Voice Activity Detection in Non-stationary Noise.

Sunil Srinivasa
Department of Electrical Engineering,
IIT Madras

A Summer Project under the guidance of
Dr. N. Rama Murthy
Scientist 'E', CAIR

Objectives of the project.

- Learning various methods for VAD proposed in literature.
- Implementation of several VAD algorithms by using MATLAB™ – successful detection of start and end points in the noisy speech signal.
- Comparison of VAD algorithms – search for the best VAD algorithm.
- Study of some Speech Enhancing methods.
- Software coding of the techniques learnt to enhance speech.

Introduction to VAD.

- What is VAD ?

  VAD stands for **VOICE ACTIVITY DETECTOR**.

  It carries out the process of distinguishing between conversational speech and silence(noise) i.e.

  It detects the beginning and ending of talk spurts.

  A VAD makes use of a set of rules incorporated in the **Voice Activity Detection Algorithms**.

Advantages of VAD.

- Saves the processing time in canceling noise from speech.
- Helps to reduce Internet Traffic.
- Achieves Band Width reduction, as high-frequency noisy frames are discarded.
- Allows for multi-data transmission.
- Causes silence compression – very important for both fixed and mobile telecommunication systems.

Few Applications of VAD.

Required in some communication applications like :

- Speech Recognition
- Speech Coding
- Hands-Free Telephony
- Echo Cancellation.

It is also required in Voice over IP systems for providing Toll Grade Voice Quality.

It is an integral part of Transmission systems.

Desirable aspects of VAD Algorithms.

- A good decision rule even at low SNR.
- Adaptability to non-stationary background noise.
- Low computational complexity.
- Toll Quality Voice Reproduction – very important for Voice over Packet Networks.
- Maximum saving in Band Width.

General features of VAD algorithms – The basic processing steps.

- Band-pass filtering done (300-3400 Hz) to remove out-of-band components in the noisy speech.
- Zero-padding of the speech signal and Segmentation of speech into 10-30ms frames.
- Windowing and Overlapping – for frequency domain algorithms.
- Feature Vector Extraction.
- Adaptation of a suitable threshold – Decision making.
- Classification of frames as “ACTIVE”(containing speech) or “INACTIVE”(noisy).

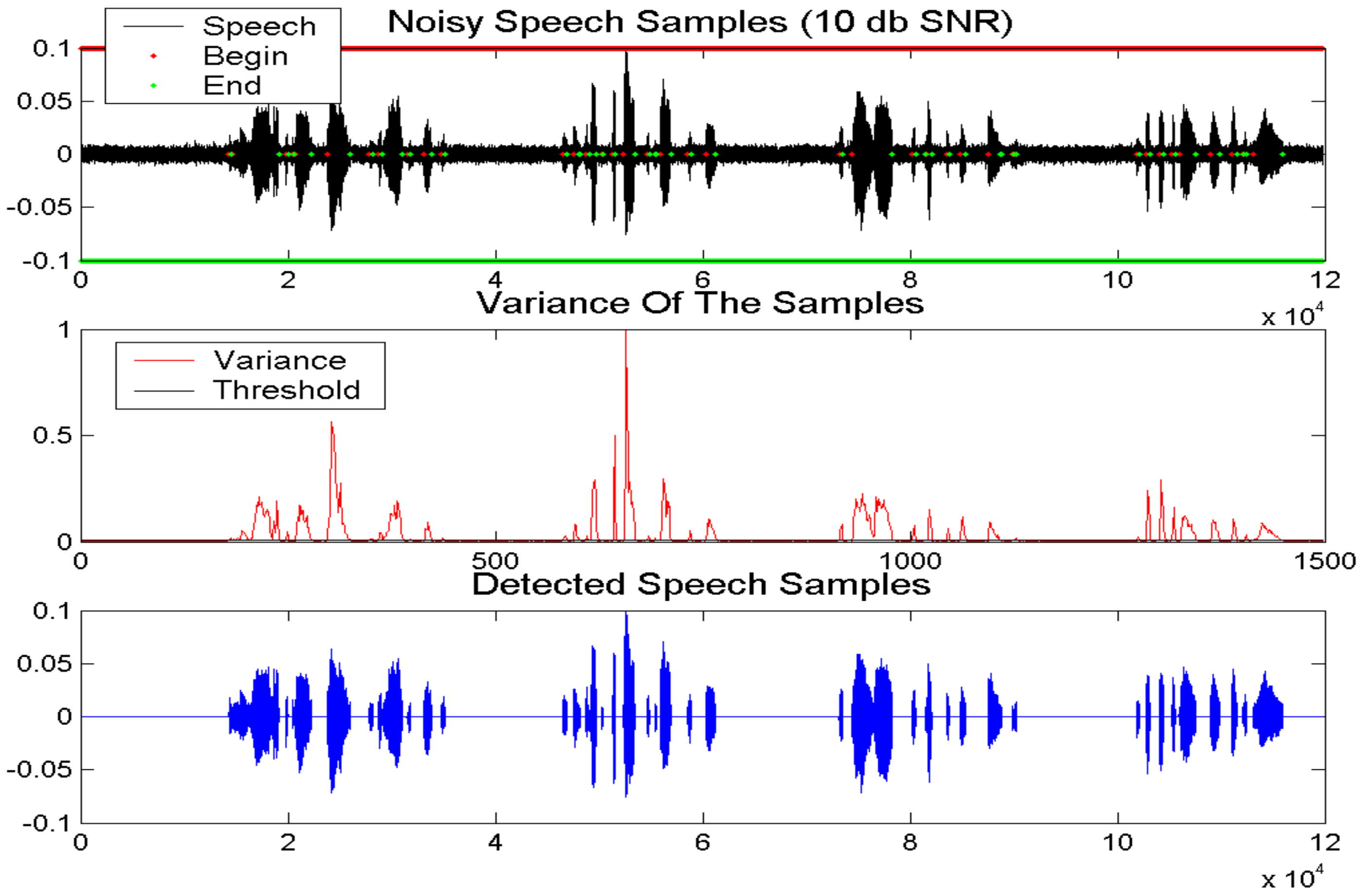
Some Feature Vectors.

- Short-time energy
- Short-time variance
- Zero-crossing
- Cepstrum
- Sub-band energies
- Periodicity measure
- Pitch and Timing
- Spectral flatness

Variance Based Detector.

- Simple and Robust.
- Time domain algorithm.
- Feature vector is the short-time variance.
- Variance of speech is considerably higher than that of noise – easy decision making.
- Threshold – completely based on variance.
- Detects low energy phonemes at considerable SNR.
- However, low SNR causes undue clippings.

Response of the Variance Detector for the TIMIT



Linear Energy-Based Detector (LED).

- Simplest and easiest to implement.
- Works completely in the time-domain.
- Feature vector is the short-time energy.
- Energy of voice is higher than that of noise – makes way for suitable filtering.
- Threshold vector based on the energy.
- Adaptation is carried on the threshold to achieve robustness.

Modified Threshold Value.

$$E_{th}(new) = E_{th}(old) \cdot (1 - p) + E_{sil} \cdot p$$

where,

$E_{th}(new)$ is the updated value of the threshold,

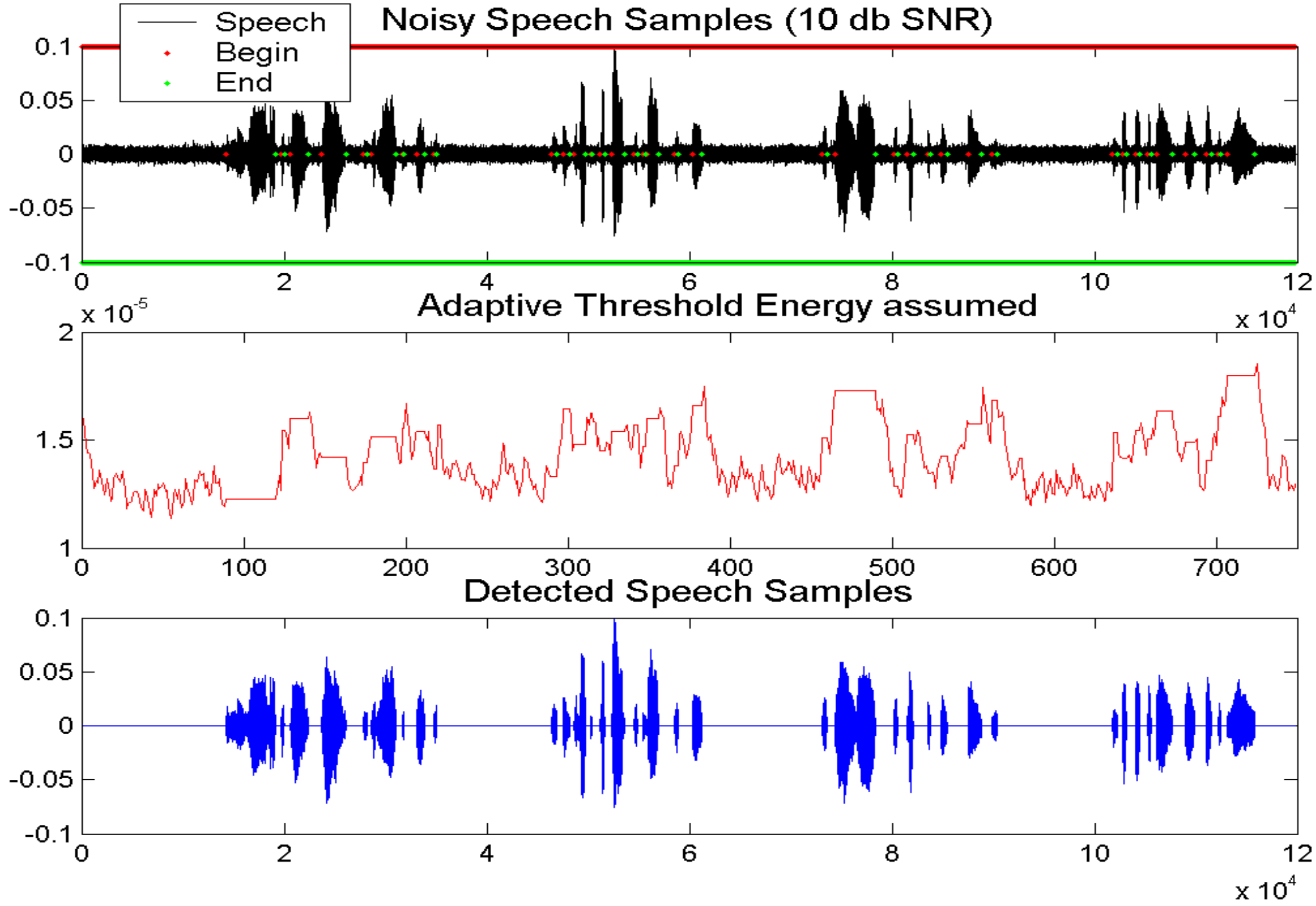
$E_{th}(old)$ is the previous energy threshold and

E_{sil} is the energy of the most recent silent/noisy frame.

Linear Energy-Based Detector (LED).

- Gives acceptable quality of speech after compression.
- However, it fails to detect correctly under varying noise conditions. This is because it cannot adapt to rapidly changing background noise.
- Also, non-plosive phonemes like ‘fish’ and ‘thief’ are clipped completely, as it is purely energy based.

Response of the LED for the TIMIT speech database.



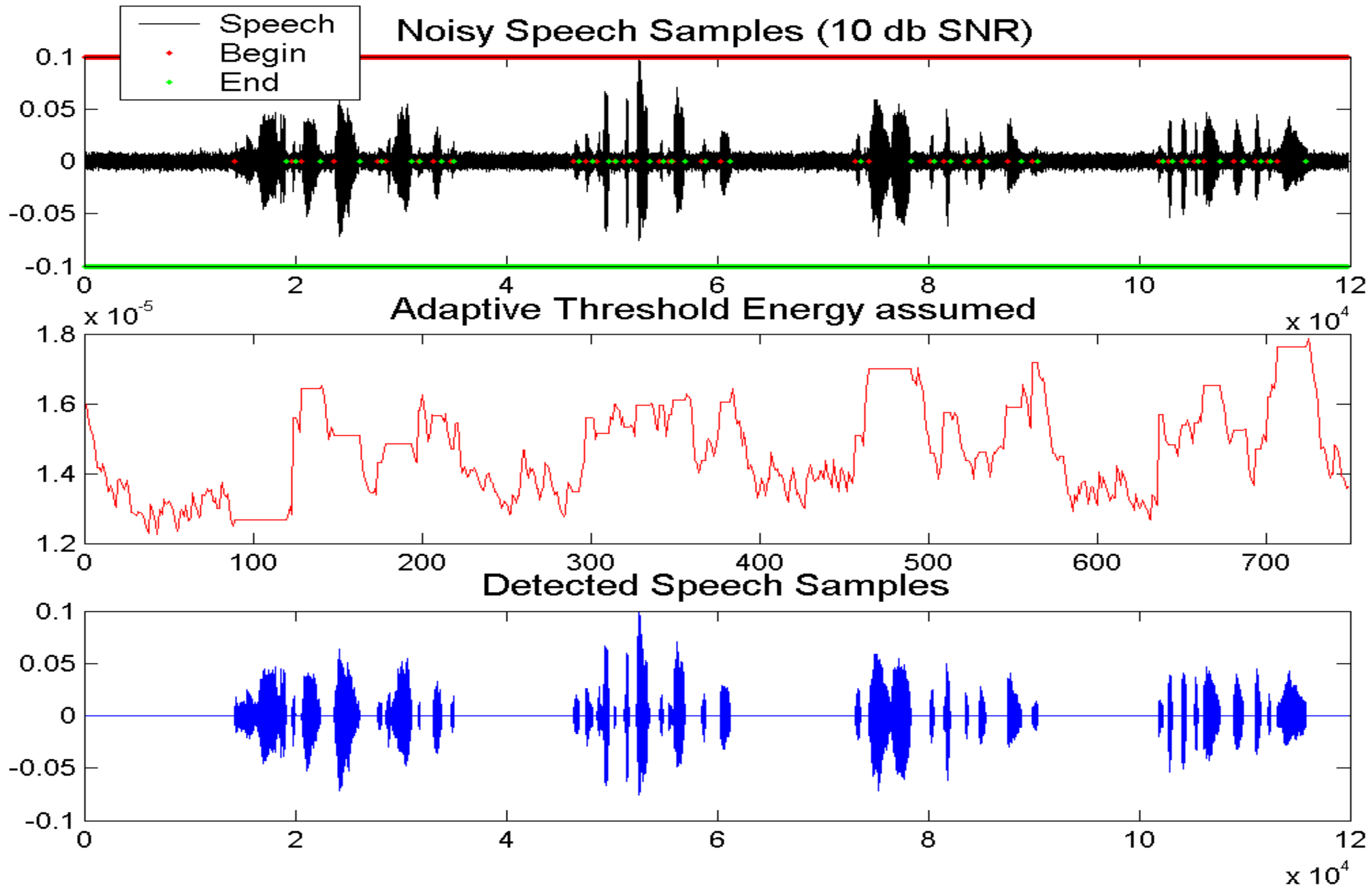
Adaptive Linear Energy-Based Detector (ALED).

- The LED is incapable for adapting to non-stationary noise.
- Hence, further adaptability is required for efficient detection.
- This works exactly on the same principle as the LED except for modifications in the threshold adaptation.
- Can robustly work even in rapidly changing background noise conditions.
- However, low energy phonemes are still found to be clipped, as in LED.

Adaptability to Non-stationary Noise.

Calculated $\frac{\sigma_{new}}{\sigma_{old}}$	p
$\frac{\sigma_{new}}{\sigma_{old}} \geq 1.25$	0.25
$1.25 \geq \frac{\sigma_{new}}{\sigma_{old}} \geq 1.10$	0.20
$1.10 \geq \frac{\sigma_{new}}{\sigma_{old}} \geq 1.00$	0.15
$1.00 \leq \frac{\sigma_{new}}{\sigma_{old}}$	0.10

Response of the ALED for the TIMIT speech database.



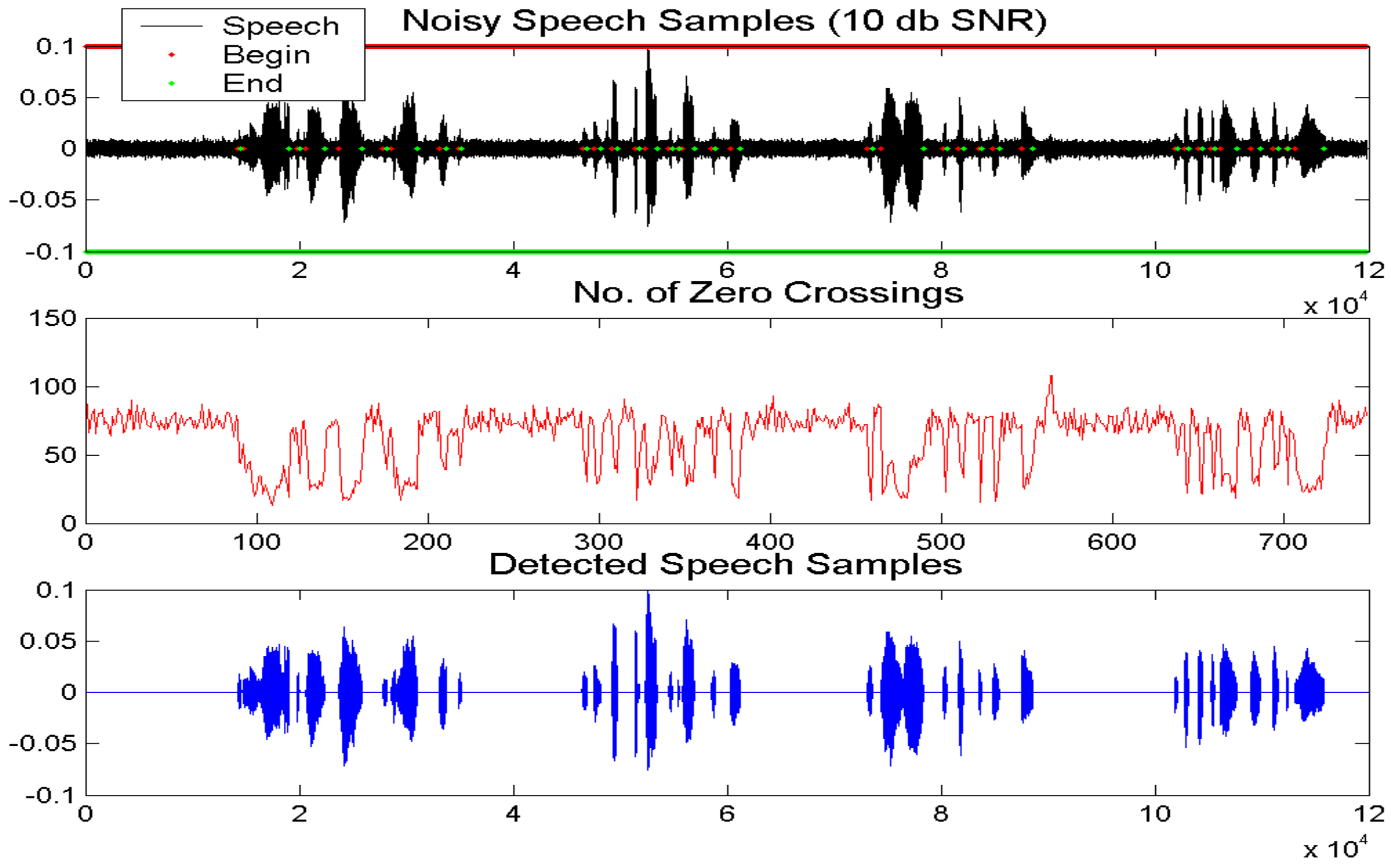
Zero Crossings Detector (ZCD) or The Weak Fricatives Detector.

- The LED and ALED were exclusively energy based. Low SNR caused unnecessary cuts.
- This algorithm is meant to detect even low energy voice segments.
- Feature vector used is the no. of zero crossings in a frame.
- Threshold is completely decided by the zero crossing count.

Zero Crossings Detector (ZCD).

- The no. of Zero crossings for a noisy frame is far higher than that of a speech frame. This is used as the basis for cutoff.
- It successfully detects even low energy phonemes.
- However, it often makes incorrect decisions as speech and noise may have the same no. of zero crossings.
- Also, unvoiced segments are totally cut off.

Response of the ZCD for the TIMIT speech database.



Least Squares Periodicity Estimate (LSPE).

- The LSPE uses a periodicity measure to locate the voiced sections of the speech.
- It works in the time domain.
- Periodicity of speech is far higher than that of noise.
- The threshold vector is “The normalized periodicity estimate”.
- Works reliably even at 5 db SNR.
- The price paid for this is the slight loss of sensitivity in the detected samples.

The Normalized Periodicity Measure.

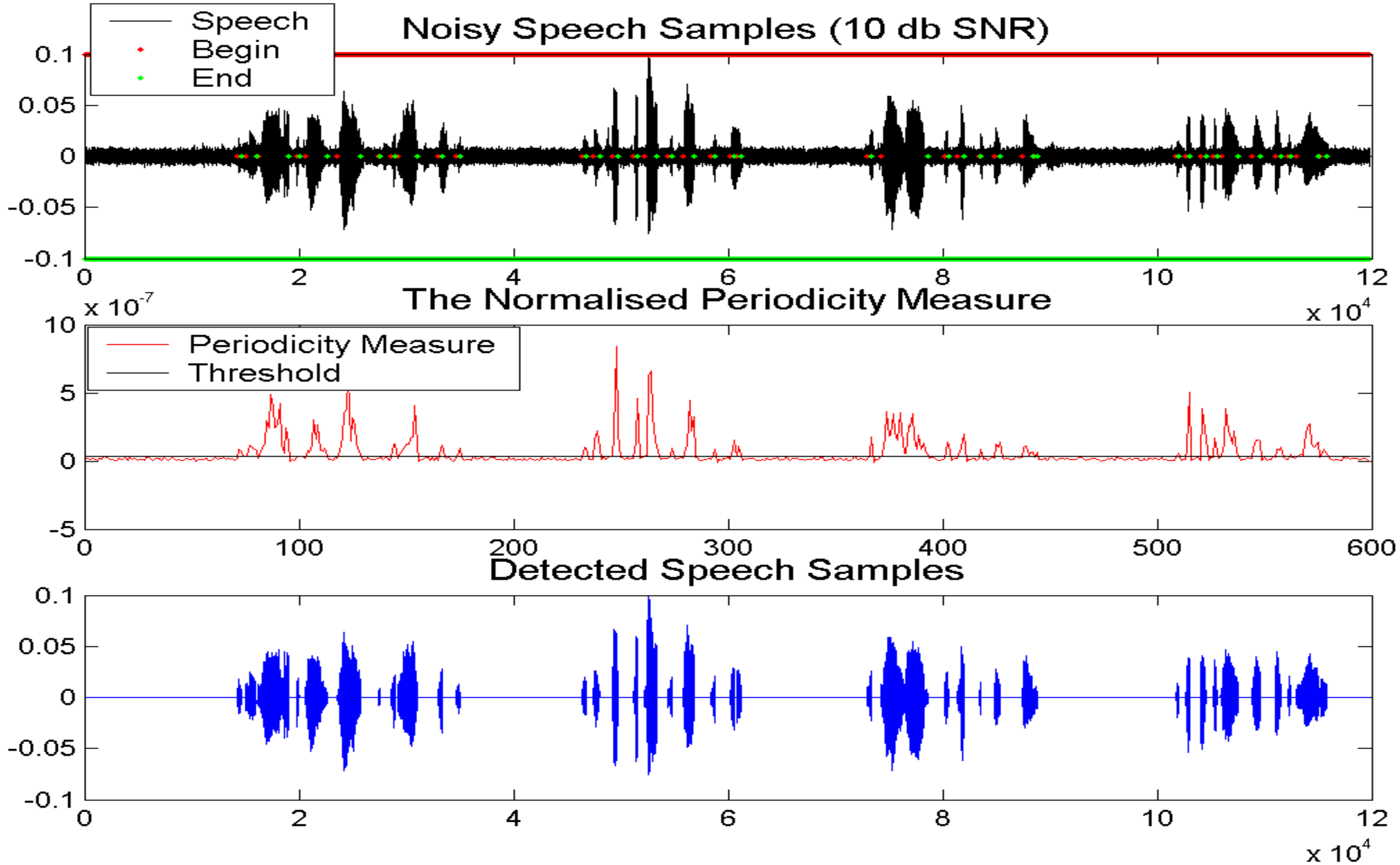
$$R_1(\hat{P}_0) = \frac{I_0(\hat{P}_0) - I_1(\hat{P}_0)}{\sum_{i=1}^{N} s^2(i) - I_1(\hat{P}_0)} \qquad P_{min} \leq \hat{P}_0 \leq P_{max}$$

where, P_{min} and P_{max} are the minimum and maximum number of samples in a pitch period,

$$I_1(\hat{P}_0) = \sum_{i=1}^{\hat{P}_0} \sum_{h=0}^{K_0} \frac{s(i + h\hat{P}_0)^2}{K_0}$$

$$I_0(\hat{P}_0) = \sum_{i=1}^{\hat{P}_0} \frac{[\sum_{h=0}^{K_0} s(i + h\hat{P}_0)]^2}{K_0} \qquad K_0 = [(N - i)/\hat{P}_0] + 1$$

Response of the LSPE-Based VAD for the TIMIT speech



Adaptive Linear Sub-Band Energy Detector (ALSED).

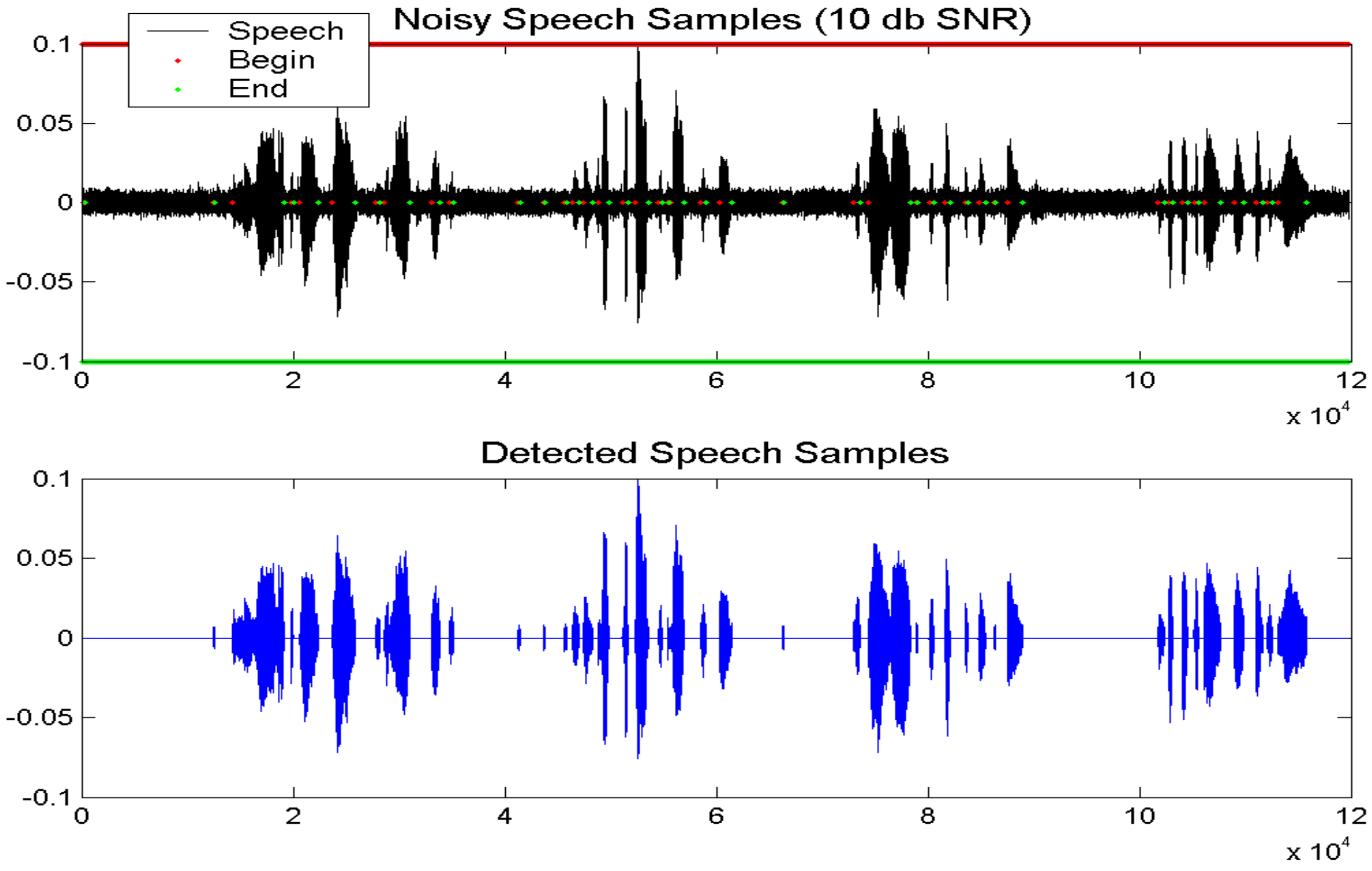
- Speech signal – split into sub-bands.
- Spectral energy is calculated for each – compared with threshold for that band.
- Adaptability as in ALED adopted.
- Selective threshold comparison in the lowest band alone provides good decisions.
- However, this detector performs poorly at low SNR. Also it doesn’t detect low energy phonemes.

Recursive computation of threshold.

$$E_{new(threshold)}(f_n) = (1-p)\cdot E_{(new)}(f_n) + p\cdot E_{old(threshold)}(f_n)$$

$E_{(new)}(f_n)$ is the energy of the most recent silent frame, $E_{old(threshold)}(f_n)$ and $E_{new(threshold)}(f_n)$ are the previous and updated values of the threshold vector for the n^{th} sub-band.

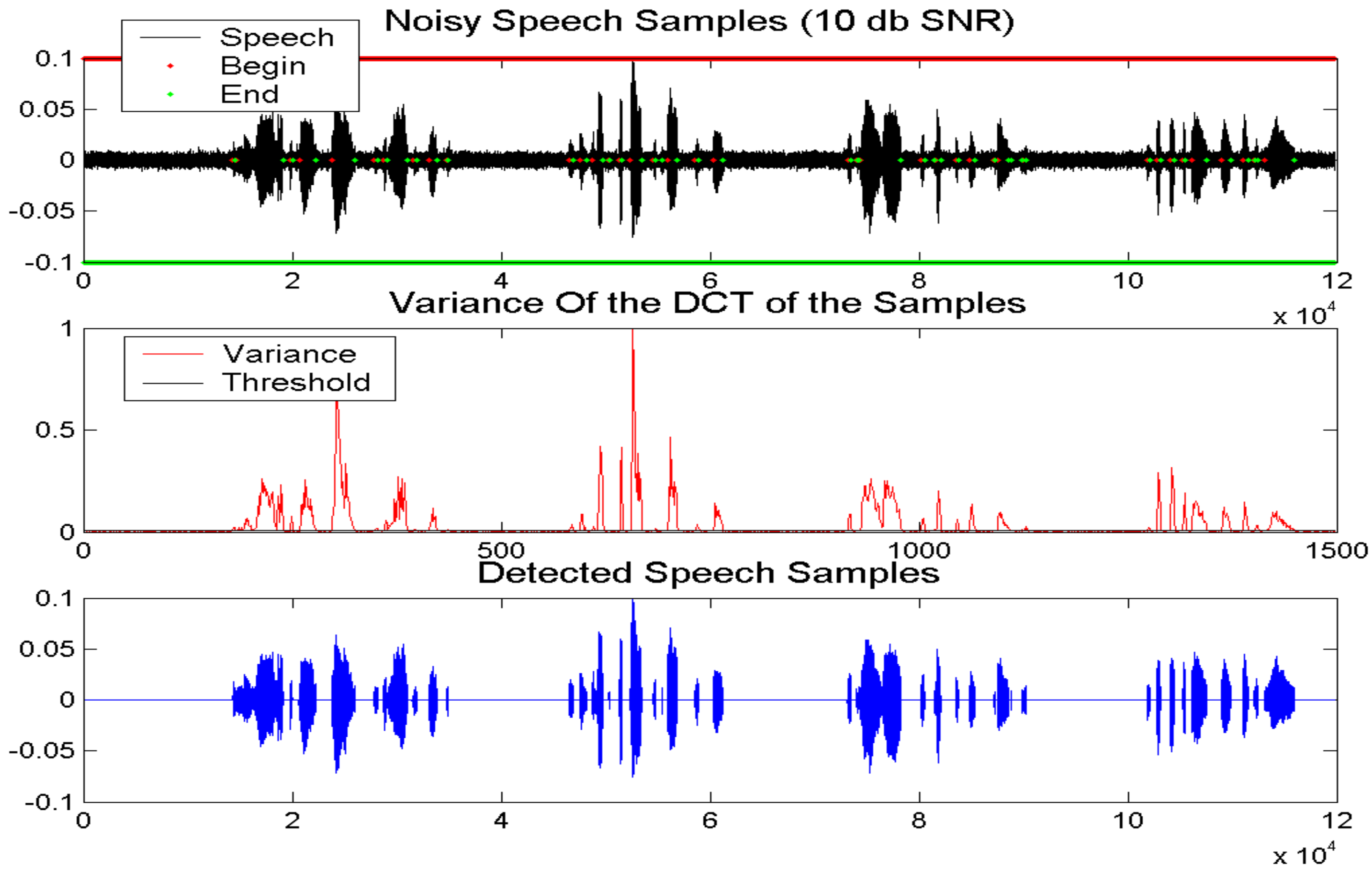
Response of the ALSED for the TIMIT speech database.



Spectral Flatness Detector (SF).

- This is a frequency domain algorithm.
- It is based on the fact that noise has a flat spectrum.
- Threshold vector assumed is the variance of the Fourier transform(DCT) of the signal.
- Works well even in low SNR as it uses a statistical approach to the energy distribution in the spectra.
- However, it requires a large no. of floating point operations.

Response of the SFD to the TIMIT speech database.



Cepstral Detector.

- It is a frequency domain algorithm.
- It makes use of cepstral qualities of speech.
- Cepstrum of speech is defined as IFT(log|FT(speech)|).
- It is suitable for VAD applications as the variance of cepstrum for speech signals is far greater than that for noise.
- Threshold is based on the variance of the cepstral coefficients.

Cepstral Detector

- Cepstral coefficients are derived from the Linear Prediction coefficients.
- This method is advantageous as cepstra can effectively model vocal tracts. Start of voice patches is detected accurately in most cases.
- However, low SNR leads to numerous false detections.
- Also, excessive fricatives in a sentence affect the VAD results.

Conversion of LPC to Cepstral coefficients.

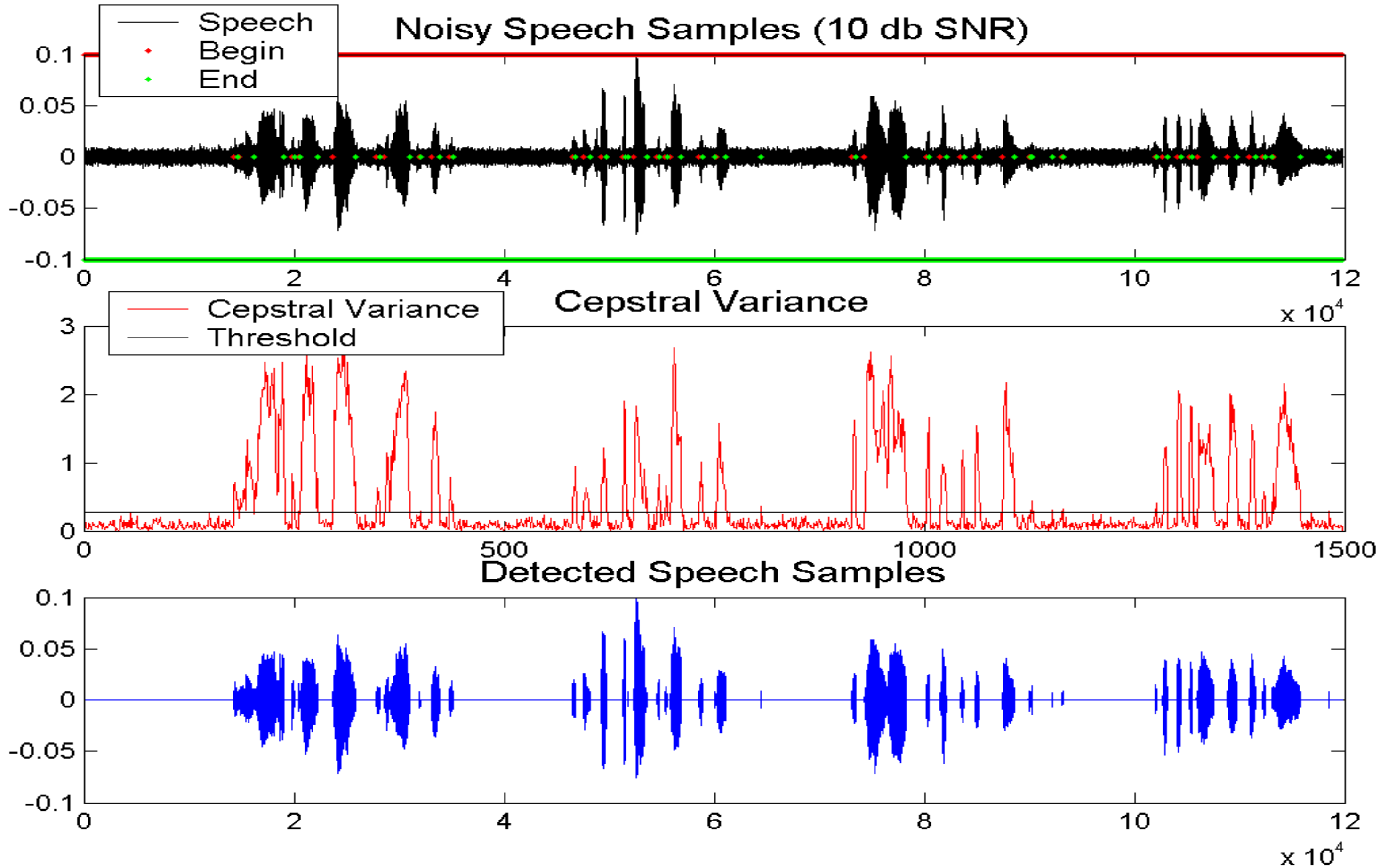
$$c_n = l_n + \sum_{k=1}^{m} (1 - \frac{k}{n}) \cdot l_k \cdot c_{n-k}$$

where,

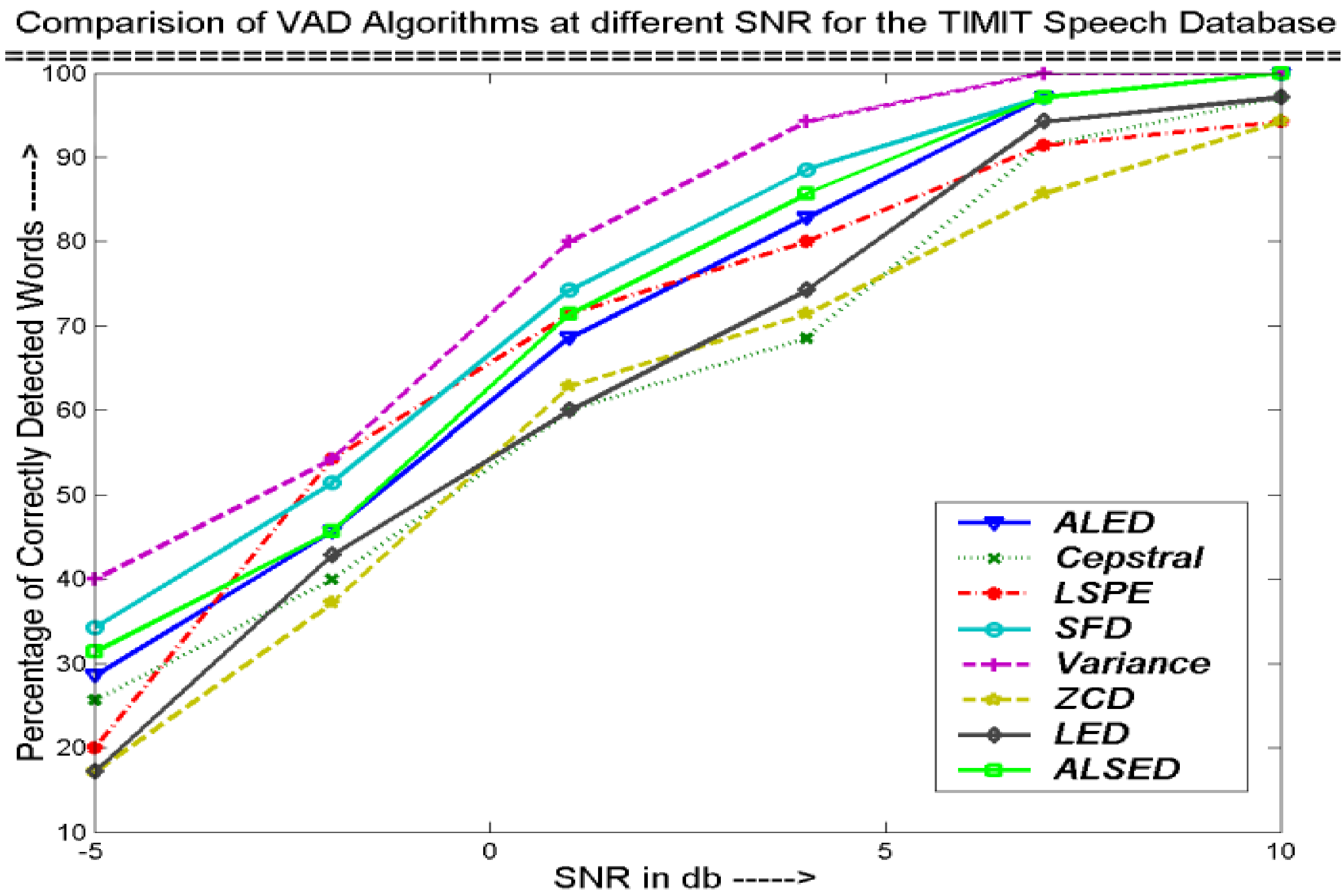
l_n represents the n^{th} LPC and

c_n is the n^{th} Cepstral coefficient.

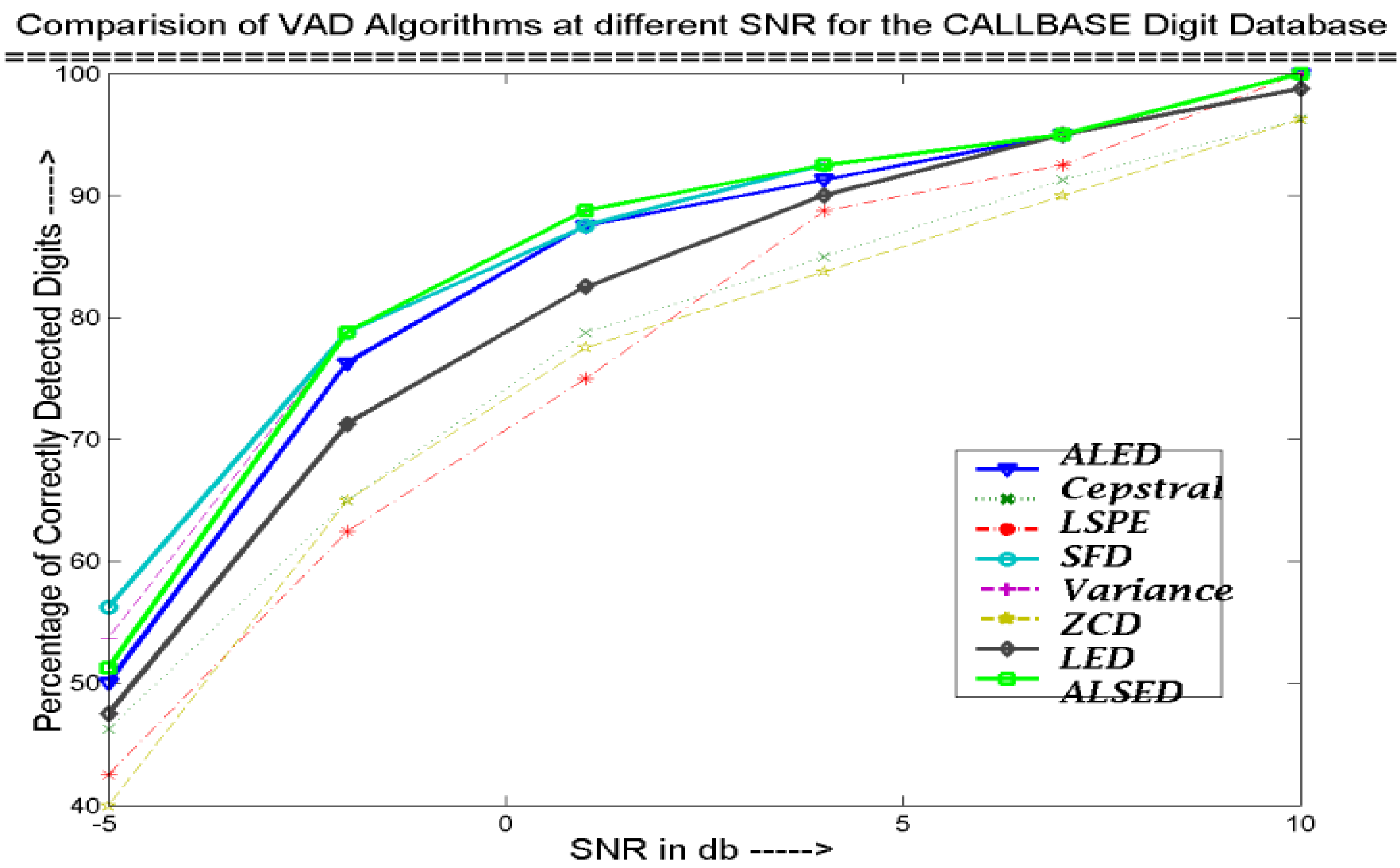
Response of the Cepstral Detector for the TIMIT



Comparison of VADs.



Comparison of VADs.



Other VAD Algorithms.

- VAD working on the fusion of two or more basic VADs.
- VAD based on the Viterbi algorithm.
- VAD using Bayesian adaptation with conjugate normal distributions.
- VAD based on a certain statistical model.
- VAD working on the principle of a hidden Markov model.
- VAD based on the higher-order statistics(HOS) of speech.
- VAD by tracking Power Envelope dynamics.
- VAD based on the perpetual wavelet packet transform.
- VAD working on the principle of CAPDM architecture.
- VAD based on time delay estimation and fuzzy activity classification.

Conclusions and Results.

- Almost all algorithms worked well even at 10 db SNR. As SNR was reduced, performance deteriorated.
- Variance-based Detector and SFD performed outstandingly. Variance – the best at low SNR.
- Performance – depended on word length.
- Threshold value – very critical.

Introduction to Speech Enhancement.

Requirements of Speech Enhancement.

- Remove background noise.
- Improve speech quality and intelligibility.
- Suppress undesired interference.

Two Speech Enhancement Algorithms.

- Using MMSE-LSAE.
- Using Adaptive Wavelet Packet.

Using MMSE-LSAE.

$$y(t) = x(t) + n(t),\ 0 \le t \le T.$$

$$X_k = A_k e^{j\alpha_k} \qquad\qquad Y_k = R_k e^{j\vartheta_k}$$

$$\hat{A}_k = exp[E(\ln A_k | y(t))], 0 \le t \le T$$

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k} exp[\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t}] \cdot R_k$$

Using MMSE-LSAE.

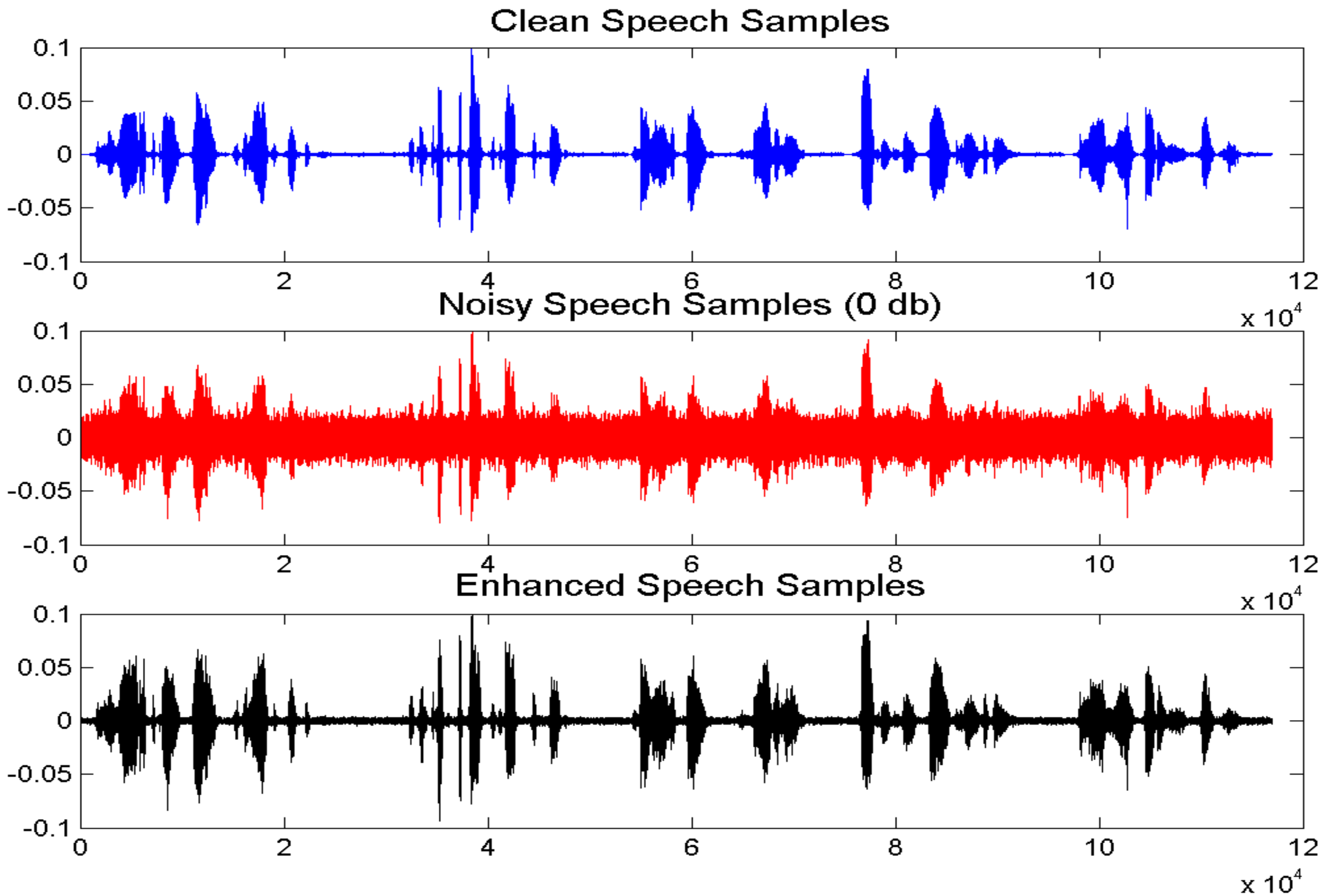
ξ_k is the priori SNR,

$$v_k = \frac{\xi_k}{1+\xi_k}\lambda_k$$

$$\frac{1}{\lambda_k} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_n(k)}$$

$\lambda_x(k)$ and $\lambda_n(k)$, denoting the variances of the k^{th} spectral component of the speech signal and noise.

Performance Evaluation.



Using Adaptive Wavelet Packet.

Noise Estimation based on Spectral Entropy using Histogram of Intensity

1. Estimate spectral pdf through histogram of wavelet packet coefficients for each node. Histogram is composed of B bins.
2. Calculate the normalized spectral entropy.

$$Entropy(n) = -\sum_{b=1}^{B} P \cdot log_B(P)$$

with,

n = 1,2,...No. of best nodes

$$P = \frac{No.\ of\ Wavelet\ Packet\ Coefficients\ c_k\ in\ bin\ b\ and\ node\ k}{Node\ size\ in\ adapted\ wavelet\ packet\ tree}$$

Using Adaptive Wavelet Packet.

3. Estimate *spectral magnitude intensity by histogram* and standard deviation of noise for node dependent wavelet thresholding.

4. Define an auxiliary threshold α.

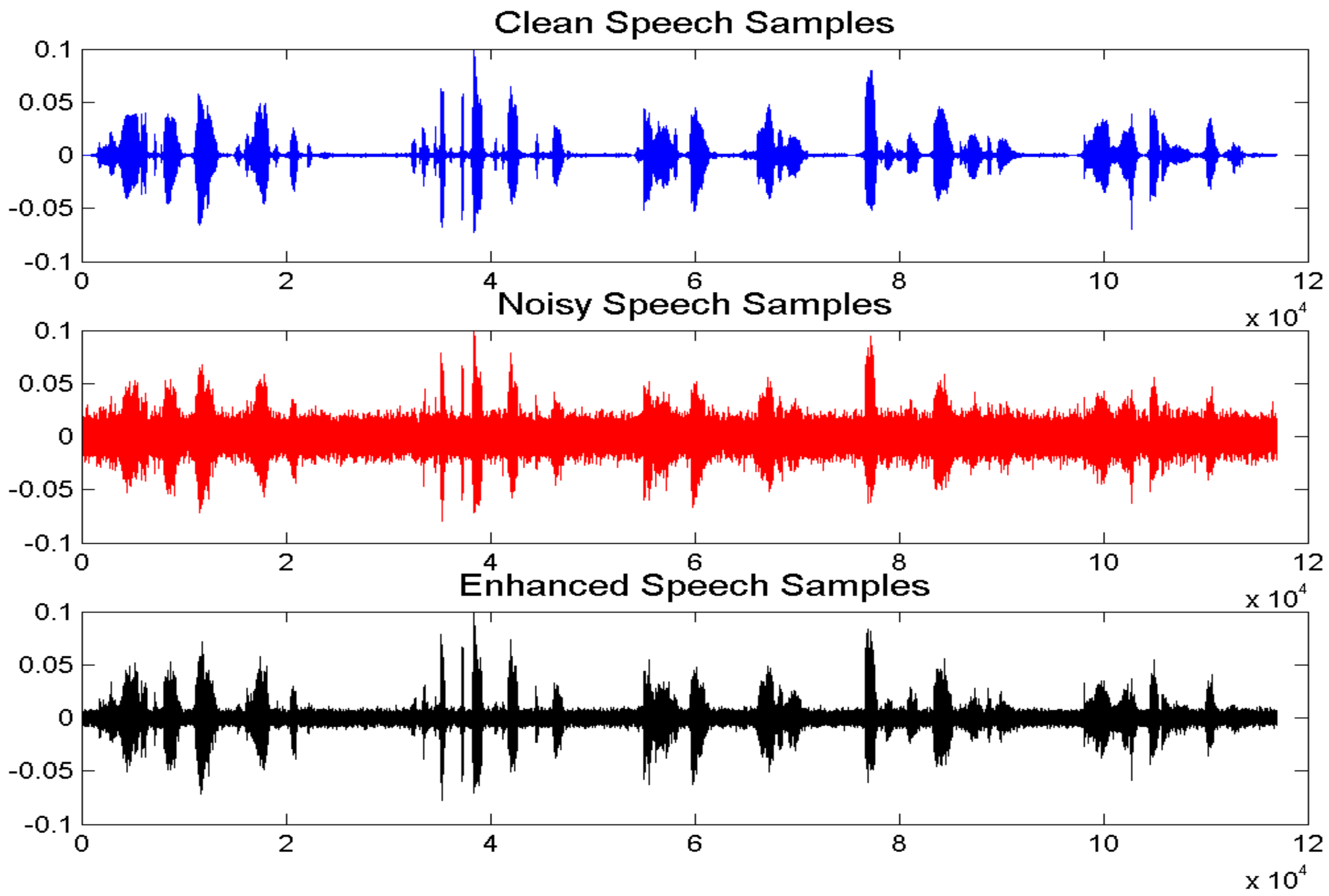
$$\alpha(n) = Entropy(n) \cdot (node\ size) \cdot \beta$$

where the range of β is from 0.7 to 0.9. It is usually taken as 0.8.

$$\hat{\sigma}_k = [No.\ of\ bins\ in\ node\ k\ bigger\ than\ \alpha(n)] \cdot bin\ width \qquad T = \hat{\sigma}_k \sqrt{2\log(N\log_2(N))}$$

$$\begin{aligned} THR_{\mu-law}(X,T) &= X,\ |X| > T, \\ &\quad T \cdot (\frac{1}{\mu}[(1+\mu^{|\frac{X}{T}|}) - 1] \cdot sqn(x)),\ |X| \le T. \end{aligned}$$

Performance Evaluation.



Sample Sounds

- Clean Speech Samples
- Noisy Speech Samples
- Output from the Variance VAD
- Enhanced Speech(MMSELSAE)

Thank You

One and All.