC-1-C baseline VAD script and ITU-T G.729 Annex B [11] were also plotted. The former was included in CENSREC-1-C for reference, and the latter is a standard VAD which is commonly used in many VoIP applications. From this figure, it is clear that the SNS algorithm improves the VAD accuracy of standard algorithms greatly.

C. Augmentation of Noise Suppression

Next, the effects of various augmentation were tested. Figure 3 shows how the VAD accuracy changes as α increases. The vertical axis represents the average of FAR and FRR. There is a very large improvement of the accuracy from $\alpha = 1$ to $\alpha = 2$, and then it becomes stable with larger α. Figure 4 shows the results obtained by $\alpha = 5.0$ and various β, in which a small but clear improvement was obtained by β larger than 1.1. Figure 5 shows the results obtained by $\alpha = 5.0$, $\beta = 1.4$ and various η. We also observed great improvement from $\eta = 0.01$ to $\eta = 0.05$, and it becomes stable with larger η, although there is a small degradation.

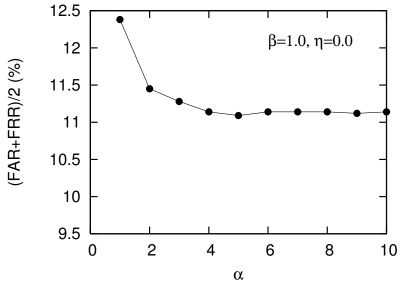

Fig. 3. Results of over-subtraction experiments.

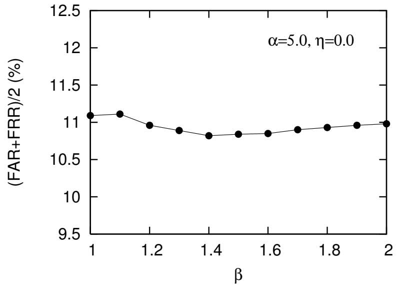

Fig. 4. Results of gain augmentation experiments.

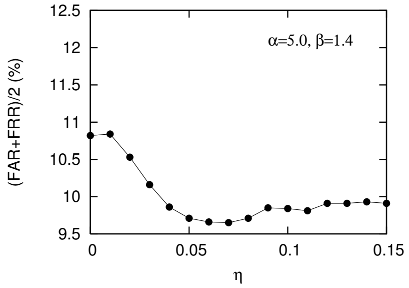

Fig. 5. Results of prominent component removal experiments.

We have investigated the augmented statistical noise suppression (ASNS) algorithm in more detail, with the best parameter setting ($\alpha = 5.0$, $\beta = 1.4$, and $\eta = 0.07$). Table II is the detailed results in the spread sheet provided by CENSREC-1-C, where the threshold P was adjusted to get approximately equal values of FAR and FRR. Although these results were obtained with the carefully adjusted parameters, Figs. 3, 4, and 5 show that the improvements are stable with large α, β, and η. For example, we obtained $\mathrm{FRR} = 9.62\%$ and $\mathrm{FAR} = 10.07\%$,

TABLE II

Frame-level evaluation of ASNS with the optimal parameter setting.

Real Data	False Rejection Rate [%]			
		Remote Microphone		
		Restaurant	Street	Average
False Rejection Rate [%]	High SNR	4.80	5.40	5.10
	Low SNR	13.60	13.70	13.65
	Average	9.20	9.55	9.38

Real Data	False Acceptance Rate [%]			
		Remote Microphone		
		Restaurant	Street	Average
False Acceptance Rate [%]	High SNR	11.00	1.20	6.10
	Low SNR	25.90	1.60	13.75
	Average	18.45	1.40	9.93

which are comparable to the results of Table II, with rather ill-tuned parameters of $\alpha = 8.0$, $\beta = 1.8$, and $\eta = 0.1$.

Figure 6 shows comparison results between ASNS and other state-of-the-art algorithms, including the ones mentioned earlier in this paper. First, the algorithm based on high order statistics (HOS) [12] is a simple unsupervised VAD algorithm. However, as reported in [12], it was designed to discriminate close-talk and far-field speech, and could not detect remote microphone speeches correctly (although it is much better than CENSREC-1-C baseline and G.729 Annex B). The algorithm based on conditional random field (CRF) [13] integrates various features using CRF. It requires a speech model to be trained beforehand, but the VAD accuracy is notably improved. The obtained combination of FAR and FRR is approximately on the ROC curve of SNS. Finally, Fujimoto's SKF-based algorithm, which also requires a pre-trained speech model and was known to outperform Sohn's algorithm when evaluated by CENSREC-1-C [5], is slightly less accurate than SNS. However, it was reported that introducing Gaussian pruning (GP) makes the SKF-based algorithm more accurate [14]. As shown in Fig. 6, SKF/GP resulted in better performance than SNS. Finally, the ROC curve of ASNS lies far from others on the lower-left side, indicating the very promising performance of the new algorithm.

V. Conclusions

In this paper, we proposed a new VAD algorithm, which is based on augmented statistical noise suppression. The noise suppression is based on the OM-LSA speech estimator, but various augmentation techniques were introduced to improve the VAD accuracy. Traditional power-based algorithm was combined with noise suppression, instead of the likelihood ratio test. The proposed algorithm is robust under various noisy conditions, and has the advantage that any speech or noise model need not be trained beforehand. Evaluation experiments using CENSREC-1-C public database demonstrated that the proposed algorithm has better performances than many known algorithms.

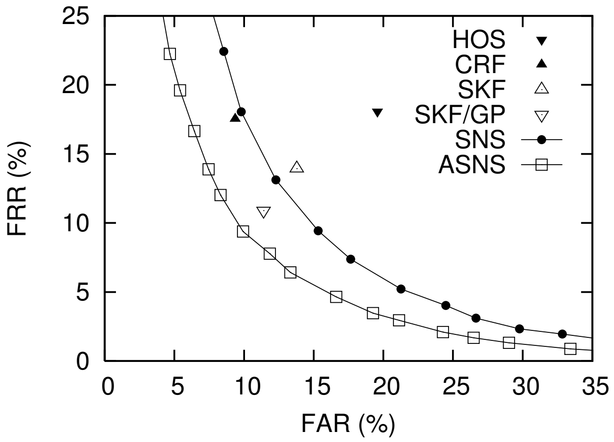

Fig. 6. Comparison of VAD accuracy. **HOS**: high order statistics. **CRF**: conditional random fields. **SKF**: switching Kalman filter. **SKF/GP**: switching Kalman filter with Gaussian pruning. **SNS**: statistical noise suppression. **ASNS**: augmented statistical noise suppression.

References

[1] S.E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," *Proc. IEEE ICASSP 2002*, Orlando, FL, USA, 2002.

[2] J.-L. Shen, J.-W. Hung, and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," *Proc. ICSLP 1998*, Sydney, Australia, 1998.

[3] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Pittsburgh, PA, USA, 2006.

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

[5] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," *IEICE Trans. Information and Systems*, Vol.E91-D, No.3, pp.467-477, 2008.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol.32, no.6, pp.1109-1121, 1984.

[7] I. Cohen and B. Berdugo, "Speech enhancement for Non-stationary Noise Environments," *Signal Processing*, vol.81, pp.2403-2418, 2001.

[8] M.W. Mak and H.B. Yu, "Robust voice activity detection for interview speech in NIST speaker recognition evaluation," *Proc. APSIPA ASC 2010*, Hong Kong, 2010.

[9] N. Kitaoka, et al., "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoustical Science and Technology*, vol.30, no.5, pp.363-371, 2009.

[10] Y. Obuchi, R. Takeda, and M. Togami, "Bidirectional OM-LSA speech estimator for noise robust speech recognition," *Proc. 2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Big Island, HI, USA, 2011.

[11] ITU-T, "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," ITU-T Recommendation G.729 - Annex B, 1996.

[12] D. Cournapeau and T. Kawahara, "Evaluation of real-time voice activity detection based on high order statistics," *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[13] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda, "Voice activity detection based on conditional random fields using multiple features," *Proc. Interspeech 2010*, Makuhari, Japan, 2010.

[14] M. Fujimoto, S. Watanabe, and T. Nakatani, "Voice activity detection using frame-wise model re-estimation method based on Gaussian pruning with weight normalization," *Proc. Interspeech 2010*, Makuhari, Japan, 2010.