A VOICE ACTIVITY DETECTOR USING THE CHI-SQUARE TEST

Beena Ahmed[1] and W. Harvey Holmes[2]

[1]RMIT University, Melbourne, VIC 3001, Australia
[2]University of New South Wales, Sydney, NSW 2052, Australia
beena.ahmed@rmit.edu.au, h.holmes@unsw.edu.au

ABSTRACT

This paper proposes a voice activity detector (VAD) that makes the speech/noise classification by applying the statistical chi-square test to each frame. It also uses a continuous update of the background noise estimate. The speech is first enhanced using a noise reduction system, with noise estimates also obtained with the help of the chi-square test. The noise-reduced signal is decomposed into sub-bands, and the chi-square test is used again in another form to compare the observed signal distribution to the estimated noise distribution. If the chi-square test determines that they are close, the frame is declared to be noise, otherwise speech. The performance of this VAD was found to be significantly superior to several benchmark VADs, with accuracies above 89% even at a SNR of 0 dB, which is up to 25% better than the others.

1. INTRODUCTION

Voice activity detectors (VADs) classify frames of a speech signal into speech (actually speech plus noise) or noise only. We assume an additive noise model in which the speech signal s(t) is corrupted by uncorrelated additive noise $w(t)$, giving the degraded composite signal $y(t)$:

$$y(t) = s(t) + w(t) . \quad (1.1)$$

There are two ways of interpreting this model:

1. The 'additive noise' point of view; i.e. the speech signal is corrupted by additive noise
2. The 'additive signal' point of view; i.e. the residual noise signal has speech added to it.

Nearly always the problems of voice activity detection, noise estimation, and noise reduction have been approached from the first point of view [1, 2].

This paper presents a unique solution to the problem based on the second point of view, which has not had application in VADs before. Since the decision made by the VAD is critically dependent on the current noise estimate, the background noise estimate is continuously updated (with the help of the chi-square statistical test). The noisy speech is then enhanced using this noise estimate. Finally, a speech/noise detector, also based on the chi-square test, is used to make the VAD decision.

2. THE CHI-SQUARE TEST

The chi-square test seeks to determine if there is a good fit between the frequencies of the observed data and the frequencies of the expected or theoretical data [3]. It uses the chi-square statistic to compare the two frequency distributions and test the hypothesis that the observations come from the same probability distribution. The expected frequencies of each class are determined on the basis of a presumed model.

In this paper, the chi-square test has been applied to the two problems of noise estimation and voice activity detection. In both of these cases, noise-only segments of corrupted speech samples need to be identified – to obtain accurate updating of the noise estimates in the first case, or for accurate identification of speech frames in the second case. The chi-square test provides a tool to compare the estimated noise probability density function (pdf) with the current signal pdf and decide whether they are the same.

3. NOISE ESTIMATION USING THE CHI-SQUARE TEST

As already remarked, noise estimation is critical for both noise reduction and voice activity detection. To estimate the noise with improved spectral sensitivity, the input signal is first passed through a bandpass filterbank H_1, ..., H_M, as shown in Figure 1. Each of these sub-banded signals is then divided into time frames of size L, where $M<<L$. The chi-square test is first applied to the outputs of this filter bank, as briefly described below.

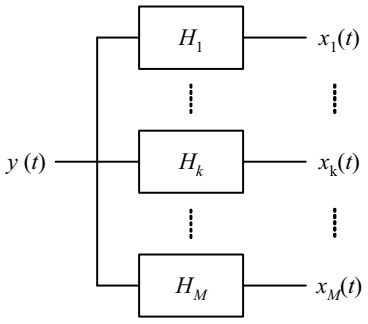

Figure 1. The division of a signal into M sub-bands

Let $\mathbf{x}_{k,p}$ be the signal in sub-band k and frame p. The samples in the noise vector $\mathbf{w}_{k,p}$ will be given by

$$\mathbf{w}_{k,p} = \left[w_{k,p}(1), ..., w_{k,p}(L) \right] \tag{3.1}$$

To identify the noise-only frames, for either updating the noise estimate or to make a decision in the VAD, the following hypothesis is proposed for each frame,

$$\begin{aligned} H_0 &: \quad \mathbf{x}_{k,p} = \hat{\mathbf{w}}_{k,p-1} && \text{: noise-only frame} \\ H_1 &: \quad \mathbf{x}_{k,p} = \hat{\mathbf{w}}_{k,p-1} + \mathbf{s}_{k,p} && \text{: noise-plus-speech frame} \end{aligned} \tag{3.2}$$

where $\mathbf{w}_{k,p-1}$ is the noise estimate from the previous frame p-1 and $\mathbf{s}_{k,p}$ is the speech signal present in frame p.

To test the above hypothesis, the chi-square test is applied to the samples in the current frame p for each sub-band k. The noise estimates $\mathbf{w}_{k,p-1}$ from the previous frame, p-1, are first grouped into N bins, whose boundaries are chosen such that the numbers in each bin, e_i, are approximately equal. These numbers define a noise histogram which approximates the noise pdf of the previous frame, p-1, and the resulting vector of expectations $\mathbf{e}$ is

$$\mathbf{e} = [e_1, \ldots, e_i, \ldots, e_N] \quad \text{for } N \text{ bins.} \tag{3.3}$$

The vector $\mathbf{o}$ of observation is obtained in the same way from the current signal $\mathbf{x}_{k,p}$, using the same N bins. If the number of observed values in bin i is o_i, we have simply

$$\mathbf{o} = [o_1, \ldots o_i, \ldots o_N] \ . \tag{3.4}$$

The chi-square test is then applied to these bins, where the chi-square statistic [3] is given by

$$\chi^2 = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i} \tag{3.5}$$

The calculated value of this chi-square statistic is compared to the appropriate threshold value of the chi-squared distribution, which depends on the allowed error probability and which is obtained from standard chi-square tables. If the obtained value is more than the tabulated value the hypothesis is rejected, otherwise it is accepted.

The hypothesis test can thus be written as

$$\chi^2 \gtrless \text{threshold} \quad \Rightarrow \quad \begin{cases} H_1 \\ H_0 \end{cases} \tag{3.6}$$

Background noise is usually assumed in noise estimation techniques to have a Gaussian distribution. Speech is (relatively) non-stationary and is non-Gaussian. Given a noisy speech signal, the chi-square test can be used to effectively identify the noise-only segments of the signal. If the chi-square test is applied on relatively long time frames of the signal, the assumption can be made that the noise pdf is Gaussian.

The chi-square test is applied to the noisy speech signal after decomposition into sub-bands. The test will determine if the observed signal distribution follows the previously estimated Gaussian noise distribution. If H_0 is accepted, i.e. if the current frame is determined to be a noise-only frame similar to the current noise estimate, the pdf of the noise is updated using a simple one-pole smoothing filter with smoothing coefficient λ. Otherwise the existing pdf is retained. This process is repeated for each sub-band and their noise estimates are updated.

4. VOICE ACTIVITY DETECTION USING NOISE REDUCTION AND A CHI-SQUARE TEST

The proposed VAD consists of three main components:

1. Chi-square based noise estimator (as in Section 3),
2. Noise reduction system, and
3. Chi-square based decision module.

A block diagram of the complete system is shown in Figure 3. This system is based upon two new principles in the context of VADs: noise reduction and chi-square detection. Testing showed that the accuracy of a VAD increased dramatically if speech frames are detected in a noise-reduced signal. The chi-square test was also found to accurately and robustly distinguish speech-plus-noise frames from noise-only frames.

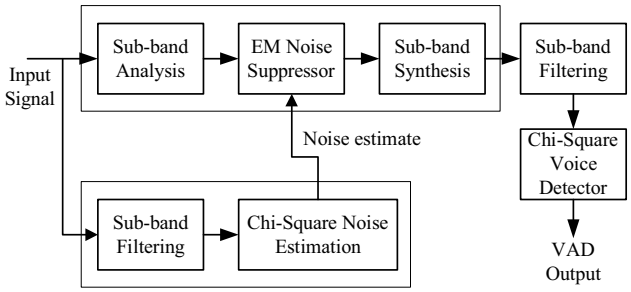

Figure 2. Block diagram of the proposed VAD

In this VAD, each frame of the input speech is first spectrally decomposed and then the noise is suppressed. Testing of noise reduction systems based on spectral subtraction showed that the Ephraim and Malah (EM) noise suppression rule [4] was the most effective for this purpose. Noise estimates, which are needed by the noise reduction system, are provided by the chi-square noise estimator described in Section 3.

After synthesis, the noise-reduced signal is again filtered into sub-bands so the chi-square test can be reapplied to obtain an accurate speech/noise decision. The

final synthesis-analysis steps could be omitted if the sub-bands were the same, since they would then cancel out.

5. IMPLEMENTATION DETAILS

In the noise reduction system, the noisy speech was bandlimited between 0.2-4 kHz, and then enhanced by passing through a DFT overlap-add filterbank with optimally chosen gains in each band. An analysis frame size of 256 samples was used, with a step size of 64 samples. The optimum band gains were estimated from the spectrally decomposed noisy signal using the Ephraim and Malah gain function with a smoothing parameter of 0.98. Noise estimates were obtained from the chi-square noise estimator described in Section 3.

The VAD decision was made using another chi-square detector on the synthesised noise-reduced signal. In it a frame size of 15 ms, i.e. 125 samples, was used with an overlap of 25 samples at a sampling frequency of 8192 Hz. The signal was divided into 8 equal sub-bands using a digital IIR filterbank of elliptic bandpass filters of order 10, with bandwidths of 487.5 Hz. The histogram of the current frame was calculated and divided into 7 classes (or bins). The first three frames were recursively averaged to provide the starting value of the noise vector. The chi-square test was applied to each sub-band signal in each current frame, comparing it to the previously estimated noise distribution. Those frames in which the null hypothesis was accepted for all 8 sub-bands were declared to be noise, whereas all others were declared to be speech-plus-noise frames.

The assumption in the noise estimator that the noise pdf is Gaussian is valid only if long frame sizes are used. Hence to estimate the noise (only), the current input frame was combined with seven previous frames, giving frame sizes of 122 ms for noise estimation. These frames were then filtered into eight sub-bands using the IIR filterbank above. The observed and expected distributions in each sub-band were divided into 7 classes and tested using the chi-square test. Only those frames in which the null hypothesis was accepted for all 8 sub-bands were declared to be noise. The noise estimate was then updated using a single-sided one-pole recursive filter. Testing found that a smoothing coefficient of 0.95 gave the best results.

6. TEST RESULTS

The VAD was applied to 12 sentences with added babble, car, pink, and white noises at SNRs of 0, 5, and 10 dB. Figure 3 shows the results obtained from a sample test sentence at 10 dB SNR. The chi-square detector manages to pick up short bursts of speech that are only a few hundred samples (20 ms) long.

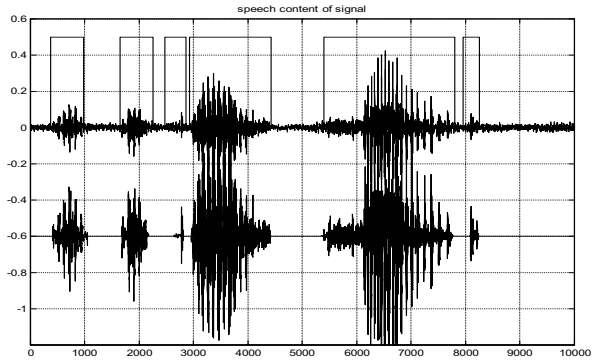

Figure 3. The performance of the chi-square VAD.
a. The output of the chi-square detector and the noisy speech signal with added babble noise at SNR = 10 dB;
b. The clean speech signal.

Figure 4 shows a noisy speech sample at an SNR of 5 dB with the noise portion at the start of the sample identified using the chi-square noise estimator by itself (without the final chi-square voice detector).

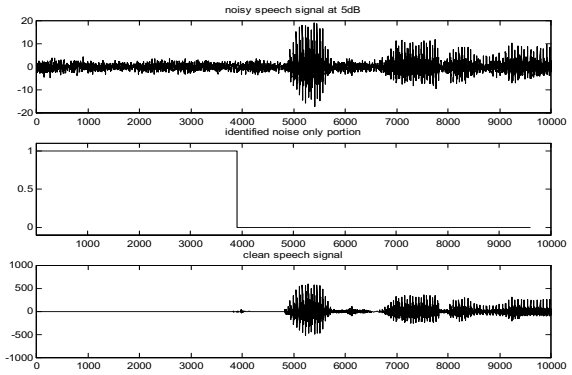

Figure 4. The performance of the chi-square noise estimator. a. The noisy speech signal (SNR = 5 dB); b. The frames declared to be noise-only; c. The clean speech signal.

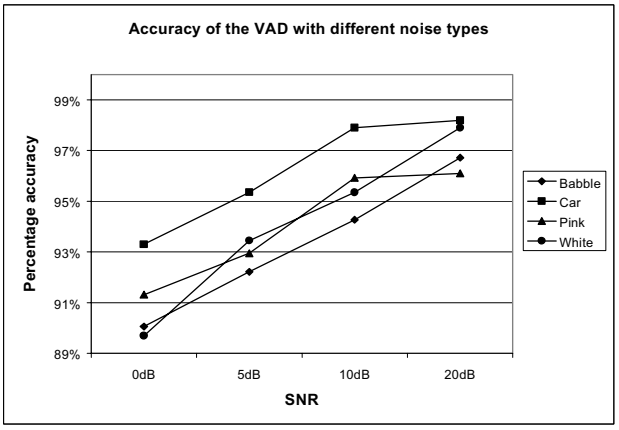

Figure 5. The accuracy of the chi-square based VAD applied to speech samples with various added noise types.

7. COMPARISON WITH OTHER VADS

The proposed VAD performance was compared to that of these existing benchmark VADs:

1. VAD in the GSM standard [1]
2. VAD in the ITU standard G.729 Annex B [2]
3. VAD proposed by Sohn et al. [5]

The VADs were run on 12 test sentences with added babble, car, pink, and white noise at SNRs of 10, 5, and 0 dB. The VADs were compared by their percentage accuracy compared to the true speech/noise classification. The results are given in Table 1.

Table 1. A comparison of the percentage accuracies of the benchmark VADs and the proposed VAD.

		Percentage Accuracy			
Added Noise	SNR dB	GSM VAD	G.729 B VAD	Sohn VAD	Proposed VAD
Babble	10	63.95	81.81	87.32	94.27
	5	58.40	73.47	84.93	91.92
	0	56.45	62.09	80.97	90.06
Car	10	79.73	77.22	96.22	97.90
	5	75.60	73.89	96.23	95.36
	0	67.32	67.34	91.21	93.30
Pink	10	72.89	81.01	84.89	95.92
	5	71.90	72.65	83.13	92.95
	0	57.50	63.90	79.08	91.32
White	10	77.14	91.85	96.83	95.35
	5	74.23	81.57	91.62	93.45
	0	66.12	65.25	85.67	89.70

Table 1 shows that the proposed VAD consistently outperforms the others, with an accuracy above 89% in all scenarios, and 97.9% with added car noise at 10 dB SNR. The GSM VAD performs best with added car noise, having an accuracy of 79.93% at 10 dB SNR. Its accuracy rapidly falls as the SNR decreases, down to 56.45% with babble noise at 0 dB. Similarly the G.729 B VAD, though performing better than the GSM, still declines rapidly with decreasing SNR. Its best accuracy is 91.85% with added white noise at 10 dB, down to 65.25% at 0 dB. The Sohn statistical model-based VAD does not perform as well as the proposed VAD, but is more reliable than the other two benchmark VADs, ranging from 96.22% with added pink noise at 10 dB to 91.21% with added pink noise at 0 dB.

Figure 6 shows that the proposed VAD clearly outperforms the other benchmark VADs, with its accuracy consistently above the others at all SNRs.

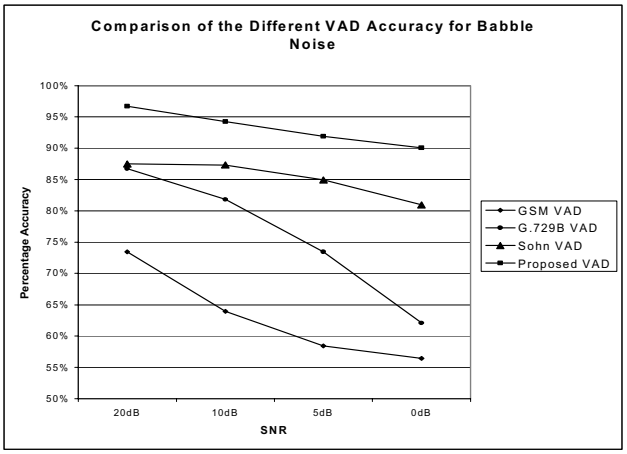

Figure 6. The percentage accuracy of the proposed VAD compared to the accuracies of the benchmark VADs used on samples with added babble noise at different SNRs.

8. CONCLUSIONS

In this paper a novel VAD based on the statistical chi-square test has been proposed. The VAD departs from the traditional heuristic nature of VADs by making a speech/noise decision based on deviations from the noise distribution. On comparison with other benchmark VADs, the proposed VAD was found to provide the most accurate speech/noise classification for a range of SNRs and noise types.

9. REFERENCES

[1] D.K. Freeman, G. Cosier, C.B. Southcott and I. Boyd, "The voice activity detector for pan-European digital cellular mobile telephone service", *Proc. ICASSP*, 1989, vol. 1, pp. 369-372.

[2] A. Benyassine, E. Shlomot and H.Y. Su, "ITU-T Rec. G.729B Annex B: A silence compression scheme for use with G.729B optimized for V.70 digital simultaneous voice and data applications", *IEEE Communications Magazine*, Sept 1997, pp. 64-73.

[3] H.J. Larson, *Introduction to Probability Theory and Statistical Inference*, Wiley, New York, 1982, 3rd ed.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. ASSP.*, vol. ASSP-32, pp. 1119-1120, Dec. 1984.

[5] J. Sohn, N.S. Kim and W. Sung, "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*, vol. 6, no. 1, Jan 1999, pp. 1-3.