EVALUATING VAD FOR AUTOMATIC SPEECH RECOGNITION

Sibo Tong[a], *Nanxin chen*[a], *Yanmin Qian*[a], *Kai Yu*[a]
Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering MOE-Microsoft Key Lab. for Intelligent Computing and Intelligent Systems Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
{supertongsibo, sjtu_cnx, yanminqian, kai.yu}@sjtu.edu.cn

ABSTRACT

Voice activity detection (VAD) plays a crucial role in speech processing, especially in automatic speech recognition (ASR). It identifies the boundaries of the speech to be recognized and the boundary accuracies may significantly affect the recognition performance. Conventional VAD evaluation criteria are mostly based on frame-level accuracy of speech/non-speech classification, which may result in weak correlation between VAD and ASR performance. Even though some VAD evaluation criteria consider boundary effects, there has not been an effective overall criterion suitable for evaluating the effect of VAD on ASR. This paper proposes an integrated VAD evaluation criterion taking into account various boundary effects. Experiments on an English Switchboard task showed that, conventional frame accuracy based VAD criterion has weak and unstable correlation with word error rate while the proposed overall criterion is much more stably correlated to word error rate.

Index Terms— evaluation metric, voice activity detection, speech recognition

1. INTRODUCTION

As is widely acknowledged to the public, the procedure of detecting presence of speech periods from a set of continuously given signal is called the voice activity detection, abbreviating for VAD. VAD is one of the most critical techniques for speech signal processing. It is broadly applied to various speech applications such as speech recognition, speech synthesis, speech coding, speech enhancement and can directly influence the performance of future work. This effect is particularly significant in the field of speech recognition. Speech recognition system processes valid signal, in which noise has been already eliminated, and gets the right text. However, fully pure speech signal can hardly be obtained in real world. Using a VAD with high performance can drastically reduce false alarms from non-speech periods and deletion errors from speech segments in input signal and consequently improve the performance of a speech recognition system. For above-stated reasons, VAD techniques have their own significance and have been probed far and widely.

A number of techniques have been proposed for VAD, including both unsupervised systems mostly based on energy [1], zero crossing rate [2], the periodicity measure [3], higher-order statistics in LPC residual domain [4] and supervised systems, including support vector machines [5], Gaussian mixture models (GMMs) [6], deep neural networks (DNNs)[7], which trains a classifier using features such as Mel-frequency cepstral coefficients (MFCCs) or perceptual linear prediction coefficients (PLPs) as input.

In order to evaluate a VAD algorithm for speech recognition, using automatic speech recognition (ASR) performance is the most direct approach. However it makes the system computationally costly and tough to be practical in our daily life. Thus, for the purpose of attaining a more objective and more precise assessment to the performance of a VAD algorithm, people traditionally calculated frame accuracy and used ROC curve [8] to get an evaluation for a VAD algorithm. Later, criteria that consider boundary effect to some extent [9, 10] were proposed. However, it fails to put forward an effective overall criterion suitable for evaluating the effect of VAD on ASR. Besides some evaluations of the subjective performance of VAD [11] has also been proposed. But it is not an effective criterion to evaluate the effect of VAD on ASR. Under this circumstance, an integrated VAD evaluation criterion taking various boundary effects into account is proposed in this paper.

The remainder of this paper is organized as follows. Section 2 reviews previous evaluation criteria for VAD algorithm briefly. In section 3, new VAD evaluation criterion based on both frame-level and boundary-level is demonstrated in detail. Experimental results are provided through experiment based on both artificial data and real data in section 4. Finally section 5 concludes the whole paper.

2. REVIEW OF PREVIOUS EVALUATION CRITERIA

VAD can be used for various purposes of speech processing. In this paper, automatic speech recognition

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

(ASR) is the only focus. In an ASR system, VAD is the pre-processing step in order to distinguish speech frames from original audio files. With the development of different models using in VAD systems, the performance of these systems need to be compared independently of the whole ASR system due to the long time cost.

Traditional VAD evaluation framework treats the whole VAD system as a classifier where criteria for classifiers can be introduced. Different frames are considered as different input points where criteria based on frames accuracy can estimate the performance of this classifier.

The frames accuracy (ACC) [12, 13] is the most basic criterion used to evaluate performance which defined as:

$$J_A = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N_r}\delta(x_i, x_i^{ref})}{\sum_{r=1}^{R} N_r} \quad (1)$$

where x and x_{ref} indicates the VAD output label (speech/nonspeech) and the ground truth respectively, i denotes the index of frame and r denotes the index of utterance, N_r is the total number of frames of utterance r, R is the total number of utterances and

$$\delta(x, y) = \begin{cases} 1 & : \quad x = y \\ 0 & : \quad x \neq y \end{cases} \quad (2)$$

Based on this, receiver operating characteristic curve (ROC) [8, 14, 15] and detection error tradeoffs (DET) curve [16] are graphical plots which illustrate the performance as its discrimination threshold is varied. ROC curves plot TPR vs. FPR and DET curves plot TPR vs. FNR. TPR is the rate of TP (True Positive) in total speech frames, FPR is the rate of FP (False Positive) in total silence frames, and FNR is the rate of FN (False Negative) in total speech frames. Single metrics such as equal error rate (EER) or area under the curve of ROC (AUC) can be calculated from these graph and evaluate the performance for different VAD algorithms. In addition to EER, single metrics such as F-score (F1, harmonic mean of precision and recall) [17, 18, 19] and HTER (average of FPR and FNR) [20, 21] can also be calculated.

It has been observed in our previous experiments that all the above frame-level criteria share similar properties in VAD evaluation. They are good for evaluating classifiers and but may not fit for automatic speech recognition systems because there aren't enough evidence to show that the correlation between these metrics and performance of ASR systems.

To address this issue, segment-level VAD evaluation criterion is proposed [10]. If a segment contains a true speech segment, it is judged as correct and segment-level accuracy is then calculated accordingly. However, if the detected segment is very long, it may contain lots of additional noise and lead to significant insertion errors. Hence, segment-level criterion may not be an ideal criterion.

In contrast, other researchers proposed a number of boundary-level ASR related metrics. In [9] and [22], four metrics are proposed to evaluate the effect of VAD on ASR, including FEC (clipping at the front of a speech burst), MSC (clipping in the middle of a speech burst), NDS (noise detected as speech) and OVER (amount of time during which the output of the VAD is on after the reference has switched off) These metrics effectively consider errors in different positions of segments because they may have different influence for tasks. For example, for ASR tasks, boundary effects are very important which directly influence the start and end for sentences [10]. Although the four criteria are all related to ASR performance, they vary disparately on different VAD results. It is then hard to say which of the four is of most importance. The lack of an overall evaluation metric limits the use of these criteria. What's more, the effectiveness of the above criteria (incl. the segment-level one) is hypothetical and there have not been experiments to show actual correlation between these criteria and the ASR recognition performance.

3. NEW VAD EVALUATION CRITERION

As indicated in the previous section, there has not been an all in-one metric which is correlated to ASR performance. In this section, a criterion evaluating both the VAD classification and the effect on subsequent ASR systems is proposed. The nature of VAD evaluation is different from a normal binary classifier in that the objects are sequences rather than single data samples. Consequently, two facts need to be considered during VAD evaluation: frame and boundary (or segment) accuracy. In particular, boundary accuracy is considered to be correlated to the final ASR performance. Therefore, in addition to frames accuracy, several criteria are proposed in this paper to evaluate boundary-level accuracy.

3.1. Start boundary accuracy (SBA)

This is a sub-criterion considering the start boundary of each speech segment. Assume that there are N speech segments in reference. For each of the reference start boundaries, calculate a start boundary score if there exist a VAD speech start boundary matching it (allowing a plus or minus error margin L). Considering an interval around each start boundary, the left part belongs to silence period while the right lies in speech segment. For the reason that frames error during speech period usually leads to more serious recognition error than that during silence period, a weighting function is used to calculate the score in every

start boundary. Different weighting function can be considered which gives a heavier weight to the frames in reference speech segment. Step function

$$f(x) = \begin{cases} 1 & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases} \tag{3}$$

is a good choice. For the start boundary s_r of utterance r, consider an interval $[s_r - L, \ s_r + L]$ around it. Here L is used to adjust to fit for different cases such as long sentences or short sentences. And we use x and x_{ref} to represent each frame in a VAD output and the ground truth respectively. Therefore start boundary score for s_r is calculated as:

$$J_S^r = \frac{\sum_{i\in[s_r-L,s_r+L]} f(i-s_r)\delta\left(x_i, \ x_i^{ref}\right)}{\sum_{i\in[s_r-L,s_r+L]} f(i-s_r)} \tag{4}$$

Here SBA is defined as:

$$J_S = \frac{\sum_{r=1}^{R} J_S^r}{R} \tag{5}$$

where R indicates the number of speech segments.

3.2. End boundary accuracy (EBA)

Similar as SBA, for each of the R end boundaries for reference speech segment, use weighting function to calculate the score for every end boundary e_r.

$$J_E^r = \frac{\sum_{i\in[e_r-L,e_r+L]} f(e_r-i)\delta\left(x_i, \ x_i^{ref}\right)}{\sum_{i\in[e_r-L,e_r+L]} f(e_r-i)} \tag{6}$$

The average of these sums will be defined as *End Boundary Accuracy* (EBA) J_E, similar as equation (5).

3.3. Border precision (BP)

Previous research shows that number of segments has crucial influence on word error rate (WER) in ASR [10]. In order to avoid the case that VAD algorithm gives too many segments, we define the coefficient correct *Border Precision* (BP) equals to the correct boundary precision, where correct boundary means speech boundary lies in the interval of reference speech boundary. We assume that there are M speech segments from VAD algorithm. SBA and EBA can be used as start and end boundary average score. So BP can be calculated as

$$J_B = \frac{R}{2M}(J_S + J_E) \tag{7}$$

Once individual metric is defined, it is desirable to combine them. For most cases average is used to combine different criteria. But due to the experiment, neither of these criteria can be small for a good VAD system. So harmonic mean is used here, which tends to mitigate the impact of large quantities and to aggravate the impact of small ones. Finally, in order to get the final score, the harmonic mean of these four quantities, which we called VACC (VAD Accuracy) are calculate to estimate performance of VAD algorithms.

$$J_{VAD} = \frac{4}{\frac{1}{J_A}+\frac{1}{J_S}+\frac{1}{J_E}+\frac{1}{J_B}} \tag{8}$$

4. EXPERIMENTS

In this section, firstly the drawback of frame-level accuracy and the efficacy of evaluation on boundary are both illustrated, then the comparison between the criterion in the past and the proposed criterion proceeded in actual VAD algorithm are demonstrated in detail. All the experiments are evaluated on a LVCSR task, Switchboard English. And to exam the robustness of the proposed criterion, additional experiment is evaluated on a Chinese continuous speech data set. 13-dimensional PLP features with per-speaker CMN and CVN, along with first and second derivatives were extracted. Cross-word triphone models with 3001 tied-states was used. State alignment for DNN training was generated using a GMM model. A trigram language model which was trained on the transcription of the 2000h Fisher corpus and interpolated with a background trigram model was used for decoding.

4.1. Experiment based on artificial data

To prove the limitation of the accuracy of frame-level, we designed the experiment as following stated. For the correct VAD labels, we first fixed all the ending boundaries of all speech segments, manually changing the beginning boundaries. And we examined the relationship between recognition performance and VAD evaluation. Then a similar procedure aiming at ending boundary, in which the only difference was fixing the beginning boundary and adapting the ending boundary simultaneously, was accomplished. For convenience, a pre-processing that roughly cutting out the speech in the switchboard was adopted, which assuring that there was only one speech segment in one audio file. For each speech segment, we varied the changing distance of boundaries in the range of -500ms (with the beginning or ending of speech segments cut off 500ms) to +500ms (similarly expanded). The result of fixing the ending boundary and modulate the beginning boundary is displayed in figure 1.

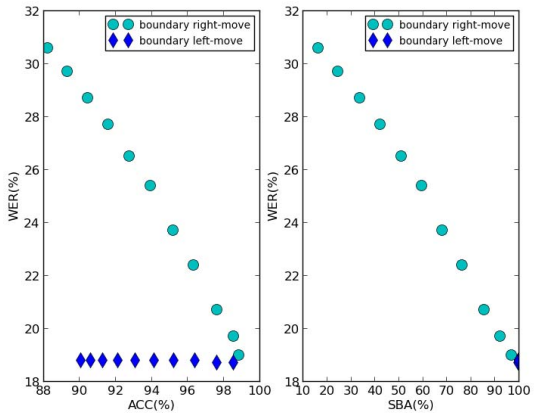

Figure. 1. Start boundary effect on ACC/SBA.

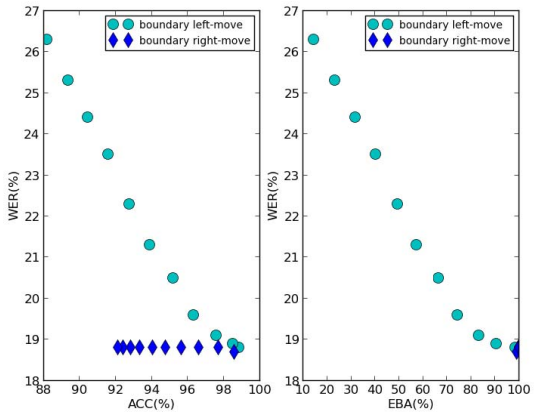

Figure. 2. End boundary effect on ACC/EBA.

In the picture, the dark rhombic dots stand for the results of moving the beginning boundary left-oriented. It can be easily found that the ACC decreases continuously, but the recognition accuracy seems constant. Since left-movement to the beginning boundary is equivalent to extension of the speech segment. Hence, the influence on the final speech recognition is so slight that it can be neglected. However, right-movement denotes the lack of speech information which leads to the declination in the final recognition accuracy. In this case, the ACC can hardly represent the effect of ASR. Later by adopting the approach we proposed, the monotonous relationship between SBA and WER is obvious.

Analogously, when it refers to fixing the beginning boundaries and regularizing the end boundaries, similar results can be obtained easily. As figure 2 shows, one value of ACC corresponds to various WER, while EBA and WER show strong connection in monotony.

On the other hand, considering the influence of the number of segment in VAD output on the performance of ASR, we designed the artificial experiment as following:

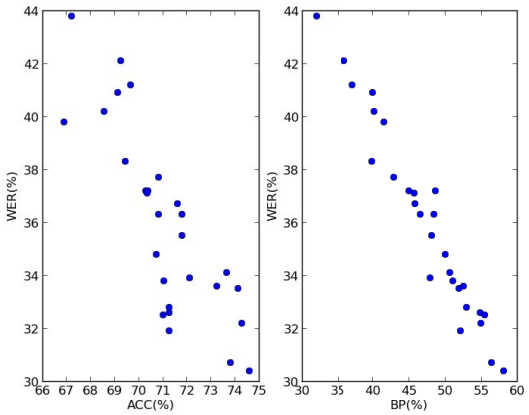

Figure. 3. Segment effect on ACC/BP.

firstly, recognize the speech segment in the reference adopting VAD algorithm, but controlling manually to guarantee that the beginning and the end part of segment are correctly recognized; Besides, stochastically selecting some of the silence segments, a small quantity of which are modified into speech segment. Therefore on premise of guaranteeing the correctness of boundaries, some false speech segments were manually added. The result is shown in figure 3.

It is obvious that both missing segment in speech and the insertion in silence will cause the rise of recognition error. In this case, it can be easily find that compared to ACC, BP owns a closer linear relationship with error rate in speech recognition.

4.2. Real VAD experiment using different parameters for post-processing

From three perspectives, artificial experiments described previously verified the rationality of the criterion proposed and the unsteadiness of frame-level accuracy. When examining the relationship between VAD procedure in real condition and the effect of ASR, the same problem was detected.

In this part, based on 309 hours data from Switchboard, several different frame-level classifiers was trained, including a binary (speech/non-speech) GMM classifier [6] with 128 mixtures, a binary ELM classifier [23] with 12000 hidden nodes and a binary DNN classifier [7] with 1500 hidden nodes in each of the two hidden layers.

To reduce the classified errors, post-processing is adopted in this paper as described following. After obtaining all decisions of frames given by a classifier, re-decided each frame and modify the class of current frame if more than M frames in N following frames are different from it. By varying the value of M and N, different VAD

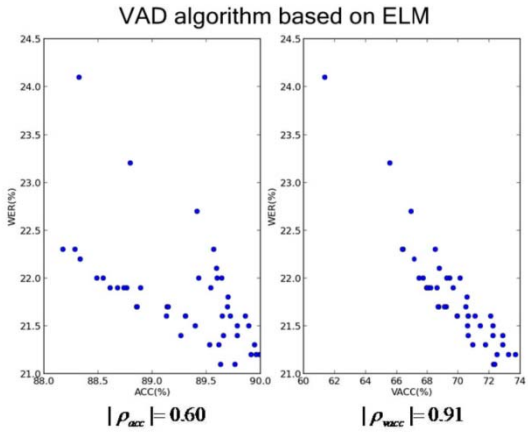

Figure. 4. Performance on ELM.

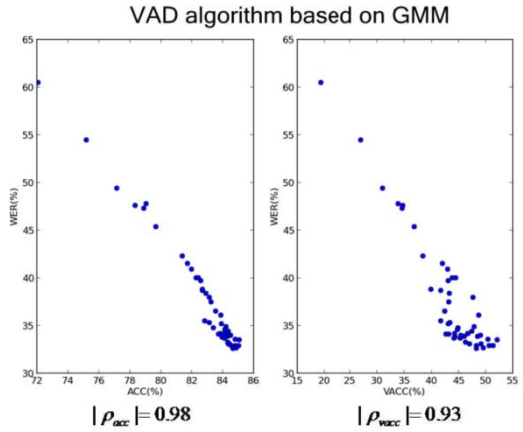

Figure. 6. Performance on GMM.

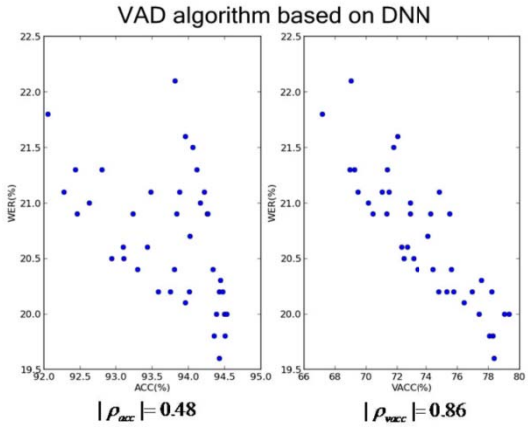

Figure. 5. Performance on DNN.

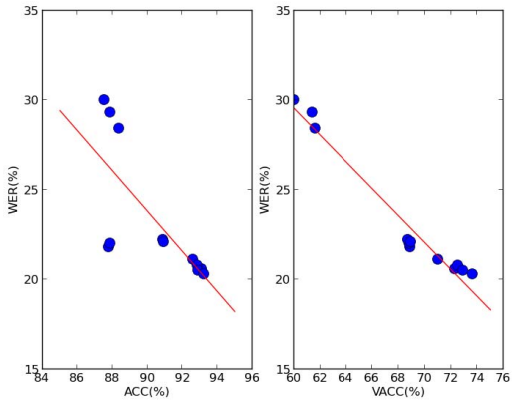

Figure. 7. Correlation-ship using different algorithm.

outputs can be obtained and relationship between VAD outputs and ASR performance was investigated. In the experiment, N varied from 24 to 50 while M was in the range of $\frac{2}{3}M$ to M. In order to investigate the correlation, the absolute value of correlation coefficient was calculated and compared. In this paper, the absolute value of correlation coefficient between frame-level accuracy and WER is defined as $|\rho_{acc}|$, and the absolute value of correlation coefficient between VACC and WER is defined as $|\rho_{vacc}|$. The results are described in the figure 4, figure 5, figure 6 and table 1.

Table 1. Correlation between WER and ACC/VACC.

Correlation	GMM	ELM	DNN
$\|\rho_{acc}\|$	**0.98**	0.60	0.48
$\|\rho_{vacc}\|$	0.93	**0.91**	**0.86**

It can be found that, concerning GMM-based VAD algorithm, frame-level accuracy can sensitively reflect the quality of WER. Tight correlation between them can be found. Meanwhile, so are the correlation between VACC and WER, but slightly inferior to frame-level accuracy. Yet when considering DNN-based VAD algorithm and ELM-based VAD algorithm, the correlation between frame-level accuracy and WER are weaker, and is hard to judge the recognition performance by ACC. However, even if under the circumstance that ACC does not work, close correlation-ship between VACC and WER can still be found. It convinced us that there is a consistent correlation between VACC and WER across different VAD algorithms.

4.3. Real VAD experiment using different algorithms

In research of VAD in real world, there are usually limited adjustments to the parameters of post-processing. More concentration is paid on the parameters on algorithm itself and the comparison between various algorithms. Hence, the following experiment was carried out.

In order to test the correlation between VAD performance and recognition error rate using different

algorithms, various but efficient models for VAD were trained and evaluated, including GMM models with different number of mixtures (128 mixtures, 256 mixtures, 512 mixtures), DNN models with different complexity (1 hidden layer with 1500 nodes, 1 hidden layer with 1500 nodes and 500 nodes pruned [24], 2 hidden layers with 1500 nodes in each, 2 hidden layers with 1500 nodes in each and 1200 nodes pruned, 2 hidden layers with 1500 nodes in each and 1500 nodes pruned), ELM models with different number of hidden nodes (5000 nodes,6000 nodes, 13000 nodes, 14000 nodes). To all of these algorithms, the empirically observed post-processing parameters were used, which valued 40 and 30 respectively.

The result is described in the figure 7. It can be found that for different VAD algorithms, $|\rho_{vacc}|$ is much higher than $|\rho_{acc}|$. The former one is up to 0.974, while the latter one is only 0.716. Hence, compared to ACC, VACC is a more credible criterion to measure the VAD algorithm in speech recognition.

4.4. VAD experiment on different noise conditions

In previous experiment, there is an assumption that little extension of the speech segment has little or no influence on the final speech recognition. It is apparently true when the data is clean with high Signal Noise Ratio (SNR). Therefore the proposed criterion works well on clean condition. In order to figure out whether the proposed criterion is robust in different SNRs, follow experiments were conducted.

In this section, a Chinese continuous speech data set was used as test set. The Chinese data set is a corpus with 132 WAVs collected in real environment through recorder in mobile phones. Each WAV lasts about 2 minutes in average. Subway noise was added to the corpora to gain test set with different SNRs (5db, 10db, 15db, 20db). A binary DNN classifier with 1024 hidden nodes in each of the three hidden layers was used as the VAD recognizer for Chinese corpus. As the same, different VAD outputs were obtained by varying the value of post-processing parameters *M* and *N*, and relationship between VAD outputs and ASR performance was investigated. $|\rho_{acc}|$ and $|\rho_{vacc}|$ were examined through the experiment and the results are shown in table 2.

Table 2. Correlation between CER and ACC/VACC in Chinese set

Correlation	No noise	20db	15db	10db	5db
$\|\rho_{acc}\|$	**0.97**	**0.96**	0.88	0.46	0.77
$\|\rho_{vacc}\|$	0.93	**0.96**	**0.94**	**0.88**	**0.84**

Table 2 lists the correlation based on Chinese speech corpus. From the table, it can be seen clearly that close correlation-ship between VACC and WER can still be found even if in noisy condition. However frame-level accuracy is not a stable measure to evaluate VAD performance for ASR. Furthermore, after analysis the ASR results, we found that more deletion error occurs, rather than insertion error at the beginning or the end of the speech segment, with the decrease of SNR. That may be a reason why the proposed criterion still works in noisy condition.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a new criterion for VAD performance evaluation is proposed considering both frame-level correctness and boundary-level accuracy. By both artificial experiment and real experiment, the new criterion shows greatness of correctness and steadily in performance and is better to use in automatic speech recognition.

In the future, we hope to apply these ideas in different areas such as speaker verification. Also more experiments should be done to figure out whether other weighting functions can get a better performance in different environment.

6. ACKNOWLEDGEMENT

This work was supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the China NSFC project No. 61222208 and JiangSu NSF project No. 201302060012.

REFERENCES

[1] K. H. Woo, T. Y. Yang, K. J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[2] J. C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer," in *Second European Conference on Speech Communication and Technology*, 1991.

[3] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, no. 4, pp. 377–380, 1992.

[4] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[5] N. Mesgarani, M. Slaney, and S. A. Shamma,

"Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[6] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesel ` y, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program." in *Proc. Interspeech*, 2012.

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

[8] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics*. Wiley New York, 1966, vol. 1974.

[9] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. ICASSP*, 1989.

[10] N. Kitaoka, K. Yamamoto, T. Kusamizu, S. Nakagawa, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, and Y. e. a. Denda, "Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance," in *Proc. ASRU*, 2007.

[11] F. Beritelli, S. Casale, and G. Ruggeri, "A psychoacoustic auditory model to evaluate the performance of a voice activity detector," *Signal processing*, vol. 80, no. 7, pp. 1393–1397, 2000.

[12] C. Breslin, M. Gasic, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young, "Continuous ASR for flexible incremental dialogue," in *Proc. ICASSP*, 2013.

[13] T. V. Pham, C. T. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *International Conference on Computing and Communication Technologies RIVF'09*, 2009.

[14] J. P. Egan, *Signal detection theory and ROC-analysis*. Academic Press, 1975.

[15] J. A. Swets, "The relative operating characteristic in psychology," *Science*, vol. 182, no. 4116, pp. 990–1000, 1973.

[16] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc.EuroSpeech*, 1997.

[17] D. Reich, F. Putze, D. Heger, J. Ijsselmuiden, R. Stiefelhagen, and T. Schultz, "A real-time speech command detector for a smart control room," in *Proc. Interspeech*, 2011.

[18] Mehta, C. K. Pham, and C. E. Siong, "Linear dynamic models for voice activity detection." in *Proc. Interspeech*, 2011.

[19] A. V. Ivanov and G. Riccardi, "Automatic turn segmentation in spoken conversations." in *Proc. Interspeech*, 2010.

[20] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, 2010.

[21] H. Ghaemmaghami, B. J. Baker, R. J. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proc. Interspeech*, 2010.

[22] Y. Mamiya, J. Yamagishi, O. Watts, R. A. Clark, S. King, and A. Stan, "Lightly supervised GMM VAD to use audiobook for speech synthesiser," in *Proc. ICASSP*, 2013.

[23] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[24] T. X. He, Y. C. Fan, Y. M. Qian, T. Tan, and K. Yu, "Reshaping deep neural network for fast decoding by node-pruning," in *Proc. ICASSP*, 2014.