High-Accuracy, Low-Complexity Voice Activity Detection Based on A Posteriori SNR Weighted Energy

Tan, Zheng-Hua; Lindberg, Børge

Published in:
Proceedings of the International Conference on Spoken Language Processing

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2009 BRIGHTON

High-Accuracy, Low-Complexity Voice Activity Detection Based on *A Posteriori* SNR Weighted Energy

Zheng-Hua Tan, Børge Lindberg

Multimedia Information and Signal Processing (MISP), Department of Electronic Systems, Aalborg University, Aalborg, Denmark

zt@es.aau.dk, bli@es.aau.dk

Abstract

This paper presents a voice activity detection (VAD) method using the measurement of *a posteriori* signal-to-noise ratio (SNR) weighted energy. The motivations are manifold: 1) the difference in frame-to-frame energy provides a great discrimination for speech signals, 2) speech segments, besides their characteristics, are accounted also on their reliability e.g. measured by SNR, 3) the *a posteriori* SNR for noise-only segments will theoretically equal to 0 dB, being ideal for VAD, and 4) both energy and *a posteriori* SNR are easy to estimate, resulting in a low complexity. The method is experimentally shown to be superior to a number of referenced methods and standards.

Index Terms: voice activity detection, SNR, energy, noise

1. Introduction

Voice activity detection is an important component in many real-world speech application systems [1]. For example, speech coding schemes for mobile communication and Voice-over-IP generally include VAD to avoid unnecessary processing and transmission of silence regions and thus save on computation and on network bandwidth. Speech enhancement and speech recognition systems can take advantage of VAD to improve performance. It is, however, a challenging task to develop VAD methods that are accurate in both clean and noisy environments.

VAD attempts to detect the presence or absence of human speech in a segment of an acoustic signal. In general, the detection is realized in two key steps: First, some features are calculated from a segment of the acoustic signal; secondly, a classifier is applied to the features to categorize the segment as speech or non-speech.

A variety of features have been used for VAD decisions. Classical features are energy and zero crossing rate. Recently more sophisticated features such as entropy [2] and Mel-filter bank outputs [3] have been proposed.

For the classification, techniques such as support vector machines [4], Gaussian mixture models [5] and decision trees [6] have been used. The simplest technique is a threshold based approach in which the decision is made by comparing the calculated value(s) against certain threshold(s).

The ETSI advanced front-end [7] includes two VAD algorithms. The first one is energy based and is used for noise estimation, while the second marks each 10 ms frame in an utterance as speech/non-speech so that the information can be used for frame dropping at the server recognizer [8]. Only the second algorithm is further analyzed in this paper. It has two stages: a frame-by-frame stage consisting of three measures (whole spectrum, spectral sub-region and spectral variance) and a decision stage analyzing the pattern of buffered measurements for making VAD decision.

The G.729 VAD algorithm uses the following features: full- and low-band frame energy, a set of line spectral frequencies and the frame zero-crossing rate [9]. The G.723.1 VAD algorithm compares the energy of the inverse filtered signal with a threshold [10].

The key parameters for a VAD technique are accuracy, latency and complexity. Complexity is important since VAD applies to various applications which often involve low-resource devices.

The ETSI advanced front-end VAD performs very well in noisy environments, but very poor in clean condition as shown in the experiments on the Aurora 2 database detailed in Section 3. A comparison in [11] also shows that the advanced front-end VAD is primarily suitable for stationary noise environments. The Mel-filter bank outputs based VAD [3] is highly accurate for clean speech, but its performance in noisy environments is worse than the advanced front-end VAD in terms of frame error rate. Both are significantly superior to the G.729 and G.723.1 VAD algorithms.

This paper proposes a high-accuracy and low-complexity VAD method that performs well in both clean and noisy environments. The method uses *a posteriori* SNR weighted energy as the feature and a threshold for decision-making. The feature was proposed by the authors and applied to frame selection in variable frame rate analysis in [12]. The feature has shown to be able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like vowels and no frames to silence, even for very low SNR signals. In this paper the selected frames are further processed for speech/non-speech classification, leading to an effective VAD method. A number of experiments are conducted to investigate the accuracy and latency of the proposed methods.

This paper is organized as follows. Section 2 presents the *A posteriori* SNR weighted energy based VAD method. Section 3 conducts experiments on the Aurora 2 database and compares the proposed method against several existing VAD methods and standards. The paper is concluded in Section 4.

2. *A posteriori* SNR weighted energy based VAD

Voice activity detection aims at classifying an input signal into speech and non-speech segments. This objective is shared by variable frame rate (VFR) analysis. In general VFR analysis selects frames according to signal characteristics with the goal of selecting more frames for speech segments, especially for segments containing fast changing speech events, and less or no frames for non-speech segments. Mostly, VFR analysis extracts features for each frame at a fixed-frame-rate first and then uses a certain criterion to retain or omit frames. The decision is done by calculating some distance measure and comparing it with a threshold.

In [12], we present a variable frame rate method based on *a posteriori* SNR weighted energy, which has been shown to be superior to cepstral feature based VFR method and has demonstrated a soft VAD property. In this work the *a posteriori* SNR weighted energy is used as the feature for a VAD method as detailed in the following subsection.

2.1. The algorithm

Note that the VAD method is not purely energy based. Rather, it is based on the combination of accumulative energy distance and SNR. Further the method first conducts frame selection as done in VFR analysis and then applies VAD decision on the frame selection results. Alternatively, the *a posteriori* SNR weighted energy can be used for VAD decision directly.

The algorithm of the method is as follows:

1. Compute the *a posteriori* SNR weighted energy distance of two consecutive frames as

$$D(t) = |\log E(t) - \log E(t-1)| \cdot SNR_{post}(t) \quad (1)$$

where log $E(t)$ is the logarithmic energy of frame t, and $SNR_{post}(t)$ is the *a posteriori* SNR value of frame t by using a 1 ms frame shift and a 25 ms frame length. *A posteriori* SNR is defined as the logarithmic ratio of the energy of noisy speech to the energy of noise.

2. Compute the threshold T for frame selection as

$$T = \overline{D(t)} \cdot f(\log E_{noise}) \quad (2)$$

where $\overline{D(t)}$ is the average weighted distance over a certain period (in this work, it is calculated over one utterance; in practice, $\overline{D(t)}$ calculated over preceding segments can be used and it is then updated frame-by-frame). $f(\log E_{noise})$ is a sigmoid function of $\log E_{noise}$ to allow a smaller threshold and thus a higher frame rate for clean speech, and it is defined as $f(\log E_{noise}) = 9.0 + \frac{2.5}{1 + e^{-2(\log E_{noise} - 13)}}$ where the constant of 13 is chosen so that the turning point of the sigmoid function is at *a posteriori* SNR of between 15 and 20 dB.

3. Update the accumulative distance: $A(t) += D(t)$ on a frame-by-frame basis and compare it against the threshold T: If $A(t) > T$, the current frame is selected and $A(t)$ is reset to zero; otherwise, the current frame is discarded. If the current frame is not the last one, the search continues, that is, go back to step 1.

4. Apply a moving average on the selected frames (as detailed in Subsection 2.2) and compare each average value $M(n)$ against a threshold T_{vad}: If $M(n) > T_{vad}$, the current frame is classified as speech; otherwise, the current frame as non-speech.

The use of *a posteriori* SNR, rather than *a priori* SNR, avoids the problem of assigning zero or negative weights to frames with $SNR_{prio} \le 0dB$ and subsequently discarding them due to their non-positive weights. As such, the *a posteriori* SNR weight for noise-only frames will be theoretically equal to 0 dB, making it ideal for VAD; negative *a posteriori* SNR values may still appear in practice and are then set to zero to prevent negative weights. In this work E_{noise} for calculating $SNR_{post}(t)$ and $\log E_{noise}$ for calculating T are both simply estimated by averaging the first 10 frames (frame shift being 1 ms) of an utterance which are considered noise only.

As only the logarithmic energy and the *a posteriori* SNR value are calculated for each frame, the method has a very low complexity.

2.2. VAD decision

Based on the frames selected in Step 3 in the VAD algorithm, there are several ways to derive VAD decisions. Two approaches are presented as follows.

2.2.1. Approach 1

As the frame selection is conducted on the basis of a 1 ms frame shift, an ordinary 10 ms frame shift is applied to the frame selection output to derive 10 ms frame labels: If there are one or more selected frames contained in the 10 ms window, the current frame is assigned as 1; otherwise, it is 0.

Next, a moving average is applied to the generated sequence of 0s and 1s, i.e. the 10 ms frame labels.

In general, for the time series $x_1, x_2, ..., x_N$, the m-point moving average replaces the value of x_n with

$$M(n) = \frac{1}{m_1 + m_2 + 1}(x_{n-m_1} + ... + x_{n-1} + x_n + x_{n+1} + ... + x_{n+m_2}), \quad n = m_1 + 1, ..., N - m_2 \quad (3)$$

It is a central moving average when $m_1 = m_2$, a prior moving average when $m_2 = 0$, and a biased moving average when $m_1 \neq m_2$.

In this work, zeros are appended to both sides so that the moving average can be calculated from $n = 1$ until $n = N$. The latency of the VAD method is controlled by adjusting m_2 and the time series $x_1, x_2, ..., x_N$ is the 10 ms frame labels.

The output of the moving average $M(n)$ is compared against a threshold T_{vad}: If $M(n) > T_{vad}$, the current frame is classified as speech; otherwise, the current frame as non-speech.

2.2.2. Approach 2

This approach applies a moving average directly to the frame selection results in order to take advantage of the distribution of the selected frames – a higher frame density in fast changing events, a lower frame density in steady regions and an even lower frame density in non-speech part, achieved by the frame selection scheme.

The moving average $M(n)$ is calculated on the basis of a 10 ms frame shift and is measured as the average number of frames within the moving average window as follows.

$$M(n) = \frac{1}{m_1 + m_2 + 1} \sum_{m=-m_1}^{m_2} frame\_selection(\alpha \times (n + m)) \quad (4)$$

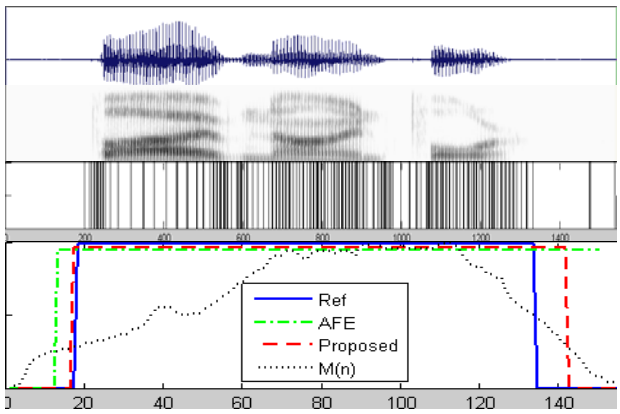
The function $frame\_selection(t)$ represents whether the tth frame is selected or not in the frame selection process: The value is 1 if selected and 0 if not. The constant $\alpha = 10$ maps the 1 ms frame shift for frame selection to the 10 ms frame shift for VAD. Again, the latency of the VAD method is controlled by adjusting m_2.

The output of the moving average $M(n)$ is compared against a threshold T_{vad} for making VAD decisions as done in Approach 1.

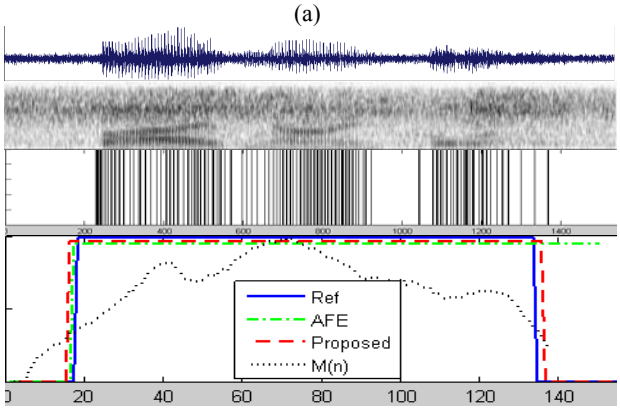
2.3. Illustrative VAD results

Figure 1 illustrates the intermediate and final VAD results for two input speech signals. In this experiment the Approach 2 using a 37-point central moving average is applied. The utterances are the English digits "five nine four" in clean and 5 dB noisy environments. The figure presents results on waveform, spectrogram, frames selected by the proposed method, and VAD results.

The results show that the VAD method performs well even for a speech signal of 5 dB in terms of both frame selection and VAD decision. Also it is observed that the VAD result of the proposed method is more precise than that of the advanced front-end.



(a)



(b)

Figure 1: *VAD experiment for the English digits "five nine four": (a) For clean speech: waveform (the 1st panel), spectrogram (the 2nd panel), frames selected by the proposed method (the 3rd panel), VAD results (the 4th panel: the solid blue for the reference VAD, the dash-dot green for the advanced front-end VAD, the dashed red for the proposed VAD and the black dotted for the moving average M(n)); (b) for 5 dB speech with the same order of panels as in (a).*

3. Experiments

Extensive experiments are conducted on the Aurora 2 database [13], which is the TI digits database artificially distorted by adding noise and using a simulated channel distortion. The database encompasses two training sets for clean and multi-condition training and three test sets. The three test sets are contaminated by different types of noise with SNR ranging from -5 to 20 dB. Test Set A includes clean speech and noisy speech corrupted by four noise types: "Subway", "Babble", "Car", and "Exhibition". The four types of noise in Test Set B are "restaurant", "station", "airport" and "street". Test Set C includes convolutional noise.

3.1. Generation of frame-based reference VAD

The frame-by-frame reference VAD is generated from forced-alignment speech recognition experiments. Whole word models are trained on clean speech data for all digits using the HTK recognizer. Each of the whole word digit models has 16 hidden Markov model (HMM) states with three Gaussian mixtures per state. The silence model has three HMM states with six Gaussian mixtures per state. A one state short pause model is tied to the second state of the silence model.

The trained word models are used for performing forced-alignment for the 4004 utterances (clean speech) from which all utterances in Test Set A, B and C are derived from by adding noise. The forced-alignment results are used to set the time boundaries for speech segments to create a frame-based reference VAD.

3.2. Performance comparison

The experimental results for the ETSI advanced front-end VAD and for the proposed method (Approach 2) on the Test Set A are shown in Table 1 and Table 2, respectively. Only results on Test Set A are presented here in order to show the performance of these methods on different types of noise in detail.

Table 1. *Percentage of frame errors obtained by the ETSI advanced front-end VAD on Aurora 2 database Test Set A across SNR values and noise types.*

SNR	Subway	Babble	Car	Exhibition	Average
Clean	18.96	18.22	17.86	18.62	18.41
20 dB	16.64	15.69	12.41	15.11	14.95
15 dB	16.60	15.61	11.78	14.81	14.69
10 dB	16.66	15.21	11.13	14.39	14.34
5 dB	16.35	15.24	11.07	14.40	14.26
0 dB	17.12	16.73	12.76	14.34	15.23
-5 dB	19.40	23.43	26.29	20.52	22.43
Average	17.39	17.16	14.76	16.03	16.33

Table 2. *Percentage of frame errors obtained by the proposed method (Approach 2) on Aurora 2 database Test Set A across SNR values and noise types.*

SNR	Subway	Babble	Car	Exhibition	Average
Clean	8.20	8.19	7.79	8.31	8.17
20 dB	8.09	7.89	7.87	7.86	7.93
15 dB	8.66	8.51	8.47	8.61	8.56
10 dB	9.93	9.49	9.68	9.65	9.69
5 dB	12.50	11.99	11.12	11.68	11.82
0 dB	18.66	17.27	13.92	16.25	16.52
-5 dB	28.58	25.34	19.62	24.81	24.57
Average	13.52	12.67	11.24	12.45	12.46

Overall frame error rates for the advanced front-end VAD and the proposed method are 16.33% and 12.46%, respectively. The proposed method gives substantially lower frame error rates than the advanced front-end VAD for all noise types. Its performance for high SNR signals is very promising and much better than that of the advanced front-end

VAD, and it is slightly worse for very low SNR signals (0 dB and -5 dB).

Additional experiments show that the proposed Approach 1 gives an overall frame error rate of 13.73% which is close to the one obtained by the Approach 2. A 37-point central moving average is applied in both methods, which, however, indicates a latency of 18 frames. Latency experiments are conducted in Subsection 3.3.

Frame error rates for several VAD methods on the Aurora 2 Test Set A, B and C are presented in Table 3. All three test sets are used here so that these experiments are compatible with those in [3]. Results for G.729, G.723.1 and MFB VAD methods are cited from [3].

The results in Table 3 verify that in terms of frame error rate, the proposed method is significantly better than other methods, including the Mel-filter bank VAD [3], and three standards: the ETSI-DSR advanced front-end VAD [7], G.729 VAD [9] and G.723.1 VAD [10].

Table 3. *Percentage of frame errors obtained by several methods on Aurora 2 database Test Set A, B and C. The results for G.729 VAD, G.723.1 VAD and, MFB VAD are cited from [3].*

SNR	G.729 VAD	G.723.1 VAD	MFB VAD	DSR AFE VAD	Proposed VAD
Clean	12.84	19.45	6.92	18.41	8.11
20 dB	24.53	21.31	15.39	15.16	8.27
15 dB	26.13	23.29	17.70	14.96	9.04
10 dB	27.38	24.44	20.12	14.59	10.59
5 dB	29.13	26.30	22.75	14.54	13.54
0 dB	32.23	26.56	26.16	15.62	19.50
-5 dB	35.21	28.58	31.09	22.08	28.19
Average	26.78	24.28	20.02	16.48	13.89

3.3. Latency experiments

In the previous experiments for the proposed method, a 37-point central moving average is applied, which gives a latency of 18 frames. To reduce the amount of delay, further experiments are conducted by using two different types of moving average: a biased moving average that uses more preceding data than succeeding data, and a prior moving average that uses preceding data only.

It is obvious that the use of biased or prior moving average will result in wrongly classifying the first several speech frames as non-speech. To handle this problem, an adaptive threshold is applied as follows.

$$T_{vad}(n) = T_{vad} - \frac{1}{3}(m_1 - m_2 - \sum_{m=m_2-m_1}^{-1} vad\_decision(n+m)) \quad (5)$$

The function $vad\_decision(n)$ is the VAD decision at frame n: 1 for speech and 0 for non-speech.

Table 4. *Percentage of frame errors obtained by the proposed method (Approach 2) with different amounts of latency on Aurora 2 database Test Set A across noise types.*

Latency	Subway	Babble	Car	Exhibition	Average
18 frames	13.52	12.67	11.24	12.45	12.46
6 frames	15.90	15.00	13.27	14.72	14.72
0 frame	17.02	16.24	14.50	16.04	15.94

The results in Table 4 show that the method with 6 frames latency still demonstrates a good performance. In the case of 0 frame latency, the performance, though still better than the referenced methods, degrades substantially as compared with the one with latency. One potential way to improve it is to apply a weighted prior moving average.

4. Conclusions

This paper presented an efficient VAD method that relies on *a posteriori* SNR weighted energy. Experiments on the Aurora 2 database show that the method is superior to a number of referenced methods including the Mel-filter bank VAD and three standards: the ETSI-DSR advanced front-end VAD, G.729 VAD and G.723.1 VAD. The method further has the advantage of having a low complexity.

5. Acknowledgements

The authors would like to thank Damjan Vlaj, University of Maribor for valuable discussions on testing the ETSI advanced front-end VAD.

6. References

[1] Ramirez, J., Segura, C., Benitez, C., Torre, A. and Rubio, A., "A new Kullback-Leibler VAD for speech recognition in noise," IEEE Signal Processing Letters, 11(2), 2004.

[2] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in Proc. EUROSPEECH 2001, Aalborg, Denmark, September 2001.

[3] Vlaj, D., Kotnik, B., Horvat, B. and Kacic Z., "A Computationally Efficient Mel-Filter Bank VAD Algorithm for Distributed Speech Recognition Systems", EURASIP Journal of Applied Signal Processing 2005:4, 487-497.

[4] Dong, E., Liu, G., Zhou, Y. and Zhang, X., "Applying support vector machines to voice activity detection," in Proc. ICSLP 2002, Denver, USA, 2002.

[5] Shah, J.K., Iyer, A.N., Smolenski, B.Y. and Yantorno, R.E., "Robust voiced/unvoiced classification using novel features and Gaussian mixture model," in Proc. ICASSP 2004, Montreal, Canada, 2004.

[6] Shin, W.-H., Lee, B.-S., Lee, Y.-K. and Lee, J.-S., "Speech/non-speech classification using multiple features for robust endpoint detection," in Proc. ICASSP 2002, Orlando, Florida, USA, 2002.

[7] ETSI, "Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm," ES 202 050 v1.1.1, 2002.

[8] Pearce, D., "Distributed speech recognition standards," in Z.-H. Tan, and B. Lindberg (eds.), Automatic speech recognition on mobile devices and over communication networks, Springer-Verlag, London, Feb. 2008, pp. 87-106.

[9] ITU, "Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear-prediction (CS-ACELP) Annex B: A silence compression scheme," ITU Recommendation G.729, 1996.

[10] ITU, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. Annex A: Silence compression scheme," ITU Recommendation G.723.1, 1996.

[11] Fujimoto, M., Ishizuka, K., and Nakatani, T., "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in Proc. ICASSP 2008.

[12] Tan, Z.-H. and Lindberg, B., "A Posteriori SNR Weighted Energy Based Variable Frame Rate Analysis for Speech Recognition," in Proc. Interspeech 2008, Brisbane, Australia, September 2008.

[13] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proc. ISCA ITRW ASR, Paris, France, 2000.