SIP (2017), vol. 6, e7, page 1 of 9 © The Authors, 2017.

doi:10.1017/ATSIP.2017.8

ORIGINAL PAPER

Combining augmented statistical noise suppression and framewise speech/non-speech classification for robust voice activity detection

YASUNARI OBUCHI

This paper proposes a new voice activity detection (VAD) algorithm based on statistical noise suppression and framewise speech/non-speech classification. Although many VAD algorithms have been developed that are robust in noisy environments, the most successful ones are related to statistical noise suppression in some way. Accordingly, we formulate our VAD algorithm as a combination of noise suppression and subsequent framewise classification. The noise suppression part is improved by introducing the idea that any unreliable frequency component should be removed, and the decision can be made by the remaining signal. This augmentation can be realized using a few additional parameters embedded in the gain-estimation process. The framewise classification part can be either model-less or model-based. A model-less classifier has the advantage that it can be applied to any situation, even if no training data are available. In contrast, a model-based classifier (e.g., neural network-based classifier) requires training data but tends to be more accurate. The accuracy of the proposed algorithm is evaluated using the CENSREC-1-C public framework and confirmed to be superior to many existing algorithms.

Keywords: Speech, Voice activity detection, Noise suppression, Convolutional neural network, CENSREC-1-C

Received 12 September 2016; Revised 5 June 2017

I. INTRODUCTION

Voice activity detection (VAD) is a task to identify active voice periods from an input audio signal. The audio signal may include stationary or transient noise, music, or background speech (babble noise). Even if the power of these interfering sounds is large, and the signal-to-noise ratio (SNR) is low, the voice activity detector must extract the signal of voice periods only and send them to a speech-communication device or speech recognizer. Accurate VAD is essential for the communication device to achieve efficiency and to avoid insertion and deletion errors by the speech recognizer.

Since human speech is approximately stable on the time scale of 20–30 ms, most VAD studies are based on framewise processing. In the framewise processing, a feature vector is extracted from each frame and used by the binary classifier to distinguish between speech and non-speech. From this perspective, we can categorize previous studies of VAD into the two following groups.

The first group focuses on identifying better framewise features. Obviously, the simplest feature is instantaneous power, which is sufficient in high-SNR environments. However, power-based VAD is vulnerable to noise and does not work effectively in many applications. To achieve robustness under noisy conditions, various framewise features have been proposed. Even in the early period of digital communication, the accuracy of VAD results were known to improve by the zero-crossing rate [1]. Recently proposed features include cepstral features [2], MFCC [3, 4], spectral entropy [5], long-term spectral envelope [6], periodic–aperiodic component ratio [7], and higher order statistics [8].

The second group of VAD studies focuses on the classifier. Although fixed or adaptive thresholding is used for one-dimensional features, various classification algorithms are applicable for multi-dimensional features. Examples of simple methods are the Euclidean distance [2] and LDA (linear discriminant analysis) [3] methods. More sophisticated approaches include the GMM (Gaussian mixture model) [9] and the support vector machine (SVM) [4, 6]. More recently, classifiers based on DNNs (deep neural networks) have been proposed [10–13].

As described above, a VAD system can be made by combining a feature extractor and a speech/non-speech classifier. However, such a simple combination does not take into account that both the human voice and various noises have temporal dependency. In contrast, VAD accuracy under noisy conditions can be improved by introducing a scheme

School of Media Science, Tokyo University of Technology, 1404-1 Katakura, Hachioji, Tokyo 192-0982, Japan

Corresponding author:
Y. Obuchi
Email: obuchiysnr@stf.teu.ac.jp

to deal with inter-frame information. The simplest example is temporal smoothing, in which a short period of speech is re-labeled as non-speech, and a short period of non-speech is re-labeled as speech. The hangover scheme, in which additional frames are re-labeled as speech on both sides of a speech period, is also used in many cases. Some features such as long-term spectral envelope and order statistics filter [14] utilize temporal information implicitly. Classifiers using HMM (hidden Markov model) [15], conditional random field (CRF) [16], and recurrent neural network [11] are more explicit examples of incorporating temporal information.

Among the methods in the second group, one of the standard is the work of Sohn *et al.* [17], which is based on the decision-directed estimation of the speech and noise distribution [18]. In their method, speech and noise processes in each frequency band are treated as independent Gaussian random variables, and their parameters are estimated recursively. This approach is widely used in speech enhancement and is known as the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator. Following the success of Sohn's algorithm, Fujimoto and Ishizuka [19] proposed a more sophisticated temporal architecture that uses switching Kalman filter (SKF), resulting in significant improvement.

In this paper, we investigate a series of new VAD algorithms by extracting and expanding the key components in the successes of Sohn's and Fujimoto's methods. Although the noise suppression part and the classification part are strongly linked in their methods, we assume that each part can produce the desired results separately. Therefore, our approach focuses on introducing state-of-the-art noise suppression and classification methods and adapting them to VAD.

In the noise suppression part, the optimally modified log spectral amplitude (OM-LSA) speech estimator [20] is used. When applying OM-LSA, some augmented parameters are used because the existence of voice can be confirmed by reliable components only; the unreliable components should be removed to avoid errors. This is quite different from the case of speech communication and recognition, in which the distortion caused by the augmented parameters is very harmful. The introduction of augmented noise suppression is the first major contribution of this paper.

In the classification part, we first investigate approaches without model training. Such an approach could be applied to any language and situation, even when no training data are available. Subsequently, we pursue even higher VAD accuracy by using model-based approaches. The model can be trained in either an unsupervised or supervised manner. We examine various training methods and show that the best performance can be obtained using convolutional neural networks (CNNs). The introduction of a CNN as a postprocessor of augmented noise suppression is the second major contribution of this paper.

This paper is an extended version of [21, 22]. The former paper proposed to use augmented statistical noise suppression, while the latter proposed to use CNNs. In addition, more details on the implementation and some additional evaluation results are presented.

The remainder of this paper is organized as follows. In Section II, we briefly review the statistical noise suppression by OM-LSA and introduce augmented parameters. In Section III, a simple classification method without pre-trained models is described. In Section IV, we discuss various model-based classification methods and describe our final and optimal CNN-based classifier in detail. Experimental results are shown in Section V, and the last section is for conclusions.

II. AUGMENTED STATISTICAL NOISE SUPPRESSION

A) Original OM-LSA

The MMSE-STSA speech estimator was based on the idea that the amplitude of each frequency component $X(k,l)$ is a random variable, and the estimate $\hat{X}(k,l)$ should minimize

$$E\{(|X(k,l)| - |\hat{X}(k,l)|)^2\} \tag{1}$$

where (k,l) is the frequency component and frame indices.

The log spectral amplitude (LSA) estimator is the improved version of MMSE-STSA, in which the cost function is defined by the difference of log X:

$$E\{(\log|X(k,l)| - \log|\hat{X}(k,l)|)^2\}, \tag{2}$$

which is known to be more suitable for speech processing.

Solving (2) is mathematically complicated but straightforward, and the solution is expressed as follows.

$$|\hat{X}(k,l)| = G_H(k,l)|Y(k,l)|, \tag{3}$$

$$G_H(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{\nu(k,l)}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{4}$$

$$\nu(k,l) = \gamma(k,l)\xi(k,l)/(1+\xi(k,l)), \tag{5}$$

where $Y(k,l)$ is the amplitude of observed signal in the frequency domain, and $G_H(k,l)$ is the gain function. Variables $\xi(k,l)$ and $\gamma(k,l)$ are called *a priori* SNR and *a posteriori* SNR, respectively. The *a posteriori* SNR is defined by the current frame observation;

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\sigma_m^2(k,l)} \tag{6}$$

where the noise process variance $\sigma_m^2(k,l)$ must be obtained separately. As in [20], we use the minima-controlled recursive averaging (MCRA) noise estimation to obtain $\sigma_m^2(k,l)$. The a priori SNR is estimated recursively using the estimated gain of the previous frame;

$$\xi(k,l) = c_1 G_H^2(k,l-1)\gamma(k,l-1) + (1-c_1)\max\{\gamma(k,l)-1, 0\} \tag{7}$$

where c_1 is an adjustable weight parameter. A more detailed derivation of the above-mentioned solution could be found in [23].

OM-LSA modifies the result of LSA by taking the weighted geometric mean of $G_H(k,l)$ and its lower boundary $G_{\min}$, where the weight is based on the speech presence probability $p(k,l)$.

$$|\hat{X}(k,l)| = G(k,l)|Y(k,l)| \tag{8}$$

$$G(k,l) = [G_H(k,l)]^{p(k,l)} G_{\min}^{1-p(k,l)}. \tag{9}$$

The speech presence probability is obtained by

$$p(k,l) = \left[1 + \frac{q_0}{1-q_0}(1+\xi(k,l))\mathrm{e}^{-v(k,l)}\right]^{-1} \tag{10}$$

where q_0 is the a priori speech absence probability.

After the noise suppression process of OM-LSA was done, there are two approaches to obtain VAD results. The first approach is to reconstruct the waveform of the noise-suppressed signal by combining the estimated amplitude $\hat{X}(k,l)$ and the original phase. Once the reconstructed waveform was obtained, any existing VAD method can be applied. The simplest example is to classify speech and non-speech frames using a power threshold, which we discuss in the next section. It is also possible to apply more sophisticated machine learning-based method, which we discuss in Section IV. The second approach is to use the speech presence probability $p(k,l)$ defined by (10) or the likelihood ratio $\Lambda(k,l)$ used in [17]. They can be integrated to the framewise score as follows.

$$P(l) = \prod_{k=0}^{K-1} p(k,l), \tag{11}$$

$$\begin{aligned}\Lambda(l) &= \prod_{k=0}^{K-1} \Lambda(k.l), \\ &= \prod_{k=0}^{K-1} \frac{1}{1+\xi(k,l)} \exp \frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}\end{aligned} \tag{12}$$

where K is the number of frequency components. VAD can be realized simply by applying a threshold for these framewise scores.

B) Augmented implementation of OM-LSA

The original OM-LSA estimator was intended to provide better speech signal for communication and recognition. Therefore, it was designed to maintain the balance between lowering the noise level and not causing the distortion. However, reducing the noise is particularly important for VAD, while the distortion is rarely harmful. In view of the priority of noise removal, we propose to use some augmentation techniques for the OM-LSA speech estimator.

The first step of augmentation focuses on (6). Although MCRA is known to be a reliable noise estimation algorithm, we believe that the noise level must be over-estimated so that any unreliable part of the spectrogram does not affect the VAD results. In fact, we formerly found that the noise level must be under-estimated to obtain more accurate results of speech recognition [24], because distortion is the main cause of misrecognition. Both under and over estimation of the noise can be realized by introducing a weight factor α as follows.

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\alpha\sigma_m^2(k,l)}. \tag{13}$$

The second step of augmentation focuses on (8) that defines how the estimated gain is applied to the speech amplitude. Augmentation is realized simply by introducing a modification parameter β as follows.

$$|\hat{X}(k,l)| = G^{\beta}(k,l)|Y(k,l)|. \tag{14}$$

The third step of augmentation is a screening step of the prominent noise component. In contrast that speech has a wideband spectrum, there are kinds of noises that have a very sharp peak in the power spectrum, such as electrical beeps and musical instrument sounds. The prominent component in the power spectrum could be removed by the process described as follows.

$$|\hat{X}(k,l)| = \begin{cases} G^{\beta}(k,l)|Y(k,l)| & \mathrm{rank}(k) \geq \eta K \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where $\mathrm{rank}(k)$ is the number of frequency components whose magnitude is larger than the kth frequency. At last, our VAD-oriented OM-LSA-based speech estimator is defined by equations (4), (5), (7), (9), (10), (13), and (15).

III. CLASSIFICATION WITHOUT MODEL TRAINING

After the noise suppressed signal is obtained, we extract framewise features and classify each frame into speech or non-speech. The simplest classification algorithm is to apply a threshold to the frame power, which is calculated as follows.

$$Q(l) = \sum_{k=0}^{K-1} w(k)|\hat{X}(k,l)|^2 \tag{16}$$

where $w(k)$ is the A-weighted filter coefficient corresponding to the kth frequency. In the above equation, $\hat{X}(k,l)$ is not necessarily the same as that of (15) because sometimes the optimal time scale is not equal between noise suppression and VAD. In this paper, the frame power for VAD is calculated using 20 ms half-overlapping frames from the reconstructed waveform, whereas the noise suppression was executed using 32 ms half-overlapping frames. In fact, we apply the modification from (14) to (15) not directly during noise suppression. Instead, we apply (15) for the reconstructed waveform.

Once the frame power is calculated, the frame is labeled as speech if the power is larger than the threshold. Although this approach is not effective in low-SNR environments by itself, it works quite effectively when it is combined with noise suppression. It should be also noted that the same

principle is applicable to other framewise features, such as $P(l)$ and $\Lambda(l)$.

A typical modification of the threshold-based VAD is to add inter-frame smoothing. Consonants at the beginning and the end of an utterance is likely to be mislabeled as non-speech due to its small power. Therefore, it is reasonable to add several-frame speech periods (hangovers) on both ends of the period labeled as speech. It is also reasonable to remove (relabel as non-speech) very short speech periods because they are likely to be transient noise, and to fill (relabel as speech) short non-speech periods because they are likely to be voiceless intervals within utterances.

Throughout this paper, we remove speech periods of 100 ms or shorter, fill non-speech periods of 80 ms or shorter, and add 80 ms hangovers on both ends of the speech period.

IV. CLASSIFICATION WITH UNSUPERVISED AND SUPERVISED MODEL TRAINING

If we use multi-dimensional features, more accurate VAD results could be obtained using more sophisticated classifiers. Such classifiers are mostly model-based, so we need to prepare a set of training data.

If we have a set of unlabeled training data, then the training process is called unsupervised. Clustering such as k-means algorithm is an example of unsupervised training. In the case of VAD, two clusters corresponding to the speech class and non-speech class are generated. The training process itself cannot decide which cluster corresponds to the speech, but the non-speech cluster can be easily found by testing a silent frame.

If we have a set of correctly labeled training data, then we can train the classifier model more precisely. It is called supervised training. In this paper, we investigate three model-based supervised classifiers: decision tree (DT), SVM, and CNN.

For all clusterers and classifiers, the same feature set is used. The feature vector for the lth frame consists of $|\hat{X}(k,s)|^2$, where $0 \leq k < K$ and $l-2 \leq s \leq l+2$. The resulting feature vector has $5K$ elements. In the rest of this paper, $K = 40$ is used.

When we use the model-based classifier, we omit the modification by (15), and use (14) instead. It is because the modification from (14) to (15) is learnable by the classifier if it improves the classification accuracy for the training data. Moreover, the machine learning algorithm may find even a better modification.

When we use the framewise clusterers (k-means) and framewise classifiers (DT, SVM, and CNN), inter-frame smoothing is applied as in the case of simple thresholding. Additional 40 ms speech periods were added on both ends of the utterance. Other criteria are the same as in the case of simple thresholding; noises of 100 ms or shorter were removed and silences of 80 ms or shorter were filled.

Table 1. Detail of CENSREC-1-C real dataset.

Number of speakers	5 female + 4 male			
Utterances	Japanese connected digits			
Sampling rate	8 kHz			
Environment	Restaurant		Street	
Noise level (dBA)	Low 53.4	High 69.7	Low 58.4	High 69.2
Estimated SNR (dB)	Low 0.86	High 4.68	Low −3.29	High −2.23
Number of files	36	36	36	36
Number of utterances	345	345	345	345

V. EXPERIMENTAL RESULTS

A) CENSREC-1-C evaluation framework

The proposed VAD algorithm was evaluated using CENSREC-1-C [25], a public VAD evaluation framework. The data part of CENSREC-1-C consists of two datasets: simulated dataset and real dataset. In this paper, we used the real dataset made of Japanese connected digit utterances. The CENSREC-1-C real dataset is further divided into four subsets based on the environment (university restaurant and vicinity of highway street) and SNR level (high and low). Its details are shown in Table 1, where the average noise levels were cited from [25] and the SNR levels were estimated using WADA-SNR [26].

CENSREC-1-C includes not only the data themselves, but also the voice activity labels and performance calculation tools. When a specific algorithm and corresponding settings are given, the false alarm rate (FAR) and the false rejection rate (FRR) are automatically calculated. FAR is the ratio of incorrectly labeled non-speech frames over the total non-speech frames. FRR is the ratio of incorrectly labeled speech frames over the total speech frames. We also use the average error rate (AER), which is the average of FAR and FRR.

B) Evaluation using classifiers without model training

The first set of experiments using CENSREC-1-C was conducted to compare three framewise scores, $P(l)$, $\Lambda(l)$, and $Q(l)$. The threshold-based classifier was applied to those three scores obtained by statistical noise suppression without augmentation ($\alpha = 1.0, \beta = 1.0, \eta = 0.0$). Other parameter setting was the same as in [21] ($C_1 = 0.99, G_{min} = 0.01, q_0 = 0.2$). Figure 1 shows the ROC curves obtained by applying various threshold values. Since CENSREC-1-C provides the baseline VAD tool, the corresponding ROC curve was plotted for comparison. It is clear that all three scores resulted in lower FARs and FRRs than the baseline. Among them, $P(l)$ is clearly less effective, and $Q(l)$ is slightly better than $\Lambda(l)$. Accordingly, $Q(l)$ is used as the framewise score in this section.

The second set of experiments was to confirm the effectiveness of the augmented statistical noise suppression. The same threshold-based classifier was applied to the output of statistical noise suppression, in which various augmentation was applied.

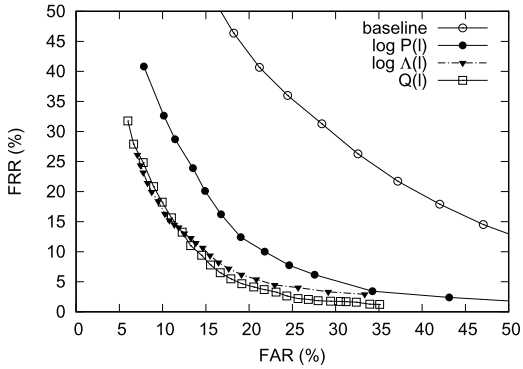

Fig. 1. Comparison of framewise scores. The same threshold-based classifier was applied.

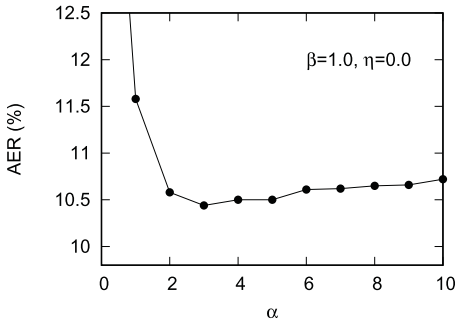

Fig. 2. Effect of over-subtraction expressed by various α.

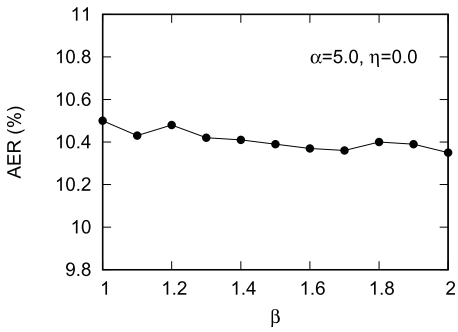

Fig. 3. Effect of gain augmentation expressed by various β.

Figure 2 shows the effectiveness of noise over-estimation, defined by (13). The values of β and η were not augmented, and the value of α was changed. The point $\alpha = 1.0$ corresponds to the original OM-LSA, and it was reported that $\alpha \approx 0.2$ is optimal for speech recognition [24]. For each value of α, various thresholds were tested, AER = (FAR + FRR)/2 was calculated, and the smallest AER was plotted. The results revealed that decreasing the value of α is only harmful for VAD, and AER drops rapidly when α increases from 1.0 to 2.0. When α is further increased, AER seems to be saturated.

Figure 3 shows the effectiveness of gain nonlinearity, defined by (14). The value of α was fixed at 5.0, and various values of β were tried. In this case, some improvements were observed with increasing β, but it is not as dramatic as in the case of increasing α.

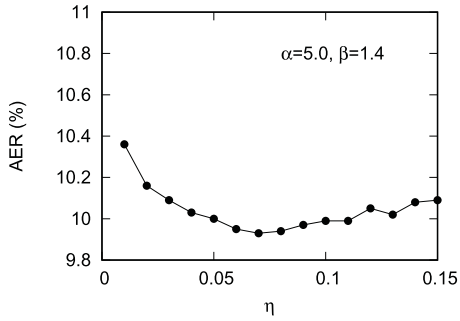

Fig. 4. Effect of prominent component removal expressed by various η.

The results obtained by various values of η are shown in Fig. 4, where β was fixed at 1.4. Again, AER drops rapidly with increasing η, and reached its smallest value at $\eta = 0.07$. The best AER was 9.93%, where FAR = 12.42% and FRR = 7.43%.

We also investigated the possibility of using adaptive parameters. Obviously, the most effective parameter is the frame power threshold. Since the dataset of CENSREC-1-C is made of four subsets, we checked the AER of our best setting ($\alpha = 5.0, \beta = 1.4, \eta = 0.07$) obtained with the optimal threshold for each subset. In this case, the AERs for Restaurant (SNR high), Restaurant (SNR low), Street (SNR high), and Street (SNR low) were 7.00, 18.03, 2.73, and 3.94% respectively. The average was 7.93%, which is much better than the AER obtained with the fixed threshold (9.93%).

Similar experiments were conducted with the adaptive α, β, and η. In the case of α (see Fig. 2), the lowest AER of 10.44% was still improved to 9.27%. In the case of β (see Fig. 3), the lowest AER of 10.35% was still improved to 9.53%. Finally, in the case of η (see Fig. 4), the lowest AER of 9.93% was improved to 9.37%.

Although these results indicate the potential value of adaptive parameters, their effectiveness are strongly dependent on the correct evaluation of the environment. We must consider the risk of performance degradation caused by the incorrect parameter setting.

C) Classifier training

More accurate VAD results could be achieved by the model-based classifiers. In this paper, we investigate three types of model-based classifiers: DT, SVM, and CNN. We also try k-means clustering to compare supervised and unsupervised training. Before applying them to CENSREC-1-C, we train the classifiers and clusterer using a separate dataset.

1) Training and development data

The dataset for classifier and clusterer training, referred to as Noisy UT-ML-JPN database in this paper, was created using UT-ML public database [27] and our proprietary noise data. The noise data were recorded in a running car and in a cafeteria; they were added to the utterances of Japanese subset of UT-ML database with SNR of 0, 5, and 10 dB. Speech/non-speech labels for Noisy UT-ML-JPN database were generated automatically by applying a

Table 2. Length and number of frames of Noisy UT-ML-JPN database.

	Length (min)	Speech frames	Non-speech frames
Training	228.4	616,566	738,168
Development	37.7	86,958	136,308

power threshold for the clean version of UT-ML database. The labeling process is completely framewise, meaning that inter-frame smoothing was not applied.

The Japanese subset of UT-ML database includes six male and six female speakers, each of which read one long article (54.8 s on average) and 50 short phrases (3.3 s on average). Original data were recorded by 16 kHz sampling rate, but they were downsampled to 8 kHz. One second silences were appended on both ends of utterances before noises were added.

After adding the noise, augmented statistical noise suppression described in Section II-B) was applied. Based on the results of Section V-B) and [21][1], the parameter setting was $\alpha = 5.0, \beta = 1.4$, and $\eta = 0.07$. Applying the same noise suppression process to the training and test data are called noise adaptive training (NAT), and known to be effective for speech recognition [28]. We also prepare the clean version of the training data for comparison.

Finally, data of one male and one female speakers were used for development, and the other data were used for model training. The size of the training and development set are shown in Table 2.

2) Toolkits

The classifiers were trained using publicly available toolkits. WEKA [29] was used for DT and SVM training, and Caffe [30] was used for CNN training. The *k*-means clustering program was prepared by ourselves.

When we use WEKA for DT and SVM, we adopt the classifier ensemble approach. Although we have plenty of training samples (more than 600 k of speech and more than 700 k of non-speech), it takes too long to execute the training using the whole data at once. Therefore, we split the training data into small chunks, and trained many classifiers using different data chunks. The final classification result can be obtained by voting of these small classifiers. CNN training and *k*-means clustering were executed using all data.

DT by WEKA (J48) runs without any specific parameter adjustment. SVM by WEKA (SMO) requires at least selection of the kernel, so we used the two-dimensional polynominal kernel. CNN by Caffe requires quite a few parameters to be set, but we only show the network topology in Fig. 5, that would be the most important information.

[1]Although the same method was applied to the same dataset by the same author, the results of Section V-B) and [21] are not exactly the same. It is because the program used in [21] belongs to the organization to which the author has no access at present.

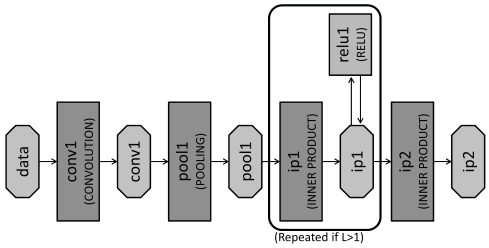

Fig. 5. CNN topology. Relu stands for rectified linear unit. The output layer (ip2) has two units corresponding to speech and non-speech.

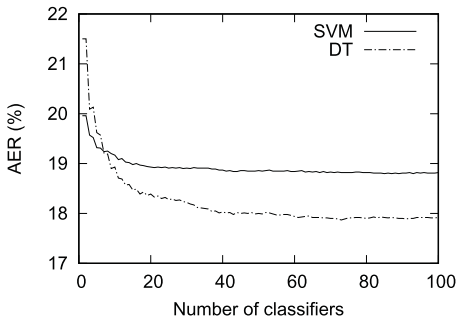

Fig. 6. Preliminary evaluation of classifier ensemble.

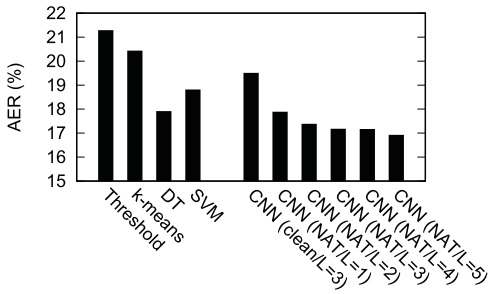

Fig. 7. Preliminary evaluation by comparing various classifiers.

3) Preliminary evaluation

Before going back to CENSREC-1-C, we checked the basic performance of three classifiers and one clusterer using the development data. The development data were processed by augmented statistical noise suppression using the same parameters as the training data. The processed data were then divided into frames and classified by either DT, SVM, CNN or *k*-means.

Figure 6 depicts the results obtained by the classifier ensemble of DT and SVM with various sizes. Inter-frame smoothing was not applied, so the graph shows the pure performance of framewise classification. It should also be noted that the classifier directly classifies the frame data, so we obtain a single pair of FAR and FRR instead of the ROC curve. The results indicated that SVM is superior to DT if only few classifiers were used. However, it was reversed if we have seven or more classifiers. AER was saturated at about 20 classifiers (SVM) or 40 classifiers (DT).

Figure 7 shows the comparison of simple thresholding, k-means clustering, and various classifiers. If we apply simple thresholding, AER was 21.29%. In the case of k-means clustering, an additional bias θ was introduced to play the role

of adjustable threshold. A frame is labeled as non-speech if $d_{ns} - d_s < \theta$, where d_s (d_{ns}) is the distance between the frame and the centroid of speech (non-speech) cluster. Although AER was as high as 27.19% when $\theta = 0$, it becomes 20.44% if the value of θ was optimized. The AERs of DT and SVM were clearly better than simple thresholding and k-means. On the right-hand side of the figure, AERs obtained by various CNNs are plotted, where L is the number of fully-connected intermediate layers (rectangular block including ip1+relu1 of Fig. 5). As we expected, CNN (NAT) provides better results than CNN (clean), indicating that NAT was effective. It is also found that the deeper the network is, the more accurate classification results were obtained.

D) Evaluation using model-based classifiers

Finally, we applied various classifiers to CENSREC-1-C. The evaluation procedure was the same as in the preliminary evaluation, except that inter-frame smoothing was applied. In addition, a gain adjustment factor was multiplied to to input signal before applying DT, SVM, and CNN, in order to compensate the unmatched data acquisition environment. The gain factor plays the role similar to the adjustable threshold; a large gain factor is equivalent to a small threshold, and vice versa. Therefore, we can obtain an ROC curve by using various gain factors. The bias of k-means clustering also serves to make an ROC curve.

Figure 8 shows the ROC curves obtained by k-menas, DT, SVM, and CNNs with various number of intermediate layers. The results of threshold-based classifier, which is exactly the same as $\eta = 0.07$ of Fig. 4, was added for comparison. It is clear that the model-based classifiers achieved better results than the threshold-based classifier. However, unsupervised training by k-means clustering did not improve the accuracy. As for CNNs, unlike the preliminary evaluation, the shallowest network achieved the most accurate VAD results. The different tendency regarding to the number of layers can be attributed to the overfitting effect. In the preliminary evaluation, the deeper CNN can learn even small details of the training data which can contribute only under the matched condition. In the CENSREC-1-C evaluation, the shallower CNN has a more generalized model, which contributes to avoid errors related

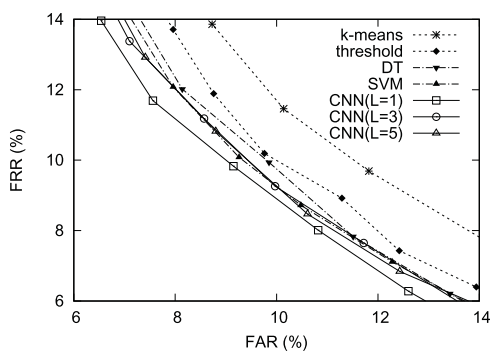

Fig. 8. ROC curves for CENSREC-1-C obtained by various classifiers. DT and SVM represent the voting results of 100 classifiers.

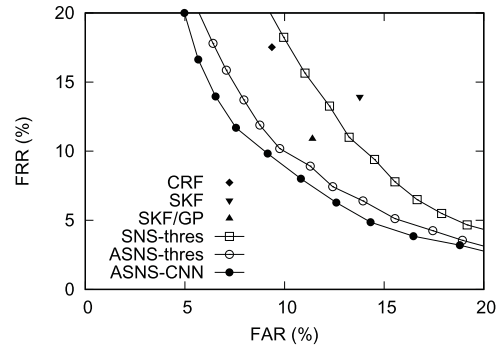

Fig. 9. ROC curves for CENSREC-1-C obtained by the proposed algorithms. Reference points of CRF, SKF, and SKF/GP were cited from the published papers.

Real Data	False Rejection Rate [%]			
		Remote Microphone		
False Rejection Rate [%]		Restaurant	Street	Average
	High SNR	2.83	5.16	4.00
	Low SNR	13.81	17.54	15.68
	Average	8.32	11.35	9.84

Real Data	False Acceptance Rate [%]			
		Remote Microphone		
False Acceptance Rate [%]		Restaurant	Street	Average
	High SNR	10.54	1.69	6.12
	Low SNR	23.13	1.25	12.19
	Average	16.84	1.47	9.15

Fig. 10. Detail of CENSREC-1-C evaluation.

to the condition-specific features. That argument is supported by the fact that the superiority of DT over SVM is not observed in the CENSREC-1-C evaluation, because DT is more likely to overfit the training data even though it includes some pruning algorithms.

So far, we have found that the combination of augmented statistical noise suppression and CNN-based classifier achieved the best VAD accuracy among considered algorithms. In Fig. 9, it is demonstrated how the ROC curve shifted toward the left-bottom corner. **SNS-thres**, which is the copy of $Q(l)$ of Fig. 1, is our starting point, and it was greatly improved by augmented statistical noise suppression, resulted in **ASNS-thres**. There was some additional improvement by introducing CNN, and the final ROC curve was plotted as **ASNS-CNN**. We also plotted some published results using CENSREC-1-C, such as using CRF [16], SKF [19], and SKF with Gaussian pruning (SKF/GP) [31], and confirmed that the proposed algorithm outperformed them.

Following the convention of CENEREC-1-C, the details of the VAD results obtained by **ASNS-CNN** are show in Fig. 10.

E) Computational complexity

In addition to the VAD accuracy, we confirmed the processing time and latency of the proposed algorithm. Table 3

Table 3. Processing Time for CENSREC-1-C.

	Time (s)	RTF
Noise suppression	236.33	0.037
Classification by CNN	93.17	0.014
Total	329.80	0.051

shows the processing time for the 144 files of CENSREC-1-C dataset (6465.33 s). The noise suppression and classification programs were implemented separately on Ubuntu 16.04 running on Intel Core i7 processor (3.6 GHz) with 16 GB RAM. Although the programs were not optimized in terms of speed, the real-time factor (RTF) was 0.051, meaning that the total processing time was about 1/20 of the real time.

Regardless of the processing speed, the proposed algorithm cannot avoid the framing latency. Noise suppression causes a half-frame (16 ms) latency and VAD causes five half-overlapping frame latency (60 ms). The total latency up to 84 ms (including 8 ms frame boundary adjustment) is not negligible, but the speech recognition system uses similar number of successive frames as the input feature. Therefore, the latency does not matter if the proposed VAD algorithm is used for speech recognition.

VI. CONCLUSIONS

In this paper, we analyzed state-of-the-art VAD algorithms from the perspective of statistical noise suppression and framewise speech/non-speech classification. Based on the analysis, we proposed two modification approaches. First, statistical noise suppression was augmented using additional parameters so that all the unreliable spectral components were removed. Second, CNN-based classifier was introduced to improve the accuracy of the framewise classifier. Evaluation experiments using the CENSREC-1-C public framework demonstrated that each of the modifications achieved noticeable improvements in VAD accuracy. The results of the proposed algorithm were also compared with results reported in the literature, indicating that the proposed algorithm is better than SKF/GP, which was shown in [31] to be superior to the standard VAD algorithms such as ITU-T Recommendation G.729 Annex B [32], ETSI Advanced Front-End [33], and Sohn's algorithm [17].

A comparison between matched and unmatched conditions indicated that it is important to avoid overfitting problem under unmatched conditions; hence a relatively shallow CNN is appropriate in general. The effect of the overfitting problem would also be related to the size and variety of the training data; however, this remains an open problem for future study.

REFERENCES

[1] Rabiner, L.R.; Sambur, M.R.: An algorithm for determining the endpoints of isolated utterances. *Bell Syst. Tech. J.*, **54** (2) (1975), 297–315.

[2] Bou-Ghazale, S.E.; Assaleh, K.: A robust endpoint detection of speech for noisy environments with application to automatic speech recognition, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, USA, 2002, IV-3808–IV-3811.

[3] Martin, A.; Charlet, D.; Mauuary, L.: Robust speech/non-speech detection using LDA applied to MFCC, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, USA, 2001, 237–240.

[4] Kinnunen, T.; Chemenko, E.; Tuononen, M.; Fränti, P.; Li, H.: Voice activity detection using MFCC features and support vector machine, in *Int. Conf. on Speech and Computer*, Moscow, Russia, 2007, 556–561.

[5] Shen, J.-L.; Hung, J.-W.; Lee, L.-S.: Robust entropy-based endpoint detection for speech recognition in noisy environments, in *Int. Conf. on Spoken Language Processing*, Sydney, Australia, 1998, 232–235.

[6] Ramirez, J.; Yelamos, P.; Gorriz, J.M.; Segura, J.C.: SVM-based speech endpoint detection using contextual speech features. *Electron. Lett.*, **42** (7) (2006), 426–428.

[7] Ishizuka, K.; Nakatani, T.: Study of noise robust voice activity detection based on periodic component to aperiodic component ratio, in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Pittsburgh, PA, USA, 2006, 65–70.

[8] Cournapeau, D.; Kawahara, T.: Evaluation of real-time voice activity detection based on high order statistics, in *Interspeech*, Antwerp, Belgium, 2007, 2945–2948.

[9] Lee, A.; Nakamura, K.; Nisimura, R.; Saruwatari, H.; Shikano, K.: Noise robust real world spoken dialog system using GMM based rejection of unintended inputs, in *Interspeech*, Jeju Island, Korea, 2004, 173–176.

[10] Zhang, X.-L.; Wu, J.: Deep belief networks based voice activity detection. *IEEE Trans. Audio Speech Lang. Process.*, **21** (4) (2013), 697–710.

[11] Hughes, T.; Mierle, K.: Recurrent neural networks for voice activity detection, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, 7378–7382.

[12] Ryant, N.; Liberman, M.; Yuan, J.: Speech activity detection on YouTube using deep neural networks, in *Interspeech*, Lyon, France, 2013, 728–731.

[13] Fujita, Y.; Iso, K.: Robust DNN-based VAD augmented with phone entropy based rejection of background speech, in *Interspeech*, San Francisco, CA, USA, 2016, 3663–3667.

[14] Ramirez, J.; Segura, J.C.; Benitez, C.; de la Torre, A.; Rubui, A.: An effective subband OSF-based VAD with noise reduction for robust speech recognition. *IEEE Trans. Speech Audio Process.*, **13** (6) (2005), 1119–1129.

[15] Kingsbury, B.; Jain, P.; Adami, A.G.: A hybrid HMM/traps model for robust voice activity detection, in *Interspeech*, Denver, CO, USA, 2002, 1073–1076.

[16] Saito, A.; Nankaku, Y.; Lee, A.; Tokuda, K.: Voice activity detection based on conditional random field using multiple features, in *Interspeech*, Makuhari, Japan, 2010, 2086–2089.

[17] Sohn, J; Kim, N.S.; Sung, W.: A statistical model-based voice activity detection. *IEEE Signal Process. Lett.*, (6) (1999), 1–3.

[18] Ephraim, Y.; Malah, D.: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, **32** (6) (1984), 1109–1121.

[19] Fujimoto, M.; Ishizuka, K.: Noise robust voice activity detection based on switching Kalman filter. *IEICE Trans. Inf. Syst.*, **E91-D** (3) (2008), 467–477.

[20] Cohen, I.; Berdugo, B.: Speech enhancement for non-stationary noise environments. *Signal Process.*, **81** (2001), 2403–2418.

[21] Obuchi, Y.; Takeda, R.; Kanda, N.: Voice activity detection based on augmented statistical noise suppression, in *APSIPA Annu. Summit and Conf.*, Holywood, CA, USA, 2012, 1–4.

[22] Obuchi, Y.: Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression, in *IEEE International Conference on Acoust. Speech Signal Process.*, Shanghai, China, 2016, 5715–5719.

[23] Ephraim, Y.; Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-33** (2), (1985), 443–445.

[24] Obuchi, Y.; Takeda, R.; Togami, M.: Bidirectional OM-LSA speech estimator for noise robust speech recognition, in *IEEE Automatic Speech Recognition and Understanding Workshop*, Big Island, HI, USA, 2011, 173–178.

[25] Kitaoka, N. *et al.*: CENSREC-1-C: an evaluation framework for voice activity detection under noisy environments. *Acoust. Sci. Technol.*, **30** (5) (2009), 363–371.

[26] Kim, C.; Stern, R.M.: Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis, in *Interspeesh*, Brisbane, Australia, 2008, 2598–2601.

[27] Speech Resources Consortium (NII-SRC): University of Tsukuba Multilingual Speech Corpus (UT-ML). http://research.nii.ac.jp/src/en/UT-ML.html.

[28] Deng, L.; Acero, A.; Plumpe, M.; Huang, X.: Large-vocabulary speech recognition under adverse acoustic environments, in *Int. Conf. on Spoken Language Processing*, Beijing, China, 2000, 806–809.

[29] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H.: The WEKA data mining software: an update. *SIGKDD Explorations*, **11** (1) (2009), 10–18.

[30] Jia, Y. *et al.*: Caffe: convolutional architecture for fast feature embedding, arXiv preprint (2014), arXiv:1408.5093.

[31] Fujimoto, M.; Watanabe, S.; Nakatani, T.: Voice activity detection using frame-wise model re-estimation method based on Gaussian pruning with weight normalization, in *Interspeech*, Makuhari, Japan, 2010, 3102–3105.

[32] 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.

[33] ETSI ES 202 050 v1.1.5, Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm, 2007.

Yasunari Obuchi received B. S. and M. S. degrees in Physics from the University of Tokyo in 1988 and 1990, respectively. He received Ph.D. in Information Science and technology from the University of Tokyo in 2006. From 1992 to 2015, he had been with Central Research Laboratory and Advanced Research Laboratory, Hitachi, Ltd. From 2002 to 2003, he was a visiting scholar at Language Technologies Institute, Carnegie Mellon University. He was a visiting researcher at Information Technology Research Organization, Waseda University from 2005 to 2010. He also worked at Clarion Co., Ltd. from 2013 to 2015. Since 2015, he has been a Professor at School of Media Science, Tokyo University of Technology. He was a co-recipient of the Technology Development Award of the Acoustical Society of Japan in 2000. He is a member of IEEE, Information Processing Society of Japan, Institute of Electronics, Information and Communication Engineers, Acoustical Society of Japan, and Society for Art and Science.