Siamak RASOOLZADEH, Mohsen RAHMANI

Department of Computer Engineering, Malayer Branch, Islamic Azad University, Malayer, Iran

Sub-band VAD Based on Continuous Noise Estimation in Wavelet Domain

Abstract. *Voice Activity Detectors (VADs) are widely used in speech processing applications such as speech enhancement and coding. In speech enhancement systems, we use VAD in order to obtain more accurate estimation of noise. Voice activity detection methods usually work in time or frequency domains. In this paper, we propose two approaches for voice activity detection in wavelet domain for continuous noise estimation: Sub-band VAD and Full-band VAD. We use the correct noise/speech classification rate to show the acceptable performance of our proposed VADs. In addition, we apply sub-band VAD to a speech enhancement system. Objective and subjective measures show that Sub-band VAD outperforms Full-band VAD in noise reduction applications.*

Streszczenie. *Analizowano detektor aktywności głosowej VAD w zastosowaniu do poprawy jakości dźwięku mowy. Celem było określenie szumów towarzyszących. Zastosowano transformatę falkową do analizy szumów. (**Określanie szumów przy wykorzystaniu transformaty falkowej i detektora VAD**)*

Keywords: website; quality metrics; evaluation.
Słowa kluczowe: dźwięk mowy, detektor głosu, transformata falkowa..

Introduction

Speech/non speech detection is an important problem in speech processing and affects numerous applications including robust speech recognition, real-time speech transmission on the Internet or combined noise reduction and echo cancellation schemes in the context of telephony [2,4]. During the last decade, numerous researchers have developed different strategies for detecting speech on a noisy signal and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [3,10].

Voice activity detection methods can be divided into two main groups: thresholding methods and modelling methods. In thresholding methods, we extract some acoustic features form a noisy speech frame such as: energy, zero crossing rate, spectral amplitude, etc [5,11]. With comparing these features values with a threshold, noise and speech frames are determined. In modelling methods, with a learning algorithm, like neural networks, hidden Markov model or fuzzy model, noise and speech segments are learned [5,9].

Thresholding methods are fast and efficient in comparison with modelling methods. So usage of thresholding methods is very common in real time speech processing applications.

Most of VAD methods are implemented in time or frequency domain. In this paper we proposed two thresholding based voice activity detection methods in wavelet domain: sub-band and full-band VAD. These proposed methods are applied in a real-time speech enhancement system.

The rest of the paper is organized as follow. In section ***Wavelet transformation***, we briefly review discrete wavelet transform. In section ***Continuous noise estimation in wavelet domain***, we describe continuous noise estimation in wavelet domain. In section ***voice activity DETECTION (VAD)***, we proposed two types of VAD. In section ***Evaluation of VAD on noise reduction system***, we reported experimental results. And finally we conclude our work in section ***Conclusion***.

Wavelet transformation

Wavelet transform (WT), proposed by Morlet and Grossman, is utilized in image processing and speech processing [6,7].The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently.

One level structure of DWT is depicted in Figure (1).

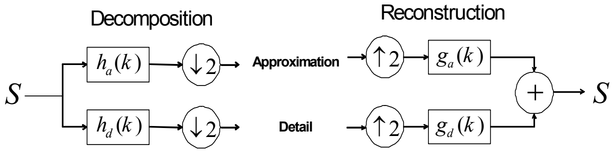

Fig. 1. Wavelet decomposition and reconstruction phases

DWT has two major phases, decomposition and reconstruction. In decomposition phase, input signal is decomposed into two sub-bands: low frequency and high frequency sub-bands. These subbands are achieved separately by applying low-pass and high-pass decomposition filters to the input signal. Low frequency and high frequency sub-bands are named as approximation and detail respectively. Wavelet tree can be extended by reapplying high-pass and low-pass decomposition filters on details and approximation sub-bands. This is known as Wavelet Packet Tree (WPT) if there is 2L sub-bands in the Lth level of wavelet tree.

In the reconstruction phase, same as decomposition, high-pass and low-pass reconstruction filters applied to approximation and detail, and then sum them together to obtain a sub-band in upper level. This process continues till output signal is reconstructed in the root of WPT.

Continuous noise estimation in wavelet domain

Continuous noise estimation works based on adaptive filtering. In adaptive filtering method, gain of filter is updated during time based on the power of noise. This noise estimation method is used in many applications [8].Most of these applications use continuous noise estimation in frequency domain. In this paper new continuous noise estimation is proposed that performs in wavelet domain.

In wavelet filter-bank, a set of coefficients are applied to present each sub-band, while in frequency filter-bank, a coefficient is used for presenting each sub-band. So we can expect that we obtain a better noise estimation using continuous noise estimation in wavelet domain.

In continuous noise estimation, we use a filter including 2 steps. In the first step, an adaptive filter is applied to noisy signal. After that, powers of input noise and clean signal are estimated in each subband. In the second step, the enhancement filter is calculated based on powers of estimated noise and estimated speech signal.

The adaptive filter, H(j,k) can be computed as in equation (1) using the estimated power of noise, P^N(j,k-1), and the estimated power of clean signal, P^S(j,k-1). These parameters are estimated in the previous frame.

$$(1)\qquad H(j,k)=\frac{\hat{P}_S(j,k-1)}{\hat{P}_S(j,k-1)+\hat{P}_N(j,k-1)}$$

The power of signal y, in k-th frame and n-th subband, is calculated as:

$$(2)\qquad P_Y(n,k)=\sum_{m=1}^{M}Y^2(m,n,k)$$

In this equation M is number of coefficients in this subband.

Based on equation 3, estimation of noise power in current frame, P^N(j,k), is calculated based on power of noisy signal in current frame, X(j,k), and estimated of noise power in previous frame in each sub-band.

$$(3)\qquad \begin{aligned}&\hat{\mathrm{N}}^2(j,n)=\mathrm{X}^2(\mathrm{j},\mathrm{k}).(1-\mathrm{H}(\mathrm{j},\mathrm{k}))\\&\hat{\mathrm{P}}_{\mathrm{N}}\ (j,k)=(\lambda_{\mathrm{n}}).\hat{\mathrm{P}}_{\mathrm{N}}(j,k-1)+(1-\lambda_{\mathrm{n}}).\hat{\mathrm{N}}^2(j,n)\end{aligned}$$

where λ_N is the forgetting factor for noise estimation ($0 < \lambda_N \leq 1$). In Equation (4) estimated power of clean signal is approximated based on power of noisy signal in current frame, *X(j,k)*, and power of estimated clean signal in previous frame ,*P^S(f,k-1)*.

$$(4)\qquad \begin{aligned}&\hat{\mathrm{S}}_1^2(j,k)=H(j,k)\times X^2(j,k)\\&\hat{\mathrm{P}}_{\mathrm{S}}(j,k)=(\lambda_{\mathrm{s}})\hat{\mathrm{P}}_{\mathrm{S}}(j,k-1)+(1-\lambda_{\mathrm{s}}).\hat{\mathrm{S}}_1^2(j,k)\end{aligned}$$

where λ_S is forgetting factor for speech estimation. In the second step, based on equation (5), main noise reduction filter in each sub-band, *A(j,k)*, is calculated. In this equation, ***α(1-H)*** is used to control the amount of remined background noise and speech artefacts [1].

$$(5)\qquad A(j,k)=H(j,k)+\alpha(1-H(j,k))$$

The advantage of using two steps filtering is that estimation and reduction filters are separately calculated. This approach is efficient because estimation filter discussed above, follows the changes of noise power in time. Furthermore, estimated filter adapts to these changes and produces better noise estimation. Then, computed filter can reduce the noise more accurately.

Voice Activity Detection (VAD)

As mentioned above, VAD using threshold values have less computational cost in comparison with VAD methods based on modelling. So these methods are suitable in order to apply to real-time applications. In thresholding methods, some parameters such as zero-crossing rate and power are compared with threshold values. In this paper, we propose a new VAD algorithm based on power of estimated speech.

As mentioned, noise estimation method, proposed in previous section, contains two steps. In the first step, an estimation of speech power is calculated. This estimation could be employed as a threshold value to distinguish between noise and speech sections in the proposed VAD algorithm.

So equation (4) is modified by combination of continuous noise estimation and this VAD. In the following, two types of VAD, Fullband VAD and Subband VAD, are proposed.

Full band VAD (FVAD)

In the first step of noise estimation, speech power is calculated using equation (4) in each sub-band. In case of FVAD, powers of speech for all sub-bands are accumulated together. This value is used as estimation of speech power in current frame. If this value is less than a determined threshold, current frame is labelled as noise frame; otherwise current frame is labelled as speech frame. In noise frames, power of estimated noise is updated based on equation (3). In speech frames, power of estimated noise is not changed. Figure (2) shows the block diagram of FVAD.

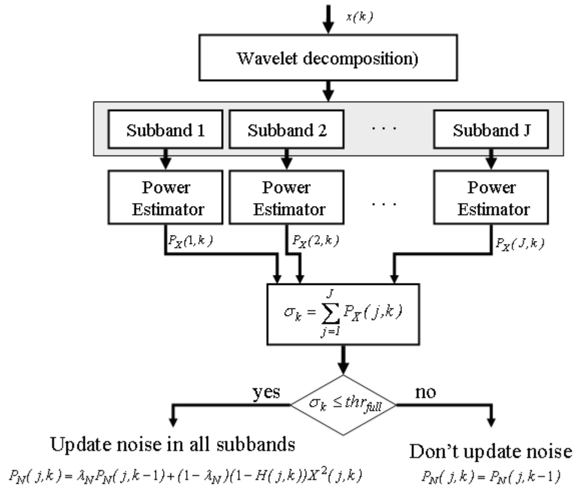

Fig. 2- Fullband VAD Diagram

Subband VAD (SVAD)

In FVAD, when a frame is considered as noise, the estimation of noise power is updated in all sub-bands. It is clear that most of speech power is not uniformly distributed in all bands range. Therefore, using identical noise estimation rules for all sub-bands is not suitable. Therefore, we propose SVAD to apply it to subband level.

In case of SVAD, different threshold values are computed for each sub-band. If estimation of speech power in each sub-band is less than these thresholds, then the subband is labelled as noise, otherwise it is labelled as speech. In sub-band VAD, a different VAD is considered for each sub-band. Consequently, SVAD employs N VAD for noise estimation (N is number of sub-bands).

If a sub-band is considered as noise, the estimation of noise power is updated in the sub-band; otherwise the estimation of noise power is not changed in the sub-band. Figure (3) shows block diagram of SVAD technique.

SVAD performs better than FVAD, when input noise is distributed in some subbands such as car and babble noises.

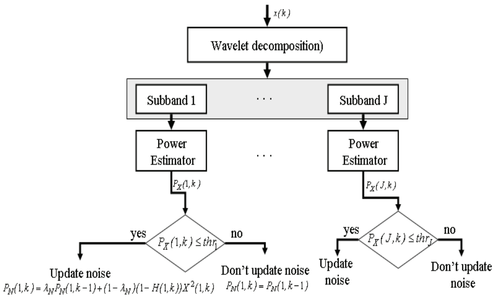

Fig. 3- Subband VAD Block Diagram

Evaluation of proposed VAD methods

We propose two different experiments to evaluate FVAD. In the first experiment, some reference signals are created (clean signal which its speech segments are labelled manually). White noise with 3 different SNR values (5, 10, 20 dB) is added to reference signals to generate noisy signals. Table (1) shows classification rate of FVAD for noisy signals.

Table 1-Correct Detection Rate (FVAD)

	SNR=20	SNR=10	SNR=5
Classification rate	97.02%	94.4%	90.1%

By decreasing speech power and fixing noise power, the SNR decreases. This fact leads us to propose another experiment. In this experiment, input SNR is 20dB.Furthermore, when speech power is decreased to 5, 10 and 15dB; three different SNR values are achieved. We want to show the sensitivity of FVAD performance to determined thresholds. Table (2) shows FVAD classification rate for the second experiment.

Table 2-Correct Detection Rate (FVAD)

	Input SNR=20 dB		
Decrease in Speech power	5 dB	10 dB	15 dB
Classification rate	94.99%	89.90%	84.65%

We propose two different experiments for evaluating SVAD. Evaluating of SVAD is more difficult than FVAD, because in SVAD there is a VAD in each sub-band. So wavelet filter-bank must decompose reference signals at first. After that, speech segments are labelled manually in each sub-band.

After creating reference signals in each sub-band, we can calculate SVAD classification rate in each sub-band. Tables (3) and (4) show sample of classification results in presence of white noise with different SNR values (5, 10 and 20dB) in sub-band numbers of 2 and 14.

In high frequency sub-bands, VAD classification rate decreases because most of speech power is distributed in lower frequency sub-bands. Therefore, detection of noise from speech signal in high frequency sub-bands becomes more difficult. Results in these tables justify this issue.

Table 3-Correct Detection Rate in 2th subband(SVAD)

	SNR= 20 dB	SNR= 10 dB	SNR= 5 dB
Classification rate	96.52%	93.9%	88.97%

Table 4-Correct Detection Rate in 14th subband(SVAD)

	SNR= 20 dB	SNR= 10 dB	SNR= 5 dB
Classification rate	78.27%	71.22%	67.04%

Evaluation of VAD on noise reduction system

In the previous section, accuracy of FVAD and SVAD are evaluated. It is shown that SVAD has higher classification rate than FVAD. In this section, we evaluate proposed VADs with applying them to a noise reduction system. In order to evaluate performance of proposed methods, we use a noise reduction system based on continuous noise estimation in wavelet domain without using VAD named WVAD.

In this experiment, white, car and babble noises with 3 different inputs SNR values of 0, 10 and 20 dB are added to 12 clean signals, selected from TIMIT database. So, 108 noisy signals are generated. SNR improvement (SNRI) criteria and listening tests are used to evaluate proposed methods. For this purpose, in equation (5), α is determined to reduce 20dB noise from noisy signal.

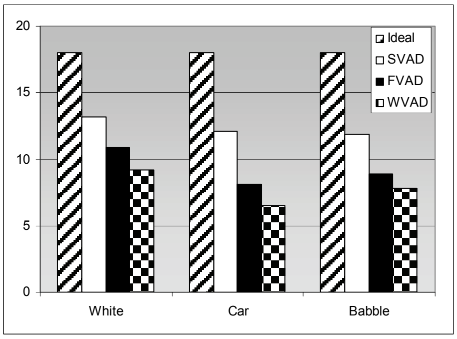

Fig. 4 - SNRI for input SNR=20dB

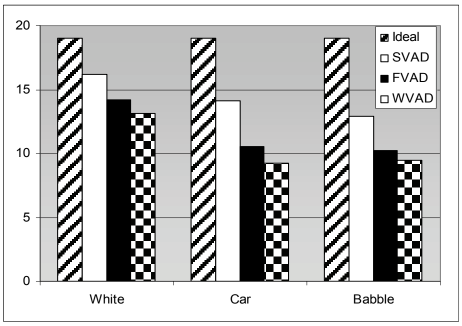

Fig. 5 - SNRI for input SNR=10dB

The average of SNRI for FVAD, SVAD and WVAD in different input SNR values, are reported in figures 4, 5, and 6. When white noise is present, difference of SNRI values for SVAD and FVAD is about 2dB. But this difference in presence of car and babble noises is about 3 to 5dB. Appropriate performance of SVAD in comparison to FVAD is clear in presence of non-white noises. This is due to that non-white noises have different behaviour in different subbands. Then, different decision making for updating noise in each subband is more efficient. In presence of white noise, the performance of SVAD is not significant because of identical distribution of noise in all subbands.

As you can see from the figures, SVAD and FVAD outperform WVAD in most cases. This is due to that noise estimation is more accurate in SVAD and FVAD methods.

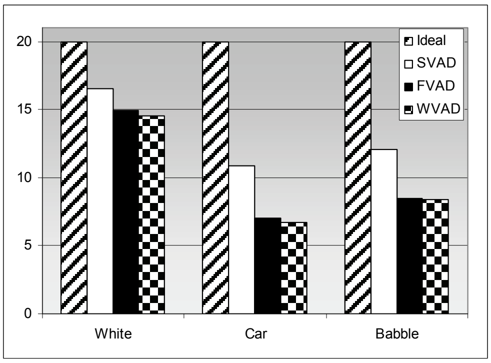

Fig. 6 - SNRI for input SNR=0dB

For listening test, each person listens to the enhanced signal and gives a score between 1 and 3 to signal. Scores 1, 2 and 3 means perfect, medium and poor speech signal quality, respectively. Authors evaluate enhanced signals based on speech artefact and remained background noise.

Figure (7) shows the average of given scores by 20 listeners .These results are shown for speech artefact in presence of white, car and babble noises with input SNR value of 10dB. Based on this figure, SVAD create less speech distortion in enhanced signals especially in presence of car and babble noises.

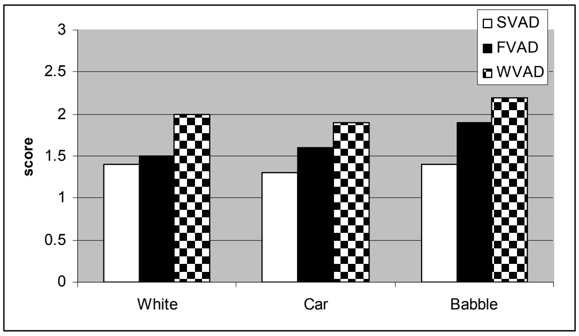

Fig. 7 - Speech Artefact Scores (input SNR=10dB)

Figure (8) shows the listening test scores for the residual noise in presence of white, car and babble noises with input SNR value of 10dB.

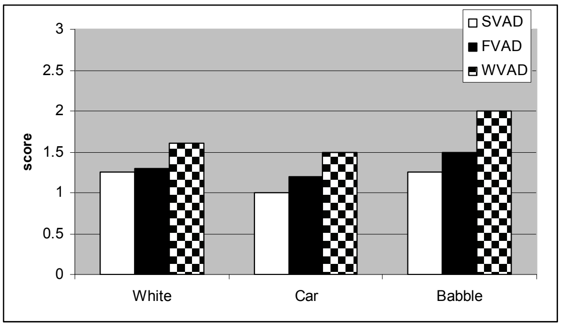

Fig. 8 - Residual Noise Scores (input SNR=10dB)

Same as previous listening test, SVAD has better score than other methods.

Conclusions and directions for future research

Two types of VAD (sub-band and full-band) are suggested in this paper. The evaluation results show that we can obtain better noise estimation using SVAD in wavelet domain.

Integrating noise reduction system and subband VAD has two advantages:

- Appropriate performance of proposed VAD help the noise reduction system to achieve a better noise estimation in continuous noise estimation method.
- Combined VAD and noise estimation method, can improve the performance of VAD classification rate and noise estimation accuracy.

Results of VAD classification rate and SNR improvement of proposed noise reduction system, justify these advantages.

REFERENCES

[1] M. Rahmani, M. Mohammadi, A. Akbari, "Back ground noise control for speech enhancement," in Proc.14th ICEE, IRAN, Tehran, 2006.

[2] F. Beritelli, S. Casale, A. Cavallaero, "A robust voice activity detector for wireless communication using soft computing," IEEE Journal, Vol. 16, Issue. 9, pp 1818-1829, December 1998.

[3] P. Renevey, A. Drygajlo, "Entropy Based Voice Activity Detection in Very Noisy Conditions," in Proc. Eurospeech 2001, Aalborgh 2001.

[4] V. Gilg, C.Beaugeant, M.Schonle, "Methodolog for the design of a robust voice activity detector for speech enhancement," International Workshop on Acoustic echo and Noise Control (IWAEN 2003)

[5] Y. Kida, T. Kawahara," Voice Activity Detection based on Optimally Weighted Combination of Multiple Features," in Proc. Eurospeech , pp. 2621-2624,2005.

[6] A. Grossman, R. Martinet., J. Morlet , "Reading and understanding continuous wavelet transform," Springer-Verlag, pp. 2-20, 1989

[7] I. Yann, S. Ngee, C. Kiat, "Wavelet for speech denoising, " IEEE TENCON vol.2, pp. 479 – 482. 1997.

[8] P. Sovka, P. Pollak, J. Kybic, "Extended Spectral Subtraction," in Proc. *EUSIPCO* 1996, Trieste, Italy, September. 1996.

[9] F. Beritelli, S. Casale, G. Ruggeri, S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," IEEE Signal Processing Letters,Vol. 9 , Issue 3 , pp.85 – 88, March 2002.

[10] C. JIA, B. XU, "An Improved Entropy-Based Endpoint Detection Algorithm," in Proc. ICSLP 2002, Beijing 2002

[11] E. Kosmides, E. Dermatas, G. Kokkinakis, "Stochastic endpoint detection in noisy speech," SPECOM Workshop, pp. 109-114, 1997.

Authors: dr Siamak Rasoolzadeh corresponding author, Department of Computer Engineering, Malayer Branch, Islamic Azad University, Malayer, Iran. Email: Siam.rasoolzade@gmail.com; dr Mohsen Rahmani, Department of Computer Engineering, Malayer Branch, Islamic Azad University, Malayer, Iran. Email: m-rahmani@araku.ac.ir