Voice activity detection based on deep neural networks and Viterbi

Liang Bai, Zhen Zhang, Jun Hu
National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029, China, hj@cert.org.cn

Abstract. Voice Activity Detection (VAD) is important in speech processing. In the applications, the systems usually need to separate speech/non-speech parts, so that only the speech part can be dealt with. How to improve the performances of VAD in different noisy environments is an important issue in speech processing. Deep Neural network, which proves its efficiency in speech recognition, has been widely used in recent years. This paper studies the present typical VAD algorithms, and presents a new VAD algorithm based on deep neural networks and Viterbi algorithm. The result demonstrates the effectiveness of the deep neural network with Viterbi used in VAD. In addition, it shows the flexibility and the real-time performance of the algorithms.

1 Introduction

Automatic speech recognition (ASR) and natural language processing (NLP) have become the key technology in human-computer science. Recent years, the ASR technology has been rapidly developed, and there are many applications emerging in normal social life which have brought great convenience to people, such as automatic dictation machine and automatic telephone answering [1]. In all the speech recognition applications, VAD plays an important role in speech processing which greatly affects the performance of the speech recognition [2].

The continuous speech is composed of a series of speech segments and noise segments [3], and the duration of speech part is less than 40% of the speech time [4]. The speech recognition system only needs to process the speech part, and the noise should be removed before the processing. In this way, both of the speed and performance of the recognition system will be improved. The purpose of VAD is to remove the noise part of the speech and retain the speech part to the speech recognition system.

With the rapid development of speech signal processing in the past decades, the VAD technology has been constantly improved. In 70’s, Rabiner used the short-time energy and the short-time zero crossing rate as the feature to detect the start and end points of speech [5]. Then, F. K. Soong proposed a segmentation method based cepstrum [6]. At the same time, Wilpon proposed a method based on the speech signal short-time spectrum change to determine the sub-word segment information [7]. In 90’s, Erdal proposed a segmentation method based on speech parameters [8], and then, Euvaldo F. Proposed a phoneme segmentation method based on the speech tracking segment [9]. Start McClellan proposed a segmentation method based on spectral entropy to improve the anti-noise performance [10].

In the past period, signal processing algorithms, such as time domain or frequency domain parameters estimate, are mainly used in VAD, and the parameters contain short-time zero-crossing rate, and harmonic noise ratio. The signal processing method has a good performance in the match environments, but not suits for different noise environments. The main reason is that the signal processing algorithm assumes that the short-term speech signal is stable, and
needs to estimate the noise condition. However, once the estimation is not accurate, the performance will be seriously affected.

With the rapid development of speech recognition industry, there are lots of new VAD algorithms arising, and the most popular method is based on statistical modeling methods such as the method based on hidden markov model (HMM) [11], the method based on support vector machine (SVM) [12] and methods based on neural networks [13-15]. The statistical modeling method is to train the model to distinguish between short-term noise and voice by probability calculations. Since there is a strong non-linear learning characteristics of statistical models, we can add the training data to improve the generalization ability of the models.

In recent years, the speech recognition based on deep neural network (DNN) has got great success. Compared to the traditional GMM-HMM models, the performance has been significantly improved, since the DNN has a stronger learning ability than the traditional models. In this paper, we applied the DNN algorithm to the VAD procedures, with the Viterbi algorithm to optimize the performance. The experimental results showed this method has better effectiveness and adaptability than the method based on Bayesian Information Criterion (BIC) in ASR tasks.

The rest of this paper is organized as follows. Section 2 introduces the structure of DNN and the training procedures of proposed VAD model. Section 3 presents the experiment results. Finally, Section 4 concludes the paper and discusses the future work.

2 VAD based on DNN

2.1 Deep neural network

Artificial neural network (ANN) is a computational model that simulates the store and process information functions of human brain. Specifically, a large number of non-linear processing units (neurons) are interconnected in an ANN, and the weights of connections between two neurons can be adjusted. The ANN can be changed according to the environment which it applied to, and can constantly adjust the weights between neurons according to the train data set. As a result, the ANN can complete complicate tasks such as classification, and recognition. Because of the good learning characteristics of ANN, it is widely used in machine learning.

DNN has a plurality of hidden layers. Compared to the traditional ANN, the DNN has stronger nonlinear learning ability. Generally, the low-level network is used for the input signal feature extraction, and the high-level network is used for category classification. When the scale of DNN is large enough, it can enhance the accuracy of classification and prediction.

The structure of DNN is constituted by input layer, hidden layers and output layer. The input layer receives the speech feature; the hidden layers process and analyze the relevance between feature and classifications; the output layer generates the posterior of classifications. As shown in [16], given the speech feature \( p(q_i) \), DNN output the posterior \( x^{\text{post}}_{t-1} \), it is calculated by Bayesian formula:

\[
p(x^{\text{post}}_{t-1} \mid q_i) = \frac{p(q_i \mid x^{\text{post}}_{t-1}) p(x^{\text{post}}_{t-1})}{p(q_i)}
\]

\( p(q_i \mid x^{\text{post}}_{t-1}) \) is the priori probability of state \( q_i \), it is the frequency approximation on training set.

The DNN training uses error back propagation (EBP) algorithm. EBP is a supervised learning algorithm, and it is first proposed by Rumelhart and McCCelland [17]. The EBP is based on the gradient descent algorithm, and it can adjust the network weights effectively. The EBP algorithm consists of two main aspects: (1) the forward propagation of information. (2) the backward propagation of error. EBP can learn the mapping relationship between input and output elements, without the prior knowledge. It uses the steepest descent method to minimize the squared error sum of the network.

In this paper, we propose to use DNN as the statistical model to separate speech/non-speech, and it only needs a few states to construct the model. Its input layer is the feature which is extracted to distinguish speech/non-speech, and its output layer generates the probability for speech/non-speech. Since the EBP algorithm is supervised, we need to prepare the training data with precise transcript. In the experiment, we use manual speech/non-speech segments to train a GMM-HMM model, and then
we use this model to do force-alignment to get the frame-level speech/non-speech classification [1]. The training procedures of DNN model for VAD need three steps, as shown in Table 1.

Table 1: The training scheme of DNN.

| The training scheme of DNN |
|---------------------------|
| 1: Train a baseline GMM-HMM model with manual speech/non-speech segments training data |
| 2: Use the GMM-HMM to do force alignments on the training data to get frame-level segment information |
| 3: Train the DNN model to do speech/non-speech classification and detect the start and end point of speech |

2.2 Speech/non-speech decision

We constructed the DNN model of speech/ non-speech. When the feature of speech is input, the DNN model calculates the probability of this speech frame and determines it is speech or not. We experiment two methods to determine the start/end point of speech:

a. Threshold determination method: experiments show that using a certain amount of training data to train DNN, the accuracy of VAD could get good performance, so the first method is to send feature of every speech frame into DNN model. When the probability of the speech frame is larger than 0.5, the frame is determined as speech, represented as "0". Otherwise the non-speech, represented as "1". With a certain length of window median filter to remove some of the jumping points, we get the final start/end points of speech.

b. Viterbi method: the Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of events implied by the observed sequence of states [18]. For VAD, The DNN’s output sequence is equivalent to the observation sequence. There is transition probability between speech and non-speech. From the starting point of the speech signal to judge the frame for the current frame, speech/non-speech probability for each frame appears before obtained by multiplying the transition probabilities between frames to give the overall probability of speech. The maximum probability path is most likely the hidden sequence.

In the Viterbi algorithm, there are two parameters need to be set: the first one is the speech/non-speech state transition probability; the second one is the prior probability of the speech/non-speech. These two parameters are usually obtained on the development set, and then fixed for the test data.

3 Experiment

3.1 Data and system set

In this paper, the training data set is Mandarin CTS dialogue, provided by LDC: Call-Home, Call-Friend and Call-HKUST [19], as shown in Table 2. In order to investigate the VAD’s adaptability in mismatch noise environment, we use the bank customer service records as the development and test sets. The development set contains 44 telephone conversations, about 3 hours. The test set contains 39 phone conversations, about 3.5 hours.

Table 2: The speech data for DNN-VAD training

| Corpus    | Length (Hour) |
|-----------|---------------|
| Call-Home | 15            |
| Call-Friend | 20          |
| Call-HKUST | 90            |

In order to improve the generalization of DNN, we add noise to the training set to make DNN able to handle different noise environments. We record nine kinds of different noise, including road noise, restaurant noise, music, background voice, wind noise and other common types (these noises are
independent of the test set). All the noises are randomly added to the train set, and the SNR is controlled in the range of 0-20db. As a result, the total amount of training data is expanded from 125 hours to 250 hours.

The DNN input feature is the normal 13-dimensional MFP/LP feature, with one-dimensional pitch feature and on-dimensional NCC feature which is represented as the credibility of pitch [20]. All features will be the third-order differential, and plus the static characteristics, constitute a total of 60 input feature dimensions. There are only 2 output states of DNN: speech or non-speech. We conduct our experiments on different scales of hidden layers to observe the recognition performance and speed.

We also compare the DNN-based VAD to the current BIC-based VAD [21]. BIC algorithm first divides the speech record into small segments by calculating the similarity between adjacent segments. The segments with the similarity that is greater than the threshold are spliced together to generate a new segment. The BIC algorithm can splice the speech/non-speech effectively when the threshold is set reasonable.

3.2 Baseline experiments

The evaluation index in our experiments are miss rate/false alarm rate and character error rate (CER). Miss rate/false alarm rate is to evaluate the accuracy of VAD directly, and CER is the speech recognition performance with the application of VAD.

a. Miss rate/false alarm rate

In this paper, we first compare the VAD result to the transcripts to get miss rate/false alarm rate of speech/non-speech classification. Since the miss rate and false alarm rate of speech and non-speech are symmetric, we only need to evaluate the speech segments. For speech segmentation task, the miss rate indicates the probability that the algorithm identifies the speech segments as non-speech segments. On the contrary, the false alarm rate indicates the probability that the algorithm identifies the non-speech segments as speech segments.

As shown in Table 3, we conduct experiments on both BIC-based and DNN-based VAD. We also compare the threshold method and Viterbi method. With the BIC algorithm, the experimental results show that there is low miss rate, but the false alarm rate is relatively high. The reason is that BIC algorithm avoids the loss of effective speech segments, and it expands the duration of speech. This leads to the phenomenon of high false alarm rate. Compared to BIC algorithm, the DNN-based VAD achieves low false alarm performance while the performance of miss rate is also satisfied. The Viterbi method gets better performance than the threshold method, since the Viterbi method calculates the historical information. Thus, in the sequent experiments, we use Viterbi method as the default algorithm to determine speech/non-speech.

Table 3: The Miss Rate and False Alarm Rate for different VAD methods

| VAD    | BIC       | DNN       |
|--------|-----------|-----------|
|        | Threshold | Viterbi   |
| Miss Rate % | 0.3       | 1.8       | 1.3       |
| False Alarm Rate % | 28.1      | 1.9       | 1.4       |

b. Character error rate

Considering the VAD is usually used as the front end of speech recognition system, its purpose is to remove non-speech part to ensure that noise does not enter the recognition system. It can effectively reduce the amount of calculation. The VAD affects the performance of recognition system significantly. If the VAD abandons too many speech segments, it will lead the high deletion error rate to the results. On the contrary, if too many non-speech segments are classified to speech mistakenly, the calculation cost will be remarkably improved, so the insertion error rate will increase. The CER is used to evaluate the speech recognition performance, and it also shows the performance of VAD.

As shown in Table 4, on the development and test sets, the DNN-based VAD gets 2% and 1% decline on CER than the BIC-based VAD. In particular, the two algorithms get the same insertion error rate,
but the deletion error rate of DNN-based VAD is lower than the other significantly. This shows that the DNN-based VAD can avoid the mistake which misjudges the speech to non-speech.

### Table 4: The CERs for Dev set and Test set

|       | Sub | Del | Ins | CER  |
|-------|-----|-----|-----|------|
| BIC   | 17.9| 8.5 | 9.1 | **35.4** |
| DNN   | 17.6| 6.5 | 9.5 | **33.6** |
|       |     |     |     |      |
| Test  | Sub | Del | Ins | CER  |
| BIC   | 17.1| 7.6 | 10.6| **35.3** |
| DNN   | 16.8| 6.8 | 10.8| **34.4** |

The scale of DNN model is closely related to the performance of classification. We compare the affection on performance and speed with different size of DNN hidden layers. We conduct experiments with small model with 125 hidden layer nodes and big model with 512 hidden layer nodes. The models have both 3 hidden layers. Compared to the small model, the big one is able to cover more information of training data, and it can also do more complicated decision. However, the speech of big model is much slower than the small model due to the increasing time of computing. On a 25 minutes data set, the small model needs 9.3 seconds and the big model needs 10 minute 2 seconds.

### Table 5: The CERs for big and small models

|                | Hidden layer nodes | Corr | CER  |
|----------------|--------------------|------|------|
| big model      | 512                | 75.9 | **33.6** |
| small model    | 128                | 75.9 | **33.7** |

Table 5 shows the performance between big model and small model on the customer service dialogue data. The results show that the performance on small model is 0.1% lower than that of big model. The affection of size of DNN model is not obvious. If the environment noise is very harsh, the big model can supply better robustness. Considering the speed of small model is faster significantly, we adopt the small DNN with 125 hidden layer nodes as our default experiment set.

### 4 Conclusion

In this paper, we proposed the VAD algorithm based on DNN, and compared it to the traditional BIC algorithm. On the customer service dialogue task, the results showed that the DNN-based VAD achieved better performance on miss rate/false alarm rate and CER. At the same time, since the data we used to test is different from the training data, it proved that the DNN-based VAD has stronger adaptability.

We introduced the DNN into VAD algorithm, and combined it with Viterbi algorithm. There are still many questions about theories and applications need to be discussed. In the further, we will focus our research on the feature exaction and choosing the feature reflecting the difference of speech/non-speech effectively. On the other hand, we will research on the structure of DNN and look for more suitable neural network models.

### Acknowledgements

This work was supported by Youth Foundation of CNCERT (2016QN-11).

### References

[1] Q. Zhang. "Bilingual English mixed acoustic modeling", Institute of Acoustics, (2010).
[2] L. Han, B. Wang and S. Duan. "Development of voice activity detection technology", Application Research of Computers, 4, pp. 1200-1226.
[3] A. Benyassine, E. Shlomot, S. Huanyu, and E. Yuen, "A robust low complexity voice activity detection algorithm for speech communication systems”, 1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding, pp.97-98, (1997).
[4] A. Vahatalo and I. Johansson. "Voice activity detection for GSM adaptive multi-rate codec", 1999 IEEE Workshop on Speech Coding Proceedings, pp.55-57, (1999).
[5] L. R. Rabiner and R. W. Schafer, Digital processing of speech signals, Prentice Hall, (1978).
[6] T. Svendsen and F. Soong. "On the automatic segmentation of speech signals", Proceedings of ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.77-80, (1987).
[7] J. Wilpon, B. Juang and L. Rabiner. "An investigation on the use of acoustic sub-word units for automatic speech recognition", Proceedings of ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.821-824, (1987).
[8] L. Karray and J. Monné. "Robust speech/non-speech detection in adverse conditions based on noise and speech statistics", ICSLP, pp.1471-1474, (1998).
[9] E. F. Cabral and G. D. Tattersall. "Trace-segmentation of isolated utterances for speech recognition", 1995 International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95, pp.365-368, (1995).
[10] S. A. McClellan and J. D. Gibson. "Spectral entropy: An alternative indicator for rate allocation", IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94., pp.201-204, (1994).
[11] L. Bing. "A new method based on HMMs for noise-robust voice activity detector", Xihua University, (2012).
[12] F. Chunhui. "A study of voice activity detection algorithm based on HMM/SVM", Taiyuan University of Technology, (2013).
[13] Q. Wang, J. Du, X. Bao, Z. Wang, L. Dai, and C. Lee. "A universal VAD based on jointly trained deep neural networks ", INTERSPEECH, pp.2282-2286, (2015).
[14] A. Hussain, S. A. Samad and L. B. Fah. "Endpoint detection of speech signal using neural network", Proceedings of TENCON 2000, pp.271-274, (2000).
[15] L. Wang, C. Zhang, P. C. Woodland, M. J. Gales, P. Karanasou, P. Lanchantin, X. Liu, and Y. Qian. "Improved DNN-based segmentation for multi-genre broadcast audio", 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5700-5704, (2016).
[16] X. Yeming, Z. Qingqing, S. Liming, P. Jielin, and Y. Yonghong. "Optimization of deep neural network in acoustic modeling for mandarin speech recognition", Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 26, (2014).
[17] D. Williams and G. Hinton, "Learning representations by back-propagating errors", Nature, 323, pp.533-538, (1986).
[18] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, 77, pp. 257-286, (1989).
[19] Q. Zhang, J. Pan and Y. Yan. "Tonal articulatory feature for mandarin and its application to conversational LVCSR", 10th Annual Conference of the International Speech Communication Association, (2009).
[20] Q. Zhang, F. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan. "Improved modeling for F0 generation and V/U decision in HMM-based TTS", 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4606-4609, (2010).
[21] L. Tantan, P. Jielin, S. Hongbin, and Y. Yonghong. "Speaker diarization algorithm based on CLR and BIC", Technical Acoustics, 26, pp. 1181-1185, (2007).