14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A MULTI-DECISION SUB-BAND VOICE ACTIVITY DETECTOR

Alan Davis [1], *Sven Nordholm* [1], *Siow Yong Low* [1] *and Roberto Togneri* [2]

[1]Curtin University of Technology
Kent St, Bentley, 6102, Australia
phone: +(61) 8 6488 4645, fax: +(61) 8 6488 7254,
email: {davisa; sven; siowyong}@watri.org.au
web: www.curtin.edu.au

[2]The School of Electrical, Electronic and Computer Engineering,
The University of Western Australia
35 Stirling Highway, Crawley, 6009, Australia
phone: +(61) 8 6488 2535, fax: +(61) 8 6488 1065,
email: roberto@ee.uwa.edu.au
web: www.ee.uwa.edu.au

ABSTRACT

In this paper, a new paradigm for voice activity detection (VAD) is introduced. The idea is to exploit the spectral nature of speech to make independent voice activity decisions in separate sub-bands, resulting in multiple decisions for any frame. A potential method to perform multi-decision sub-band VAD is proposed then evaluated with a small test set. The evaluations illustrate the concept and potential benefit of multi-decision sub-band VAD.

1. INTRODUCTION

Voice activity detection (VAD) is a topic of significant practical importance. The applications for VAD techniques are diverse and far reaching from power saving in mobile devices to estimation of noise and speech statistics in speech enhancement schemes [1]. Here we introduce the concept of multi-decision sub-band VAD and propose a multi-decision sub-band VAD scheme.

Traditionally, VAD schemes operate by partitioning a set of sampled data into small periods (frames), typically in the order of 20ms. Over this period speech can be considered short-term stationary, and often the frames overlap. The frames are then analyzed to determine the presence of speech, and each frame is classified 'speech-active' or 'speech-inactive'. Therefore for any sequence with N frames, there will be N speech activity decisions, one for each frame.

This full-band definition of VAD fails to exploit the spectral 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 frequency bands at a given time, i.e. a given frame may be 'speech-active' however not all frequency bands are 'speech-active'. Here we introduce the concept of a multi-decision sub-band voice activity detector.

A multi-decision sub-band voice activity detector makes an independent speech activity decision for each sub-band. Therefore for any sequence with N frames, there will be $N \cdot K$ decisions, where K is the number of sub-bands, i.e. K decisions per frame.

This extension to the traditional definition of VAD is especially useful in speech enhancement schemes. Usually speech enhancement schemes such as spectral subtraction [2] heavily depend on voice activity detectors. Enhancement schemes utilize VAD to estimate noise statistics by excluding periods of speech activity from noise statistics updates. Use of a multi-decision sub-band voice activity detector allows such schemes to track noise variations during speech periods, in non-active sub-bands. This results in better estimates of noise statistics and increased performance of speech enhancement algorithms. This action is similar to that of minimum statistics [3], however this is more explicit.

Several recently proposed VAD structures lend themselves easily to multi-decision sub-band VAD. For instance [4], [5] and [6] are all capable of making independent decisions in separate sub-bands, however currently sub-band decisions are collapsed into a traditional full-band decision. Here we propose a scheme that is based around the core mechanism in [6]. The proposed scheme utilises an over-sampled polyphase filter bank [7] to decompose the full-band signal into multiple sub-band signals. Each individual sub-band signal is then independently tested for speech activity, with an individual decision for each sub-band being the final result.

2. PROPOSED STRUCTURE

The proposed structure utilizes a polyphase over-sampled filter bank [7]. The design of the filter bank directly influences the characteristics of the VAD scheme. Here we use a K band filter bank that decimates by a factor of $\frac{K}{2}$ to perform the sub-band decomposition.

Figure 1 illustrates the proposed structure. Initially, the full-rate time domain data $x(n)$ is decomposed into sub-bands and decimated. A new sample index m is used to represent the reduced rate. The resulting reduced rate complex sub-band signal $x_k(m)$ in the k^{th} sub-band is then analyzed by a decision device (DD).

The decision device analyzes each sub-band independently to produce an initial decision of speech activity. The result is a boolean output $D_k(l)$ for each sub-band k. The decision device does this by framing the complex sub-band signal and analyzing the frequency content, hence the new index l.

Finally, the initial speech activity decision, $D_k(l)$ is analyzed by further logic to reduce false-alarms. The decision analysis modifies the initial decisions $D_k(l)$ to rule out unlikely scenarios, and outputs a final boolean decision in the k^{th} sub-band $V_k(l)$. The set of K sub-band speech activity decisions, $\{V_0(l), V_1(l), ..., V_{K-1}(l)\}$ form the set of final speech activity decisions for the l^{th} frame.

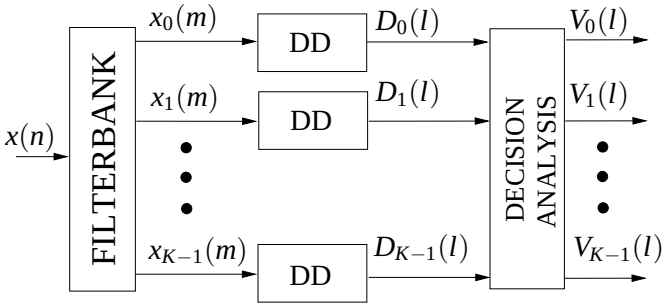

Figure 1: *Proposed Subband VAD structure*

3. FILTER BANK

In order to decompose the full-band signal $x(n)$ into K sub-bands, a polyphase filter bank is utilized [7]. Such a structure uses a single low-pass prototype filter and a modulator to generate a set of sub-band filters. The filters all have the same characteristics as the low-pass prototype filter. This process may be done extremely efficiently through the use of an FFT algorithm coupled with the polyphase representation.

In this implementation a 64 sub-band polyphase filter bank was used. The prototype filter was designed using the windowing method with a normalized cut-off frequency of $\frac{2\pi}{K} = \frac{2\pi}{64}$. In order to reduce in-band aliasing, the sub-bands were over sampled by a factor of 2, i.e. each sub-band signal is decimated by a factor of $\frac{K}{2}$. The number of coefficients in the prototype filter was 256.

The polyphase filter bank used in this implementation is not the only possible method. An overlapping FFT filter bank [8] can also be used with success, however it is important to determine the amount of delay that can be tolerated, along with the desired sub-band characteristics.

4. DECISION DEVICE

As earlier introduced in section 2, a decision device is used to determine the presence of speech activity in each sub-band independently. The decision device utilises the core statistical mechanism in [6] to make a preliminary speech activity decision in each sub-band. As a summary, the decision device essentially treats each sub-band k as a separate band-limited signal that can be tested for speech activity.

The decision device operates by framing a sub-band signal $x_k(m)$ into frames of length L, then analyzing the frequency content of the particular frame. In this implementation, the frames were chosen to overlap 50%. The length of the frame was $L = 8$ sub-band samples and the overlap was thus 4 sub-band samples.

We model the k^{th} framed noisy sub-band speech signal as,

$$x_k(m,l) = s_k(m,l) + v_k(m,l), \tag{1}$$

where $s_k(m,l)$ is the framed sub-band clean speech signal and similarly $v_k(m,l)$ is the framed sub-band noise signal in the k^{th} sub-band and l^{th} frame. It is assumed the speech and noise are uncorrelated.

We define a signal to noise ratio (SNR) measure as,

$$\psi_k(f,l) = \frac{P_{xx,k}(f,l)}{\hat{P}_{vv,k}(f)} - 1, \tag{2}$$

where $P_{xx,k}(f,l)$ is the power spectral density (PSD) of l^{th} frame in the f^{th} frequency bin of the noisy sub-band speech signal $x_k(m,l)$. $\hat{P}_{vv,k}(f)$ is the expected noise PSD in the f^{th} frequency bin and k^{th} sub-band. For clarity, the term frequency bin consistently does *not* refer to a sub-band. Here we are analyzing the spectral content of the sub-band signal $x_k(m)$.

The PSD is estimated using the Welch method of overlapping windows. This was implemented by averaging over adjacent frames. The expected noise PSD is estimated during an initial silence period, where it is assumed that there is no speech present.

As per [6], the detection mechanism is a statistical mechanism that considers two distinct hypotheses,

$$H_0 : \psi_k(f,l) = \frac{P_{vv,k}(f,l)}{\hat{P}_{vv}(f)} - 1,$$
$$H_1 : \psi_k(f,l) = \frac{P_{vv,k}(f,l) + P_{ss,k}(f,l)}{\hat{P}_{vv}(f)} - 1,$$

where H_0 represents that null hypothesis that no speech is present and H_1 represents the alternative hypothesis that speech is present. $P_{vv,k}(f,l)$ represents the PSD of the noise in the l^{th} frame in the k^{th} sub-band and f^{th} frequency bin, similarly $P_{ss,k}(f,l)$ represents the PSD of the speech. In order to determine between speech activity and non-speech activity, we wish to determine which hypothesis is more likely for each frame and sub-band.

We assume during non-speech periods that the distribution of the SNR measure is Gaussian [6]. We thus model the pdf of the SNR measure during non-speech periods as,

$$p(\psi_k(f,l)|H_0) = \frac{1}{\sqrt{2\pi\sigma^2_{v,k}(f)}} exp\left(\frac{-\psi^2_k(f,l)}{2\sigma^2_{v,k}(f)}\right), \tag{3}$$

where $\sigma^2_{v,k}(f)$ is the variance of the SNR measure during periods of non-speech activity in the f^{th} frequency bin and k^{th} sub-band.

We now wish to determine a threshold with which to compare the SNR measure to, and thus determine if the null or alternative hypothesis is more likely. We represent this comparison as,

$$\psi_k(f,l) \gtrless^{H_1}_{H_0} \eta_k(f), \tag{4}$$

where $\eta_k(f)$ is a threshold in the f^{th} frequency bin and k^{th} sub-band.

As per [6] a threshold in terms of the probability of false alarm can be determined. This threshold may be derived as,

$$\eta_k(f) = \sqrt{2\sigma^2_{v,k}(f)} \cdot erfc^{-1}(2P_{FA}), \tag{5}$$

where P_{FA} is the probability of false-alarm and $erfc^{-1}(u)$ is the inverse complementary error function.

A decision is made in each sub-band by comparing the average of the SNR measure over frequency, to the average of the threshold over frequency,

$$\frac{1}{L}\sum_{f=0}^{L-1} \psi_k(f,l) \gtrless^{H_1}_{H_0} \frac{1}{L}\sum_{f=0}^{L-1} \eta_k(f), \tag{6}$$

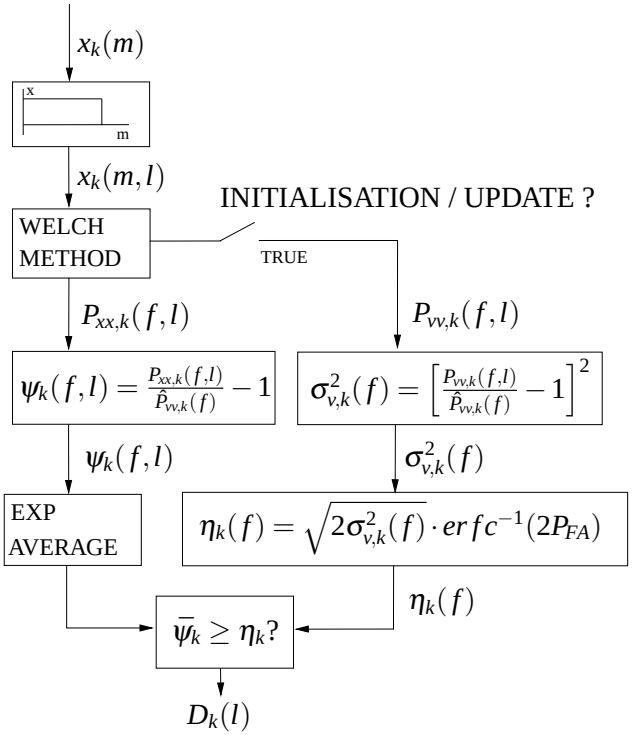

Figure 2: *Block Diagram of Decision Device*

where H_1 is decided if the average SNR is larger than or equal to the average threshold, otherwise H_0 is decided. As per Fig. 1, $D_k(l)$ is decided as,

$$D_k(l) = \begin{cases} 1, & \sum_{f=0}^{L-1} \psi_k(f,l) \geq \sum_{f=0}^{L-1} \eta_k(f), \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

4.1 Implementation

Figure 2 illustrates the implemented decision device. This is identical for all sub-bands k. Initially, the k^{th} sub-band signal is divided into overlapping frames of length L. By combining multiple frames, the Welch method of overlapping windows is then used to determine the PSD.

Next, the SNR measure is calculated as per (2), where $\hat{P}_{vv}(f)$ is estimated during an assumed initial non-speech period. The SNR measure is then smoothed with an exponential average. The smoothed SNR measure is calculated as,

$$\bar{\psi}_k(f,l) = \alpha\, \psi_k(f,l) + (1-\alpha)\, \bar{\psi}_k(f,l-1), \quad (8)$$

where α is the smoothing coefficient. A value of $\alpha = 0.95$ was found to give good results for this implementation.

The smoothed SNR measure is then averaged over all frequency bins f and compared to the averaged threshold η_k to determine the initial decision.

During the assumed initial non-speech period the variance of the SNR measure is estimated. The variance is used to compute the threshold for each frequency bin as per (5). This threshold is then averaged over all frequency bins f and compared to the averaged SNR measure to determine speech activity. If the average SNR measure is larger or equal to the average threshold, then $D_k(l)$ is set to '1', otherwise $D_k(l)$ is set to '0' as per (7).

In this implementation the probability of false alarm (P_{FA}) was set to 0.05 (5%) and the number of overlapping frames used to estimate the PSD was 2. Further, the threshold $\eta_k(f)$ was updated during periods of non-speech activity as determined by the sub-band speech activity decisions $V_k(l)$. The variance of the SNR measure during non-speech periods $\sigma^2_{v,k}(f)$ was also updated by this method.

5. DECISION ANALYSIS

The decision analysis block is incorporated to help reduce false alarms. It does this in two distinct ways, firstly, if a single sub-band indicates there is speech, when no adjacent sub-bands indicate speech, then this is considered to be a false alarm. The initial speech-active decision is then set to speech-inactive. Secondly, if less than Q sub-bands are active at one instance, it is assumed that there is no speech present. All speech-active decisions are then set to speech-inactive.

Note, the action of the decision analysis depends highly on the assumption that adjacent sub-bands are uncorrelated. Further, during periods of speech activity, it is assumed that more than Q sub-bands will be active.

The first action of the decision analysis can be expressed by initially estimating a metric,

$$M_k(l) = \sum_{p=-1}^{1} D_{k+p}(l) \quad k = 1,2,...,K-2. \quad (9)$$

The final decision $V_k(l)$ is then determined in the following way,

$$V_k(l) = \begin{cases} 0, & M_k(l) \leq 1, \\ D_k(l), & M_k(l) > 1, \end{cases} \quad (10)$$

for $k = 1,2,...,K-2$. The final decision for the cases $k = 0$ and $k = K-1$ are set as, $V_k(l) = D_k(l)$.

The final decision $V_k(l)$ is then subject to the secondary test. A second metric P is calculated as,

$$P(l) = \sum_{k=0}^{K-1} V_k(l). \quad (11)$$

The final decision is then modified as,

$$V_k(l) = \begin{cases} 0, & P(l) \leq Q, \\ V_k(l), & P(l) > Q, \end{cases} \quad (12)$$

for all k. Through experimentation a value of $Q = 8$ was found to be appropriate for the particular implementation, however this is dependent on the sampling frequency of the original data and the number of sub-bands K, and should be tuned appropriately.

6. EVALUATION

In order to evaluate the scheme, two short sequences were chosen to illustrate the concept of multi-decision sub-band VAD. A formal analysis such as in [9] was not undertaken. The primary reason for this is that the traditional VAD metrics do not directly apply to this scheme given its multi-decision sub-band nature, and the scheme cannot be compared to current schemes in the usual manner.

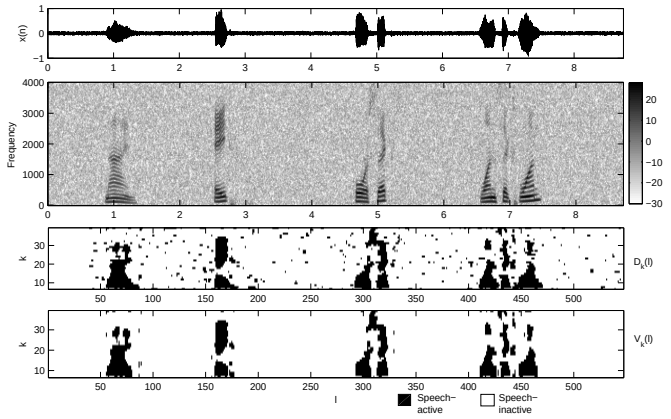

Figure 3: *First sequence time domain data, spectrogram, decision device output $D_k(l)$, and final VAD output $V_k(l)$ (white Gaussian noise)*

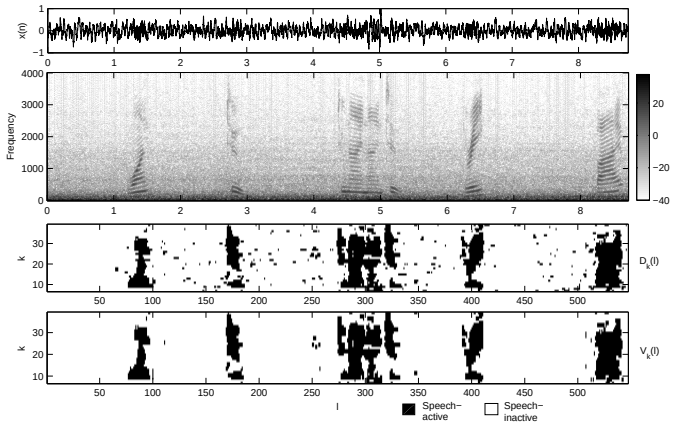

Figure 4: *Second sequence time domain data, spectrogram, decision device output $D_k(l)$, and final VAD output $V_k(l)$ (vehicle noise)*

Each sequence was approximately 9 seconds in length and was created by concatenating spoken digits from the TIDIGITS database (both connected and discrete). The first spoken sequence was "nine, eight, four-six, one-six-one". The second sequence was "one, two, two-nine-nine-two, three, five". White Gaussian noise was added to the first sequence with an average SNR of approximately 15dB. Vehicle noise from the NOISEX-92 database was added to the second sequence with an average SNR of approximately -10dB. The sampling frequency of the data was 8000Hz.

Figure 3 illustrates the multi-decision sub-band VAD concept. The figure shows the original time domain data of the first sequence, a spectrogram of the data, the decisions as decided by the decision device for each sub-band and the final decision as decided after the decision analysis. As can be seen, the decision analysis removes many false alarms, with negligible impact on the detection of speech. Further, the merits of the concept of multi-decision sub-band VAD can be clearly seen. From the final decisions $V_k(l)$, it is clear that noise statistics may be updated during what would traditionally be referred to as speech periods.

Figure 4 further illustrates the utility of the multi-decision sub-band voice activity detector. The vehicle noise is primarily low frequency in nature. Examining the results, we see the multi-decision sub-band voice activity detector is unreliable in the lower few sub-bands, where it is unable to discriminate speech activity from noise. However in all other sub-bands, you can clearly see the scheme operates as expected, thus illustrating the effectiveness of multi-decision sub-band VAD. Traditionally, a voice activity detector would have trouble detecting speech active periods due to the extremely low SNR of this sequence.

7. CONCLUSION

In conclusion the concept of multi-decision sub-band VAD was introduced. The concept hinges on the fact that speech will not simultaneously encompass all frequencies at a given moment. Bearing this in mind, it was proposed that speech activity decisions could be made independently in sub-bands, resulting in a set of multiple speech activity decisions for any instance.

A method for making multi-decision sub-band speech activity decisions was proposed. The method utilized an oversampled polyphase filter bank to decompose the original data into multiple parallel sub-bands. These sub-bands were then independently tested for speech activity using a statistical mechanism.

Finally, evaluations indicated that multi-decision sub-band VAD can be useful, especially in speech enhancement applications. This usability comes from the ability of the scheme to allow noise statistics to be updated during what would traditionally be labeled speech active regions.

REFERENCES

[1] A. Davis, S.-Y. Low, S. Nordholm and N. Grbic, "A subband space constrained beamformer incorporating voice activity detection", in *Proc. IEEE ICASSP'05*, Philadelphia, 2005, pp. 65-68.

[2] H. Gustafsson, S. E. Nordholm and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging", *IEEE Trans. on Speech and Audio*, vol. 9, no. 8, pp. 799-807, Nov. 2001.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Trans. on Speech and Audio*, vol. 9, no. 5, pp. 504-512, July 2001.

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

[5] Y. D. Cho, K. Al-Naimi and A. Kondoz, "Improved statistical voice activity detection based on a smoothed statistical likelihood ratio", in *Proc. IEEE ICASSP'01*, Salt Lake City, 2001, pp. 737-740.

[6] A. Davis, S. Nordholm and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold" *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 412-424, Mar. 2006.

[7] P.P. Vaidyanathan, *Multirate system and filter banks*, Prentice Hall, 1993

[8] R.E. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis", *IEEE Trans. on Acous., Speech and Signal Processing*, vol. 28, no. 2, pp. 99-102, Feb. 1980.

[9] F. Beritelli, S. Casale and G. Ruggeri, "Performance evaluation and comparison of ITU-T/ETSI voice activity detectors", in *Proc. IEEE ICASSP'01*, Salt Lake City, 2001, pp. 1425-1428.