Proc. 1996 Australian New Zealand Conf. on Intelligent Information Systems, 18-20 November 1996, Adelaide, Australia. Editors, Narasimhan and Jain

SPEECH DETECTION METHOD ANALYSIS AND INTELLIGENT STRUCTURE DEVELOPMENT

Anthony LITTLE Dr. Leon REZNIK
Department of Electrical and Electronic Engineering
Victoria University of Technology
P.O. Box 14428 MCMC
MELBOURNE VIC 8001 AUSTRALIA
Fax: +(613) 9688 4908
Email: alittle@cabsav.vut.edu.au

Abstract
Making a decision about the presence/absence of the speech signal in a noisy environment is a very important problem which must be solved in many application areas including speech recognition.
This paper presents a methodology analysis of several algorithms, identifying both their advantages and disadvantages to examine their ability to detect certain properties of speech. From these results, a speech detection system is designed using algorithms which complement each other. The system includes an application of 4th order cumulants and zero crossing methods to determine the presence of speech. The system is then tested upto noise levels of -4dB while maintaining responsive detection. An adaptive fuzzy rule base is then proposed for implementation and operation under varying environmental conditions.

Keywords: VAD, VOX, Speech Detection

1.0 Introduction

Speech is comprised of many components, which in combination, produce an identifiable sound to the human ear. Detection under ideal conditions is relatively straight forward, however a typical situation comprises both speech and associated background noise. The type and level of noise introduced to a speech signal present a difficult problem when attempting to ascertain if speech is present. Even when the noise level is negligible, the problem of speech detection is not simple because of the complex components of the speech signal.

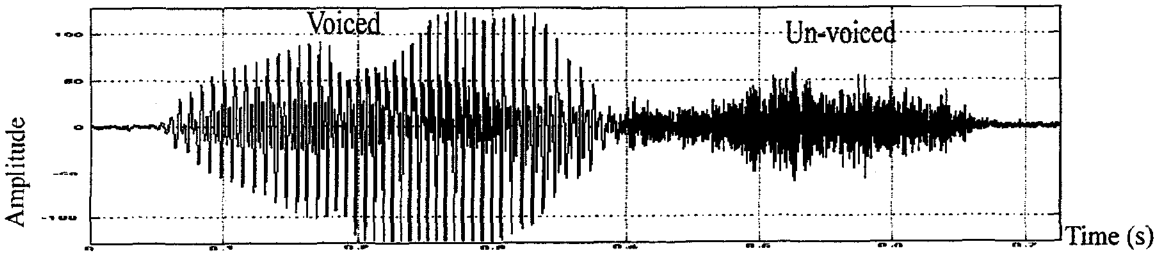
Most of the methods applied have been oriented towards detection of speech by identifying one property of the signal. Such methods have been strongly limited in reliable detection and practical application due to the complexity of a speech signal. This paper presents a new approach to the development of an adaptive robust speech detection method.

2.0 Speech Classification

Speech can be classified into two distinct categories, 'voiced speech' and 'un-voiced speech'.

Voiced speech is associated to those sounds which have propagated from the vocal cords. The vocal cords vibrate causing a sound with frequency components, which is then shaped via the vocal tract, mouth, tongue and lips. Voiced speech can be modelled as a slowly time-varying linear system that is excited by a quasi-periodic pulse signal. (See Figure 1) Spectral analysis reveals that the frequency components are spread mostly in the lower half of the speech bandwidth.

Un-voiced speech is associated to those sounds which do not require the vibration of the vocal cords and are termed 'fricatives'. These sounds are generated by forming a constriction at some point in the vocal tract and air forced through the constriction at high velocity to produce turbulence. This turbulent air flow is shaped by the vocal tract, mouth, tongue and lips to produce sounds such as the 's' in 'sign', or 'ch' in 'watch'.



The word "Moose"
Figure 1

Observation identifies fricatives by the difference in density and the lack of a recognisable pattern, (See Figure 1). Spectral analysis reveals a consistantly spread spectrum in the upper half of the speech bandwidth, indicating non-uniformity and randomness of the signal. As a result of not using the vocal cords to produce such sounds, the energy content of fricatives is significantly lower. Due to the nature of fricative signals, they can be difficult to identify when concealed with noise possessing similar properties.

3.0 Analysis of Speech Attribute Detection

Voiced and un-voiced speech do not posses similar properties, rather they appear to be the converse of each other. Voiced speech contains a small subset of frequencies and is highly structured, whereas un-voiced speech typically contains a large even portion of upper frequencies and appears somewhat uncertain and un-structured. Therefore, for a VAD algorithm to become efficient and accurate, it must employ methods to detect for both types of speech sounds.

[1] suggested a combination of simple algorithms to detect different attributes, rather than one complicated all-encompassing algorithm. The following results have been compiled to identify methods which are simple and exhibit robust attribute detection. The information will then be used to select the methods needed to extract a sufficient amount of data to determine if speech is present.

The methods listed in Table 1 have been implemented and simulated using several different techniques and variations to obtain a more concentrated focus on their prospective ability. The ability has been statistically measured to define which speech attributes can be extracted with a good confidence margin.

	Method Analysis	
Property	**Voiced Speech**	**Un-voiced Speech**
Energy	Voiced speech contains a high amount of energy, which may be detected by an adaptive threshold. As the noise level is increased, word clipping becomes more evident. Short time energy algorithms working on a time decaying principle produce the most responsive results.	Detection by energy is not a robust method due to the high frequency, low amplitude nature of the signal.
Derivative	The excitation of the vocal cords causes a discernible rate change. Speech which contains an initial low amplitude vocal sound is usually short in duration and does not cause appreciable clipping. This method oes not provide a robust responsive output when noise is significant.	Fricatives usually contain lower amplitudes and smaller derivatives, which restricts the detection ability when noise is present. An improvement can be obtained by calculating the accumulating derivative of the signal. All methods employed were ineffective when applied to a signal with significant noise levels.
Cepstral	Voiced speech contains quasi-periodic oscillations which typically exist within the range of 300 750Hz (formant frequency). The Cepstrum is a measurement between the peaks of this formant frequency. The algorithms employed concentrated on the presence of the cepstrum and not the value. Although the output is reliable, it does not provide rapid response for a real time VAD system.	[2] developed a cepstral VAD that had a high degree of noise immunity for voiced signals, however un-voiced speech could not be reliably detected when appreciable noise was present.
Zero Crossing	Voiced speech exists mostly in the lower half of the speech spectrum. The zero crossing rate is similar or lower than most forms of noise, althouth the period of the signal is more correlated.	Due to the high frequency nature of the majority of fricatives, this algorithm is particularly suitable for un-voiced detection. More notably, this method is not dependant on amplitude and therefore less effected by noise levels. Best results have been obtained by a short time summation of the zero crossing positions. High frequency noise can cause momentary peaks and false triggering when the adaptive threshold is set too low.

Table 1

Property	Voiced Speech	Un-voiced Speech
3rd Order Cumulants	This technique utilises the Gaussian similarity to noise to effectively reject a broad range of noise, whereas the voiced signals have a more skewed signal and less symmetrical in the short time duration and are not rejected. The simulated results have demonstrated excellent robust responsive detection upto a 50% level increase in additional noise.	The nature of the method rejects signals which are linear processes such as fricatives and Gaussian noise. A smaller sample window detects the presence of fricatives, however the output is not a robust indicator and becomes less responsive to un-voiced signals with additive noise.

Table 1 - continued

As can be seen from the results, there is no one method capable of detecting both voiced and un-voiced speech to an acceptable quality, however in combination a more accurate VAD may be developed.

4.0 Results

[3] applied high order statistics to determine the presence of noise or noise + speech. The algorithm output showed excellent immunity to noise and reliable detection of voiced speech, however fricatives were attenuated and careful thresholding was required. A VAD system has been developed to incorporate both zero crossing (un-voiced detection) and higher order cumulants (voiced detection). Both algorithms are less susceptible to a broad range of noise from effecting the results and are not entirely dependant on the amplitude of the signal.

Investigation into fourth order cumulants reveals a more reliable source of voiced speech detection, with a threshold of low variance to noise. A small 2ms delayed output (look ahead) for both algorithms provides sufficient information for end-point detection with excellent response times, in most cases just prior to speech. Zero crossing is less reliable in responsive detection when some high frequency forms of noise are present. To decrease the sensitivity to different noise types and improve the reliability, the author is currently developing an envelope detector for high frequency fricative signals. Fricatives may then be effectively detected by the slow time envelope variation in the upper frequency domain, while flat envelopes representing noise will be rejected.

Figure 3a depicts results obtained from a VAD system using both fourth order cumulants and zero crossing with a SNR>30dB (no speech clipping). Figure 3b shows the same VAD system operating when the environment contains a SNR= -4dB. It can be clearly observed that both algorithms are required collectively to detect speech.

5.0 Adaptive VAD Structure

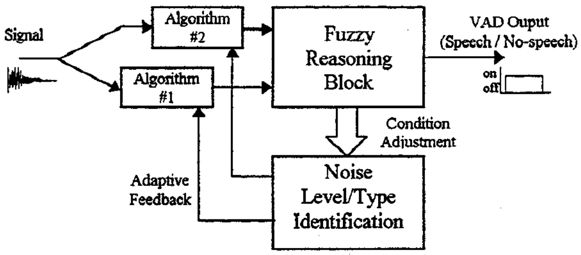
The Adaptive VAD structure should contain computationally efficient algorithms, as above, which are utilised by a central reasoning structure to make a collective decision of speech or no-speech. Combined with this structure is any additional calculations required by the algorithms to become adaptive to the changing conditions of the signal. This approach has been taken to increase the overall computational efficiency. Spectral analysis has not been used as a method of detection due to the efficiency and relative higher computational requirements.

A fuzzy logic approach has been taken as the structure which is most suitable to derive a final decision based on the components and to realise a non-linear structure of the process, (See Figure 2). For example;

(a) if energy is high and zero crossing low, then the sample contains a voiced signature.

(b) if zero crossing is lower than average noise and energy is marginally higher, then the sample most likely contains a voiced signature.

Such rules have been developed by processing simulation results.



Adaptive VAD Structure
Figure 2

6.0 Conclusion

Several algorithms have been analysed for their robustness in detecting properties of speech in a signal of unknown noise content.

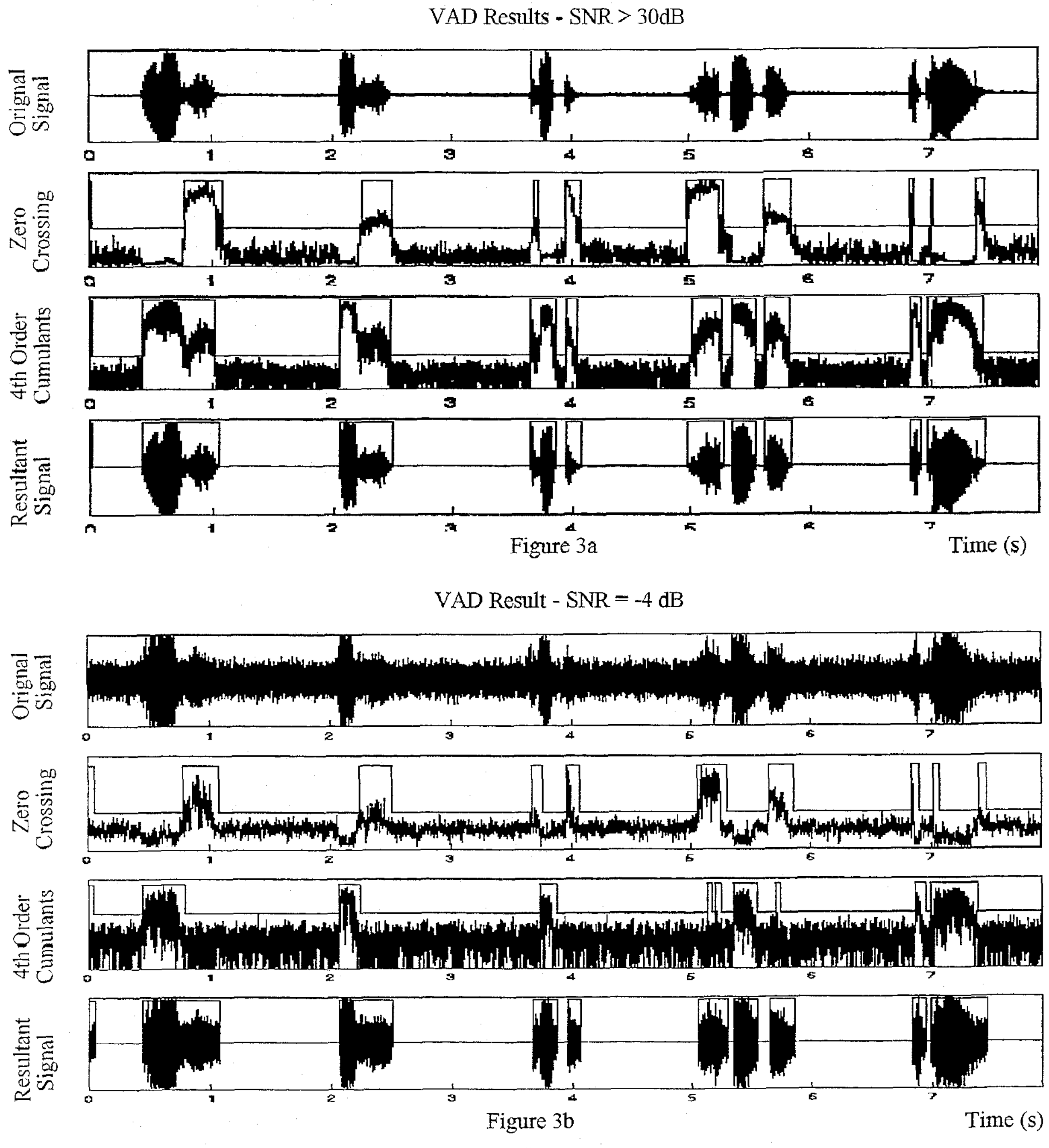
Fourth order cumulants produced the most reliable and responsive detection of voiced signals and zero crossing for un-voiced signals. Both are less vulnerable to noise due to the low dependence on signal amplitude. A fuzzy rules based adaptive VAD is the most suitable structure for implemention due to the complementing nature of the processing results. Simulated results demonstrated in combination the VAD system was robust to noise and responsive in switching times.

7.0 References

[1] BISHOP, A.; FEAKES, K.L.; PETTITT, A. An Intelligent HF Squelch. Sixth International Conference on 'HF Radio Systems and Techniques'. p31-5, IEE, London UK, 1994

[2] HAIGH, J.A.; MASON, J.S. Robust Voice Activity Detection using Cepstral Features. 1993 IEEE Region 10 Conference, TENCON '93. p321-4 vol.3, IEEE, New York, NY, USA, 1993.

[3] RANGOUSSI M.; CARAYANNIS G. Higher Order Statistics Based Gaussianity Test Applied to On-Line Speech Processing. Conference Record of the 28th Asilomar Conference on Signals, Systems and Computers, p303-7 vol.1, IEEE Comput. Soc. Press, Los Alamitos, USA, 1994.



VAD Results - SNR > 30dB

Figure 3a

VAD Result - SNR = -4 dB

Figure 3b

Figure 3 - "Moose, Bush, Cut, Speech, Depend". Both figures depict the orignal speech sample, each algorithm output and then the resultant signal generated by the VAD switching times. In both cases, (a) SNR >30dB & (b) SNR= -4dB, the switching times are almost identical in their high quality performance switching.