A NEW APPROACH FOR ROBUST REALTIME VOICE ACTIVITY DETECTION USING SPECTRAL PATTERN

M. H. Moattar[1] *and M. M. Homayounpour*[1] *and Nima Khademi Kalantari*[2]

[1] Laboratory for Intelligent Signal and speech Processing, Computer Engineering and Information Technology Department, Amirkabir University of Technology, Tehran, Iran
Email: {moattar, homayoun}@aut.ac.ir
[2] Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran
Email: nimakhademi@aut.ac.ir

ABSTRACT

In this paper a Voice Activity Detection approach is proposed which applies a voting algorithm to decide on the existence of speech in audio signal. For this purpose, the proposed approach uses three different short time features along with the pattern of spectral peaks of every frame. Spectral peaks pattern is appropriate for determining vowel sounds in speech signal even in the presence of noise. Therefore this measure can be applicable in voice activity detection in which the vowels characterize the speech signal. Experiments show that incorporating this measure along with our recently proposed approach for VAD, will improve the results of the algorithm considerably while imposing little computational overhead. The proposed approach is evaluated on different datasets with various noises and SNR levels and satisfying results are achieved.

Index Terms— Voice Activity Detection, Spectral Peaks Pattern, Spectral Flatness

1. INTRODUCTION

Voice Activity Detection (VAD) is a critical task in many speech/audio applications. According to [2], the required characteristics for an ideal voice activity detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no prior knowledge of the noise. The most challenging characteristics of an ideal VAD algorithm are robustness against noisy environment and its computational complexity especially when a real-time application is targeted. The performance of all the VAD algorithms degrades to a certain extent with the decrement in SNR. Most of the previously proposed methods have partially overcome this problem but in exchange have resulted in higher computational complexity. Simplicity and robustness against noise are both essential characteristics of voice activity detection.

Many different VAD approaches have been proposed and the main concern of these methods was robustness of the approach. The difference between most of the previous methods is the features used. Various kinds of robust acoustic features, such as autocorrelation function based features [3], spectrum based features [4], the power in the band-limited region [1, 5], Mel-frequency cepstral coefficients [3], delta line spectral frequencies [5], and features based on higher order statistics [6] have been proposed for VAD. Using multiple features in parallel has lead to more robustness against different noises. In some previous works multiple features are applied in combination with some modeling and decision algorithms such as CART [7] or ANN [8].

Some of the most widely proposed voice activity detection (VAD) methods are statistical pattern classification approaches [9]. These methods require the noise model to be trained in advance using a set of corresponding noisy speech data. This limits their use in unknown noisy environments. Also, most of these methods assume the noise to be stationary during a certain time period, thus they are sensitive to changes in SNR of observed signal or the nature of noise which is common in real world applications. To overcome this shortcoming, some previous works propose noise estimation and adaptation for improving VAD robustness [10], but these kinds of methods are computationally expensive.

In this paper an easy-to-implement and real-time VAD algorithm is proposed which applies a set of short-time features along with spectral pattern of speech frames to discriminate between speech and non-speech frames. The approach is briefly introduced in Section 2. In Section 3, the proposed VAD algorithm is explained in details. Section 4 explains the experiments and discusses the observations. Finally, Section 5 includes a brief conclusion on the achievements of this paper.

2. PROPOSED APPROACH

In a recently published paper by the authors, a Voice Activity Detection algorithm has been proposed which applies a voting scheme on a set of three short-time features for speech/non-speech discrimination [11]. The proposed feature set includes short-term energy, spectral flatness measure and the most dominant spectral component of audio frame. The resulted performance of the proposed algorithm was satisfactory in various noises and SNRs while maintaining low complexity of the total VAD process.

In [12] it is suggested that spectral peaks of audio frames are good measure for discriminating vowel sounds from other sounds including non-speech. In [12] it is claimed and evaluated by experiments that spectral peaks of vowel sounds are robust against the noisy environments even in severe noise conditions and can be successfully applied in voice activity detection. Using this approach, the problem of voice activity detection changes to the problem of detecting the presence of vowels which are very rare in non-speech signal. Figure 1 illustrates the spectrum of the vowel sound in three different SNR and noise conditions. As it can be seen in the figure, the positions of the peaks in the frame spectrum are relatively unchanged for all three cases.

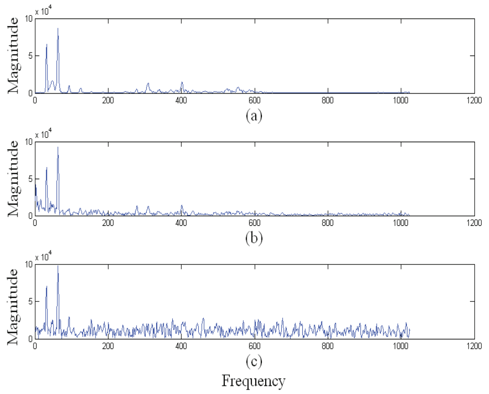

Figure 1. Spectrum of a vowel sound
(a) Clean (b) Corrupted by pink noise (SNR= 5)
(c) Corrupted by white noise (SNR= -5)

Inspired from the mentioned paper, this paper proposes a combinational approach to improve the performance of VAD algorithm in presence of various noises and SNR conditions. Positions of spectral peaks are the most important factor in discriminating vowel sounds from others. In this approach a huge set of patterns of the locations of spectral peaks of various vowel sounds is extracted from a set of training data. In test stage, for every incoming audio frame s, the relevance of the frame to the vowel family V is calculated with the following measure which is called the peak-valley difference (PVD):

$$PVD_s = \max_{X \in V}(PVD(S,X)) \tag{1}$$

$$PVD(S,X) = \frac{\sum_{i=0}^{N-1}\left(X[i]*S[i]\right)}{\sum_{i=0}^{N-1} S[i]} - \frac{\sum_{i=0}^{N-1}\left(X[i]*(1-S[i])\right)}{\sum_{i=0}^{N-1}\left((1-S[i])\right)} \tag{2}$$

In which S is the spectral peaks signature of the input frame and X is a sample pattern of the spectral peaks. Each pattern X is a binary vector containing the peak positions of a vowel sound. In X the positions of spectral peaks is marked by 1 while the other parts are 0. To extract this information, a simple peak detection algorithm can be applied. The PVD calculates the average energy difference between peak and valley bands for an input sound [12]. After computing the PVD for the incoming frame, the frame is marked as a vowel if PVD_s exceeds a decision threshold.

3. VOICE ACTIVITY DETECTION

The proposed approach includes two stages. In the first step (training step), the spectral peaks signatures of vowel sounds (V) are extracted from a set of training data. The training phase of the proposed approach is summarized in the following table.

Table 1. Training phase of the proposed approach

1- A set of vowel segments is extracted from labeled training data.
2- Each segment is divided into 30 ms length frames at 10ms frame rate.
3- The Fourier transform is applied on every frame s.
4- The average spectrum of each sound is calculated.
5- The average spectra of vowel sounds are grouped using *k*-means clustering algorithm.
6- Peak detection is applied on each resulted cluster centriod.
7- The peak signature vectors of vowel sounds, which are described in the pervious section, are extracted and stored for future reference.

The VAD algorithm starts with framing the audio signal. A hamming window is applied on the current speech frame. First N frames are used for threshold initialization. For each frame the features including energy, spectral flatness, the most dominant spectral component and spectral peak-valley difference are computed. The audio frame is marked as a speech frame if more than one of the feature values is beyond the pre-computed threshold. The complete procedure of the proposed method is described below:

Table 2. The proposed voice activity detection algorithm

```
1) Divide the input signal to 30 ms duration frames with 20 ms overlap. Compute the number of frames $nbFrames$.
2) For $i$ from 1 to $nbFrames$
    a) Compute frame energy, $E(i)$.
    b) Apply FFT on each speech frame and compute the frame spectrum, $|S(i)|$.
        I) Find $F(i) = \arg\max_k(|S(k)|)$ as the most
```

```
          dominant frequency component.
      II) Compute the PVD of $|S(i)|$, $PVD(i)$.
      III) Compute the absolute value of Spectral Flatness measure, $SFM(i)$.
  $SFM(i) = 10\log_{10}(G/A)$                                  (3)
          Where $A$ and $G$ are arithmetic and geometric means of $|S(i)|$ respectively.
    c) Supposing that the first N frames are non-speech, find the minimum value for
       $E\,(Min\_E)$, $F\ (Min\_F)$, $SFM\ (Min\_SFM)$ and $PVD\ (Min\_PVD)$.
    d) Set Decision threshold for $E$, $F$ and $SFM$ and $PVD$.
    e) Set $Counter = 0$.
      I) If $E(i) > Thresh_E$ then $Counter++$,
      II) If $F(i) > Thresh_F$ then $Counter++$,
      III)If $SFM(i) > Thresh_{SFM}$ then $Counter++$,
      IV)If $PVD(i) > Thresh_{PVD}$ then $Counter++$,
    f) If $Counter > 1$ then mark the current frame as speech else mark it as silence
    g) If current frame is marked as silence, update the energy minimum value:
  $Min\_E = \frac{(Silence\_Count * Min\_E) + E(i)}{Silence\_Count + 1}$     (4)
    h) Compute the new value of $Thresh_E$
  $Thresh_E = Param_E * \log10(Min\_E)$                        (5)
4- Ignore silence run less than 5 successive frames.
5- Ignore speech run less than 5 successive frames.
```

The decision thresholds in the proposed approach are defined as follows:

$$Thresh_{Feature} = Min\_Feature + Param_{Feature} \qquad (6)$$

In which, $Feature \in \{E, F, SFM, PVD\}$. The optimal values of $Param_{Feature}$ can be found on a set of evaluation speech data so that the total performance of the algorithm on the evaluation data maximizes.

4. EXPERIMENTS

For evaluating the proposed method, we used two different speech corpora. The first one is TIMIT Acoustic-Phonetic Continuous Speech Corpus which is used in training phase as well. In training phase 5000 vowel segments were extracted from the train set of this corpus to determine the spectral signature of the vowel frames. Also we used the test data of this corpus in our test evaluations. The second dataset which is commonly used for evaluating VAD algorithms is the Aurora2 Speech Corpus. The Aurora2 speech corpus includes clean speech data as well as noisy speech. To show the robustness of the proposed method against noisy environments, we added different noises with different SNRs to the clean speech signals. The same evaluations are also performed on the original approaches described in [11] and [12] and the G. 729B VAD algorithm as a reference method.

Two common metrics for VAD performance evaluation are Silence Hit Rate (HR0) and Speech Hit Rate (HR1). To have a better metric for comparing two different VAD algorithms, the mean of HR0 and HR1 is considered as the final evaluation metric (T).

The proposed method is first evaluated on TIMIT database with five different noises. For these evaluations, white, babble, pink, factory and Volvo noises with 25, 15, 5 and -5 db SNRs are added to the original speech data. The result of these evaluations is summarized in Tables 3 and 4.

Table 3: VAD performance on TIMIT database for different noises

Noise Type	Method proposed in [11] (%)	Method proposed in [12] (%)	Proposed method (%)
white	86.27	81.28	87.77
babble	85.40	65.27	79.46
pink	83.22	76.77	84.34
factory	77.21	76.73	88.68
Volvo	72.29	66.38	77.19

In Table 3 the results of evaluations is mentioned according to the noise type. As seen in this table, the performance of the proposed algorithm is relatively higher than the performance of the two other approaches. The only exception is for the babble noise in which the accuracy of the algorithm degrades dramatically. This performance reduction is also observable for the original PVD based method proposed in [12]. The reason of this low performance can be easily described. The two mentioned approaches apply a vowel detection criterion which is vulnerable to the existence of speech like signals such as babble noise, in the original speech signal.

Table 4: VAD performance on TIMIT database for different SNR levels

SNR Level (db)	Method proposed in [11] (%)	Method proposed in [12] (%)	Proposed method (%)
25	96.09	87.41	96.14
15	88.13	83.53	90.73
5	78.26	68.17	80.13
-5	61.02	54.03	64.95

Table 4 shows the average performance of different approaches in various SNR levels independent from the noise type. The proposed approach performs almost the same as the method proposed in [11] in higher SNRs, but outperforms the other methods in the lower SNRs.

The proposed method is also evaluated on Aurora2 speech dataset which contains speech data corrupted with three different noises in 6 SNRs as well as the clean speech signal. The two other methods besides the G. 729B VAD are also evaluated which results are listed in Table 5.

Table 5: VAD performance on Aurora2 speech dataset in %

Method	Process Time per Second of Speech	Subway noise	Babble Noise	Car Noise
G. 729B	0.05s	67.4	67.3	71.5
Method proposed in [11]	0.02s	74.4	79.28	80.2
Method proposed in [12]	0.12s	72.12	79.44	75.04
Proposed method	0.14s	75.67	81.22	82.16

The above table shows an improvement in VAD detection performance when the voting scheme in [11] is applied in parallel with the spectrum peak-valley detection approach in [12] to decide the silence/non-silence parts of an audio segment. It also shows that, the proposed approach is robust against threshold selection because of the parallel fusion of decision rules. Also Table 5 depicts the fact that using the two approaches in parallel causes the resulted method to be more independent from training speech which is used as the reference spectral patterns.

Table 5 also shows the average process time of the evaluated approaches for each second of input audio signal. As shown in the above table the method proposed in [11] is much faster than the other methods especially the methods which apply the spectral patterns of vowel sounds. The processing time difference between these approaches is due to the pattern matching procedure that the later one requires.

5. CONCLUSIONS

This paper has presented an easy-to-implement VAD algorithm which is appropriate for online processing and is robust against the noisy conditions in real world application. The proposed VAD approach is in fact a voting algorithm on 4 different speech/non-speech discriminating features which are used in parallel. From these features peak-valley difference of the speech frame spectrum is a good measure for detecting vowel sounds in audio signal even in presence of noise. As vowel sounds are rarely observable in non-speech segments this feature is a good criterion for detecting speech parts of an audio signal. The main flaw of this approach arises when the audio signal is corrupted by babble noise in which vowel-like frames does corrupt the main speech. This deficiency is partially covered by our proposed voting approach on all four features but the total performance will still be unsatisfactorily lower than the expectations.

6. ACKNOWLEDGMENTS

The authors would like to thank Iran Telecommunication Research Center (ITRC) for supporting this work under contract No. T/500/14939.

7. REFERENCES

[1] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin and J. P. Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," IEEE Communications Magazine 35, pp. 64-73, 1997.

[2] M. H. Savoji, "A robust algorithm for accurate end pointing of speech," Speech Communication, pp. 45–60, 1989.

[3] T. Kristjansson, S. Deligne and P. Olsen, “Voicing features for robust speech detection,” Proc. Interspeech, pp. 369-372, 2005.

[4] R. E. Yantorno, K. L. Krishnamachari and J. M. Lovekin, “The spectral autocorrelation peak valley ratio (SAPVR) – A usable speech measure employed as a co-channel detection system,” Proc. IEEE Int. Workshop Intell. Signal Process, Hungary, pp. 193-197, 2001.

[5] M. Marzinzik and B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” IEEE Trans. Speech Audio Process, 10, pp. 109-118, 2002.

[6] K. Li, N. S. Swamy and M. O. Ahmad, “An improved voice activity detection using higher order statistics,” IEEE Trans. Speech Audio Process., 13, pp. 965-974, 2005.

[7] W. H. Shin, "Speech/non-speech classification using multiple features for robust endpoint detection," In Proceeding of ICASSP 2000, pp. 1399-1402, 2000.

[8] G. D. Wuand and C. T. Lin, "Word boundary detection with Mel scale frequency bank in noisy environment," IEEE Trans. Speech and Audio Processing, vol. 8, no. 5, pp. 541-554, 2000.

[9] A. Davis, S. Nordholm, and R. Togneri, “Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold,” IEEE Trans. Audio, Speech, and Language Process., vol. 14, no. 2, pp. 412-424, 2006.

[10] B. Lee and M. Hasegawa-Johnson, "Minimum Mean Squared Error A Posteriori Estimation of High Variance Vehicular Noise," in Proc. Biennial on DSP for In-Vehicle and Mobile Systems, Istanbul, Turkey, June 2007.

[11] M. H. Moattar and M. M. Homayounpour, “A Simple but Efficient Real-Time Voice Activity Detection Algorithm,” Eusipco 2009, Glasgow, Scotland, pp. 2549-2553, 2009.

[12] I.C. Yoo and D. Yook, “Robust Voice Activity Detection Using the Spectral Peaks of Vowel Sounds,” ETRI Journal, Volume 31, Number 4, pp. 451-453, August 2009.