The NP Speech Activity Detection Algorithm

Joseph Pencak and Douglas Nelson

Dept. of Defense, 9800 Savage Rd., Ft. Meade, Md.

Abstract

This paper describes a new algorithm, the NP algorithm, for detecting speech signals of varying quality and gain levels. NP operates in the frequency domain and renders speech/no-speech decisions based on a signal-to-noise ratio (SNR) derived from a sorted power spectrum. In addition to the SNR estimates, a spectral whitening process and an estimate of the variance in the ratio of the signal power to total energy are also used to identify and reject signals that are stationary or nearly stationary

The key features of this algorithm are:

- Detection is based on a single FFT
- Decisions are independent of signal gain
- Process has 3 dB/octave processing gain from the transform
- Frequency domain processing permits exploiting structure of signal

Design Constraints of Initial Implementation

The NP algorithm was originally developed to be used in a DSP based system as a front-end to sort speech signals from non-speech signals. The front end requirements for that system were that it must be able to operate in a severe noise environment containing stationary and slowly varying interference. In addition, the system required acceptance of all understandable speech in a near/far environment which could exhibit 20 dB instantaneous variations in power. Two conventional speech detection methods were considered and rejected since they could not be readily adapted to the constraints of the problem. Conventional speech detection based on energy detection can not adapt to sudden fluctuations in signal power, and the methods based on autocorrelation/pitch-detection could not perform in the interference environment. The method described in this paper does meet the original system requirements and has generally outperformed other methods to which it has been compared in head to head tests. In addition, the DSP implementation uses approximately 25 - 30% of the resources of a DSP32C chip to process a voice-grade channel. Precise quantitative test results measuring performance are not currently available since there is no marked database available which establishes truth.

Since the original implementation provided a data stream of 8-bit mu-law data sampled at 8 kHz, the description of all processes in the algorithm and all figures are based on these parameters. The algorithm has been adapted to other data sources without significant problems.

Computing the SNR

In the NP algorithm, signal detection is based on a SNR calculation which is estimated from a sorted spectrum. The transform size is not critical, but must provide enough resolution to separate pitch bars and must be short enough that pitch, if present is reasonably stationary. In the original implementation, a frame size of 0.1 seconds was specified. This equates to 800 samples at an 8 kHz sample rate. The frames were augmented to 1024 padding with data and allowing the successive data frames to overlap by 224 samples. These parameters provide very good performance for a wide variety of speakers, both male and female.

In calculating the spectrum, the data are converted to floating point numbers normalized to assume values in the interval (-1,1). The frames of data are then windowed, by multiplying each frame of data by a raised cosine window. A power-spectrum is computed from a normal FFT. DC, power related components and other undesirable low and high frequency spectral artifacts are removed by discarding all spectral components below 195 and higher than 3843 Hz. In effect, this process removes the first three harmonics of the 60 Hz carrier used in domestic electrical power. In addition, the spectrum above 3843 Hz contains little information of use in speech detection.

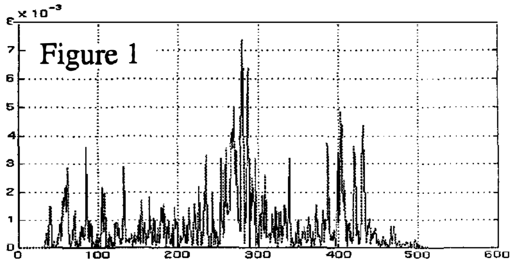


Figure 1

Typical 195-3843 Hz Speech Spectrum

0-7803-2431-5/95 $4.00 © 1995 IEEE

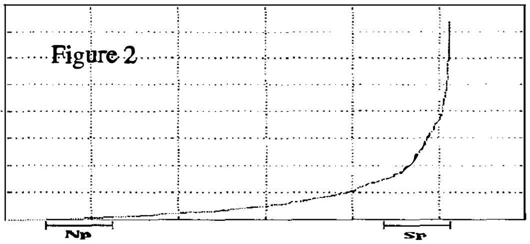


Figure 2

Sorted Speech Spectrum Showing Regions
Used to Compute Noise and Signal Power

The total spectral energy is calculated by summing the power spectrum,

$$E_T = \sum_{\omega} \hat{x}(\omega)^2 \text{, where}$$

$\hat{x}$ is the Fourier transform of x. Finally, the power spectrum is sorted in increasing order, resulting in the high order elements containing the maximum energy and the low order elements containing the minimum. Figure 2 illustrates this sorted spectrum. The objective of this sort is to isolate the speech energy occurring in the pitch bars and frication bands from the noise energy, residing between the pitch bars and/or away from the frication bands.

This process, as implemented does not take advantage of the periodic structure of voiced speech. Ideally, the bin locations could be sorted along with their associated power values so that pitch structure could be exploited. In principal, this would make the algorithm perform better but would increase the computational overhead. In view of the quality of the results obtained in initial testing, the additional complexity did not appear to be necessary.

Given this sorted spectrum, we can now estimate the SNR which is the basis of our decisions to mark a frames as speech or non-speech. Bands used to estimate signal and noise power are shown in Figure 2. We define Np as the estimated noise power density and Sp as the estimated signal power density. Np is computed as

$$\mathrm{Np} = \frac{1}{100}\left(\sum_{i=45}^{145} \hat{x}_i\right)$$

where $\hat{x}_i$ is the i-th element of the sorted power spectrum.

To compute Sp, we do not average power over a fixed number of elements as we did in computing Np. Computing the signal power density in this manner results in poor performance, possibly because the percentage of the spectrum contributing to speech energy dynamically changes. Sp is computed as

$$\mathrm{Sp} = \frac{1}{512-L}\left(\sum_{i=L}^{512} \hat{x}_i\right),$$

where L is the minimum integer satisfying

$$\sum_{i=L}^{512} \hat{x}_i \geq 0.4E_T$$

and E_T is the total spectral energy.

The 40% cutoff is an experimentally derived parameter and is approximately the energy contributed by the pitch bars around the first formant. With the calculation of Sp and Np, the SNR can be estimated as

$$SNR = Sp/Np$$

Clearly, SNR calculated this way is gain and frame independent. To complete speech detection, we compare the SNR to a threshold. If SNR is above a threshold, a speech signal is declared to be present. The value of the threshold normally used is 90 (about 20 dB SNR for the strongest spectral components of speech). The sensitivity of the detector can be decreased or increased by adjusting the threshold up or down respectively.

Spectral Whitening and Variance Tests

Although the technique described does well in finding speech in clean as well as noise-degraded environments, and immediately eliminates signals with uniform spectral distributions (e.g., white noise), the algorithm will false alarm on signals whose spectral distributions are biased (e.g., colored noise, tones, narrowband FSK, etc.). To compensate for this problem, two additional features are incorporated into the detection process to eliminate nearly stationary signals. These are adaptive spectral whitening and a secondary test based on the variance in the ratio of signal power spectral density to total power.

The spectral whitening is performed using a leaky integrator which tracks the long-term average power of each frequency in the spectrum. The objective is to identify which frequencies continually have high energy, indicating the presence of stationary signals. The integrating function is defined as

$$E_i(\omega) = 0.99E_{i-1}(\omega) + 0.01C_i(\omega)$$

where $E_i(\omega)$ is the estimate of the average power at frequency ω at the i^{th} frame and $C_i(\omega)$ is the power at frequency ω observed in the powerspectrum of the i^{th} frame. The values $E_i(\omega)$ track the long-term growth or decay in power for each frequency. The whitened spectrum is computed as

$$W_i(\omega) = \frac{C_i(\omega)}{E_i(\omega)}$$

In the implementation described here, the whitening is not applied unless $E_i(\omega)$ exceeds a threshold for some i and ω. The default value for this threshold is set at an experimentally derived value of 20,000. If applied, whitening is applied prior to the calculation of SNR, and SNR is calculated using the whitened spectrum. For stationary signals, the computed whitened SNR is near 1, well below the set threshold indicating possible speech. resulting in the input frame being marked as not speech. The whitening process typically takes 4-5 seconds to stabilize and reject tones. This time may be shortened by lowering the threshold or by changing the coefficients of the leaky integrator (e.g., 0.75 and 0.25 rather than 0.99 and 0.01).

Spectral whitening works well for eliminating high-energy stationary signals, but will fail to lock to low-energy stationary signals which do not exceed the threshold and thus trigger the filter. To handle this case, the algorithm uses a secondary test based on the variance in the ratio of the signal density and the total spectral energy. This is accomplished by computing

$$V = E[(X-\mu)^2],$$

where

$$X = \log_2(Sp/E_T).$$

and μ is estimated using a leaky integrator

$$\mu_i = 0.75\times\mu_{i-1} + 0.25\times|X_i|,\quad \mu_0 = 0$$

The log function in the variance calculation is essentially companding and helps to normalize high and low-energy signals.

In practice, the variance is again approximated using a leaky integrator

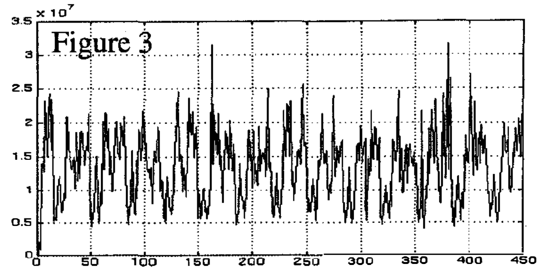
$$V_i = 0.75\times V_{i-1} + 0.25\times(X-\mu_i)^2,\quad V_0 = 0$$

Finally, the value of V is smoothed using a final leaky integrator

$$\tilde{V}_i = 0.75\times\tilde{V}_{i-1} + 0.25\times V_i,\quad \tilde{V}_0 = 0.$$

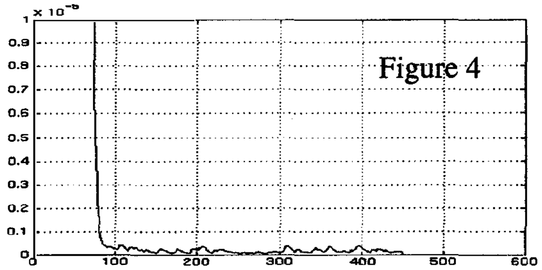
For stationary signals, the value of $\tilde{V}$ will be close to 0 and the algorithm will typically adapt to such signals in about 3 seconds. To complete the process, a threshold is set, typically at approximately 0.1. If the variance is less than this threshold, the signal is determined to be non-speech. As with the whitening process, the rate of adaptation can be affected by changing the threshold level or the coefficients on the integrators.

The variance test is independent of signal gain; however, some signals, such as multi-tone FSK have a variance which is nearly as large as that of speech. It is impractical to attempt to distinguish these signals from speech using variance alone since the threshold would have to be set so high that some readable noise-degraded speech would be rejected. As a result, the variance test complements, but cannot replace, the whitener and vice versa



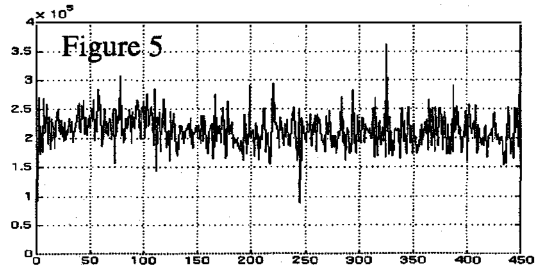
Time Varying SNR of Quantized Weak Tone

Figure 3 is a plot of the computed SNR for each 0.1 second frame of a 45-second long 1 kHz sine wave at 20 millivolts peak to peak. The signal was generated with sufficient SNR to trigger the basic SNR based detector but with too little power to trigger the adaptive whitener

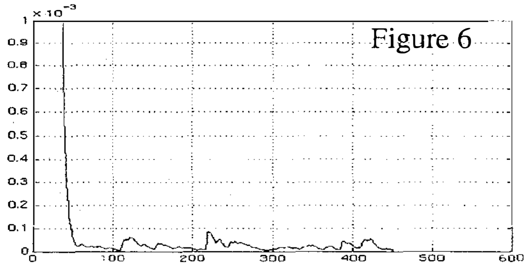


Time Varying Variance of Weak Tone

Figure 4 shows the estimated variance of the weak sine wave. After approximately 8 seconds, the variance settles down and attains its steady state. The threshold of 0.1 is met after about 2-3 seconds and is well above this steady state value.



Time Varying SNR of Multi-tone Noise



Time Varying Variance of Multi-tone Noise

Figure 5 and Figure 6 show plots for the SNR and variance computations respectively for a signal consisting of multi-tone background noise.

Detector Output (DSP Implementation)

In the DSP implementation, NP renders speech/no-speech decisions for each 0.1 second frame of input based on the current transform and information accumulated in the whitener and variance integrators. The device marks each frame as speech or non-speech for use in data compression and follow-on processing.

In practice, the device maintains a history buffer of these decisions. The purpose of this history buffer is to enable the processor to accept several frames prior to the onset of speech and as many as two frames after the offset of speech. Currently the system accepts 2 frames prior to onset and 1 frame after offset of speech. By accepting a few frames before and after the detected speech, transitions from silence to speech are smoother, and the intra-word gaps are virtually eliminated.

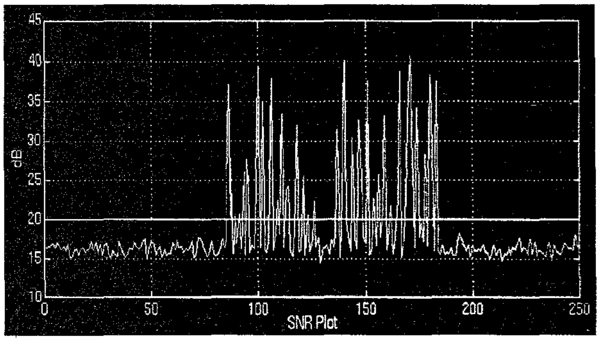


Figure 7 Time Varying NP-SNR 0-dB SNR(approx) (0-dB SNR is estimated over 4kHz bandwidth spectrum. NP-SNR estimates signal power using only strongest pitch bars and realizes an approximate 30-dB processing gain from the 1024 point FFT used as the front end)

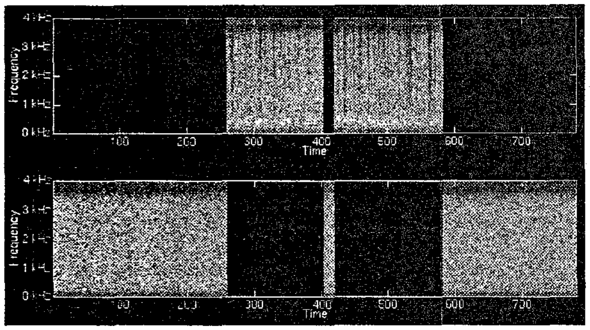


Figure 8 Output of Speech Detector 0-dB SNR(approx)
Top: Detected Speech, Bottom: Rejected Noise

Figures 7-8 show performance of the NP detector on a speech signal at approximately 0-dB SNR. The detector has correctly identified two segments of speech and three segments of non-speech.

Future Directions

There are two modifications which are being implemented in the DSP hardware. One of these changes represents an improvement to the algorithm, and the other id being implemented to provide reliability estimates for machine speech recognition processes. The first improvement is implementation of a pitch based SNR estimate. In this process, the signal power for voiced speech is measured at the pitch bars, and the noise power is estimated between the pitch bars. This is a relatively simple modification since the pitch bars form a a harmonic structure which is an even function. The process of estimating the power in the pitch bars is essentially a cosine transform of the spectrum, with the signal power estimated where the cosines are positive and the noise power estimated where the cosines are negative.

The second improvement is that the SNR estimates and the SIR estimated using the interference estimated using the whitening filter may be used as signal quality estimates. The intent is to estimate the expected reliability of machine recognition processes from these signal quality estimates.

REFERENCES

[1] Corman, Leiserson, Rivest, Introduction to Algorithms, McGraw-Hill, 1990.

[2]J.L.Flanagan, Speech Analysis Synthesis and Perception, Springer-Verlag, Berlin, 1965.

[3] B. Mak, et.al. "A Robust Speech/Non-Speech Detection Algorithm Using Time and Frequency-Based Features".Proc IEEE-ICASSP Conf., 1992.