Voice Activity Detection Using Source Separation Techniques

Nikos Doukas, Patrick Naylor and Tania Stathaki

Signal Processing Section, Dept of Electrical Engineering,
Imperial College, UK. e-mail: n.doukas@ic.ac.uk

Abstract

A novel Voice Activity Detector is presented that is based on Source Separation techniques applied to single sensor signals. It offers very accurate estimation of the endpoints in very low Signal to Noise ratio conditions, while maintaining low complexity. Since the procedure is totally iterative, it is suitable for use in real-time applications and is capable of operating in dynamically adapting situations. Results are presented for both White Gaussian and Car Engine background noise. The performance of the new technique is compared with that of the GSM Voice Activity Detector.

1. Introduction

Voice Activity Detection (VAD) is important in many areas of speech processing technology, such as noise reduction, voice recognition, speech coding etc, and has been extensively studied ([7], [5], [1]). Most of the existing techniques focus on relatively mild noise conditions (small positive SNR, for example the conditions found in an office environment). The work presented in this paper focuses on much more adverse conditions, with SNR's in the range of -5 to -10 dB, with particular attention paid to the case of car noise. The VAD is based on a recently presented noise reduction technique [2], [3]. The optimisation procedure is totally iterative and therefore suitable for use in real-time, dynamically adapting situations. This paper reviews the Single Sensor Source Separation (SSSS) noise reduction technique, applies it to VAD and presents the results. The performance of the new technique is compared with that of the Voice Activity Detector for the GSM standard [9].

2. Single Sensor Source Separation (SSSS)

The problem of Independent Source Separation has recently attracted considerable attention (see e.g [6]). A Signal Enhancement technique based on separating the signals from a single sensor has recently been presented in [2] and [3]. The signal received from just one sensor is passed through filters that model the distortions normally undergone in the channel. This is followed by an optimisation procedure which is implemented via a Lagrange Programming Neural Network (LPNN, see [8]). The setup is shown in Figure 1.

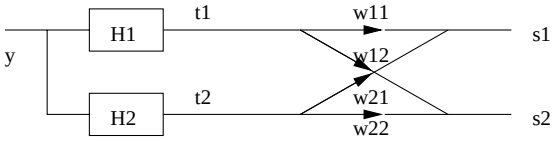


Figure 1: Single Sensor Source Separation Block Diagram

The input filters were chosen to be:

$$H_1 = \delta + z^{-1} + \delta z^{-2}$$

and

$$H_2 = -\delta + z^{-1} - \delta z^{-2}$$

with δ typically 0.5.

The LPNN allows constraints to be imposed which substitute, in part, information normally obtained from a second sensor. With reference to the signals shown in Figure 1, the objective function to be minimised is: $J = \sum_{i,j} \left(E[s_1^{2i+1} s_2^{2j+1}] \right)^2$ subject to the constraint that $s_1 + s_2 = y$ where y is the received signal. This gives the following Lagrange function to be minimised:

$$J = \sum_{i,j} \left(E[s_1^{2i+1} s_2^{2j+1}] \right)^2 + \lambda(s_1 + s_2 - y) \qquad (1)$$

A full account of this technique can be found in [2] and [3].

The error signal of the LPNN, calculated as $\epsilon = \sum_{i,j} \left(E[s_1^{2i+1} s_2^{2j+1}] \right)^2$, is of particular interest to this work. When the input signals remain stationary, ϵ will remain at a constant value. When the input signal statistics change, ϵ will suddenly increase and then drop back down to its new equilibrium point. Unbiased estimates $\hat{\epsilon}$ of ϵ are produced using: $\hat{\epsilon}_n = (1-\alpha)\hat{\epsilon}_{n-1} + \alpha \sum_{i,j} \left(s_1^{2i+1} s_2^{2j+1} \right)^2$. The parameter α controls the speed at which the algorithm will reconverge after the statistics of the inputs have changed.

3. Application to Voice Activity Detection

For the purposes of the VAD, the parameter α used in the calculation of $\hat{\epsilon}$ is chosen to be relatively small (typically 0.9) so that the response to changes is relatively swift. The setup used is shown in Figure 2. The error signal is first smoothed. The adaptive threshold is then calculated and continuously updated to account for changes in the background noise statistics. During silence the threshold T_n is updated as

$$T_{n+1} = (1-\beta)T_n + \beta f_n,$$

where f_n is the smoothed error signal. During speech

$$T_{n+1} = (1-\gamma)T_n + \gamma f_n,$$

with $\gamma << \beta$ (typically $\frac{\beta}{\gamma} = 100$) so that the algorithm does not start to track speech. The decision changes from silence to speech when the signal f exceeds $\lambda_0 \times T_n$ and conversely, from speech to silence when f falls below $\lambda_1 \times T_n$ (typically $\lambda_0 = 1$ and $\lambda_1 = 1.4$). As an additional heuristic, the threshold is not allowed to go below $2 \cdot 10^{-6}$. The complexity of the overall VAD algorithm is of the order of 50 operations per sample.

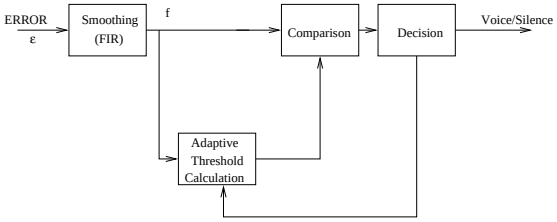


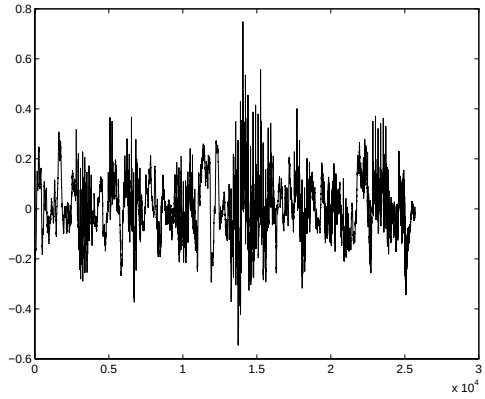
Figure 2: SSSS based Voice Activity Detector Block Diagram

Stability considerations for the overall sytem can be found in [4]. The method is found to be stable, provided the power of the input signal is below a certain level determined by the choice of the various parameters of the system, such as the input filter coefficients, γ, β etc. For the values shown above the bound obtained is 0.2. In practice this is found to be tight, and the input power can be higher.

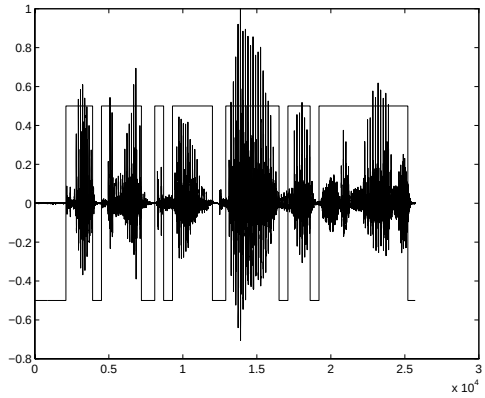
4. Results

Tests were carried out using the phonetically labeled TIMIT ([10]) database and both white and car noise. White noise is the worst case because the SSSS algorithm performs worst in wideband noise. Results are given in the form of example segmented phrases and as statistics of the error in the determination of the endpoints, calculated from 1600 tests. All tests were carried out at approximately $-8dB$ SNR, and the results of the tests are given in Figures 3, 4 and 7.

The speech waveform shown in the examples is from a male speaker and the sentence "Coconut cream pie makes a nice desert".



(a)



(b)

Figure 3: (a)Speech Contaminated with Car Noise at -8dB and (b) original, clean speech and VAD decision.

Experience showed that the GSM VAD [9] completely failed to identify any silence periods for experiments where the SNR was below 10dB (the TIMIT database was again used). For this reason, experiments were performed for both techniques at 20 dB SNR. Results are once more given both as example segmented phrases of male speech and as statistics of the error. The speech waveform shown, is the same as before. Figure 6 shows the error distribution obtained for the new method at 20dB over 1600 speech files of the timit database. Figure 8 shows the error distribution obtained for the GSM VAD, with the same data and for the same noise conditions.

At 20 dB SNR,the GSM VAD classifies the whole sentence as speech whereas the SSSS VAD is able to segment the sentence into segments of speech and pauses. This shows the SSSS VAD to be a promising method for use in, for example, methods of noise re-

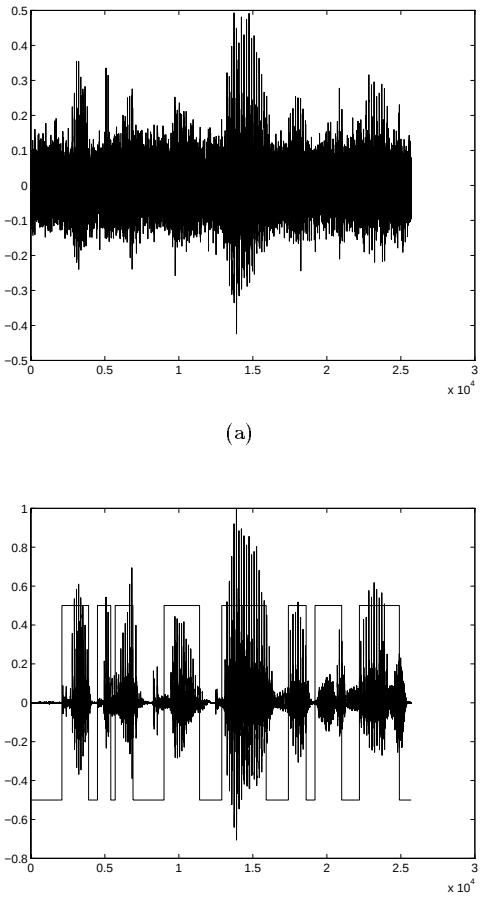


Figure 4: (a)Speech Contaminated with White Gaussian Noise at -8dB and (b) original, clean speech and VAD decision.

duction based on spectral subtraction which require noise model estimates to be updated in speech pauses. Some speech activity is falsely classified as silence in this example but tuning of the parameters of the method can be used to avoid this. At -8 dB SNR, the SSSS method gives results practically identical to those obtained at 20 dB SNR for car noise. For white gaussian noise (which is the worst case of the SSSS method) at -8 dB SNR, the results are not as good as those obtained at 20 dB, but a meaningful and useful segmentation is still obtained.

5. Conclusions

A new VAD is presented based on the Single Sensor Source Separation signal enhancement technique. The performance of the new method was shown to be superior to the one obtained from the GSM VAD and much more robust in the presence of extremely high noise levels. The algorithm is of very low computational complexity, and does not contain any division operations and is therefore suitable in this respect for practical realisations.

We would like to thank our collegue, Mr D.M. Brookes for supplying the implementation of the GSM VAD used in these tests.

6. REFERENCES

[1] Evangelos S. Dermatas, Nikos D. Fakotakis, and George K. Kokkinakis. Fast endpoint detection algorithm for isolated word recognition in office environment. *International Conference on Acoustics, Speech and Signal Processing*, 1991.

[2] N. Doukas, P. Naylor, and T. Stathaki. A single sensor source separation approach to noise reduction. In *CESA 96 IMACS Multiconference*, 1996.

[3] N. Doukas, T. Stathaki, and P. Naylor. Speech enhancement through nonlinear adaptive source separation methods. In *International Conference on Statistical Signal and Array Processing*, 1996.

[4] N. Doukas, T. Stathaki, and P. Naylor. Stability of a voice activity detector based on source separation. In *13th Intrnation Conference on Digital Signal Processing*, 1997.

[5] A. Ganapathiraju, L. Webster, J. Trimble, and P. Bush, K. andd Kornman. Comparison of energy-based endpoint detectors for speech signal processing. In *IEEE Southeastcon*, pages 500–3, 1996.

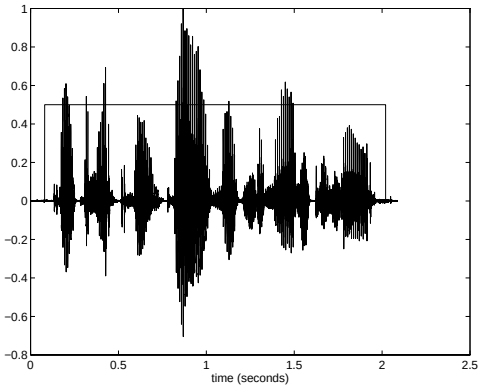
[6] P.Common, C. Jutten, and J. Herault. Blind separation of sources, part 2: Problems statement. *Signal Processing*, 24(1):11–20, July 1991.

[7] N.B. Yoma, F. McInnes, and M. Jack. Robust speech pulse detection using adaptive noise modelling. *Electronics Letters*, 32(15):1350–2, July 1996.

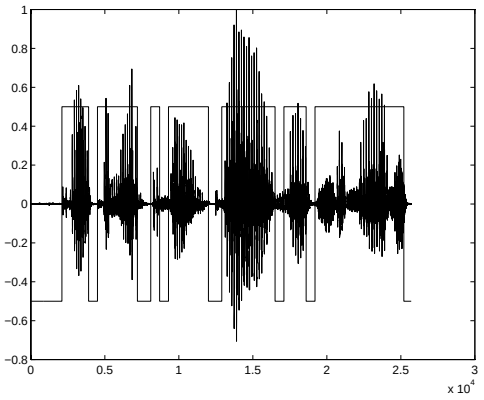
[8] S. Zhang and A. G. Constantinides. Lagrange programming neural networks. *IEEE Transactions on Circuits and Systems*, 39(7):441–52, July 1992.

[9] Global System for Communications. Digital Cellular Telecommunications System; Voice Activity Detector for Enhanced Full Rate Speech traffic channels (GSM 06.82), November 1996. European Telecommunications Standards Institute.

[10] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, October 1996. Prepared at the National Institute of Standards and Technology (NIST).



(a)



(b)

Figure 5: VAD decision made at 20 dB (Gaussian Noise)plotted over the original speech signal (a) GSM VAD and (b) SSSS VAD.

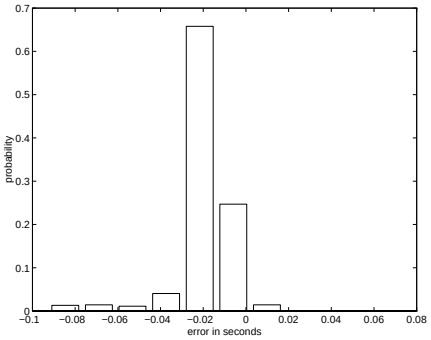


Figure 6: SSSS VAD: Probability distribution of the error (in seconds), computed over 1600 sentences from the timit database (SNR = 20dB, White Gaussian Noise).

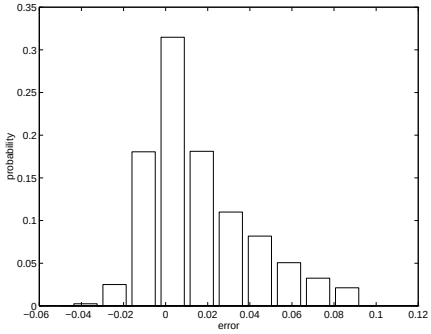


Figure 7: SSSS VAD: Probability distribution of the error (in seconds), computed over 1600 sentences from the timit database (SNR = -8dB, White Gaussian Noise).

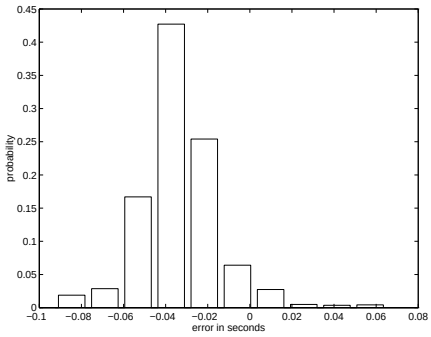


Figure 8: GSM VAD: Probability distribution of the error (in seconds), computed over 1600 sentences from the timit database (SNR = 20dB, White Gaussian Noise).