INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Robust DNN-based VAD augmented with phone entropy based rejection of background speech

Yuya Fujita[1], *Ken-ichi Iso*[1]

[1]Yahoo Japan Corporation
yuyfujit@yahoo-corp.jp

Abstract

We propose a DNN-based voice activity detector augmented by entropy based frame rejection. DNN-based VAD classifies a frame into speech or non-speech and achieves significantly higher VAD performance compared to conventional statistical model-based VAD. We observed that many of the remaining errors are false alarms caused by background human speech, such as TV / radio or surrounding peoples' conversations. In order to reject such background speech frames, we introduce an entropy-based confidence measure using the phone posterior probability output by a DNN-based acoustic model. Compared to the target speaker's voice background speech tends to have relatively unclear pronunciation or is contaminated by other types of noises so its entropy becomes larger than audio signals with only the target speaker's voice. Combining DNN-based VAD and the entropy criterion, we reject speech frames classified by the DNN-based VAD as having an entropy larger than a threshold value. We have evaluated the proposed approach and confirmed greater than 10% reduction in Sentence Error Rate.

Index Terms: Voice Activity Detection, Deep Neural Network, Entropy

1. Introduction

Voice Activity Detection (VAD) is an important component of front-end processing in speech recognition systems because it can reduce recognition errors and also the computational cost by segmenting input audio into background speech and non-speech. It is also important in high-quality hands-free radio communication and speech codecs. Conventional VAD methods can be categorized into 5 different types. The first is based on raw acoustic features such as energy or zero-crossing rate of audio signals[1, 2]. The second one is statistical models in which speech and non-speech frames are modeled by Gaussian distributions and the log-likelihood ratio is used to decide whether a frame is speech or noise. The other types use some kind of classifier: Support Vector Machine (SVM) is one of the most popular classifiers in many machine learning tasks and is also used in VAD[3]. State-space models such as HMM or Kalman-filters have also been applied to VAD[4, 5]. Finally, Deep Neural Network (DNN) based VAD is becoming popular inspired by its success in acoustic modeling[6, 7, 8].

In this paper we focus on improving DNN-based VAD performance of our speech recognition system. We are running an internally developed speech recognition system for mobile voice search. We chose DNN-based VAD for our system because it is easy to implement and train since the source code and training data are easily derived from that used for acoustic modeling. However, there are some issues that degrade speech recognition accuracy due to the failure of VAD. We analyzed some misrecognized speech and found that our system is very sensitive to speech so there are many false alarms caused by background speech from nearby peoples' conversations or TV / radio. We can categorize utterances collected through our system into three major domains according to which smartphone application the utterances come from – typical voice search application (Search), personal assistant application (Dialogue), and voice search for map application which is typically used inside a car for car-navigation (Vehicle). Utterances from the Vehicle domain are the most affected by such background speech which we try to overcome in this paper.

We propose a method that utilizes the entropy of the posterior probability output by the acoustic model DNN. We observe that most background speech comes from the conversations of surrounding people or a TV / radio speaker. Speech from a TV or radio's loudspeaker tends to be contaminated by noise or reverberation because the location of the loudspeaker is further from the microphone than the target speaker. When clear utterance frames are fed into the acoustic model, it is easy to decide which state is most likely at each frame and there is little ambiguity so the posterior probability of one state takes a higher value than the other states. In this case entropy of the posterior probability is small. On the other hand when background speech is fed into the same acoustic model, it is hard to say that only one state is most likely because of contamination by noise so many states' posteriors have higher values. In this situation, the probability distribution function of the posterior will be closer to a uniform distribution so its entropy value becomes larger. Therefore, we hypothesize that such background speech can be rejected by adding a decision based on the entropy value.

As far as we know, there are no articles about the classification of background speech although there are some methods in the literature which utilize the entropy of the spectrum of speech for VAD[9, 10] and the entropy of the posterior of the acoustic model is utilized in classifying speech and music in [11]. However, this work is different from above-mentioned work in terms of its purpose and the method of utilizing the entropy value.

2. Proposed Method

Conventional DNN-based VAD decides whether each frame is speech or non-speech by comparing the sum of speech states' posterior probabilities and the sum of non-speech states' posterior probabilities output by a DNN. A typical way of building DNNs for VAD is to train a DNN with two output states (speech/non-speech). An alternative way is to use an acoustic model DNN directly by assuming that all states assigned to non-silence tri-phones are speech states. Non-speech states are the ones that are assigned to the silence tri-phone. We chose to use the acoustic model as our DNN for VAD because we can reuse

http://dx.doi.org/10.21437/Interspeech.2016-136

its output in the entropy calculation which is a crucial part of our proposed method. We conducted preliminary experiments and confirmed that there was little difference in the performance of these two approaches.

We now describe in detail our VAD algorithm. Suppose $\boldsymbol{x}(t)$ is an acoustic feature vector at the t-th time frame and $\boldsymbol{W}_l, \boldsymbol{b}_l$ are respectively the l-th layer's weight matrix and bias vector of an acoustic model DNN with L layers, then the posterior probability is calculated as follows:

The 1-st hidden layer's output is calculated by

$$\boldsymbol{h}_1(t) = \boldsymbol{W}_1\boldsymbol{x}(t) + \boldsymbol{b}_1, \quad (1)$$

$$\boldsymbol{o}_1(t) = g_1\left(\boldsymbol{h}_1(t)\right), \quad (2)$$

and the $l = \{2, \cdots, L\}$-th layers' output is calculated by

$$\boldsymbol{h}_l(t) = \boldsymbol{W}_l\boldsymbol{o}_{l-1}(t) + \boldsymbol{b}_l, \quad (3)$$

$$\boldsymbol{o}_l(t) = g_l\left(\boldsymbol{h}_l(t)\right), \quad (4)$$

where $g_l(\cdot)$ is a non-linear activation function for the l-th layer. We used the sigmoid function for $l = \{1, \cdots, L-1\}$-th layers defined by

$$g_l(y) = \frac{1}{1+\exp(-y)}, \quad (5)$$

and the identity function for the L-th layer. The final L-th layer's output is converted to posterior probabilities using the softmax function:

$$p(i|\boldsymbol{x}(t)) = \frac{\exp(o_L^i(t))}{\sum_{i'}\exp(o_L^{i'}(t))}, \quad (6)$$

where $o_L^i(t)$ represents the i-th component of vector $\boldsymbol{o}_L(t)$. Then, the posterior probability of the speech hypothesis H_1 and non-speech hypothesis H_0 is calculated as follows:

$$p(H_1|\boldsymbol{x}(t)) = \sum_{i\in S} p(i|\boldsymbol{x}(t)), \quad (7)$$

$$p(H_0|\boldsymbol{x}(t)) = \sum_{i\in N} p(i|\boldsymbol{x}(t)), \quad (8)$$

where S denotes the set of indices representing speech states and N represents the set of indices of silence states. If the following condition is met, we decide the t-th frame is a speech frame:

$$p(H_1|\boldsymbol{x}(t)) > p(H_0|\boldsymbol{x}(t)). \quad (9)$$

In our method, the entropy based decision is also applied to speech frames classified by the above criterion. The entropy of each frame is calculated by

$$e(t) = \sum_{i\in S\cup N} p(i|\boldsymbol{x}(t))\log p(i|\boldsymbol{x}(t)), \quad (10)$$

so if the following condition is met, the t-th frame is identified as target speech and passed to the decoder:

$$e(t) < \tau. \quad (11)$$

A diagram of this algorithm is shown in Fig.1

As we mentioned in the introduction, the posterior probability of background speech could become close to a uniform distribution because of contamination by noise or reverberation so its entropy value becomes larger than clear utterances. We show the waveform, manually labeled voice regions, posterior probability of speech and the entropy value of two utterances in Fig.2 and 3. Fig.2 is a plot of a clean and clear utterance. We can see that the entropy values do not become large. Fig.3 is an utterance corrupted by speech from the radio in a car environment. Both before and after the correct voice region, the posterior probability of speech becomes larger because of background speech. In that region, the entropy value becomes larger than those of the correct voice region.

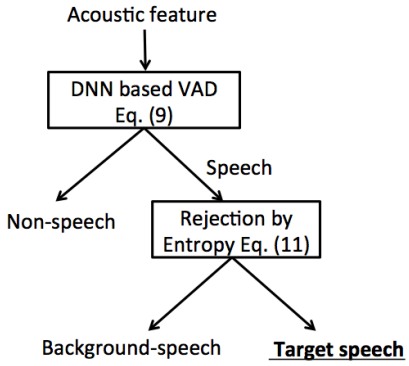

Figure 1: *Diagram of proposed VAD method.*

We plot the histograms of entropy of our development set in Fig.4 in order to see whether it is possible to classify background speech using the entropy value. Each frame of the development set is tagged as true positive, true negative, false alarm or false rejection by comparing labels generated by forced alignment. By manually checking several utterances from the development set, we confirm that most false alarms are caused by background speech. It is clear that the entropy value of frames tagged as false alarm are larger than other frames. We also plot the histogram of the moving average of entropy in Fig.5 and 6 because in [11] it is shown that averaging entropy over multiple frames makes it easier to discriminate a frame of speech or music. We expect that it works well in our background speech classification scenario too. However, averaging over multiple frames makes the histogram of false alarm frames close to true positive frames so we add the frame-wise entropy-based decision criterion to reject background speech frames. If $e(t)$ is greater than some threshold, the t-th frame is classified as background speech.

3. Experiment

3.1. Experimental setup

We evaluated the conventional and proposed VAD method using the acoustic model DNN trained on 1200 hours of transcribed speech collected through our mobile voice search system. The conventional baseline method uses only Eq. (9) and the proposed method uses Eq. (9) and (11) for speech classification as shown in Fig.1. We select 20k utterances from map applications (Vehicle domain we defined in the introduction) which are different to those used in training. Then, we divide them equally into development and evaluation sets in such a way that each set does not contain utterances from the same period of time and from the same smartphone (each set has 10k utterances). In addition to these two sets, we prepared two reduced test sets (each a subset of the above 10k evaluation and development sets, respectively) to see the contribution of the VAD method to recog-

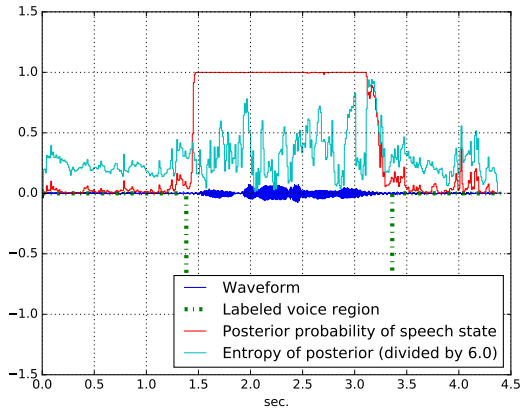

Figure 2: *Waveform, manually labeled voice region, posterior probability of speech state and entropy and of an utterance by a single speaker without background noise.*

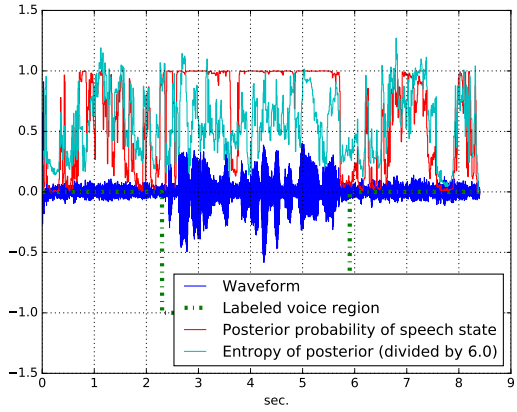

Figure 3: *Waveform, manually labeled voice region, posterior probability of speech state and entropy of an utterance with background speech.*

nition accuracy. Speech recognition errors are caused by VAD errors or ASR decoder errors, and it is not trivial to separate these causes in general. In the reduced test sets, we chose utterances from the original test sets that are correctly recognized using manually labeled VAD boundaries. With these reduced test sets, we can estimate the contribution of our VAD method to recognition accuracy.

We use two metrics to analyze performance. The first one is VAD frame error rate (FER) which is the number of frames misclassified divided by the total number of frames. The second one is phone Sentence Error Rate (SER). The reason for choosing SER is that our system is designed for mobile voice search in Japanese where the commonly used Word Error Rate (WER) metric does not always reflect the subjective performance by a user. This is because an error of one word may result in a completely different search result. The reason for using only phone information is because Japanese has 4 alphabets (kanji, hiragana, katakana and romaji) and one sentence can have multiple surface forms while having the same meaning therefore we normalized all surface forms to phones.

The audio signal of each utterance in the test set is first sent

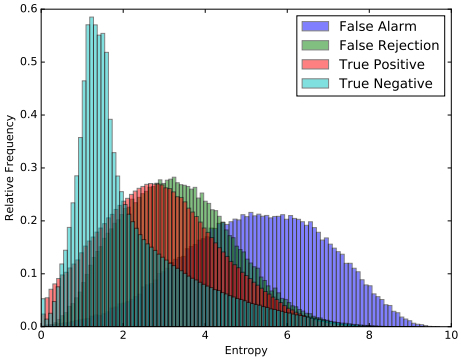

Figure 4: *Histogram of entropy of development set.*

to a VAD process and classified frame-wise into speech or non-speech. Then, the frame-wise VAD results are smoothed using a manually tuned finite-state automaton. After that, the segmented speech regions are passed to the decoder. Our decoder is an internally developed single-pass WFST decoder[12]. The language model is a tri-gram model trained using text queries of the Yahoo Japan search engine and transcriptions of mobile voice search queries. Other parameters are detailed in Table 1.

Table 1: *Parameters of the speech recognition system.*

name	value
Acoustic feature	40ch Filter Bank
Splicing	-5/+5
Number of units in hidden layers	1024
Number of hidden layers	5
Output state numbers	4003
Vocabulary size	1.3M

3.2. Results

VAD FER of the development set is shown in Table 2. The best FER is observed when we set the entropy threshold to 7.0. At that operating point, the relative reduction in FER was 5.5%. VAD FER of the evaluation set is shown in Table 3 where the relative improvement was 2.4%.

Table 2: *VAD FER of the development set.*

Method	Entropy threshold	FER %
baseline	-	4.54
proposed	6.0	4.50
	7.0	4.29
	8.0	4.46
	9.0	4.54

The SER of the reduced test set is shown in Table 4. A relative reduction in SER of more than 10% was achieved. These results show that our proposed method can correctly recognize sentences that the baseline system could not. The SER of the whole test set is shown in Table 5. The reduction in SER on the development set was 4% and on the evaluation set was 2.2%. Note that the whole test set contains mis-recognized sentences

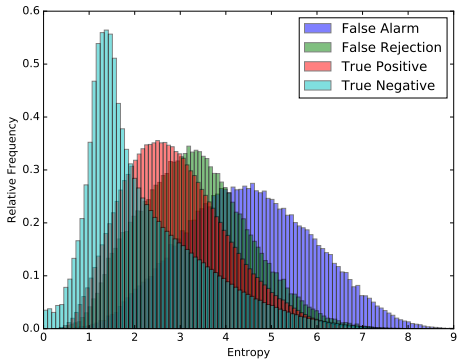

Figure 5: *Histogram of averaged entropy over 10 frames of development set.*

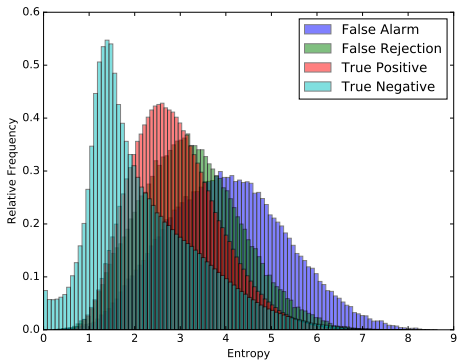

Figure 6: *Histogram of averaged entropy over 20 frames of development set.*

Table 3: *VAD FER of the evaluation set.*

Method	Entropy threshold	FER %
baseline	-	4.60
proposed	7.0	4.49

which might not have been caused by VAD failure so the improvement appears smaller than on the reduced test set.

We also checked the performance in non-target domains. Table 6 shows the results in two other domains: utterances collected through a personal assistant smartphone application (Dialogue) and a typical voice search application (Search). There was little difference in performance even though the threshold of entropy was optimized for the Vehicle domain. Therefore our method can improve recognition accuracy in the Vehicle domain without any degradation in performance in other domains.

4. Conclusion

We augmented a DNN-based VAD in order to suppress false alarms caused by background speech from TV / radio or surrounding peoples' conversations. Background speech tends to be contaminated by other noises and reverberation because the location of such sound is further from the microphone than the target speaker's voice. If utterances with such background speech are fed to the acoustic model, the posterior probability of each HMM state becomes close to a uniform distribution which results in larger entropy. Hence we utilized the entropy of the posterior probability output by the DNN acoustic model to reject background speech frames.

Experimental results showed that the FER of our proposed method was reduced by 5.5% on the development set and 2.4% on the evaluation set. The reduction in phone SER on the reduced test set in which we estimate the contribution of VAD improvement to recognition accuracy was 13.9% on the development set and 10.8% on the evaluation set. Reduction in SER on the whole test set was 4% on the development set and 2.2% on the evaluation set without any degradation in the performance in other domains.

Table 4: *Speech recognition results on the reduced test set. SER improvements in this table indicate estimated value of how much contribution is made to recognition accuracy due to VAD improvement.*

	Condition	#Utts.	SER %	#Cor.	Red. %
dev.	baseline	8554	5.46	8087	
	proposed	8554	4.70	8152	13.9
eval.	baseline	8330	3.95	8001	
	proposed	8330	3.52	8037	10.8

Table 5: *Speech recognition results on the whole test set. The whole test set contains utterances that cannot be recovered by improving the VAD.*

	Condition	#Utts	SER %
dev.	baseline	10000	16.78
	proposed	10000	16.10
eval.	baseline	10000	17.66
	proposed	10000	17.26

Table 6: *Recognition results in non-target domains.*

domain	system	#Utts	SER %
Search	baseline	10000	24.09
	proposed	10000	23.98
Dialogue	baseline	10000	23.79
	proposed	10000	23.71

5. References

[1] *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*, ITU-T Recommendation G.729, 06/2012.

[2] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, ETSI ES 202 050, 2007.

[3] J. Wu and X. L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 466–469, Aug 2011.

[4] Y. Liang, X. Liu, Y. Lou, and B. Shan, "An improved noise-robust voice activity detector based on hidden semi-markov models," *Pattern Recognition Letters*, vol. 32, no. 7, pp. 1044 – 1053, 2011.

[5] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching kalman filter," *IEICE transactions on information and systems*, vol. 91, no. 3, pp. 467–477, March 2008.

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

[7] Q. Wang, J. Du, X. Bao, Z. Wang, L. Dai, and C. Lee, "A universal VAD based on jointly trained deep neural networks," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2282–2286.

[8] I. Hwang, J. Sim, S. Kim, K. Song, and J. Chang, "A statistical model-based voice activity detection using multiple dnns and noise awareness," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2277–2281.

[9] J.-l. Shen, J.-w. Hung, and L.-s. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments." in *ICSLP*, vol. 98, 1998, pp. 232–235.

[10] C. Yang and M. Hsieh, "Robust endpoint detection for in-car speech recognition," in *Sixth International Conference on Spoken Language Processing, ICSLP 2000, Beijing, China, October 16-20, 2000*, 2000, pp. 1061–1064.

[11] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, no. 3, pp. 351 – 363, 2003.

[12] K. Iso, E. Whittaker, T. Emori, and J. Miyake, "Improvements in Japanese Voice Search," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 2109–2112.