MULTI-DECISION SUB-BAND VOICE ACTIVITY DETECTION FOR SPEECH ENHANCEMENT

Alan Davis, Siow Yong Low and Sven E. Nordholm

Western Australian Telecommunications Research Institute (WATRI),
Crawley, WA 6009, Australia

ABSTRACT

This paper compares the performance of the recently developed multi-decision sub-band voice activity detector (MDSVAD) [1], with the ITU G.729B voice activity detector (VAD) scheme [2] and the SNRVAD scheme [3] for a speech enhancement application. The study shows the importance of more detailed VAD decisions in the time-frequency plane to better maintain speech features. This is in keeping with the observation that typically a speech signal will not simultaneously excite all frequency components at any one time instant. Here, the MDSVAD exploits the spectral structure of speech versus background noise to make independent voice activity decisions in separate subbands, resulting in multiple decisions for any frame. Results show that the decisions in separate sub-bands reduce the infamous musical tones significantly in the conventional spectral subtraction algorithm compared to the other two VADs.

1. INTRODUCTION

Single channel speech enhancement techniques are limited by accurate knowledge of what can be considered as speech and noise. As such the enhancement problem is usually posed as a time-frequency problem where the corrupted speech signal is analysed in time frequency blocks. Typically, the noise is modeled as additive to the speech and the output is formed using a time varying gain function. The gain optimisation criterion is usually the mean square error estimator (MMSE) or log spectral mean square error estimator MMSE-LSA [4]. The main focus in that work has been focused on finding the optimal weight according to various criteria. However, an important observation made by Abramson *et al*. reiterates the importance of accurate noise tracking and speech detection particularly under non-stationary noise conditions [5]. This follows closely with an observation made in [1]. In that paper it was found that it is important to track the speech features as speech has different duration for different frequency bands and it is also important to improve the noise estimate update that forms the gain function. This paper revisits this point by evaluating the performance of a speech enhancement scheme when using two traditional voice activity detector (VAD) schemes [2, 3] versus the multi-decision sub-band voice activity detector (MDSVAD) [1]. One of the most interesting aspects of the paper is that the speech enhancement performance is shown to vary quite dramatically by simply changing the VAD scheme and making no changes to the speech enhancement technique itself.

WATRI is a joint venture between The University of Western Australia and Curtin University of Technology. The authors also acknowledge assistance from Sensear Pty. Ltd.

2. EVALUATION APPROACH

The goal of this paper is to show how different VAD schemes can impact on speech enhancement performance and how the MDSVAD can improve speech enhancement performance. To that end the following evaluation approach was taken. Initially, a simplistic speech enhancement scheme was designed. The scheme was then evaluated using a noisy speech utterance and each of the G.729B VAD scheme [2], the SNRVAD scheme [3] and the MDSVAD scheme [1]. The three resulting enhanced speech signals were then compared qualitatively to investigate the performance of the speech enhancement scheme for each of the three VAD cases. In this way the impact that each of the three VAD schemes has on the speech enhancement scheme can be investigated.

3. VAD STRUCTURES

Here it is prudent to mention some fundamental differences in the VAD structures being evaluated. Traditional VAD schemes operate by partitioning a set of sampled data into small periods (frames), typically in the order of 20ms. These frames are then analyzed to determine presence of speech, and are classified 'speech-active' or 'speech-inactive'. Traditionally, there is one common decision for all frequency components for each frame, both the SNRVAD and the G.729B VAD operate in this manner. Such a VAD fails to exploit the spectral time varying nature of speech. For example, a spoken phoneme will often not encompass all frequencies simultaneously. Upon examining the spectral content of phonemes it becomes obvious that often speech is not present in all

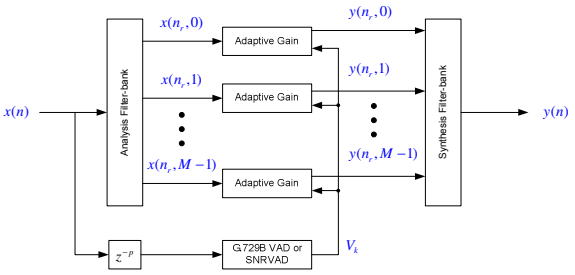

Fig. 1. Speech enhancement scheme using the G.729B or SNRVAD options

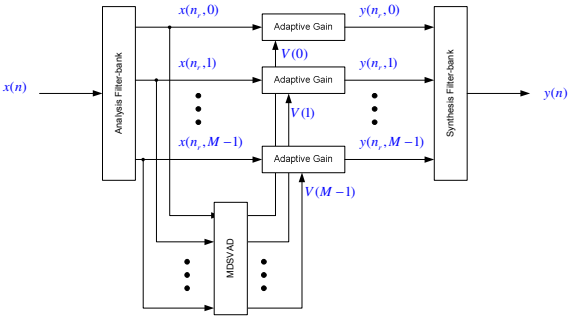

Fig. 2. Speech enhancement scheme using the MDSVAD option

frequency bands at a given time, i.e. a given frame may be 'speech-active' however not all frequency bands are 'speech-active' at the same time. The MDSVAD [1] independently determines speech activity for each sub-band. Therefore for each frame there will there M decisions per frame, where M is the number of sub-bands. This is the fundamental difference of the MDSVAD when compared to the traditional G.729B VAD and the SNRVAD schemes.

4. SPEECH ENHANCEMENT SCHEME

In order to determine the impact VAD performance can have on speech enhancement performance, a basic speech enhancement scheme was developed then implemented with each of the three VAD options previously discussed. The scheme is essentially a spectral subtraction scheme like that of Boll [6], with the speech active decisions controlled by one of the ITU G.729B VAD scheme, the SNRVAD scheme or the MDSVAD scheme.

Figure 1 shows the block diagram of the implemented speech enhancement system when using the G.729B VAD scheme or the SNRVAD scheme, and likewise Figure 2 shows the implemented system using the MDSVAD scheme. Figure 3 presents the block diagram of the 'adaptive gain' block illustrated in both Figures 1 and 2, and is common to both of the structures. Initially, the gain for a specific sub-band signal

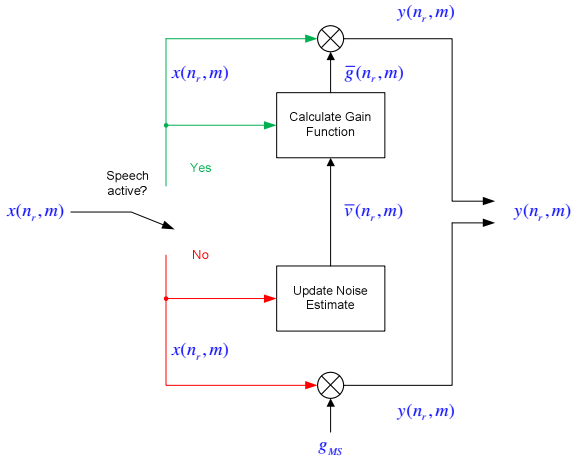

Fig. 3. Gain estimation system

$x(n_r, m)$ is found as,

$$g(n_r,m) = \begin{cases} \max\{1-\beta\dfrac{\bar{v}(n_r,m)}{|x(n_r,m)|}, g_{MS}\}, & \text{SA,} \\ g_{MS}, & \text{SIA,} \end{cases} \quad (4.1)$$

where β is the over-subtraction factor, $\bar{v}(n_r, m)$ is the estimated noise magnitude in the m^{th} sub-band and g_{MS} is the maximum allowable attenuation in a similar manner to [7], SA stands for speech active and SIA stands for speech inactive. The gain function is then averaged with an attack and release exponential average,

$$\bar{g}(n_r,m) = \begin{cases} \hat{g}(n_r,m), & \hat{g}(n_r,m) > \bar{g}(n_r-1,m), \\ \alpha_g \bar{g}(n_r-1,m) & \\ +(1-\alpha_g)\hat{g}(n_r,m), & \text{otherwise,} \end{cases} \quad (4.2)$$

where α_g is an exponential averaging constant. This means that the gain follows a rising signal amplitude, but slowly decreases in the case of reducing signal amplitude.

The noise magnitude estimate is found as follows

$$\bar{v}(n_r,m) = \begin{cases} \bar{v}(n_r-1,m), & \text{SA,} \\ \alpha_v \bar{v}(n_r-1,m) & \\ +(1-\alpha_v)|x(n_r,m)|, & \text{SIA,} \end{cases} \quad (4.3)$$

where α_v is an exponential averaging constant.

Finally, the enhanced sub-band signal is found as

$$y(n_r,m) = \bar{g}(n_r,m) \cdot x(n_r,m), \quad (4.4)$$

meaning the calculated gain is simply multiplied with the sub-band signal. The aggressiveness of the scheme can be simply controlled by the maximum allowable attenuation parameter g_{MS}. The parameter values are as follows, number of sub-bands $M = 256$, sub-band sample rate $= 125$Hz, analysis filter length $= 1024$, analysis filter cut-off $= 62.5$Hz, $\beta = 2.0$, $g_{MS} = -25$dB, $\alpha_g = 25$ms and $\alpha_v = 250$ms.

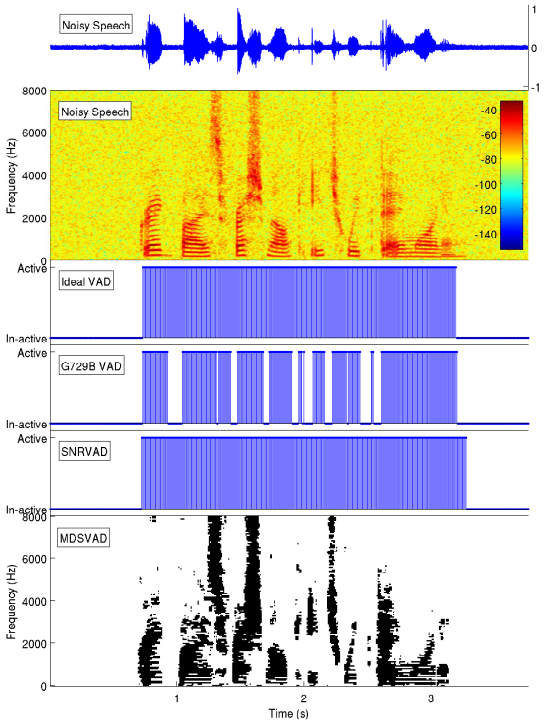

Fig. 4. Noisy speech sequence with VAD decisions in white Gaussian noise

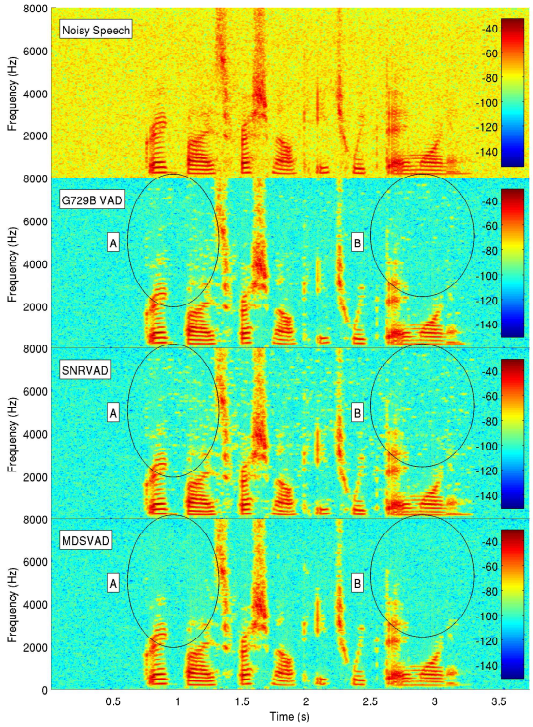

Fig. 5. Enhanced speech spectrograms in white Gaussian noise

5. INVESTIGATING THE EFFECT OF VAD PERFORMANCE ON SPEECH ENHANCEMENT PERFORMANCE

The evaluation was undertaken using white Gaussian noise and a female speaker. The white Gaussian noise was mixed with female speech taken from the TIMIT speech database [8] to a global signal-to-noise ratio (SNR) of 15dB. White Gaussian noise is an interesting scenario because it is both wide-band and stationary. The presented example is of the utterance "Greg buys fresh milk each weekday morning". The intent of this example is to show how the MDSVAD can decrease artifacts during speech periods due to its fundamentally different approach to voice activity detection.

Figure 4 shows the signal waveform and spectrogram of the noisy speech sequence and the VAD decisions for each of the three VAD options under test. Reference hand-labeled decisions are also shown so as to gain some insight into the performance of each of the three VAD options. Starting from the top of the figure, the signal waveform and spectrogram are first shown, following this the reference VAD decisions, the decisions made by the G.729B VAD scheme and then the SNRVAD decisions are presented. Finally the MDSVAD sub-band VAD decisions are shown. As can be seen, the G.729B VAD scheme makes some miss-detections, whereas the SNRVAD is quite accurate when compared to the reference decisions. Comparing the MDSVAD decisions to the spectrogram, it is clear the scheme captures the majority of the spectrum where speech energy is present.

Figure 5 shows the noisy speech spectrogram along with the enhanced speech spectrogram for each of the three VAD options tested. Starting from the top of the figure, the noisy speech spectrogram is first presented followed by the enhanced speech spectrogram where the G.729B VAD is used. Following this the enhanced speech spectrogram is shown where the SNRVAD is used and finally the enhanced speech spectrogram is shown where the MDSVAD is used. As can be seen, the noise is significantly attenuated with the speech spectrum being maintained. Regions A and B circled in the spectrograms highlight some artifacts that remain in the enhanced speech signal for the G.729B VAD and SNRVAD options. The MDSVAD however labels these regions as predominately non-speech as shown in Figure 4, this is further reinforced by the spectrogram. The traditional VAD schemes however do not have individual decisions for each sub-band and thus cannot distinguish this area as non-speech and therefore attempt to enhance the spectrum in this region by calculating a gain function and applying it. Given the variance of the background noise this results in artifacts in the spectrum which are annoying to the listener. An important note is that this reduction in artifacts is achieved without compromising the speech enhancement capabilities of the scheme as can be

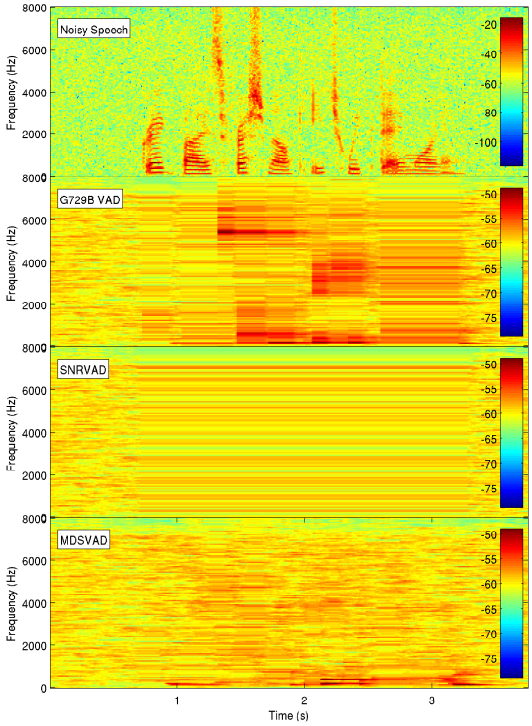

Fig. 6. Noise magnitude estimate in white Gaussian noise

seen in the spectrograms. In fact the MDSVAD spectrogram appears significantly cleaner than the other two examples.

Figure 6 shows the noise magnitude estimate made by the speech enhancement scheme over the sequence. Working down from the top of the figure, initially the noisy speech spectrogram is shown, followed by the noise magnitude estimate $\bar{v}(n_r, m)$ where the G.729B VAD scheme is employed. Following this the noise magnitude estimate is shown where the SNRVAD is employed and finally the noise magnitude estimate where the MDSVAD is used is shown. As can be seen the miss-detections created by the G.729B VAD scheme result in erroneous updates of the noise spectrum, whereas both the SNRVAD and MDSVAD schemes have more accurate noise magnitude estimates. Further, it is clear the MDSVAD is able to update the noise magnitude during what would traditionally be classified as speech periods and this results in more accurate noise estimates due to better tracking capabilities.

6. CONCLUSION

The evaluation of the three different VAD schemes have been presented. The study shows that with a standard spectral subtraction gain function it is possible to improve the enhancement results significantly by using more detailed VAD decisions in the time-frequency plane to better maintain speech features and obtain improved noise estimates. The result of improved noise estimates and better maintaining speech features result in significantly lower artifacts in both noise and speech. Also low energy portions of the speech is better preserved resulting in improved intelligibility.

7. REFERENCES

[1] Alan Davis, Sven Nordholm, Siow Yong Low, and Roberto Togneri, "A multi-decision sub-band voice activity detector," *Proc. EUSIPCO, Florence, Italy*, September 2006.

[2] ITU G.729 Annex B, *Coding of speech at 8kbit/s using conjugate structure algebraic code - excited linear prediction. Annex B: A silence compression scheme for G.729 optimised for terminals conforming to recommendation V.70*, International Telecommunication Union, 1996.

[3] Alan Davis, Sven Nordholm, and Roberto Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE. Trans. on Audio, Speech and Language Process.*, vol. 14, no. 2, pp. 412–424, March 2006.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[5] Ari Abramson and Israel Cohen, "Simultaneous detection and estimation approach for speech enhancement," *IEEE Trans. on Speech, Audio and Lang. Process.*, vol. 15, no. 8, pp. 2348–2359, November 2007.

[6] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 27, no. 2, pp. 113–120, April 1979.

[7] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 4, pp. 208–211, April 1979.

[8] John S. Garofolo, et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, 1993.