A Novel and Efficient Voice Activity Detector Using Shape Features of Speech Wave

Qiming Zhao, Yingchun Yang*, and Hong Li

Zhejiang University,
College of Computer Science and Technology, Hangzhou, China
{zqm1111,yyc,lihong}@zju.edu.cn

Abstract. A voice activity detector (VAD) is the prerequisite for speaker recognition in real life. Currently, we deal with the VAD problem at the frame level through short time window function. However, when tackling with the VAD problem manually, we can easily pick out the speech segments containing several words. Inspired by this, we firstly use IIR filter to get the envelope of the waveform and divide the envelope into separate sound segments. And then we extract shape features from the obtained segments and use K-means to cluster the data featured by the amplitude of the wave crest to discard the silent part. Finally, we utilize other shape features to discard the noise part. The performance of our proposed VAD method has apparently surpassed the energy-based VAD and VQVAD with a relative 20% decrease in error rate, While the computation time of the proposed VAD method is only 30% less than that of VQVAD. We also get an encouraging result utilizing our VAD method for speaker recognition with about 3% average decrease in EER.

Keywords: Speaker recognition, Shape feature, GMM-UBM, VAD, Speech wave.

1 Introduction

Voice activity detection (VAD) is essential in speech processing system, which is to locate the speech segments in an utterance. Currently the VAD methods can be grouped into two types. One type is energy-based VAD, which is intuitively simple. Zero-crossing rate and short-time energy are computed with the assumption that speech frames have relatively higher energy than nonspeech frames. Assigned a threshold relative to maximum or average energy of the utterance, speech frames can be distinguished from the nonspeech frames. But this type of VAD has a well-known shortcoming of sensitivity to additive noise. So before using the energy-based VAD, some speech enhanced processing is necessary, e.g spectral subtraction and Wiener filter. The other type is model-based VAD, using statistical model to do the speech nonspeech classification. Speech model and noise model should be separately trained beforehand and classification is done by comparing the score of frame given speech

[*] Corresponding author.

Z. Sun et al. (Eds.): CCBR 2014, LNCS 8833, pp. 375–384, 2014.

model and noise model. The main bottleneck is that there are too many kinds of noise and other unknown factors that affect the speech and noise modeling results. In these methods, some frequency statistics features are extracted through Fourier transform. The utilized models include SVM, HMM, GMM, etc. [5].

All these methods deal with the VAD problem at the level of frame [1], which is computed by the window function. With the frame, many kinds of acoustic features can be extracted for speech / non-speech frame discrimination. But how do we human do the VAD job manually? We never do it at the level of frame.

We just treat some adjacent frames as a unit which we call the speech segment. The speech segment may contain several words which are very close (liaison phenomenon), or just one, such as “hello”, “bye”. The speech segment can be represented by the envelope of voice part which reflects the way people speak. For example, higher the volume of speaking is, the higher the peak of the envelop is. There may be several waves in a speech segment which we call `slice'. One of the reasons that we do the VAD job in the level of segments is that the interval between words is important which can reflect a person’s speech custom in some way.

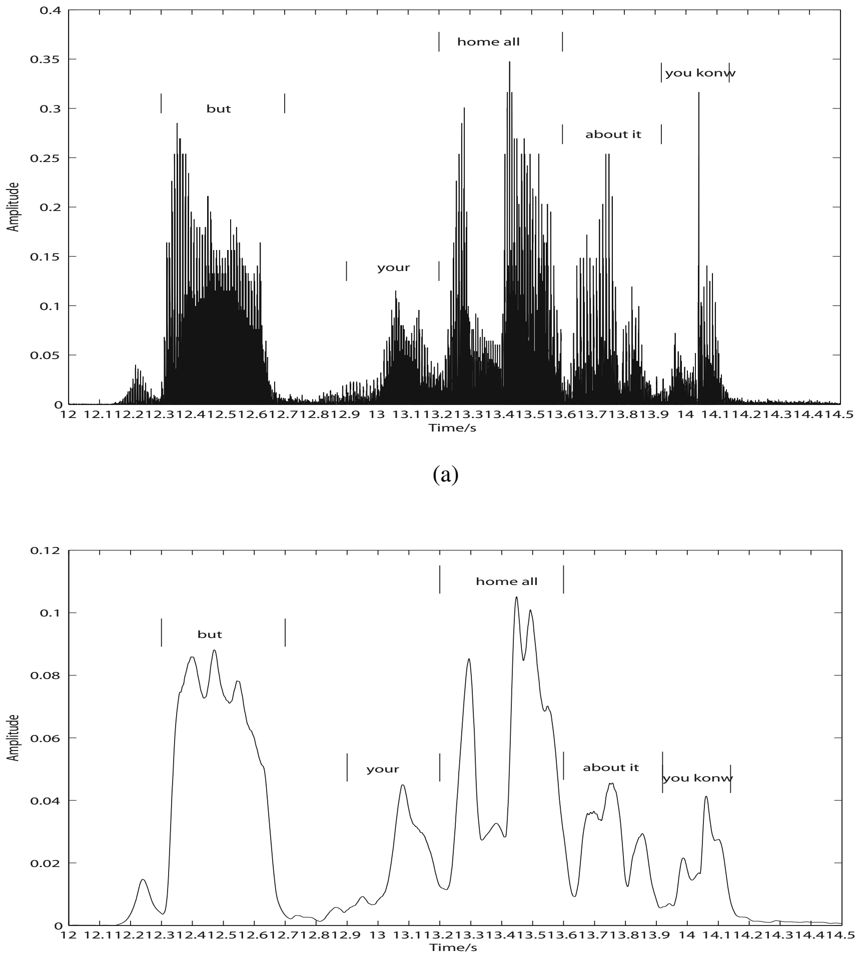
Here is an example of the speech segment. The speech content is “But your home all about it you know”, and there are two liaisons in it (“home all” and “about it”). So we have five speech segments in this speech, which are “But”, “your”, “home all”, “about it”, and “you know”.

When we see the waveform of a speech recording, in our intuition we recognize the segment with high amplitude of the wave crest as speech segment in the case of quiet condition, and it is in truth. We don't need any exercise before, and we can be a good voice activity detector easily.

But if you just look at local part of a whole speech waveform, we can't distinguish whether it is a speech segment or not. The waveform of a speech segment and a noise segment is similar. In fact, they both contain a certain amount of sample points which constitute a waveform. We are confused by the details.

As the way human do the VAD problem, on the basis of shape and trend of the envelope, the algorithm can locate the segments and find the differences between speech segments and noise segments. So the proposed VAD method is designed to use the shape feature to represent speech segment, and use statistic method to find the difference between the form characteristic of speech segments and noise segments.

When we get the envelope of the speech, if the peak of one segment is high, this segment has big chance to be a speech segment, or a noise segment, which both have high energy. Anyway, we detect the segments with high energy at first. Then we find the differences between noise segments and speech segments, for example, the length of the sound of a ring is short and the height-width ratio of it is also different from that of a speech segment. We can use these useful information to detect the noise segments. Besides, if the area of the sound segment is big or the average amplitude of the sound segment is high, this waveform is possible to be a speech segment. The length of speech interval is also useful. Although these shape features are similar, the differences with each other can help to discard the noise segments.



(a)

(b)

Fig. 1. (a) original speech wave (b)envelop of speech wave

Because an absolute quiet environment is always seldom, the amplitude of the `silent' part is not zero. Therefore, we may get a lot of tiny speech interval. We can combine the adjacent speech segments together, on the base that a complete segment has similar amplitudes in the beginning and end of the segment. Suppose the speaking rate of human is 200 words per minute which is very quick, the time of speaking one word is 0.3s. We can discard the segments with too short length.

2 Voice Activity Detectors

The ideal environment where we do the VAD problem is that all speech signals are clear without background noise. In fact, we can't avoid the interference of noise. So we will need speech enhancement methods to reduce the influence of noise. We use both Wiener filter and Spectral subtraction to compare their results.

The Wiener filter method designs a digital filter base on a minimum mean-square error criterion. After collecting noise and speech signal with noise, we subtract the noise component of the amplitude spectrum from the amplitude spectrum with noise, add the noise speech spectrum phrase, and get the enhanced speech signal after inverse Fourier converting. The basic idea of the Spectral subtraction is to suppress the additive noise in the corrupt speech signals under low signal-to-noise (SNR) condition. The estimate of the original and clean signal spectrum is obtained by subtracting an estimate of the noise power (or magnitude) spectrum from the noisy signal. The detailed methods can be found in [2][3].

Despite the background noise usually appears in the speech, there are still some clean speech signals which don't need speech enhancement. So we use a SNR threshold to judge whether this speech signal is clean or not.

We get the speech signal's SNR through an estimation method. Taking Wiener Filter for example. The formula is below. The speech signal after Wiener filter (sy_i) is taken as the clean speech signal. The noise signal is the difference between the original speech signal (sx_i) and the speech signal after Wiener filter.

$$SNR = 10 \times \log_{10} \frac{\sum_{i=1}^{m} sx_i^2}{\sum_{i=1}^{m} (sx_i - sy_i)^2} \tag{1}$$

This SNR threshold is relative to the method of speech enhancement. The value of SNR threshold T is chosen as 9.2dB when using Wiener Filter. When using spectral subtraction, the SNR threshold is more appropriate to be 22db.

If the SNR of a speech signal is below T, we use the speech signal after Wiener filter to replace the original speech signal. After this step, we get the clean speech signal without background noise. The shape features are extracted from envelop of the clean speech signals. The first feature is the biggest amplitude of peak among the sound segments (E_{peak}). The second one is the area of the sound segment (Area). The length of a sound segment (Len) and the ratio between the length and height ($R_{\frac{Len}{E_{peak}}}$) of a sound segment are the third and the fourth feature, respectively. And the mean amplitude of all sample points in a sound segment is the last feature (MeanE).Then we can use K-means to classify the five shape feature and pick out the speech segments. If the voting number of the five K-means result is bigger than the voting threshold, this segment is a speech segment.

Here is the pseudocode of the proposed VAD method.

Input: Speech signal sx_i Outputs: VAD lables

```
1. // Speech enhancement of the speech signal
   $sy_i$ ←speech enhancement
2. // Compute SNR of the speech signal
   SNR=Eq(1)
   If SNR<T //T is the SNR threshold
   $sx_i \leftarrow sy_i$
3. //Calculate the envelope of the speech signal
   $S_{env} \leftarrow$ IIR filter( $sx_i$ )
4. //Get slice of the envelop, the part between two minimum amplitudes is a slice
   $P_{min} \leftarrow$ FindMinimum( $S_{env}$ ) //positions of all minimum amplitudes
   $Seg_{slice} \leftarrow [\, P_{min}[j], P_{min}[j+1] \,]$
5. //Combining adjacent slices, get segments
   $E_{peak} \leftarrow S_{env}[P_{peak}]$ //Amplitude at the peak of a slice, $P_{peak}$ is position of the peak of a slice
   $E_{start} \leftarrow S_{env}[P_{start}]$ //Amplitude at the the start position of a slice, $P_{start}$ is start position of a slice
   While (abs( $E_i$ - $E_{start}$ )> $E_{peak}$*0.1)//amplitudes of start and end of a segment isn't symmetry
   $E_i \leftarrow S_{env}[P_{end_i}]$ //Amplitude at the end position of adjacent slice, $P_{end_i}$ is the position of the end of the ith adjacent slice
   i++ //next adjacent slice
   $P_{end} \leftarrow P_{end_i}$ //the position of the end of this segment
   $Seg \leftarrow (P_{start}, P_{end})$
6. //Extract shape features from segments
   X ←ExtractShapeFeature( $E_{peak}$, Area, Len, $R_{Len/Height}$ , MeanE)
   $E_{peak} \leftarrow$ MaxPeak( $Seg$ ) //biggest peak of the slices in a segment
   Area ← $\int_{P_{start}}^{P_{end}} S_{env} dt$ //using difference to compute the area of a segment
   Len ← $P_{end} - P_{start}$ //length of a segment
   $R_{Len/Height} \leftarrow \frac{Len}{E_{peak}}$ //ratio of length versus height of a segment
```

```
    MeanE ← $\frac{1}{n}\sum_{P_{start}}^{P_{end}} S_{env}$ //mean of all amplitudes in a segment, n is the number
    of sample points in the segment
7.  //For all segments, pick out the speech segments
    $X_i$ ←Sort the ith shape feature from large to small, i=1,2,3,4,5
    $lable_i$ ←K-means( $X_i$ ,5)//classify $X_i$ into 5 parts
    $lable_i$ =0 if the centroid of the cluster which this segment belongs to is smallest
    $lable_i$ =1 if the centroid of the cluster which this segment belongs to isn't smallest
     $Voting$ ←Sum( $lable_i$ )
     if $Voting[j] > VotingThreshold$ , $Seg_j$ is speech segment
8.  //Combine the adjacent speech segments with interval less than 0.1s
    $Seg$ ←Combine( $Seg$ )
```

3 Experiment Setup

We conduct two experiments. First, we get the optimized parameters for the proposed VAD algorithm. Then, we use the endpoint information from the optimized VAD in a speaker verification system.

3.1 VAD Development Set

We use the data of the NIST 2004, 2006 speaker recognition evaluation (SRE), with supplementary automatic speech recognition (ASR) transcripts provided by NIST, as our datasets. We only conduct VAD experiments in three parts of these data. They are the telephone and microphone recordings of the male part of the NIST data. The three parts are telephone recordings of the train part in the NIST 2004 speaker recognition evaluation (04-train-tel), telephone recordings of the train part in the NIST 2006 speaker recognition evaluation (06-train-tel) and telephone recordings of the test part in the NIST 2006 speaker recognition evaluation (06-test-tel). To ensure that the reproduction of your illustrations is of a reasonable quality, we advise against the use of shading. The contrast should be as pronounced as possible.

The VAD experiments are conducted in three methods, the proposed VAD, VQVAD [4] and Energy VAD. The Energy VAD is from open source platform ALIZE [10]. The feature in VQVAD and Energy VAD is 16 MFCCs and 1 energy followed by delta (34 dimensions) extracted with 20ms frame length at a 10ms frame rate.

The methods to evaluate the performance of VAD method is in [4]. The accuracy of a VAD is evaluated by comparing the predicted VAD labels with a clean reference segmentation obtained from the ASR transcripts provided by NIST. Let

$y_t(n) \in \{0,1\}$ and $Y_t(n) \in \{0,1\}$, respectively, denote the predicted and ground truth VAD label of frame t in file n, and let $T(n)$ denote the total number of frames in utterance n. $S1(n)$ denote the total number of points of the predicted endpoint in utterance n. $S2(n)$ denote the total number of points of the ground truth VAD endpoint in utterance n. Our primary metric for VAD tuning is average total error rate (ε),

$$\varepsilon = \frac{1}{N_{utt}} \sum_{n=1}^{N_{utt}} \frac{1}{T(n)} \sum_{t=1}^{T(n)} I\{y_t(n) \neq Y_t(n)\} \tag{2}$$

where $I\{\bullet\}$ is an indicator function and N_{utt} is the number of utterances.

3.2 Speaker Recognition Experiments

We use the base GMM-UBM model to perform the speaker recognition experiment. If the EER of these experiment using different kinds of endpoint provided by VQVAD, Energy VAD and proposed VAD is different, the endpoint information which results in lower EER is better.

To test the robustness of the proposed VAD method, we conduct Speaker recognition experiments in three different datasets. The first dataset is the data used in the VAD performance test before. The second dataset is telephone data of the train part in the NIST 2008 speaker recognition evaluation (08-train-tel), telephone data of the train part in the NIST 2010 speaker recognition evaluation (10-train-tel) and telephone data of the test part in the NIST 2010 speaker recognition evaluation (10-test-tel). The third dataset is microphone data of the test part in the NIST 2008 speaker recognition evaluation (08-test-mic), microphone data of the train part in the NIST 2010 speaker recognition evaluation (10-train-mic) and microphone data of the test part in the NIST 2010 speaker recognition evaluation (10-test-mic). The feature is as same as that in VAD experiment using VQVAD and Energy VAD. We use gender-dependent UBM containing 512 Gaussians by Expectation Maximize (EM) algorithm.

The experiment is conducted in personal computer, in which the CPU is Core i3-2130 (3.3GHz), and the RAM is 8GB DDR3.

Table 1. Experiment dataset: Column 1 is the name of the experiment, Column 2 is the data trained for universal background models (UBMs), Column 3 represents the data trained for the speaker UBM model, column 4 gives the data for testing.

Experiment	UBM	Train	Test
04-06-tel	04-train-tel	06-train-tel	06-test-tel
08-10-tel	08-train-tel	10-train-tel	10-test-tel
08-10-mic	08-test-tel	10-train-mic	10-test-mic

4 Result

We get good results in the following two kinds of experiments.

4.1 VAD Experiment

In table 2, with the increasing of the number of the voting, the amount of segments which are seen as noise segments is bigger. If the speech is long enough, even we can set a bigger voting number, we can get enough speech segments to train and test. But if the speech is relatively short, a small voting number is more appropriate.

And in table 3, we can see clearly that the proposed VAD method gets an encouraging result.

Although the time cost of Energy VAD is the least, its performance is the worst. The processing time of the proposed VAD method with Wiener Filter is 16% less than Energy VAD, and its' error rate outperforms the latter by a relative 25% decrease. And the processing time of the proposed VAD method with Spectral Subtraction is 45% less than Energy VAD, while its error rate outperforms the latter by a relative 31% decrease.

Table 2. Proposed VAD (voting number ranging from 1 to 4) with Wiener Filter (WF) and Spectral Subtraction (SS), average error rate (%) in three datasets.

voting	04-train-tel		06-train-tel		06-test-tel	
	WF	SS	WF	SS	WF	SS
1	11	11	10	10	11	10
2	11	11	10	10	11	10
3	11	11	10	10	11	10
4	13	13	12	12	13	12

Table 3. Proposed VAD (voting number:4) with Wiener Filter(WF) and Spectral Subtraction(SS) , VQVAD and Energy VAD , time cost (s) and average errors (%) in three datasets.

Type	Prop.(WF)		Prop.(SS)		VQVAD		Energy VAD	
	Time	Error	Time	Error	Time	Error	Time	Error
04-train-tel	3.52	13	2.38	**11**	4.32	16	**0.86**	20
06-train-tel	3.70	12	2.37	**10**	4.36	16	**0.82**	19
06-test-tel	3.64	13	2.39	**10**	4.38	17	**1.04**	19

4.2 Speaker Recognition Experiment

The speaker recognition results (male trials only), in terms of equal error rate (EER, %) using the proposed VAD are shown in Table 4.

The results of the proposed VAD using Wiener Filter (WF) and Spectral Subtraction (SS) have similar trend. We get better performance with the increasing of voting number threshold.

We choose the result of number of voting, 4, as our final result. Now we seemly get two different conclusions in table 2 and 4. But with the increasing of voting number, more and more segments are seen as noise segments which may be speech segments in fact. To guarantee the purity of the speech segments, we choose the result of voting number as 4 as the best result. For the rest of the experiments, we use wiener filter to enhance energy only.

The reason why the result of 04-06-tel is best is because the speech data in 2004 and 2006 is very clean, with little noise.

The comparison of speaker recognition results (male trials only), in terms of equal error rate (EER, %) using the proposed VAD's, VQVAD's and Energy VAD's endpoint is shown in Tables 5.

Except that the result of 08-10-mic is not good enough as that of VQVAD with little gap, the other two results of EER are both less than these of VQVAD and Energy VAD. Especially as we get about 4.5% decrease in the EER of 04-16-tel experiment.

Table 4. Speaker recognition EER: proposed VAD (voting number ranging from 1 to 4) with Wiener Filter (WF) and Spectral Subtraction (SS), average errors (%) in three datasets.

voting	04-06-tel		08-10-mic		08-10-tel	
	WF	SS	WF	SS	WF	SS
1	10.78	**10.46**	**16.87**	17.48	**15.10**	15.44
2	10.78	**10.40**	**16.72**	17.64	15.77	**15.10**
3	1065	**10.33**	**17.18**	18.10	**15.10**	16.11
4	**9.66**	10.46	16.41	16.10	**14.09**	14.43

Table 5. Speaker recognition result (EER, %) on three datasets with three VAD algorithms: proposed VAD with Wiener Filter (Prop.(WF)),VQVAD and Energy VAD

Experiment	Prop.(WF)	VQVAD Energy	VAD
04-06-tel	**9.66**	14.18	18.26
08-10-tel	**16.41**	17.24	20.40
08-10-mic	14.09	**13.76**	17.79

5 Conclusions

The proposed VAD method utilizes the shape features of the speech envelop to detect the endpoint information and somehow discard the noise part. The performance of the method in VAD, and the speaker recognition experiment with endpoint information are both encouraging. Although we get better score in the VAD experiments, we have further work to do when applying the endpoint information to the speaker recognition baseline experiment. The way we get the envelope of the speech record is not the best, we are testing a new method to get the envelope which can better represent the trend of the speech.

Acknowledgment. This work was supported by the National Basic Research Program of China (2013CB329504)

References

1. Haigh, J.A., Mason, J.S.: Robust voice activity detection using cepstral features. In: IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering, pp. 321–324. IEEE Press, New York (1993)
2. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing 27, 113–120 (1979)
3. Vaseghi, S.V.: Spectral Subtraction. In: Advanced Signal Processing and Digital Noise Reduction, pp. 242–260. Springer, Heidelberg (1996)
4. Kinnunen, T., Rajan, P.: A practical, self-adaptive voice activity detector for speaker verification with noise telephone and microphone data. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7229–7233. IEEE Press, New York (2013)
5. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE on Signal Processing Letters 6, 1–3 (1999)
6. Sun, H.W., Ma, B., Li, H.Z.: Frame selection of interview channel for NIST speaker recognition evaluation. In: 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 305–308. IEEE Press, New York (2010)
7. Burget, L., Matejka, P., Schwarz, P.: etal: Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Transactions on Audio, Speech, and Language Processing 15, 1979–1986 (2007)
8. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing 9, 504–512 (2001)
9. VOICEBOX: Speech Processing Toolbox for MATLAB, `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.htmlAppendix`
10. ALIZE: Open Source platform for biometrics authentification, `http://mistral.univ-avignon.fr/index_en.html`