A Robust Voice Activity Detection based on Noise Eigenspace Projection

Dongwen Ying[1], Yu Shi[2], Frank Soong[2], Jianwu Dang[1], and Xugang Lu[1]

[1]Japan Advanced Institute of Science and Technology, Nomi city, Ishikawa, Japan, 923-1292
[2]Microsoft Research Asia, Beijing, China
[1]{dongwen, jdang}@jaist.ac.jp
[2]{yushi, frankkps}@microsoft.com

Abstract A robust voice activity detector (VAD) is expected to increase the accuracy of ASR in noisy environments. This study focuses on how to extract robust information for designing a robust VAD. To do so, we construct a noise eigenspace by the principal component analysis of the noise covariance matrix. Projecting noise speech onto the eigenspace, it is found that available information with higher SNR is generally located in the channels with smaller eigenvalues. According to this finding, the available components of the speech are obtained by sorting the noise eigenspace. Based on the extracted high-SNR components, we proposed a robust voice activity detector. The threshold for deciding the available channels is determined using a histogram method. A probability-weighted speech presence is used to increase the reliability of the VAD. The proposed VAD is evaluated using TIMIT database mixed with a number of noises. Experiments showed that our algorithm performs better than traditional VAD algorithms.

Keywords: Voice activity detection, Principal component analysis, Auto-segmentation, Local noise estimation

1 Introduction

The performance of speech processing systems such as Automatic Speech Recognition (ASR) systems, speech enhancement and coding systems, suffers substantial degradations in noise environments. By applying a robust Voice Activity Detection (VAD) algorithm to those systems, their performances can be improved in the adverse environments. In clean conditions, the VAD systems using short-term energy or zero-crossing features work fairly well [1], but in noisy conditions, a traditional VAD is no longer robust when speech signal is seriously contaminated by noise. It is still a challenging problem to design a robust VAD for noise environments.

In the past twenty years, many researches have been conducted to obtain a robust VAD in adverse environments. Some of the researches paid attention to the intrinsic speech features such as periodic measure [2]. The other methods focused on the

statistical model of speech and noise signals, such as the Gaussian statistical model based VAD [3] [4], Laplacian model based VAD [5] and high-order statistical VAD [6]. However, in low Signal-to-Noise Ratios (SNR) condition, speech features and speech statistical characteristics were not easy to be obtained. To reduce the noise effect, recently, a method combining speech enhancement with VAD was proposed [8]. Their method, however, has the two problems in the speech enhancement stage: residual noise and speech distortion, which brought error to VAD.

In this paper, we propose a novel approach to realize a robust VAD. The basic consideration is that speech usually has a different distribution from noises in the energy domain. If we can sort the components that have low power for noise and high power for speech, it is possible to extract more reliable information for speech even if the average SNR of the noisy speech is low. For this purpose, first, a noise eigenspace is constructed based on an estimated covariance matrix of noise observations using Principal Component Analysis (PCA). Projecting the noisy speech onto the noise eigenspace, the reliable information can be found out in the sub-eigenspace with smaller eigenvalues. Thus, a robust VAD can be realized based on the reliable information. Section 2 introduces the principles of noise eigenspace projection. Section 3 shows the implementation of the algorithm. In Section 4, we give the experimental evaluation, and compare our algorithm with some leading algorithms.

2 Projection in noise eigenspace

This section first investigates the SNR distribution property in a noise eigenspace. Then, we describe how to obtain the noise eigenspace in real application.

2.1 SNR Distribution in Noise Eigenspace

The noise eigenspace is used to describe the property of noise energy distribution. It is constructed from by principal component analysis of noise covariance matrix. Using eigenvalue decomposition, we can get the following relationship between eigenvalues and eigenvectors:

$$C\varphi_k = \lambda_k \varphi_k, \quad k = 1, 2, ..., K \tag{1}$$

where C is the covariance matrix of a zero mean noise signal n, $\varphi(k)$ is the eigenvector corresponding to eigenvalue λ_k. By sorting the eigen-coordinates based on eigenvalues order $\lambda_1 > \lambda_2 > ,..., > \lambda_K$, we get the corresponding eigenvectors $\{\varphi_k | k = 1,2,..., K\}$. The projection of a noisy speech frame x on the kth eigen-coordinate then is written as:

$$y_k = x \cdot \varphi_k \tag{2}$$

Since the noise energy centers on some coordinates, when projecting noisy speech into the noise eigenspace, it is possible to find a sub-eigenspace with few noise energy, hence higher SNR, where we can extract available information. Here, we use a specific noise to demonstrate the idea how to extract available information

from noisy speech based on the noise eigenspace. We construct a noise eigenspace from a period of destroyer-engine noise. A speech sentence is mixed with the period of noise at 0dB. Both the speech and noise are respectively projected into the eigenspace. Since covariance matrix is calculated from the whole period of mixed noise, noise projection energy is actually the noise eigenvalue of the corresponding eigen-coordinate. The results of this processing are shown in Fig. 1. The left panel of Fig. 1 illustrates the initial distribution of projection energy in the original eigenspace. The blue curve is noise projection energy and the red is the projection energy of the clean speech. We sort eigenvalues in a descending order and rearrange the coordinate of the eigenspace according to the sorted order, where speech projections will move with the noise eigenvector in pair. For example, the channel with the maximum noise and the projected speech, shown by the dashed line in the left panel, are transferred to the lowest channel in the sorted noise eigenspace. Thus, a monotonically descending curve of the noise energy is obtained as shown in middle panel of Fig. 1, and the corresponding speech projections are shown in red curve with non-monotonic changes. In the rearranged space, one can see that in the high coordinates the speech's energy is higher than that of noise even though the average SNR is equal to zero or lower. Especially in last coordinates, the SNRs are much larger than the original SNR, as shown in right panel of Fig.1.

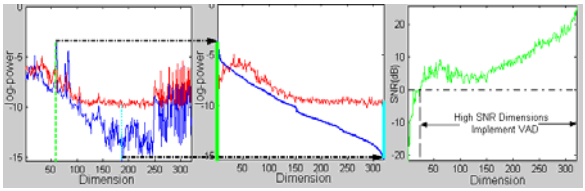

Fig. 1. Energy distributions in a noise eigenspace.

For investigating the generality, the noisy speech projections are testified using eigenspaces of other types of noises out of the NOISEX'92 database. We mixed the noises with clean speech sentences from TIMIT database at given SNR levels. In real application, it's impossible to calculate the noise covariance matrix from the whole period of mixed noise. So, we estimate the covariance matrix by the non-speech period at each sentence beginning (as described in section 2.2).

Then, we project the noise and speech onto the sorted eigenspace and measure the SNR at each coordinate. Here we define the projection SNR ξ_i of the ith coordinate as the difference between the ith coordinate SNR and the mixture SNR, as described in formula (3):

$$\xi_i = 10\log_{10}(S_i / N_i) - 10\log_{10}(S / N) \qquad (3)$$

where S and N are the total energy of a speech sentence and the mixed noise respectively. S_i and N_i are the projection energy of speech and noise at the ith coordinate respectively. The energy in the original space equals the summation of projected energy at each coordinate:

$$S = \sum_{k=1}^{K} S_k \qquad \text{and} \qquad N = \sum_{k=1}^{K} N_k$$

Thus, we further rewrite the formula as:

$$\xi_i = 10\log_{10}(S_i / \sum_{k=1}^{K} S_k) - 10\log_{10}(N_i / \sum_{k=1}^{K} N_k) \qquad (4)$$

From formula, we can find out that projection SNR ξ_i is only concerned with the percentage of energy distribution at the i[th] coordinate. Since, projection SNR has no relationship with the global average SNR, we can easily represent the relationship among projection SNR, eigen-coordinate index and distribution probability by a three-dimension color image.

The color image is constructed by this way. For each sentence, we can calculate its projection SNR at each coordinate. At a given coordinate, we construct a histogram to describe the projection SNR distribution of all noisy sentences, and represent the value as probability of occurrence. So, the probability summation of each coordinate equals to 1. We combine the histograms at all coordinates into a colored image. In this algorithm, the speech sampling rate is 16 kHz, frame length 0.02s and frame shift 0.01s. Thus, the full eigenspace has 320 eigen-coordinates.

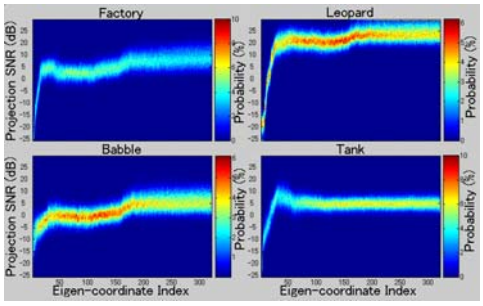

Fig. 2. Projection SNR distribution in noise eigenspace. Vertical axes describe the projection SNR. The color represents its distribution probability.

From the figure, it's easy to understand that the SNR of the projected signal on high dimensional coordinates is greater than that of projection on low dimensional coordinates. In another word, the SNR have an increasing tendency from the low to high coordinates. The statistics experiment shows the projections on eigen-coordinates with smaller eigenvalues always associate with high SNR. Therefore, it's possible to utilize the information of coordinates with smaller eigenvalues and ignore the coordinates with larger eigenvalues to carry out robust VAD.

2.2 Noise Eigenspace Estimation

Noise covariance matrix is the basis of eigenspace calculation. Before implementing VAD in eigenspace, it is necessary to obtain a reliable estimation of noise covariance matrix from noisy speech. Suppose there is somewhat a non-speech period in the

beginning of each sentence, an initial covariance matrix can be estimated from this period. Then, the covariance matrix is updated stepwise using the detected noise.

To obtain a credible estimation of the initial noise covariance matrix, the frame shift is reduced to 0.375ms so that we can obtain 350 noise frames within 140ms at the beginning of sentences. The noise eigenspace is updated based on a time-varying estimation of the covariance matrix $\hat{C}(n)$ ($K \times K$). Giving an initial estimation $\hat{C}(0)$, it is successively updated as:

$$\hat{C}(n) = \alpha \hat{C}(n-1) + (1-\alpha) x(n) x^T(n) \tag{5}$$

where n is time (frame) index, α is a low-pass, forgetting factor with value 0.98, $x(n)$ is the observed noisy signal vector.

As known, eigenvalue decomposition is a time-consuming operation. Since noise is much more stationary comparing to speech signal, it's possible to doing eigenvalue decomposition periodically. On one hand, a longer period for eigenvalue decomposition can save computation time. On the other hand, a shorter period will benefit to an accurate estimation of noise eigenspace. So, a tradeoff is made between computation time and the accuracy of eigenspace.

3 Voice Activity Detection in Noise Eigenspace

In this section, we address how to detect the voice activity in the sub-eigenspace with high SNR. Before the noisy speech projected into noise eigenspace, the input signal is partitioned into homogenous segments as units for VAD decision. We construct channels using high-SNR coordinates and realize a sub-VAD at each channel. At last, the reliable channels with greater SNR will give a voting. The processing block diagram is shown in Fig. 3.

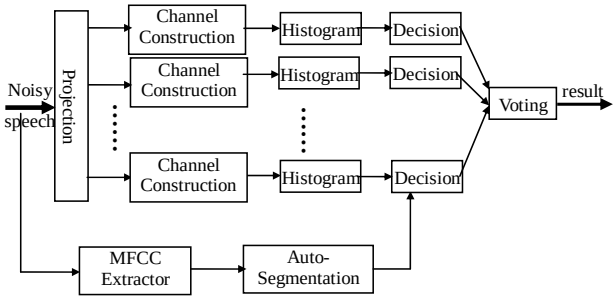

Fig. 3. Block diagram of the proposed VAD

3.1 Auto-segmentation and Channel Construction

Firstly, we use auto-segmentation to partition the frame sequence into homogeneous segments. It is based on the consideration that, in noisy speech signal, the voiced and

unvoiced blocks usually occur as segments consisting of several consecutive frames. The decision results should not transfer between speech and noise frame by frame. Here, homogeneous segments are taken as units for VAD decision, which reduces the problem of spurious changes of speech detection and limits speech-noise transfer times in the decision. The algorithm is a dynamic programming based procedure to minimize the segmentation cost [9]. In our algorithm, eight-dimension MFCC features including the log-power energy are used for auto-segmentation.

Secondly, the noisy speech frames are projected onto the noise eigenspace. Then, the every 10 adjacent projections are grouped into one channel by using the logarithm of the absolute magnitude summation to form a smoothed envelope. There are totally 32 grouped, projected channels in our algorithm.

The constructed channels located at the low dimensional coordinates have low SNR. Those channels bring much speech false alarm and contribute a little to speech hit rate. Therefore, those channels should be ignored in decision. Here, the channel SNR is used to evaluate each channel's reliability. It is estimated based on eigenvalues (average noise energy) and observed projection energy. According to experiments, the channels with SNR less than 2dB should be ignored in VAD decision. The left channels are used for VAD.

3.2 Histogram based Local Noise Estimation

For making a correct final VAD decision, we carried out a sub-VAD decision at each channel. To do so, an appropriate threshold for each channel should be given. We propose a histogram-based method to estimate the sub-VAD threshold. The sub-VAD threshold is decided by noise level and variance of noise log-power. Suppose that the noise log-power of each channel obeys a Gaussian distribution, the problem arrived at estimation of the mean (noise level) and variance of the Gaussian function.

Many approaches such as clustering [9] and GMM fitting [7] have been proposed to estimate noise level in noisy speech. All these methods are based on the following observations in the histograms of log-power energy of noisy speech [10]:

a. In the two peak mode of the histogram, the peak in lower region is usually contributed by background noise, while the peak in lower region is contributed by speech.
b. In general, the noise mode has a salient peak and its variance is smaller than that of speech. The reason is that, as commonly assumed, the energy of the background noise is more stationary than that of speech.
c. The two modes are clearly separated in high SNR conditions. As SNR is decreasing, the two modes are getting closer and eventually merge into one mode.

However, in most situations, the two-peak mode assumption is not kept well. There may be only one peak model in speech pause duration or the mode with more than two peaks on the histogram. Traditional ways for estimation of noise level can not deal with those situations. It is necessary to design a local noise estimation method to deal with one peak, two peaks, and several peaks cases. Our estimation method only concern with noise mode, since noise mode is more salient than speech mode. Based on the basic observation in (a), (b) and (c), we present a local noise estimation method, as following steps:

i. Taking a dynamic range (0~9dB relative to minimum power) to construct the 40-bin histogram. This range is wide enough to include the noise level.
ii. Using a 3-point median filter to smooth the occurrence number, and taking the first peak at left side as the noise level location.

The noise level is the average of noise log-power Gaussian model. It is also assumed that noise log-power less than the noise level is affected little by speech as shown by shadow in Fig. 4. Then, its variance is estimated by the data less than the noise level. Based upon the local noise Gaussian model, we can define a sub-VAD threshold:

$$\text{Threshold} = \mu + \gamma\sigma \qquad (6)$$

where μ is the noise level, σ is the estimated variance, γ is the coefficient for tuning the threshold. Fig. 5 illustrates the sub-VAD threshold estimation of noisy speech at 5dB in factory noise situation using the histogram method. The thick curve in the upper panel is noisy speech power envelope; the thin curve is clean speech power envelope. The dark segments in the middle panel are the detected speech segments. The threshold is calculated using formula (6). The centroids of homogenous segments partitioned by auto-segmentation are compared with sub-VAD threshold. Our local noise estimation method can deal with all cases, whether in speech pause, high or low SNR conditions.

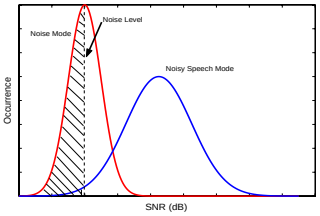

Fig.4. Noise and noisy speech mode

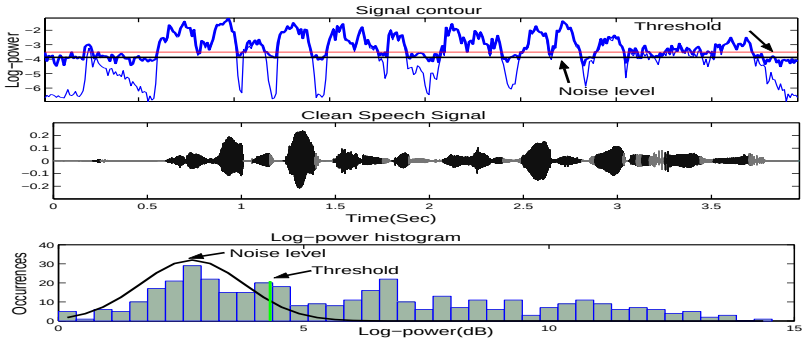

Fig.5. Noise estimation by histogram.

The coefficient γ tuning the sub-VAD threshold in formula (4) should adapt to channel SNR. In high SNR channels, γ should be smaller to make the sub-VAD sensitive to speech and be larger in low SNR channels to avoid speech false alarm.

According to experiments, when γ is linearly interpolated 1.3~1.1 between 2dB~8dB, it achieves better tradeoff between speech false alarm rate and hit rate. If channels' SNR is higher than 8dB, γ equals 1.1.

3.3 Voting and Parameter Adaptation

As mentioned in section 3.1, the channels with SNR less than 2dB are ignored. Only those channels with SNR larger than 2dB take part in the voting. So, the numbers of voting channels varies with average SNR conditions, it's necessary to normalize the votes by channel numbers. If the normalized votes exceed the threshold δ, the homogenous segments will be decided as speech. Fig. 6 is the voting result of a speech sentence mixed with babble noise at SNR=0dB. There are 30 channels with SNR larger than 2dB, taking part in the voting. In the middle panel, the red part is the detected speech segments.

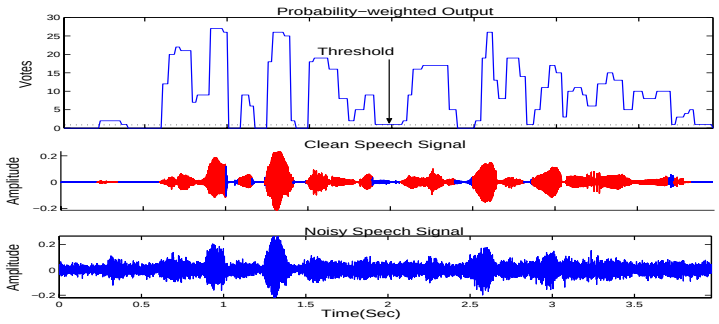

Fig.6. Detection results of 0dB babble noise

Considering the tradeoff between noise and speech hit rate, in real application, we adapt the voting threshold δ to the average SNR level as:

$$\delta = \begin{cases} \delta_l & SNR < SNR_l \\ round\left(\frac{\delta_h - \delta_l}{SNR_h - SNR_l}(SNR - SNR_l) + \delta_l\right) & SNR_l < SNR < SNR_h \\ \delta_h & SNR > SNR_h \end{cases} \quad (7)$$

where SNR_h and SNR_l are the highest and lowest SNR levels respectively in real applications; δ_h and δ_l are the voting thresholds corresponding to the highest and lowest SNR levels. For SNR between lowest and highest levels, the voting threshold is linearly interpolated between δ_l and δ_h; *round* is the nearest integer function.

4. Experimental Evaluation

To evaluate the effectiveness of our VAD algorithm, we measured the detection probability (including speech hit rate HR1 and noise hit rate HR0) for a number of noisy speech paragraphs. The experiment data were taken from the TIMIT database. We connected every ten sentences from one speaker into a speech paragraph and

mixed it with noise taken from NOISEX'92 database at variant SNR situations. Our experiment data consisted of 168 paragraphs with duration of about half a minute. The VAD references were labeled based on energy envelopes of clean speech signals.

In the detection, the paragraphs were chopped into 4-second segments. The noise eigenspace was estimated as described in section 2.2. For every 4 seconds, the noise eigenspace was updated by the detected noise. The adaptive voting threshold was calculated using formula (7), where the parameters were set as $\delta_l = 1$ for $SNR_l = -5dB$ and $\delta_h = 6$ for $SNR_l = 20dB$.

Table 1 shows the experiment results of our proposed algorithm with the traditional VAD algorithms. The values in the table are the noise hit rate (HR0) and speech hit rate (HR1) averaged over noisy speech different SNR from -5dB to 20dB. In this table, one can see that, in noisy environments, our algorithm works much better than G.729B [1] and AFE [11] algorithms.

Table 1. Experimental results

		G.729B	AFE (Wiener Filtering)	Proposed VAD
Factory	HR1	77.91%	89.91%	94.77%
	HR0	84.43%	40.76%	58.48%
Babble	HR1	74.79%	86.43%	91.18%
	HR0	74.99%	45.30%	55.81%
Tank	HR1	77.21%	90.86%	94.62%
	HR0	85.25%	36.75%	64.74%

5. Conclusions

In this paper, we proposed a noise eigenspace based VAD algorithm. A local noise estimation method was implemented in the proposed method to increase the robustness of the detection. The experiments showed that our algorithm were much more robust than traditional VAD algorithms, such as G.729 and AFE VAD algorithms.

6. Acknowledgements

Partial of this research was done when the first author worked for MSRA as an intern. It is supported by the program for the "Fostering Talent in Emergent Research Fields" in Special Coordination Funds for Promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology.

7. References

1. A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, 1996. ITU, ITU-T Rec. G.729-Annex B.
2. R. Tucker: Voice activity detection using a periodicity measure. Proc.Inst. Elect. Eng., 139(4):377-380, 1992.

3. Y. Ephraim and D. Malah: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoustic., Speech, Signal processing, ASSP-32:1109-1121, 1984.
4. J. Sohn and W. Sung; A voice activity detector employing soft decision based noise spectrum adaptation. Proc. ICASSP, pp. 365-368, 1998.
5. S. Gazor and W. Zhang: A soft voice activity detector based on a Laplacian-Gaussian model. IEEE Trans. Speech Audio Process. 11(5): 498-505, 2003.
6. E. Nemer, R. Goubran, and S. Mahmoud; Robust voice activity detection using higher-order statistics in the lpc residual domain. IEEE Trans. Speech Audio Process. 9(3):217-231, 2001.
7. Q. Li, J. Zheng, A. Tsai, and Q. Zhou: Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Trans. Speech Audio Process., 10(3):146-157, 2002.
8. J.Ramirez, J.C. Segura and et al.: An effective subband osf-based VAD with noise reduction for robust speech recognition. IEEE Trans. Speech Audio Process., 11(5):498-505, 2003.
9. Yu Shi, Frank K. Soong, and Jian-Lai Zhou: Auto-segmentation based partitioning and clustering approach to robust end pointing. Proc. ICASSP2006 .
10. Ris C, Dupont S.: Assessing local noise level estimation methods: application to noise robust ASR. Speech Communication, 34:141-158, 2001.
11. ETSI ES 2011 08 recommendation, 2000. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms.