Voice Activity Detection for Speech Enhancement Applications

E. Verteletskaya, K. Sakhnov

Abstract

This paper describes a study of noise-robust voice activity detection (VAD) utilizing the periodicity of the signal, full band signal energy and high band to low band signal energy ratio. Conventional VADs are sensitive to a variably noisy environment especially with low SNR, and also result in cutting off unvoiced regions of speech as well as random oscillating of output VAD decisions. To overcome these problems, the proposed algorithm first identifies voiced regions of speech and then differentiates unvoiced regions from silence or background noise using the energy ratio and total signal energy. The performance of the proposed VAD algorithm is tested on real speech signals. Comparisons confirm that the proposed VAD algorithm outperforms the conventional VAD algorithms, especially in the presence of background noise.

Keywords: voice activity detection, periodicity measurement, voiced/unvoiced classification, speech analysis.

1 Introduction

An important problem in speech processing applications is the determination of active speech periods within a given audio signal. Speech can be characterized as a discontinuous signal, since information is carried only when someone is speaking. The regions where voice information exists are referred to as 'voice-active' segments, and the pauses between talking are called 'voice-inactive' or 'silence' segments. The decision on the class to which an audio segment belongs is based on an observation vector. This is commonly referred to as a 'feature' vector. One or many different features may serve as the input to a decision rule that assigns the audio segment to one of these two classes. An algorithm employed to detect the presence or absence of speech is referred to as a voice activity detector (VAD).

VAD is any important component of speech processing techniques such as speech enhancement, speech coding, and automatic speech recognition. In speech enhancement applications, for example in spectral subtractive type noise reduction algorithms, VAD is used for noise estimation, which is then used in the noise reduction process. Speech/silence detection is necessary in order to determine frames of noisy speech that contain noise only. Speech pauses or noise only frames are essential to allow the noise estimate to be updated, thereby making the estimation more accurate. In speech coding, the purpose is to encode the input audio signal in such a way, that the overall transferred data rate is reduced. Since information is only carried when someone is speaking, clearly knowing when this occurs can greatly aid in data reduction. Another example is speech recognition. In this case, a clear indication of active speech periods is critical. False detection of active speech periods will have a direct degradation effect on the recognition algorithm. Other examples include audio conferencing, echo cancellation, VoIP applications, cellular radio systems (GSM and CDMA based) [1] and hands-free telephony [2].

Generating an accurate indication of the presence or absence of speech is generally difficult, especially when the speech signal is corrupted by background noise or by unwanted impulse noise. Voice activity detection algorithm performance trade-offs are made by maximizing the detection rate of active speech while minimizing the false detection rate of inactive segments. Various techniques for VAD have been proposed [3, 4, 5, 6, 7]. In the early VAD algorithms, short-time energy, zero-crossing rate and linear prediction coefficients were among the features commonly used in the detection process [3]. Cepstral coefficients [4], spectral entropy [5], a least-square periodicity measure [6], and wavelet transform coefficients [7] are examples of recently proposed VAD features. Signal energy remains one of basic components of the feature vector. Most of the standardized algorithms use signal energy and other parameters to make a decision. For voice activity detection, the proposed algorithm utilizes the total signal energy, which is compared with the dynamically calculated threshold. Besides the total energy measure, the algorithm is supplemented by using a signal periodicity measure and a high frequency to low frequency signal energy ratio for more accurate decisions on voice presence.

2 Voice activity detection principle

The basic principle of a VAD device is that it extracts measured features or quantities from the input signal and then compares these values with thresholds usually extracted from noise-only periods. Voice activity $(\mathrm{VAD}=1)$ is declared if the measured values exceed the thresholds. Otherwise, there is no speech activity or noise, and silence (VAD = 0) is present. A general block diagram of a VAD design is shown in Fig. 1.

VAD design involves extracting acoustic features that can appropriately indicate the probability of target speech signals existing in observed signals. Based on these acoustic features, the latter part decides whether the target speech signals are present in the observed signals, using a computed well-adjusted threshold value. Most VAD algorithms output a binary decision on a frame-by-frame basis, where the frame of the input signal is a short unit of time 5–40 ms in length. The accuracy and reliability of a VAD algorithm depends heavily on the decision thresholds. Adapting the threshold value helps to track time-varying changes in the acoustic environments, and hence provides a more reliable voice detection result.

2.1 VAD algorithms based on energy thresholding

In energy-based VAD, the energy of the signal is compared with the threshold depending on the noise level. Speech is detected when the energy estimation lies above the threshold.

$$\begin{array}{ll} \mathit{IF}\ (E_j > k \cdot E_r),\ \mathit{where}\ k > 1, & \mathit{frame\ is\ ACTIVE} \qquad (1) \\ \mathit{ELSE} & \mathit{frame\ is\ INACTIVE} \end{array}$$

In the equation, E_r represents the energy of the noise frames, while $k \cdot E_r$ is the threshold used in the decision-making. Having a scaling factor, k allows a safe band for adapting E_r, and, therefore, adapting the threshold. Different energy-based VADs differ in the way the thresholds are updated. The simplest energy-based method, the Linear Energy-Based Detector (LED), was first described in [8]. The rule for updating the threshold value was specified as,

$$E_{rnew} = (1 - p) \cdot E_{r\,old} + p \cdot E_{silence} \qquad (2)$$

Here, E_r new is the updated value of the threshold, $E_{r\,old}$ is the previous energy threshold, and $E_{silence}$ is the energy of the most recent unvoiced frame. The reference E_r is updated as a convex combination of the old threshold and the current noise update. Parameter p is constant $(0 < p < 1)$.

2.2 Energy of a frame

The most common way to calculate the full-band energy of a speech signal is a short-time energy calculation. If $x(i)$ is the i-th sample of speech, N is the number of samples in a frame, then the short-time energy of the j-th frame of a speech signal can be represented as

$$E_j = \frac{1}{N} \cdot \sum_{i=(j-1)\cdot N+1}^{j\cdot N} x^2(i). \qquad (3)$$

Another common way to calculate the energy of a speech signal is the *root mean square energy (RMSE)*, which is the square root of the average sum of the squares of the amplitude of the signal samples (3).

$$E_j = \left[\frac{1}{N} \cdot \sum_{i=(j-1)\cdot N+1}^{j\cdot N} x^2(i)\right]^{\frac{1}{2}} \qquad (4)$$

Fig. 2 shows that the power estimate of a speech signal exhibits distinct peaks and valleys. While the peaks correspond to speech activity, the valleys can be used to obtain a noise power estimate. Therefore, RMSE is more appropriate for thresholding, because it display valleys in greater detail.

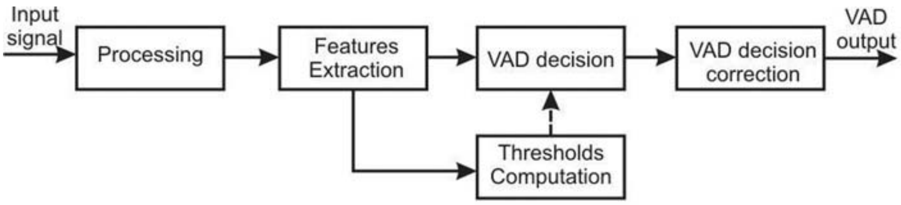

Fig. 1: Block diagram of a basic VAD design

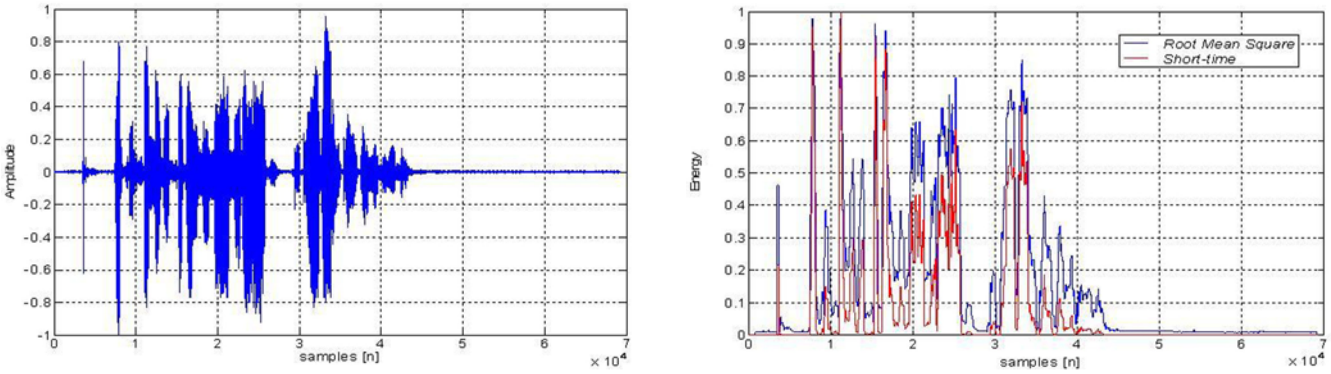

Fig. 2: Short-time vs. root mean square energy

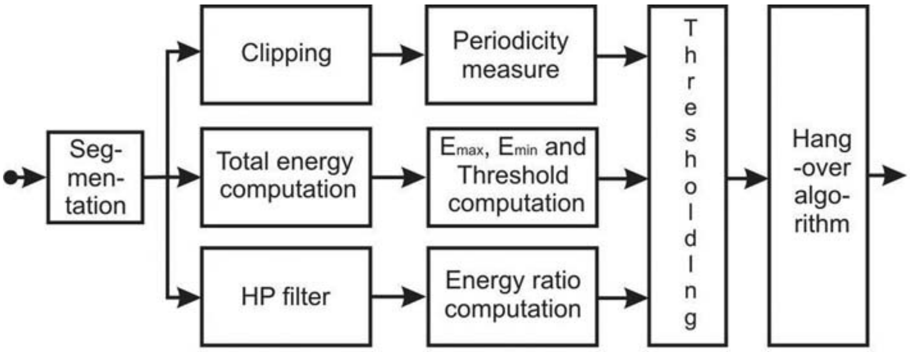

Fig. 3: Logic flowchart of the proposed VAD

3 The proposed voice activity detector

For voice/silence detection, the proposed algorithm uses a periodicity measure of the signal, as well as the high-frequency versus low-frequency signal energy ratio and full-band energy computation. A simplified flowchart of the whole algorithm is given in Fig. 3.

3.1 Feature extraction

Signal periodicity C is determined by estimating the pitch period of the signal. To reduce the computational complexity, the input signal is first center clipped [9], then the normalized autocorrelation function $R(\tau)$ given by (5) is used for pitch estimation.

$$R(\tau) = \frac{\sum_{n=0}^{N-m-1} x(n)\cdot x(n+\tau)}{\sqrt{\sum_{n=0}^{N-m-1} x^2(n+\tau)}}, \quad (5)$$

$$T_{\min} \le \tau \le T_{\max}$$

where $x(n)$ $n = 0, 1, \ldots, N$ is the input signal frame. The autocorrelation function is calculated for values of lag τ from $T_{\min}$ to $T_{\max}$. The constants $T_{\min}$ and $T_{\max}$ are the lower and upper limits of the pitch period, respectively. The pitch period of a voiced frame is equal to the value of τ that maximizes the normalized autocorrelation function. The periodicity C of the frame is given by maximum value of $R(\tau)$.

The total voice band energy E_f is computed for the voice band frequency range from 0 Hz to 4 kHz. The total voice band energy is given by (4). The computation of the threshold for total voiceband energy is based on the energy level $E_{\min}$ and $E_{\max}$, obtained from the sequence of incoming frames. These values are stored in memory and the threshold is calculated as,

$$Threshold = (1-\lambda)\cdot E_{\max} + \lambda \cdot E_{\min} \quad (6)$$

$$\lambda = \frac{E_{\max} - E_{\min}}{E_{\max}}. \quad (7)$$

Here, λ – a scaling factor controlling the estimation process. The voice detector performs reliably when λ is in the range of $[0.950, \ldots, 0.999]$. For different types of signals the value of λ cannot be the same, so it must be set up properly. Computing the scaling factor λ by (7) makes it independent and resistant to the variable background environment.

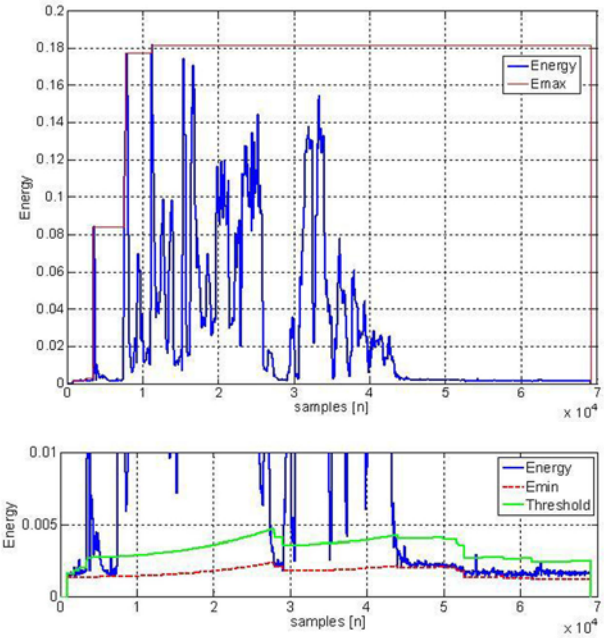

Fig. 4: Threshold computation for total band signal energy

Energy ratio E_r is computed as the ratio of the energy above 2 kHz to the energy below 2kHz in the input voice band signal. To obtain a high-frequency signal, the input signal is passed through a high-pass filter that has a cut-off frequency of 2 kHz. The high frequency to low frequency energy ratio E_r is calculated as

$$E_r = E_h/(E_f - E_h) \quad (8)$$

Where E_f and E_h are the full band and high band signal energy, respectively, calculated by (2) and expressed in dB.

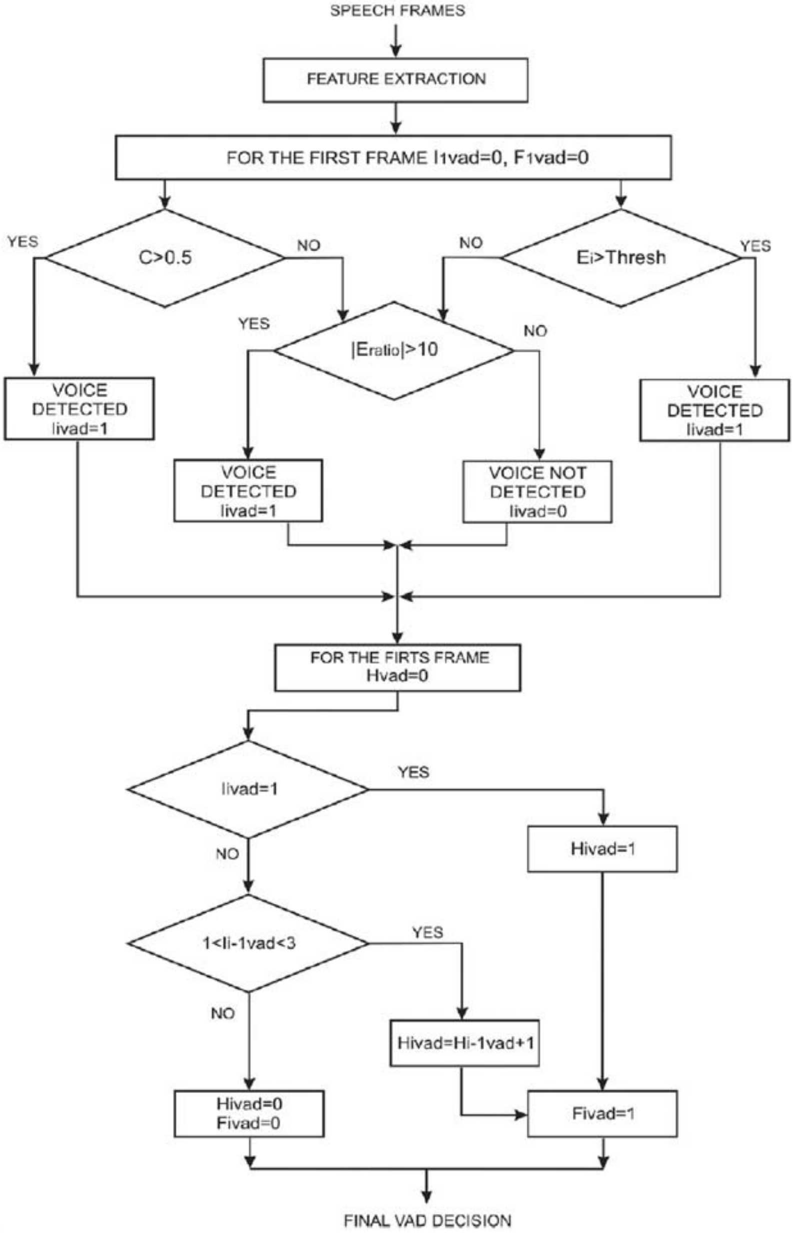

Fig. 5: Detailed flowchart of the proposed VAD

3.2 Thresholding and the hang-over algorithm

After feature extraction, the parameters are compared with several thresholds to generate an initial VAD decision (I_{VAD}) (see Fig. 5). After the thresholds have been compared to determinate the value of I_{VAD}, a final output decision is made according to the lower part of the algorithm flowchart. Output decision F_{VAD} is performed anew for each value of I_{VAD} produced by threshold comparison. The final output decision involves usage of a smoothing hang-over algorithm to ensure that detection of either the presence or the absence of speech lasts for at least a minimum period of time and does not oscillate on-and-off. Upon startup of VAD, the values of a hangover flag H_{VAD} and a final VAD flag F_{VAD} are initialized to zero. The output decision block checks whether the received I_{VAD} value is one. If so, it means that speech has been detected. The output decision therefore sets H_{VAD} and F_{VAD} to one. If the value of I_{VAD} is found to be zero, speech has not been detected. However, the output decision checks whether the value of H_{VAD} is set to one from the previous frame. If so, the output decision checks whether the smoothed value E_{fs} less the value of $E_{\min}$ is greater than 8 dB. If so, holdover is indicated, and so the output decision maintains F_{VAD} set to one, even though speech has not been detected.

4 Experimental results

The MATLAB environment was used to test the algorithms on thirty speech signals from the Czech Speech database. The test templates varied in loudness, speech continuity, background noise and accent. Both male speech and female speech in Czech language were used for the experiments. Fig. 6 shows the voice/silence classification results of the proposed VAD algorithm. The performance of the algorithm is compared to the performance of the LED algorithm [8]. A comparison is performed on real clean speech and on speech degraded by additive noise. It is clear from the figures that the proposed VAD outperformed the LED algorithm in extent of misdetection. In contrast to the LED algorithm, the proposed VAD results in correct detection of unvoiced speech regions. The proposed algorithm is able to detect the beginnings and ends of active speech segments accurately even on noisy speech signals.

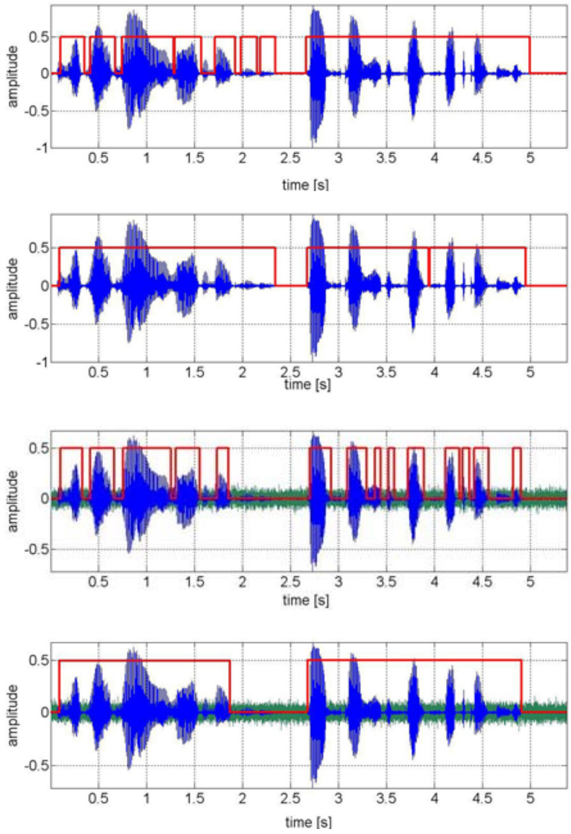

Fig. 6: Performance comparison of VAD algorithms: (a) LED algorithm clean speech, (b) proposed algorithm clean speech, (c) LED algorithm noisy speech (SNR = 5 dB), (d) proposed algorithm noisy speech (SNR = 5 dB)

5 Conclusion

This paper has presented voice activity detection algorithms employed to detect the presence/absence of speech components in an audio signal. An alternative VAD based on periodicity detection and the high-frequency to low-frequency signal energy ratio has been presented. The aim of the paper was to show the principle of the proposed VAD algorithm, and to compare it with the known linear energy-based detector (LED). The results consistently show the superiority of the proposed VAD scheme over the LED algorithm. It is easy to recognize that the algorithm has low computational complexity, and can be easily integrated into speech coders and other speech enhancement systems.

Acknowledgement

The research described in this paper was supervised by Prof. Ing. B. Simak, CSc., FEL CTU in Prague and was supported by Czech Technical University grant SGS No. OHK3-108/10 and by the Ministry of Education, Youth and Sports of the Czech Republic research program MSM 6840770014.

References

[1] ETSI TS 126 094 V3.0.0 (2000-01), 3G TS 26.094 version 3.0.0 Release 1999, Universal Mobile Telecommunications System (UMTS); Mandatory Speech Codec speech processing functions AMR speech codec; Voice Activity Detector (VAD), 2000.

[2] Benyassine, A., Shlomot, E., Su, H.-Y.: 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 application, *IEEE Commun. Mag.*, 1997, Vol. **35**, p. 64–73.

[3] Atal, B. S., Rabiner, L. R.: A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. **24**, p. 201–212, June 1976.

[4] Haigh, J. A., Mason, J. S.: Robust voice activity detection using cepstral features, in *Proc. of IEEE Region 10 Annual Conf. Speech and Image Technologies for Computing and Telecommunications*, (Beijing), p. 321–324, Oct. 1993.

[5] McClellan, S. A., Gibson, J. D.: Spectral entropy: An alternative indicator for rate allocation, in *IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Adelaide, Australia), p. 201–204, Apr. 1994.

[6] Tucker, R.: Voice activity detection using a periodicity measure, *IEE Proc.–I*, Vol. **139**, p. 377–380, Aug. 1992.

[7] Stegmann, J., Schroder, G.: Robust voice-activity detection based on the wavelet transform, in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, (Pocono Manor, PN), p. 99–100, Sept. 1997.

[8] Pollak, P., Sovka, P., Uhlir, J.: Noise System for a Car, *proc. of the Third European Conference on Speech, Communication and Technology – EUROSPEECH'93*, (Berlin, Germany), p. 1 073–1 076, Sept. 1993.

[9] Verteletskaya, E., Šimák, B.: Performance Evaluation of Pitch Detection Algorithms. Access server [online]. 2009, roč. 7, č. 200906, s. 0001. ISSN 1214-9675.

About the authors

Ekaterina VERTELETSKAYA was born in Uzbekistan. She was awarded an MSc degree in Telecommunication and Radio Engineering from the Czech Technical University, Prague in 2008. She is currently a PhD student at the Department of Telecommunication Engineering of CTU in Prague. Her current activities are in the area of digital signal processing, focused on speech coding algorithms for mobile communications.

Kirill SAKHNOV was born in Uzbekistan. He was awarded an MSc degree from the Czech Technical University in Prague in 2008. He is currently a PhD student at the Department of Telecommunication Engineering of CTU in Prague. His current activities are in the area of adaptive digital signal processing, focused on problems of acoustical and network echo cancellation in telecommunication devices.

Ekaterina Verteletskaya
Kirill Sakhnov
E-mail: verteeka@fel.cvut.cz,
sakhnkir@.fel.cvut.cz
Czech Technical University in Prague
Technická 2, 166 27 Praha, Czech Republic