6th Congress of Alps-Adria
Acoustics Assosiation
16.-17. October 2014
Graz, Austria

Austrian Acoustics Association
Österreichische Gesellschaft für Akustik

Alps Adria Acoustics Association

ISBN: 978-3-200-03921-6

COMPARISON OF VOICE ACTIVITY DETECTION METHODS IN REALISTIC NOISE SCENARIOS

Christina Leitner

JOANNEUM RESEARCH Forschungsgesellschaft mbH (christina.leitner@joanneum.at)

Abstract: *In this paper, we compare several voice activity detection (VAD) methods on noisy data. The objective is to find methods that are suitable for improving the speech communication over mobile phones in loud environments, e.g., when traffic noise is present. First, the methods are applied on synthetic data, where traffic noise is added to recordings taken in a quiet environment. To account for the Lombard effect, which is typical for speech in loud environments, we use recordings of a specially designed database. Second, real-world examples with highway traffic noise are tested. For these examples, short dialogs over the telephone were recorded with one person situated near the street. The different VAD methods are evaluated by the true positive rate, the false positive rate, and by plotting the resulting receiver operator characteristic (ROC) curves. In the case of synthetic data, the reference VAD is derived by applying an energy-based VAD algorithm on the clean data. We test recordings with 5, 0, and -5 dB signal-to-noise ratio (SNR). In the case of real-world data, the reference is derived using a laryngograph, which detects voiced portions within the speech signal. The experiments show that an algorithm based on vowel detection achieves good results in all SNR conditions.*

Keywords: Voice activity detection, high noise, mobile environments, traffic applications

1. INTRODUCTION

Voice activity detection is applied to find out where speech is present in a signal. This information can be used in speech coding and speech transmission, speech enhancement or speech recognition. For example, in speech coding and transmission signal frames without speech can be encoded with reduced accuracy or not be transmitted at all. Many VAD methods are based on statistical measures that are derived from the speech signal such as the short-term energy, the zero crossing rate, the periodicity, the pitch, cepstral features or certain distances measures [1, 2].
In our application, we apply voice activity detection on signals with background traffic noise. We assume the following emergency scenario: After a car break-down a person is calling at the traffic control room. VAD is applied on the received signal and the signal frames without speech are attenuated to reduce the perceptual load imposed on the operator answering the call. The calling speaker is situated in a tunnel or near a highway, due to this fact the incoming signal is corrupted by traffic noise. Furthermore the signal is narrowband because of the transmission over the telephone line. In order to retain the speech intelligibility, as few speech frames as possible should be attenuated. Therefore, the VAD must achieve a true positive rate as high as possible while keeping the false positive rate low.
We compare four different algorithms on synthetic and realistic data. The tested algorithms are the CrossCorr algorithm [3], the MaxPeak algorithm [3], the SubbandSNR [4], and the PeakValleyDifference algorithm [5]. The CrossCorr and the MaxPeak algorithm are based on the correlation in time domain. The SubbandSNR detects speech based on the average *a posteriori* SNR derived in frequency domain. The PeakValleyDifference algorithm detects if vowels are present in the signal to derive the VAD.
For the synthetic data, utterances of the CMU_SIN database [6] are corrupted by additive traffic noise at 5, 0, and -5 dB SNR. This database contains recordings taken while noise is played to the speaking person through earphones to simulate the Lombard effect typical for speech in noisy environments. For the realistic data, we designed a dialog scenario where spontaneous speech of a person near a highway is transmitted over the telephone and recorded at the remote end.
The traffic noise is non-stationary, consequently the used detection measures vary. We extend the algorithms by tracking these measures to adapt to different instantaneous SNRs. Furthermore, we apply smoothing as a post-processing step to improve the detection result. The results show, that the PeakValleyDifference algorithm is most robust on the synthetic and the realistic data and achieves the highest true

positive rate at the lowest false positive rate in all SNR conditions.

This paper is organized as follows: Section 2 provides a summary of the tested algorithms. In Section 3, the creation of the used datasets is described, including the recording process for the realistic data. Section 4 provides details on the experimental setup like the setting and adaptation of the detection threshold. In Section 5, the results are presented and discussed. Section 6 concludes the paper and gives an outlook on future work.

2. ALGORITHMS

2.1. MaxPeak

The MaxPeak algorithm [3] distinguishes between speech and noise by exploiting the periodicity of voiced sounds in speech. For voiced sounds, the maximum peak of the normalized autocorrelation is expected to be higher than for noise. Therefore, the autocorrelation is computed and the maximum peak value is compared against a threshold. As pre-processing step a DC-removal and pre-emphasis are applied. The computation of the autocorrelation is restricted to lag ranges corresponding to the expected pitch periods of 2 to 20 ms – which is equivalent to a pitch of 50 to 500 Hz. For each frame $x_l[n]$ of the noisy signal $x[n]$ the normalized autocorrelation is computed by

$$R_l[z] = \frac{\sum_{n=1}^{L-z} x_l[n]x_l[n+z]}{\sum_{n=1}^{L} x_l^2[n]}, \tag{1}$$

where n is the sample index, z is the autocorrelation lag and L is the number of samples in $x_l[n]$. For detection the measure

$$M'[l] = -\log(1 - M[l]) \tag{2}$$

is derived, where

$$M[l] = \max_z R_l[z] \tag{3}$$

is the so-called MaxPeak score. To detect speech, $M'[l]$ is compared against a detection threshold.

2.2. CrossCorr

The CrossCorr algorithm [3] is also based on the autocorrelation. Due to the periodic nature of voiced speech the autocorrelation function is also periodic within the 2 to 20 ms lag range. The CrossCorr algorithm provides a measure for this periodicity. The computation of the normalized autocorrelation $R_l[z]$ is equivalent to the MaxPeak algorithm above, however, without the pre-emphasis filtering. The autocorrelation is segmented into periods P_y by using every two zero-crossing points as boundary. To derive a measure for the degree of periodicity, each segment P_y is cross-correlated with its posterior segment P_{y+1} by applying

$$\hat{R}_l[z'] = \sum_{m=1}^{L'-z'} P_y[m]P_{y+1}[m+z'], \tag{4}$$

where m is the sample index, z' is the correlation lag, and L' is the number of samples in P_y. The maximum values of the autocorrelations $\hat{R}_l[z']$ are summed up and this results in the so-called CrossCorr score

$$C[l] = \sum_{y=1}^{M-1} \max \hat{R}_y[z']. \tag{5}$$

For each frame, the value of $C[l]$ is compared against a threshold to estimate if voice is active or not.

2.3. SubbandSNR

The SubbandSNR algorithm [4, 7] compares the estimated average *a posteriori* subband SNR against a detection threshold. The *a posteriori* SNR $\gamma(k)$ in one frequency band is derived from the DFT coefficient $X[k]$ of noisy speech, such that

$$\gamma_k = \frac{|X[k]|^2}{\sigma^2_{k,\text{noise}}}, \tag{6}$$

where k is the index of the frequency bin and $\sigma^2_{k,\text{noise}}$ denotes the power of the noise in the k^{th} frequency band. The average *a posteriori* SNR evaluates to

$$\bar{\gamma} = \frac{1}{N/2-1} \sum_{k=1}^{N/2-1} \gamma_k, \tag{7}$$

where N is the number of DFT coefficients. In frames where the voice is not active the expectation of γ_k is equal to 1. For detection, $\bar{\gamma}$ is therefore compared against the threshold $\gamma_{\text{thr}} = 1 + a\sqrt{\frac{1}{N/2-1}}$, which is derived from the theoretical variance of the frequency bins and where a is in the range $2 \le a \le 4$.

2.4. PeakValleyDifference

The algorithm of Yoo and Yook [5] is based on the observation that vowels have a specific spectrum that is different from the spectrum of most noise types. Therefore the detection of vowels can be used to discriminate between speech and background noise. In a training phase, speech utterances are segmented and vowels are extracted. The spectra of the frames of each vowel instance are averaged and the averages are clustered. Then, the peak signature $S_i[k]$ is extracted from the centroid of each vowel cluster. In the test phase, the peak valley distance between the spectrum $X_l[k]$ of the noisy test utterance and the peak signatures is computed by applying

$$\begin{aligned}\text{PVD}(X_l, S_i) = &\frac{\sum_{k=0}^{N-1} X_l[k] \cdot S_i[k]}{\sum_{k=0}^{N-1} S_i[k]} \\ &- \frac{\sum_{k=0}^{N-1} X_l[k] \cdot (1 - S_i[k])}{\sum_{k=0}^{N-1}(1 - S_i[k])}.\end{aligned} \tag{8}$$

For each frame, the maximum distance of all peak signatures

$$P[l] = \max_i \text{PVD}(X_l, S_i) \tag{9}$$

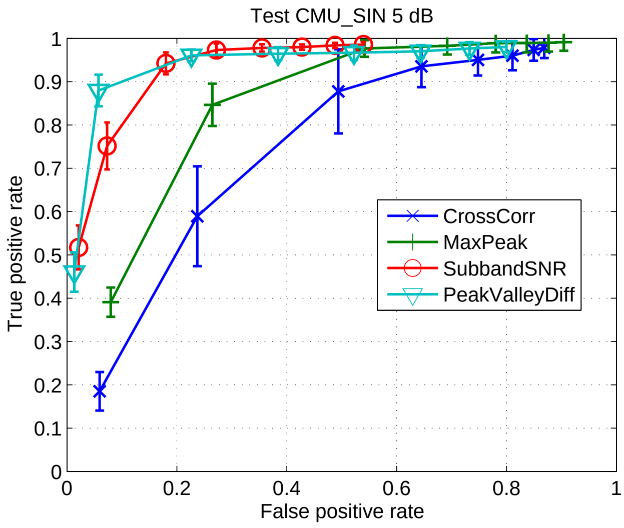

Fig. 1. Receiver operator characteristic (ROC) curve for the results on the test set of the CMU_SIN database for 5 dB SNR and smoothing with windows of 1, 5, 15, 25, 35, 45, 55, and 65 frames length. The bars mark the standard deviation of the true positive rate.

is computed and compared against a threshold to decide whether speech is present or not. The advantage of this algorithm is that it is not necessary to know the type of noise. The only restriction is that the noise should not contain speech as in babble noise. In this case, the speech can be hardly distinguished as the noise also contains vowel spectra.

3. DATA

3.1. Synthetic data: CMU_SIN

To create the synthetic data, we used the utterances of the CMU_SIN database [6]. This database contains recordings of speech uttered by a speaker exposed to a noise playback wearing headphones. This way, the Lombard effect typical for speech in loud environments is provoked. The CMU_SIN database contains 497 utterances of a male speaker of American English. The database is divided into a training, development, and test set containing, 525, 22, and 50 utterances, respectively. The training set is used to extract the vowel peak signatures for the PeakValleyDifference algorithm. For the development and test set, five utterances of the CMU_SIN database are concatenated leaving a random pause of 1 to 2 seconds between the utterances. The active speech level [8] is computed for each test sequence. The noise is cut out at a random position from a recording performed in a tunnel while vehicles are passing. The speech and the noise signal are filtered by the modified intermediate reference system filter used in [9] to simulate the telephone characteristic. The signals are added at 5, 0, and -5 dB SNR.

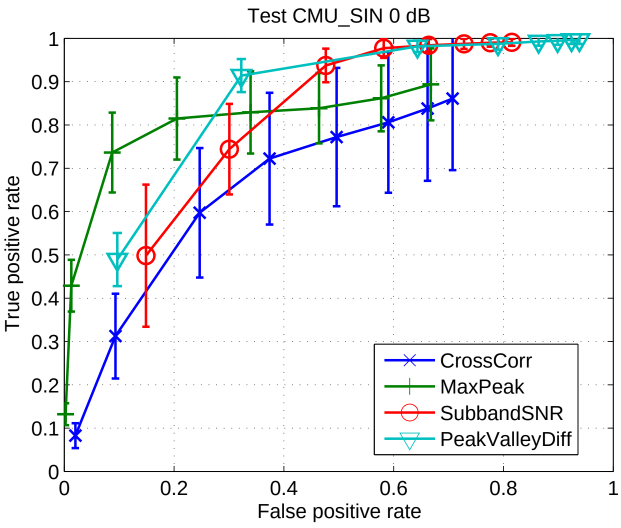

Fig. 2. ROC curve for the results on the test set of the CMU_SIN database for 0 dB SNR.

3.2. Realistic data: PHONE

For the realistic data, we set up a telephone dialog scenario where one person is located near a highway and the other in an office environment. The first person calls the second via the mobile phone. The telephone at the receiving end is equipped with an audio jack socket where the speech signal is taken off and recorded.

In order to create a stress situation, a kind of guessing game is set up. The first person is asked to explain given terms to the second person who must guess them. As many terms as possible should by guessed in a limited amount of time. To make it more challenging, certain terms related to the guessing term are forbidden. Listening to the recorded data confirms that the speech during the guessing game sounds more stressed than the speech before starting the game, e.g., at the beginning of the call.

To derive a reference for the VAD, the calling person wears laryngograph electrodes. The laryngograph detects segments of voiced speech. Unvoiced speech segments are not detected, however, this is a minor problem as (i) we only need a coarse VAD and (ii) we close gaps up to 200 ms in between speech segments by applying the closing operation known from image processing [10]. In addition to the laryngograph signal, the speech signal of the first person is recorded by a standard microphone. This speech signal is aligned with the telephone speech signal using the correlation and the same time stamps are used to align the VAD reference with the telephone signal. We recorded two persons, one male and one female who each talked approximately 10 minutes in German.

Analysis of the recorded data showed that the SNR is rather high in comparison to the synthetic data. The SNR estimate based on the reference VAD varies between 13 dB and 26 dB and is on average 20 dB. Considering only the SNR, this dataset is therefore less challenging.

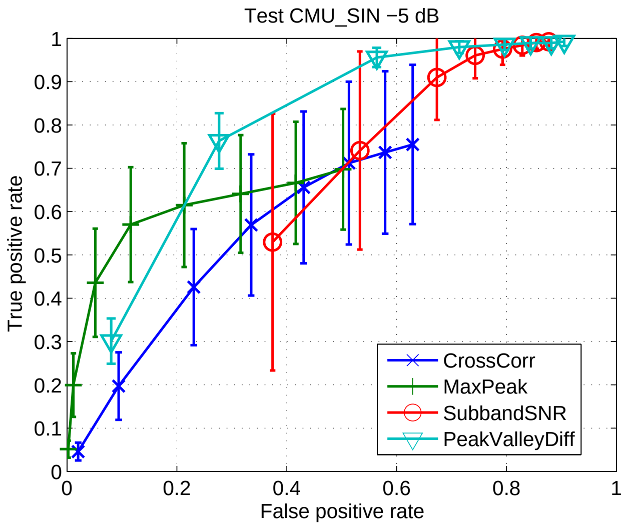

Fig. 3. ROC curve for the results on the test set of the CMU_SIN database for 0 dB SNR.

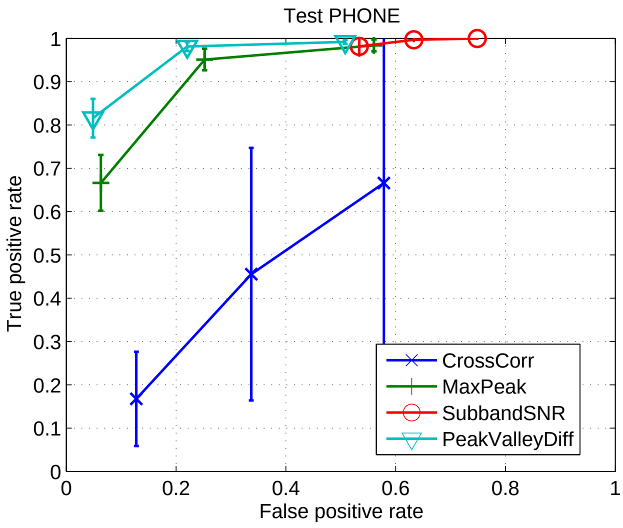

Fig. 4. ROC curve for the results on the PHONE data with smoothing with windows of 1, 5, and 15 frames length.

4. EXPERIMENTAL SETUP

4.1. Peak extraction for the PeakValleyDifference algorithm

For the PeakValleyDifference algorithm peak signatures need to be extracted from speech data. For this purpose, a standard speech recognizer is trained using HTK and the utterances of the CMU_SIN training set. Forced alignment is applied and the vowel segments are cut out based on the phoneme boundaries. Using these vowel segments, the procedure described in 2.4 is applied to derive the peak signatures.

4.2. Threshold parameter setting and adaptation

All described algorithms make use of the comparison against a detection threshold. For the SubbandSNR algorithm, the threshold is theoretically motivated. For the other algorithms, the threshold needs to be derived experimentally. Now, let $d[l]$ be the generalized notation for the detection measures $M[l]$, $C[l]$, and $P[l]$. For both datasets, we assume that the first frames are free of speech and we use $d[l]$ for defining an initial detection threshold t_{init}.

For the synthetic data, the initial threshold is defined as

$$t_{\text{init}} = \beta \cdot \bar{d}_{\text{init}}, \qquad (10)$$

where $\bar{d}_{\text{init}}$ is the mean of $d[l]$ in the first five frames and β is set according to the SNR, which is assumed to be known. For each SNR, several values for β are tested on the development set and the value resulting in the highest correlation with the VAD reference is chosen.

The used traffic noise is non-stationary and the instantaneous SNR changes considerably over time. The detection measures vary in a similar way. Therefore, we implemented a tracking mechanism to adapt the detection threshold based on this variation. For all algorithms – except of the SubbandSNR – the threshold is adapted in segments where the voice is not active using the update equation

$$t[l+1] = \alpha \cdot t[l] + (1-\alpha)(\beta \cdot d[l]), \qquad (11)$$

where $t[l]$ is the threshold in frame l and α is the update parameter set to 0.95.

For the PHONE data, two possibilities were considered: (i) estimating the SNR and using the according parameter setting from the synthetic data or (ii) using a modified tracking mechanism with SNR-independent initial setting. Experiments showed, that the approach (i) resulted in a high true positive rate at the expense of a high false positive rate. Therefore with preferred approach (ii). We compute the initial threshold t_{init} by

$$t_{\text{init}} = \bar{d}_{\text{init}} + \beta \cdot \text{std}(d_{\text{init}}), \qquad (12)$$

where $\text{std}(d_{\text{init}})$ denotes the standard deviation of d_{init}. Based on this threshold, the following frames are classified into frames of active and inactive voice. In each frame of inactive voice, the detection threshold is adapted according to

$$t[l+1] = \alpha \cdot t[l] + (1-\alpha)(\beta \cdot d[l] + 0.1 \cdot (\bar{d}_{\text{speech}} - \bar{d}_{\text{noise}})), \qquad (13)$$

where $\bar{d}_{\text{speech}}$ is the mean of the detection measure for frames where speech is detected and $\bar{d}_{\text{noise}}$ is the mean of the measure for the other frames. Experiments show that the detection adapts to a threshold leading to robust results even if the initial estimate is poor.

Due to the high SNR in the PHONE data, the γ_k of the SubbandSNR algorithm is considerably higher for some frequency bins when speech is present than with the CMU_SIN data. This results in a high true positive rate with high false positive rate. To account for this, we increased the detection threshold by increasing the parameter a to 10.

	5 dB	0 dB	-5 dB
CrossCorr	2	2.5	2.5
MaxPeak	1.5	2	2
SubbandSNR	3	2.5	2.5
PeakValleyDifference	2	1.5	1.5

Table 1. Parameter values for β – or a for the SubbandSNR algorithm – for threshold computation for the CMU_SIN data.

4.3. Smoothing

The VAD results from simple application of one of the algorithms are sparse, i.e., only a low number of speech frames is correctly recognized as speech. To achieve a higher true positive rage we apply smoothing as a post-processing step. To find a proper length for the smoothing window, several lengths are tested and evaluated as described in Section 5. For the experiments with the CMU_SIN data we tested the window lengths 1 (no smoothing), 5, 15, 25, 35, 45, 55, and 65. For the experiments with the PHONE data we tested shorter windows with length 1, 5, and 15, because the true positive rate was already relatively high.

5. RESULTS AND DISCUSSION

We evaluated the algorithms in terms of true positive rate and true negative rate and additionally plotted a ROC curve. Figures 1 to 3 show the ROC curve for the test set of the CMU_SIN data. Over all SNR conditions and both datasets the MaxPeak, SubbandSNR and the PeakValleyDifference algorithm perform better than the CrossCorr algorithm. The SubbandSNR algorithm performs especially well in the 5 dB condition. This is reasonable, as this algorithm is relatively simple and relies on a good noise estimate, which can be easier achieved in high SNRs. In the -5 dB condition the MaxPeak and the PeakValleyDifference algorithm perform better than the SubbandSNR algorithm. The MaxPeak algorithm, however, does not achieve a true positive rate as high as the PeakValleyDifference algorithm. We therefore prefer the PeakValleyDifference algorithm, as we favor having false positives over missing true positives, i.e., we do not want to lose frames where the voice is active.

The results on the PHONE data are similar to the results of the CMU_SIN data with 5 dB SNR for the PeakValleyDifference algorithm. This is especially interesting, as the peak signatures are the same as used for the experiments with the CMU_SIN database, which is in English, while the PHONE data is in German. Apparently, the peaks signature from English data generalize sufficiently well to German data.

The MaxPeak algorithm performs better then on the CMU_SIN data. The performance of the CrossCorr algorithm is worse. Presumably the reason is a poor detection of the zero crossings, which was noticed during the experiments. The SubbandSNR leads to a high true positive rate, however, at the expense of a high false positive rate. This might be caused by a poor noise estimation and requires further investigation.

6. CONCLUSION AND OUTLOOK

In this paper, we compared four VAD algorithms on synthetic and realistic data. For the experiments with realistic data, we recorded speech in a noisy environment over the telephone line.The PeakValleyDifference performs best among all SNR conditions and on both datasets. It is also robust when using vowel peak signatures extracted from English data for VAD on German data. Three of the four algorithms make use of a detection threshold which needs to be adapted according to the SNR. The setting of this threshold is crucial for the performance, therefore, we will investigate further possibilities to derive a robust setting in different SNR conditions and for different datasets. Furthermore, we plan to extend our PHONE database by recordings taken in a tunnel where the noise is louder than with the present setup near the highway. This way, we want to study the robustness of the algorithms in more challenging realistic scenarios.

REFERENCES

[1] S. Gazor and W. Zhang, “A soft voice activity detector based on a laplacian-gaussian model,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, 2003.

[2] M. Stadtschnitzer, “Reliable voice activity detection under adverse environments,” Master’s thesis, Graz University of Technology, 2008.

[3] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan, “Noise robust voice activity detection using features extracted from the time-domain autocorrelation function,” *Proceedings of Interspeech*, 2010.

[4] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons, 2006.

[5] I.-C. Yoo and D. Yook, “Robust voice activity detection using the spectral peaks of vowel sounds,” *ETRI journal*, vol. 31, no. 4, 2009.

[6] B. Langner and A. W. Black, “Creating a database of speech in noise for unit selection synthesis,” *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[7] J. Häkkinen and M. Väänänen, “Background noise suppressor for a car hands-free microphone,” pp. 300–307, 1993.

[8] ITU-T, “Objective measurement of active speech level,” *ITU-T Recommendation P.56*, 1993.

[9] ——, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” *ITU-T Recommendation P.862*, 2000.

[10] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson Prentice Hall, 2008.