Performance Analysis of Wavelet-based Voice Activity Detection

Marco Jeub[1], Dorothea Kolossa[2], Ramon F. Astudillo[2], and Reinhold Orglmeister[2]
[1] *RWTH Aachen University, Germany, jeub@ind.rwth-aachen.de*
[2] *TU Berlin, Germany, d.kolossa@ee.tu-berlin.de*

Abstract

The objective of this paper is to analyze the performance of wavelet-based voice activity detection algorithms (VAD) and to contrast it with that of the VAD standardized for the AMR-WB (Adaptive Multi-Rate Wideband) speech codec. Experimental results in clean, noisy and reverberant environments show that wavelet approaches lead to good results with respect to speech clipping and offer a much lower computational complexity. Integration of these algorithms into a Hidden Markov model (HMM) speech recognizer shows that the recognition performance using the AMR VAD can also be obtained or improved upon by wavelet based approaches, again at a notably reduced computational effort.

Introduction

Voice activity detectors (VAD) are common algorithms in digital speech processing and offer a wide variety of applications. The main fields of application are the use in speech transmission systems with discontinuous transmission to reduce the activity on the channel and for noise estimation techniques applied to noise reduction and dereverberation. In this contribution we also investigate these algorithms to improve systems for automatic speech recognition (ASR). Due to the special demands of ASR, the VAD should be able to detect all speech frames correctly and should be very robust even in environmentally difficult conditions. The experiments described in this paper are based on speech recognition using Hidden Markov models (HMM) and mel-frequency cepstral coefficients (MFCC).

Wavelet Transform

Fundamental parameters of a signal are given by its distribution over time and frequency which are obtained by the short-time Fourier transform (STFT). For this purpose the signal $f(t)$ is multiplied by a window function $w(t)$, then, the Fourier transform of this product is computed. This process is repeated by shifting the window over the signal and computing the Fourier transform for every windowed block:

$$F(t,\omega) = \int_{-\infty}^{\infty} f(\tau) w(\tau - t) e^{-j\omega\tau} d\tau \qquad (1)$$

$$= \int_{-\infty}^{\infty} f(\tau) w_{t,\omega}(\tau)^* e^{-j\omega\tau} d\tau, \qquad (2)$$

assuming $w(t)$ to be centered at $t = 0$. The resulting spectrogram provides an estimation of the frequency components within a certain time interval. However, the time resolution is the same for all frequency bands. A central advantage of wavelet analysis is the opportunity to obtain variable sized time-frequency regions. For this purpose, the continuous wavelet transform (CWT) is defined as the integral over the signal multiplied by scaled and shifted versions of a wavelet function

$$\gamma(s,\tau) = \int_{-\infty}^{\infty} f(t) \psi^*_{s,\tau}(t) dt, \qquad (3)$$

where * denotes complex conjugation. The wavelets $\psi_{s,\tau}(t)$ are generated from a mother wavelet $\psi(t)$ by

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right) \qquad (4)$$

with the scaling and translation factors s and τ. The resulting coefficients are functions of scale and position. A low scale (*detail part*) indicates a compressed wavelet which detects rapidly changing details or high frequencies, whereas a high scale (*approximation part*) stretches the wavelet, showing slow changes or low frequencies.

A very efficient method to use the discrete wavelet transform (DWT) for digital signals was developed by Mallat in [1]. Based on multiresolution analysis, it can be shown that the DWT can be obtained using a simple digital filter-bank. The downsampled outputs of the specific highpass and lowpass filters are named detail coefficients $c_1(p)$ and approximation coefficients $d_1(p)$ respectively. A multi-level decomposition can be achieved by an iterated filter-bank or so-called wavelet decomposition tree.

Voice Activity Detection

In this section, an overview of wavelet-based voice activity detectors and the AMR-WB VAD, used for comparison in the experiments, are presented. Early studies on voice activity detection where carried out e.g. in [2] and demonstrated the difficulties and challenges in adverse environments. Classical VAD algorithms extract parameters (e.g. short-term energy, zero crossing rate, autocorrelation coefficients, LPC distance features) for segments of fixed length. This leads to a limited flexibility in the time-frequency resolution. In contrast, the wavelet transform shows a low frequency but high time resolution at high frequencies and low time but high frequency resolution at low frequencies. These features of the time-scale analysis of the wavelet-decomposition are well suited for speech analysis and have already been utilized in many applications.

Wavelet-based VAD

In recent years, several authors have investigated the advantages of the wavelet transform for voice activity detection. Most of these methods are based on computing the wavelet decomposition of the input speech and extracting features from the wavelet decomposition. In [3], four different energy-based parameters are computed which are basically difference measures between the detail coefficients or between the detail and the approximation coefficients. [4] uses a similar coefficient comparison and introduces the nonlinear Teager energy operator (TEO) [5] to expand the differences between specific coefficients. Furthermore, [6] applies several different functions to enhance the features (e.g. sigmodial function, hyperbolic tangent function). The cross-correlation between different decomposition stages is used in [7]. In [8, 9], the signal is divided into 17 critical subbands by using the wavelet packet transform. The following two approaches have been identified to be the most promising candidates and are used in the upcoming experiments.

Wavelet VAD by Shaojun et al.

The first algorithm we explain in detail is based on [10] with modifications in terms of different parameters and wavelet bases. These adjustments lead to considerably better results compared to the original implementation. Since for large scales, the detail coefficients are mainly determined by speech and usually show smaller values for noise [1], at a proper scale, the average energy of speech is greater than the one of noise. This relation can be written as

$$\frac{1}{N}\sum_{p=1}^{N} c_j^y(p)^2 > \alpha \frac{1}{N}\sum_{p=1}^{N} c_j^w(p)^2, \qquad (5)$$

where N and α represent the number of wavelet coefficients and a scaling factor $\alpha > 1$. Moreover $c_j^y(p)$ and $c_j^w(p)$ denote the detail coefficients of noisy speech and noise at scale j. The algorithm uses the DWT to decompose the signal into subbands and compares the energy of detail coefficients at scales $j = 3$ and $j = 4$. During the first 5 frames ($m = 1, .., 5$), which are assumed to contain noise only, the root-mean square (RMS) is calculated by

$$\bar{c}\,_j^w = \sqrt{\frac{1}{5N}\sum_{p=1}^{N}\sum_{m=1}^{5} c_j^w(p,m)^2}, \qquad (6)$$

where $c_j^w(p,m)$ represents $c_j^w(p)$ of frame m. For each of the following frames, $\bar{c}\,_j^y = \mathrm{RMS}\left\{c_j^y(p)\right\}$ is calculated and $\bar{c}\,_j^w$ is updated during silence periods. The VAD flag is set if $(\bar{c}_3^y + \bar{c}_4^y) > (\alpha\,\bar{c}_3^w + \beta\,\bar{c}_4^w)$. We propose the use of 1*st* order Daubechies wavelets (DB1) and the scaling factors $\alpha = 0.8$ and $\beta = 0.6$.

Wavelet VAD by Pham et al.

The second wavelet VAD algorithm under consideration is based on [4, 11]. It uses the TEO and several further enhancement mechanisms. The wavelet coefficients $c_j(p)$ and $d_j(p)$ are obtained by the use of a 3-stage wavelet decomposition tree with DB3 wavelets. To improve the discrimination of speech classes under severely noisy conditions, the TEO coefficients $\widetilde{c}_j(p)$ and $\widetilde{d}_j(p)$ are calculated by

$$\widetilde{c}_j(p) = \mu(c\,_j^2(p) - c_j(p+1)\cdot c_j(p-1)) \qquad (7)$$

$$\widetilde{d}_j(p) = \mu(d\,_j^2(p) - d_j(p+1)\cdot d_j(p-1)) \qquad (8)$$

with the scaling factor $\mu = 10$. From these enhanced coefficients, the power ratio λ and power difference δ are extracted, i.e.,

$$\lambda = \frac{\frac{1}{N_{c1}}\sum_{p=1}^{N_{c1}} \widetilde{c}_j(p)^2}{\frac{1}{N_d}\sum_{p=1}^{N_d} \widetilde{d}_j(p)^2} \qquad (9)$$

$$\delta = \frac{1}{N_d}\sum_{p=1}^{N_d} \widetilde{d}_j(p)^2 - \frac{1}{N_c}\sum_{p=1}^{N_c} \widetilde{c}_j(p)^2 \qquad (10)$$

where N_d, N_c, N_{c1} are the total numbers of approximation, detail coefficients and detail coefficients at scale $j = 1$. The feature enhancement consists of two steps. First, a nonlinear scaling is applied to λ and δ and secondly, the Savitzky-Golay polynomial filter is applied for smoothing. The scaled and filtered features λ' and δ' are compared against the thresholds $\theta = Q_{33}(\lambda')$ and $\vartheta = Q_{50}(\delta')$ where Q_x denotes the x-quantile. Silence is detected when the frame is neither classified as voiced nor as unvoiced. A similar approach in [6] reduces the classification problem to the power difference feature. As in the description above, δ is calculated after wavelet transform and Teager energy computation. The feature enhancement shows better results if a sigmoidal function is applied to δ' as

$$\delta'_s = \frac{1 - e^{-2\delta'}}{1 + e^{-2\delta'}} \qquad (11)$$

and smoothed by median filtering over the duration of five frames. The adaptive threshold is computed by the quantile $Q_{50}(\delta'_s)$ of the enhanced and median filtered power difference. Additionally, a hangover scheme is adopted to smooth the results and prevent detection faults. Pauses smaller than 200ms are relabeled as speech and short talk-spurts (< 100ms) are excluded. In our experiments, replacing the median filter by a mean filter of four frames and shortening the minimum pause duration to 100ms has notably improved the results.

Adaptive Multi-rate Codec VAD

The AMR-WB VAD [12] is based upon spectral estimation and periodicity detection. It uses the short-time energy levels of different frequency bands to distinguish whether a frame of 20ms contains speech or not. More precisely, the speech frame is divided into 12 subbands using a critically decimated IIR-filter-bank consisting of allpass filters. For each of these bands, the background noise level is estimated based on the level of previous frames, the last VAD decisions and the signal stationarity as well as a tone-flag. The tone detection indicates strongly periodical signals (e.g. signaling tones, voiced speech), which are calculated using the open-loop pitch

		AMR		Shaojun		Pham	
Noise	SNR	SAN	VAR	SAN	VAR	SAN	VAR
Babble	20	0.0	100	0.0	98.3	1.7	92.5
	10	0.0	100	0.8	97.5	3.0	91.7
	0	0.0	100	1.7	96.7	5.0	91.7
Reverb and Babble	20	0.0	100	0.0	98.3	0.0	89.2
	10	0.0	100	0.0	98.3	0.0	89.2
	0	0.0	100	0.0	98.3	1.7	90.0
Car	20	0.0	86.0	0.8	82.5	0.3	94.2
	10	0.0	94	2.2	77.5	0.8	93.3
	0	0.0	100	4.5	74.2	2.5	89.2
Clean speech		0.0	65.3	0.2	87.5	0.0	94.2

Table 1: VAD performance under various conditions.

	AMR	Shaojun	Pham
real time factor	0.41	0.05	0.06

Table 2: Computational effort to perform VAD, measured as real time factor on 20min of 16kHz speech for a standard Linux PC running Matlab on a dual core 2.4GHz processor.

	clean	15 dB	10 dB	5 dB	0 dB
no VAD	**1.3%**	8.2%	22.0%	51.8 %	85.1%
AMR	1.5%	7.9%	17.6%	39.9%	74.2%
Shaojun	1.6%	8.0%	19.4%	44.1%	77.7%
Pham	1.7%	**7.6%**	**17.5%**	**37.2%**	**66.1%**

Table 3: Word error rate WER for clean training (purely additive noise), bold face indicates best performance

gain parameter of the speech encoder. A signal-to-noise ratio is obtained for every frame, dividing the calculated sub-band levels by the estimated background noise levels. Finally, this SNR is compared with an adaptive threshold to determine an intermediate VAD decision subsequently augmented by a hangover addition.

Experiments and Results

In the first experiment, we evaluate the detection performance as well as the computational complexity of the VAD algorithms. The second one evaluates the VAD in conjunction with a speech recognition system.

VAD Evaluation

The implemented VADs are evaluated on speech files from the GRID database [13]. The test is performed using 100 sentences by four different speakers (female and male). The speech signals are artificially corrupted by additive *car* and *babble noise* out of the NOISEX-92 database and convolved with a measured room impulse response of an office room ($R_{60} = 0.37\,\mathrm{s}$) [14]. The reference voice activity labels from the database are compared with the labels computed by each algorithm. The measurements are the missed detection rate or speech-as-noise (SAN) defined by SAN $= \frac{N_{sn}}{N} \times 100$ [%] and the voice-activity-rate (VAR) VAR $= \frac{N_s}{N} \times 100$ [%], where N_{sn}, N_s and N are the numbers of speech frames detected as noise, all frames detected as speech and the overall frame count. The SAN measurement is an important indicator for the subjective quality of a speech signal and determines the performance of a speech recognition system. This parameter is usually required to be zero or very low. Table 1 shows the performance characteristics. It is evident from the experiments that the AMR algorithm shows the best performance in terms of a low SAN rate, which is higher for the tested wavelet VADs. Nevertheless, the VAR values increase very fast with decreasing SNR for all methods and reach the 100% bound very rapidly so that the entire segment is detected as speech. It can further be seen that in the reverberant case, the SAN values become lower since room reverberation smears the signal energy into the silence periods and prevents from clipping in low energy regions. We conclude that the wavelet methods are favorable for non-stationary noise if a low degree of false detections is acceptable. As a second performance measure, the computational effort of the three VAD algorithms is also determined. As Table 2 shows, the effort is significantly reduced when a wavelet-based VAD is applied. It must however be said that the AMR algorithm requires around 45% of the computing time for the linear prediction and open-loop pitch analysis.

ASR Evaluation

Speech recognition is performed by means of word-based HMMs, consisting each of 16 states in a left to right model. The silence model is equipped with three states in the form of a Bakis model. The models are trained on 13 mel-frequency cepstral coefficients with first and second derivatives, which are obtained after first-order preemphasis and an STFT with a window size of 32ms. Utterance-wise cepstral mean subtraction is performed to improve robustness. The output densities of the HMMs are modelled as 39-dimensional mixture of Gaussian distributions with 4 mixtures each. The training of the HMMs is carried out using the Hidden Markov Model Toolkit (HTK) [15] and recognition takes place using a Matlab-based Viterbi decoder developed at TU Berlin [16]. In order to analyze the application of VAD for speech recognition, the AURORA-5 database is used. It is a downsampled version of the TI-DIGITS impaired by partly non-stationary background noise and artificially convolved with impulse responses with reverberation times R_{60} between 0.3 and 0.5 seconds.Two tests are conducted, one with matched and one with clean condition training. In both cases, the SNR is varied between 0dB and clean, and 8700 utterances of adult speakers are used for each test case. The resulting word error rate is determined by $WER = \frac{D+I+S}{N} \cdot 100\%$, with N as the number of reference labels, D the deletions, S the substitutions and I the insertions.

Clean Training

When training is carried out on clean data and tests are performed on distorted data, the use of appropriate noise reduction techniques would be required as seen from the baseline performance in Table 3. However, in order to assess the performance of VAD in itself, the tests here were carried out without additional signal preprocessing. Among the tested VAD algorithms, best performance is obtained by the AMR standard VAD for clean speech, but the performance of the algorithm by Pham et al. significantly exceeds that of the standardized

	clean	15 dB	10 dB	5 dB	0 dB
no VAD	4.6 %	6.3 %	10.9 %	21.0%	41.1%
AMR	**4.0%**	**5.9%**	**10.1%**	20.1%	40.9%
Shaojun	**4.0%**	**5.9%**	10.2%	**19.7%**	**39.6%**
Pham	5.3%	6.9%	11.1 %	21.1%	41.6%

Table 4: Word error rate WER for matched training (office room and living room, mean $R_{60} = 0.4s$)

VAD especially at low signal to noise ratios.

Matched Training

The recognizer is trained on the noisy, convolved training set of the Aurora 5 database, which is preprocessed by the VAD under investigation before training. Results are shown in Table 4. Again, VAD is helpful to improve recognition results, but when matched training is performed, much of the VAD can also be performed implicitly by the silence model of the HMMs. However, there remains a performance improvement throughout the range of tested SNRs, for both the AMR VAD and the algorithm proposed by Shaojun et al. Here, this first of the presented wavelet based VADs outperforms all other strategies, but the margin of improvement is not large, giving a relative error rate reduction of 5.4% compared to no VAD and 2.0% compared to the AMR VAD. Moreover, it can be seen that the VAD by Pham, while superior for clean training, is not a good candidate for matched training. This may be due to its large speech as noise count under instationary noise conditions, precluding a reliable training of the HMM silence model. Overall, for matched training the wavelet VAD by Shaojun is clearly the superior algorithm both in terms of computational effort and recognition accuracy.

Conclusions

In this paper, the application of wavelet-based voice activity detectors was studied. Two wavelet-based detectors were compared to the standardized AMR-WB VAD in terms of error rates, computational effort and recognition error rate. The performed experiments show that the VAD algorithms differ in their robustness against background noise and reverberation. While the wavelet based algorithms show good performance in terms of a low voice-activity-rate, they detect more speech frames as silence than the AMR VAD. From the implementation perspective, the wavelet transform has a relatively low computational complexity, about 15% of the AMR VAD, due to the efficient filter-bank integration. Concerning speech recognition, VAD algorithms are most helpful when training is performed on clean data, whereas for matched training, the performance gains by VAD are modest. However, in all cases, performance can be gained by applying a VAD, and in each case, wavelet based voice activity detection can outperform the AMR VAD if the appropriate algorithm is chosen. Thus, we see wavelet-based VAD as a useful enhancement of automatic speech recognition systems. Taking the computational effort into consideration, its behavior can be preferable to that of AMR-specified VAD, especially when it is needed to be used under mismatched conditions.

References

[1] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.

[2] L. Rabiner and M. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *The Bell Systems Tech. Journal*, vol. 54, 1975.

[3] J. Stegmann and G. Schroder, “Robust voice-activity detection based on the wavelet transform,” in *Proc. IEEE Workshop on Speech Coding*, 1997.

[4] T. Pham and L. Weruage, “Time-frequency analysis for voice activitiy detection,” in *Proc. of SPPRA*, Innsbruck, Austria, 2006.

[5] J. F. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc.*, 1990.

[6] T. Pham, M. Neffe, and G. Kubin, “Robust voice activity detection for narrow-bandwidth speaker verification under adverse environments,” in *Proc. of Interspeech*, Antwerp, Belgium, 2007.

[7] N. S. A. Kadel and A. M. Refat, “End points detection for noisy speech using a wavelet based algorithm,” in *Proc. NRSC*, 1999.

[8] S. Chen and J. Wang, “A wavelet-based voice activity detection algorithm in noisy environments,” in *Proc. Int. Conf. on Electronics, Circ. and Systems*, 2002, vol. 3, pp. 995–998.

[9] S. Chen and H. Wu, “Robust voice activity detection algorithm based on the perceptual wavelet packet transform,” in *Proc. ISPACS*, 2005.

[10] J. Shaojun, G. Haitao, and Y. Fuliang, “A new algorithm for voice activity detection based on wavelet transform,” in *Proc. Int. Symposium on Intelligent Multimedia, Video and Speech Proc.*, 20–22 Oct. 2004, pp. 222–225.

[11] T. Pham, *Wavelet Analysis For Robust Speech Processing and Applications*, Ph.D. thesis, Graz University of Technology, Austria, 2007.

[12] TS 26.194, *Adaptive Multi-Rate - Wideband speech codec, Voice Activity Detector*, V6.0.0, 3GPP, 2004.

[13] M. Cooke and J. Barker, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Amer.*, vol. 120, pp. 2421–2424, 2006.

[14] M. Jeub, M. Schäfer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proc. Int. Conf. on Digital Signal Proc.*, Santorini, Greece, to appear.

[15] S. Young, *The HTK Book*, Cambridge University Engineering Department, UK, 2002.

[16] D. Kolossa, *Independent Component Analysis for Environmentally Robust Speech Recognition*, Ph.D. thesis, Technical University of Berlin, 2008.