FRAMEWISE SPEECH-NONSPEECH CLASSIFICATION BY NEURAL NETWORKS FOR VOICE ACTIVITY DETECTION WITH STATISTICAL NOISE SUPPRESSION

Yasunari Obuchi

School of Media Science, Tokyo University of Technology,
1404-1 Katakura, Hachioji, Tokyo 192-0982 Japan
obuchiysnr@stf.teu.ac.jp

ABSTRACT

A new voice activity detection (VAD) algorithm is proposed. The proposed algorithm is the combination of augmented statistical noise suppression (ASNS) and convolutional neural network (CNN). Since the combination of ASNS and simple power thresholding was known to be very powerful for VAD under noisy conditions, even more accurate VAD is expected by replacing the power thresholding with the more elaborate classifier. Among various model-based classifiers, CNN with noise adaptive training presented the highest accuracy, and the improvement was confirmed by the experiments using CENSREC-1-C public database.

Index Terms— Voice activity detection, noise suppression, neural network, noise adaptive training, CENSREC-1-C

1. INTRODUCTION

Even with today's technology that is capable of recognizing complex spoken sentences, simple two-class classification between speech and non-speech is still very difficult in noisy environments. The task is called voice activity detection (VAD), and has been studied for many years with intent to apply for telecommunication and automatic speech recognition.

Since most VAD systems work as a classifier for short-term periods (frames) followed by inter-frame smoothing, two major approaches have been studied thoroughly to improve the VAD accuracy. The first approach tried to find better framewise features. The proposed features include instantaneous power, zero-crossing rate [1], cepstral features [2], spectral entropy [3], periodic-aperiodic component ratio [4], and higher order statistics [5]. The second approach focused on the classifier. Gaussian mixture model (GMM) [6] and Support vector machine (SVM) [7] are typical classifiers, and then many types of deep neural networks (DNNs) are being proposed these days [8, 9, 10].

Combination of feature extraction and classification is a simple and effective VAD framework. However, if the stationary noise is the main cause of degraded VAD accuracy, it is necessary to deal with inter-frame relationship of noise signals. From this viewpoint, Sohn et al. [11] proposed to use the decision-directed estimation [12] of the noise estimation parameter to integrate inter-frame information into their likelihood ratio test (LRT) based classifier. Using hidden Markov model (HMM) [13], conditional random field (CRF) [14], and order statistics filter (OSF) [15] are similar approaches to consider inter-frame dependencies. Recently, Fujimoto [16] introduced another temporal model of switching Kalman filter (SKF), and showed significant improvement from Sohn's method. In the case of DNN, a recurrent neural network could be an example of explicit temporal processing, as in [9].

Although there have been many ways to combine framewise classification and inter-frame estimation, the author recently showed that even a simple framewise classifier of power thresholding could outperform other VAD algorithms, if it is combined with the augmented implementation of the state-of-the-art noise suppression algorithm [17]. In other words, it is not optimal to design the best classifier and then introduce temporal processing scheme within the constraint of the classifier. It is optimal to design the best temporal processing scheme (noise suppression) and then find the best classifier for the noise suppressed signal. Based on this idea, this paper tries to contribute to the classifier finding part, which was not investigated enough in our previous paper [17]. The proposed algorithm is a sequence of augmented statistical noise suppression (ASNS), framewise speech/non-speech classification using convolutional neural networks (CNNs), and inter-frame smoothing. By enforcing the classification capability of the second part, whereas keeping the precise noise suppression capability of the first part, the proposed algorithm shows still more accurate VAD results than others, which is confirmed by the experiments using a public evaluation framework CENSREC-1-C [18].

The remainder of this paper is organized as follows. In the next section, augmented implementation of the noise suppression algorithm is described. In Section 3, CNN and other framewise classifiers are introduced, together with a brief description of inter-frame smoothing. In Section 4, we discuss more about noise adaptive training of classifiers. Experimental results are presented in Section 5, and the last section is for conclusions.

2. NOISE SUPPRESSION

In the proposed VAD framework, the input signal is first fed into the noise suppression module. The noise suppression (ASNS) algorithm used in this paper is based on the augmented implementation of optimally modified log spectral amplitude (OM-LSA) speech estimator [19]. More precisely, ASNS is defined as the gain function in the time-frequency domain as follows.

$$|\hat{X}(k,l)|^2 = G(k,l)^{\beta}|Y(k,l)|^2 \tag{1}$$

where $Y(k,l)$ denotes the k-th frequency component of the observed noisy signal at the l-th frame. β is the gain augmentation parameter. $G(k,l)$ is the gain to be estimated, and $\hat{X}(k,l)$ is the corresponding noise-suppressed signal. Before obtaining $G(k,l)$, we first calculate the LSA gain $G_H(k,l)$ by the following equations using the subtraction coefficient α.

$$G_H(k,l) = f(\xi(k,l),\gamma(k,l)) \tag{2}$$

$$f(\xi,\gamma) = \frac{\xi}{1+\xi}\exp(\frac{1}{2}\int_{\gamma\xi/(1+\xi)}^{\infty}\frac{e^{-t}}{t}dt) \tag{3}$$

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\alpha\sigma_m^2(k,l)} \tag{4}$$

$$\xi(k,l) = 0.99\,G_H^2(k,l-1)\gamma(k,l-1) + 0.01\max\{\gamma(k,l)-1,0\} \tag{5}$$

The OM-LSA gain $G(k,l)$ is obtained by modifying LSA gain $G_H(k,l)$.

$$G(k,l) = [G_H(k,l)]^{p(k,l)}G_{min}^{1-p(k,l)} \tag{6}$$

$$p(k,l) = [1+0.25(1+\xi(k,l))e^{-\nu(k,l)}]^{-1} \tag{7}$$

$$\nu(k,l) = \gamma(k,l)\xi(k,l)/(1+\xi(k,l)) \tag{8}$$

where G_{min} was set as 0.01.

In the case of speech recognition, the optimal value of the subtraction coefficient α is less than 1.0 because the larger α we use, the more distorted signal we get [20]. However, such distortion is less harmful for VAD, and a larger value of α makes the VAD system more noise-tolerant.

Once the amplitude $|\hat{X}(k,l)|$ is estimated for each half-overlapping frame, the phase of the observed signal is combined to reproduce the complex spectrum. Finally, inverse fast Fourier transform and overlap addition are applied to make the noise suppressed signal waveform.

3. FRAMEWISE SPEECH/NON-SPEECH CLASSIFICATION AND SMOOTHING

After noise suppression, the signal is divided into frames again, but this time the frame rate and frame width are optimized for the VAD purpose. Waveform of each frame is converted to the spectrum (or spectrum-based feature vector), and fed into the speech/non-speech classifier.

In this paper, four different classifiers are investigated. The simplest classifier is power thresholding, defined by

$$\hat{H}(l) = \begin{cases} H_0 & (p(l) < \theta) \\ H_1 & (p(l) \geq \theta) \end{cases} \tag{9}$$

where $\hat{H}(l)$ is the hypothesis for the l-th frame, which takes either H_0 (non-speech) or H_1 (speech), $p(l)$ is the frame power, and θ is the threshold. In [17], we proposed to use the augmented frame power calculation,

$$p(l) = \sum_{k=0}^{K} w(k)|\hat{X}(k,l)|^2 \tag{10}$$

$$w(k) = \begin{cases} w_A(k) & \text{rank}(k) \geq \eta K \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where rank(k) is the number of frequency components in the same frame whose magnitude is larger than the k-th component. K is the total number of frequency components, η is another augmentation parameter, and $w_A(k)$ is the A-weighting filter coefficient.

The other three classifiers are decision tree (DT), SVM, and CNN. For those three classifiers, the spectrum of each frame is converted to 40 log mel filterbank energies. Five successive frames are concatenated, and the resulting 200 dimension feature vector represents the central frame. Each frame of the training data has the correct label $H(l)$, so that two-class classifiers of DT, SVN, and CNN can be trained.

Since speech segments often include short silent (or very soft) periods, results of the framewise classification are modified by inter-frame smoothing. First, speech segments shorter than two frames are eliminated, and then non-speech segments shorter than 80 frames are re-labeled as speech. Finally, eight (power thresholding) or four (DT, SVM, and DNN) margin frames on the both sides of speech segments are re-labeled as speech. The shorter margin was used for DT, SVM, and DNN because the feature vector itself includes adjacent frames.

4. NOISE ADAPTIVE TRAINING OF CLASSIFIERS

While the speech/non-speech classifier can be trained using a clean speech corpus, it becomes more robust under noisy conditions if it is trained using a noisy speech corpus and the noise suppression module. Such strategy is called noise adaptive training (NAT), and proved to be effective in the case of speech recognition [21].

Although it is difficult to predict the noise type and signal-to-noise ratio (SNR) beforehand, we can use some typical noises and SNRs. In this paper, in-car noise and cafeteria noise are added to the clean training corpus with 0dB, 5dB, and 10dB SNR. Prior distribution for those noise conditions is flat, meaning that all the noisy data are simply piled up, cleaned by noise suppression, and used for classifier training.

Table 1. Numbers of frames in Noisy UT-ML-JPN database

	speech	non-speech
Training	616,566	738,168
Test	86,958	136,308

5. EXPERIMENTAL RESULTS

5.1. Experimental Setup

The proposed VAD framework was evaluated using two sets of databases: original synthesized database and CENSREC-1-C [18]. The original database, which is referred to as "Noisy UT-ML-JPN" in this paper, was made from Japanese subset of UT-ML database [22], by concatenating one second silence at the both sides of each utterance, and then adding our proprietary noise data (in-car and cafeteria, 0/5/10dB). Speech/non-speech labels were generated automatically by applying a power threshold for clean data. Inter-frame smoothing was not applied for this database in order to focus on the framewise classification only. The Japanese subset of UT-ML includes six male and six female speakers. Each speaker read one long article (54.8sec on average) and 50 short phrases (3.3sec on average). Five male and five female speakers were used for training, and one male and one female were used for test. All data were downsampled to 8kHz, and framed using 32ms (noise suppression) or 20ms (VAD) half-overlapping window. The numbers of speech and non-speech frames in the training and test sets are shown in Table 1.

CENSREC-1-C is a public evaluation platform for VAD. Only the **real** subset is used in this paper. It includes concatenated Japanese digit utterances recorded in a crowded university restaurant and in the vicinity of a highway mainline. Total 160 utterances (sampled by 8kHz) by five male and five female speakers are provided with the hand-labeled endpoint information.

Classifiers were evaluated with the help of publicly available tools. WEKA [23] was used for DT and SVM, and Caffe [24] was used for CNN.

5.2. Classifier training

As described in Section 4, noise adaptive data were used to train DT, SVM, and CNN classifiers. ASNS ($\alpha = 5.0$, $\beta = 1.4$) was applied to the training set of Noisy UT-ML-JPN database, and the noise-suppressed signals were used to train classifiers. The clean UT-ML-JPN training data were also reserved for comparison.

Although we have plenty of training data, the training process of WEKA is relatively slow for 1.4 million training samples. Therefore, we adopted "ensemble of classifiers" approach. The training data were divided into 100 subsets, each of which was used to train a separate classifier. In the evaluation phase, each frame of the test data is classified by the ensemble of the classifiers, and the obtained labels are used for voting.

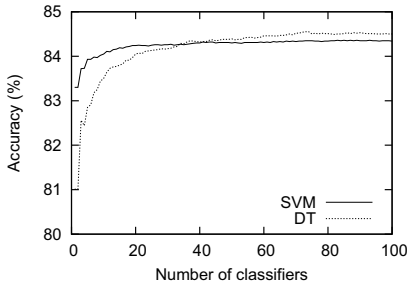

Fig. 1. Voting by ensemble of classifiers.

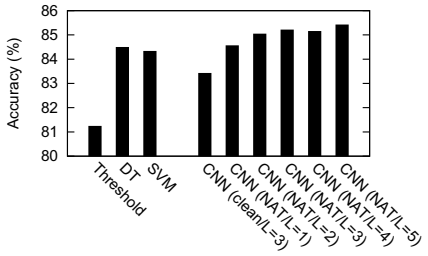

Fig. 2. Classification accuracy for matched data.

5.3. Evaluation of classifiers using matched data

The first set of evaluation experiments were conducted using Noisy UT-ML-JPN database. Since the training and test data contain the same type of noises, it should be regarded as the matched condition evaluation. All test utterances were processed by ASNS in the same way as the training data, divided into frames, and fed into the framewise classifier.

Figure 1 shows the classification accuracy of ensemble voting by DT and SVM. The horizontal axis represents the number of classifiers. The vertical axis represents the accuracy, which is the ratio of correctly classified frames to the total frames. Although a DT classifier tends to suffer from overfitting, it could be compensated by ensemble voting and the accuracy rises when multiple classifiers are used. A similar trend can be observed for SVM, but the overfitting problem is not as serious as DT. In both cases, the accuracy almost saturates by 30 or 40 classifiers.

Figure 2 shows the classification accuracy for the test set of Noisy UT-ML-JPN, obtained by various classifiers. In the case of power thresholding, the value of η was fixed at 0.07. The plotted accuracy is the best value among those obtained with various threshold values. For DT and SVM, the accura-

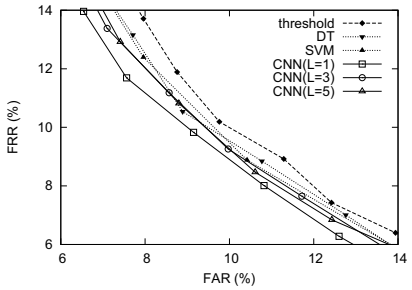

Fig. 3. Comparison of classifiers built in the ASNS-based VAD system. FAR and FRR were calculated using the script distributed with the CENSREC-1-C package.

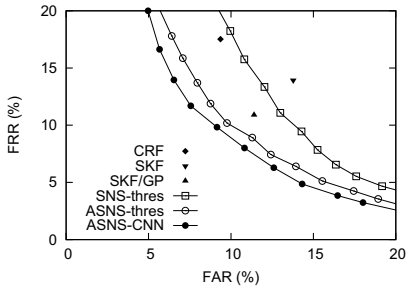

Fig. 4. Comparison of total VAD algorithms using CENSREC-1-C. The values of CRF [14], SKF [16], and SKF/GP [25] were obtained from the literature. SNS-thres corresponds to $\alpha = 1.0$, $\beta = 1.0$, and $\eta = 0.0$.

cies of 100 classifier voting were plotted. From these results, it was confirmed that a large ($> 3\%$ absolute) improvement was realized by DT and SVM.

Several CNNs with different depth were also evaluated. After the convolution (3x3, 20 filters) and pooling (2x2) layers, single or multiple fully-connected layers (100 nodes each) are prepared, and the final layer has two output nodes corresponding to the speech and non-speech hypotheses. In Fig. 2, L represents the number of fully-connected layers. Another CNN with $L = 3$ was trained using the clean data, and the corresponding accuracy was also plotted. From these results, the effectiveness of NAT was confirmed, and it was observed that the deeper network achieves the higher VAD accuracy, although even the shallowest CNN outperformed DT and SVM slightly.

5.4. Evaluation of total VAD system using real data

Next, various classifiers were built into the total VAD system, and evaluated using CENSREC-1-C. The noise condition is open for these experiments. However, the training and test data both include babble noises and traffic noises (although recorded in car and outside of car), so the condition is favorable to the model-based classifiers.

Since the audio gain was not calibrated between the training and test data, at least one adjustable parameter is necessary to keep the reasonable accuracy. In the case of power thresholding classifier, the threshold plays that role. In the case of model-based classifiers, an adjustable gain is applied to the input signal instead. As the results, receiver operating characteristic (ROC) curves are obtained by plotting false acceptance rates (FARs) and false rejection rates (FRRs) corresponding to those threshold or gain parameters. Figure 3 shows the ROC curves of power thresholding, DT, SVM, and three types of CNNs. Considering the saturation trend found in Fig. 2, DT and SVM were evaluated using 40 classifier voting to save the execution time.

Although Fig. 3 presents the similar tendency as Fig. 2, a noticeable difference is that a deeper CNN no longer achieves the better results. In fact, the ROC curve nearest to the lower-left corner was obtained with the shallowest CNN ($L = 1$). It indicates that the deeper CNN has learned the condition-specific nature of the Noisy UT-ML-JPN database, and it did not help to improve the accuracy of the shallower CNN under different conditions.

To confirm the improvement by introducing CNN as the framewise classifier, the ROC curves of various VAD algorithms were plotted in Fig. 4. It was already mentioned in [17] that the VAD accuracy can be greatly improved by introducing ASNS. That finding was re-confirmed in these experiments as the difference between SNS-thres and ASNS-thres. Moreover, it can be observed that the combination of ASNS and CNN achieves additional improvement[1], and the resulted FAR (9.15%) and FRR (9.83%) are the lowest among the published works for CENSREC-1-C.

6. CONCLUSIONS

In this paper, a new framewise speech/non-speech classifier was introduced to the augmented statistical noise suppression-based VAD system. The new classifier is realized as a convolutional neural network, and outperforms simple power thresholding as well as DT and SVM classifiers. A deeper CNN achieved higher accuracy under the matched condition, but the shallower CNN was slightly more accurate under the unmatched condition. Evaluation experiments using CENSREC-1-C public database showed that the proposed system provides higher VAD accuracy than other state-of-the-art algorithms.

[1]The reader may notice that the ROC curve of ASNS-thres in Fig. 4 is not exactly the same as that of [17]. It is because the author has changed the affiliation and wrote the ASNS program from the scratch, which is not identical to the older version. However, the improvement from ASNS-thres to ASNS-CNN using the same ASNS program was clearly confirmed as in Fig. 4.

7. REFERENCES

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[2] S. E. Bou-Ghazele and K Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," in *Proc. ICASSP*, May 2002.

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

[4] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," in *Proc. ISCA Tutorial and Research Workshop on Statistical And Perceptual Audiotion*, Sep 2006.

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

[6] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, and K. Shikano, "Noise robust real world spoken dialog system using GMM based rejection of unintended inputs," in *Proc. Interspeech*, Oct 2004.

[7] J. Ramirez, P. Yelamos, J. M. Gorriz, and J. C. Segura, "SVM-based speech endpoint detection using contextual speech features," *Electoronics Letters*, vol. 42, no. 7, pp. 426–428, 2006.

[8] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[9] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. ICASSP*, May 2013.

[10] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on YouTube using deep neural networks," in *Proc. Interspeech*, Aug 2013.

[11] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[12] 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.

[13] B. Kingsbury, P. Jain, and A. G. Adami, "A hybrid HMM/traps model for robust voice activity detection," in *Proc. Interspeech*, Sep 2002.

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

[15] J. Ramirez, J. C. Secura, C. Benitez, A. de la Torre, and A Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 6, pp. 1119–1129, 2005.

[16] 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.

[17] Y. Obuchi, R. Takeda, and N. Kanda, "Voice activity detection based on augmented statistical noise suppression," in *Proc. APSIPA Annual Summit and Conference*, Dec 2012.

[18] 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.

[19] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403–2418, 2001.

[20] Y. Obuchi, R. Takeda, and M. Togami, "Bidirectional OM-LSA speech estimator for noise robust speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Dec 2011.

[21] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Interspeech*, Oct 2000.

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

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

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