End Points Detection Widget

Amr M. Gody*
Fayoum University

Abstract

The problem of speech modeling is handled here using a new technique. The features are extracted from wavelet packets decomposition of speech signal. Wavelet packet bands are chosen to match the natural hearing response. Widgets are fixed probabilistic patterns that may be utilized to monitor specific physical phenomena. Many widgets may be designed to express some phenomena like speech end points, voiced/unvoiced and phone boundaries. In this paper the point chosen is end points widget. Signal with almost S/N = 0 (db) are still discriminate using EPD widget.

1. Introduction

Wavelet packets indicate high efficiency to express the non-homogeneous information along frequency spectrum of speech signal. It gives high flexibility to choose natural bands for feature extraction [1]. Figure 1 represents the heart of wavelet packets mechanism.

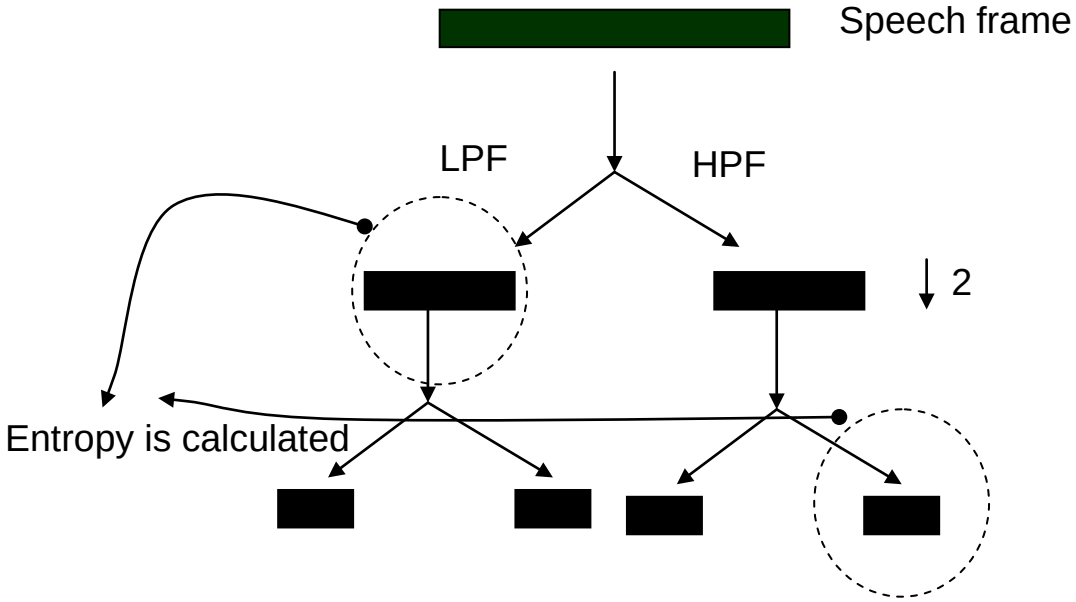


Figure 1 Wavelet packets filters structure

The speech signal is exposed to the Wavelet packets filters. Symmetric filters in all levels are always splitting the band into two equal half bands. A symmetric tree like that one in figure 1 is created. This tree gives us the chance to be more flexible in choosing the analysis bands than the dyadic wavelets that hide the high frequency bands.

* Department of Electrical Engineering, Faculty of Engineering, Fayoum University, El-fayoum , EGYPT., Email: amr.m.gody@gmail.com.

In the research point introduced by Hai Jiang [1], the idea of band selection is illustrated. Her choosed to try fixing the bandwidths in Mel scale. For the Mel scale, the edge bandwidths were calculated based on

$$M = \frac{1000}{\log 2}\log(1+\frac{f}{1000})\,(\text{mels}) \qquad \textbf{(1)}$$

where M is the bandwidth in (mels). Equation 1 relates the frequency to the mel scale. The mel scale, proposed by Stevens, Volkman and Newman in 1937 [2] is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 (mels). Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments [2].

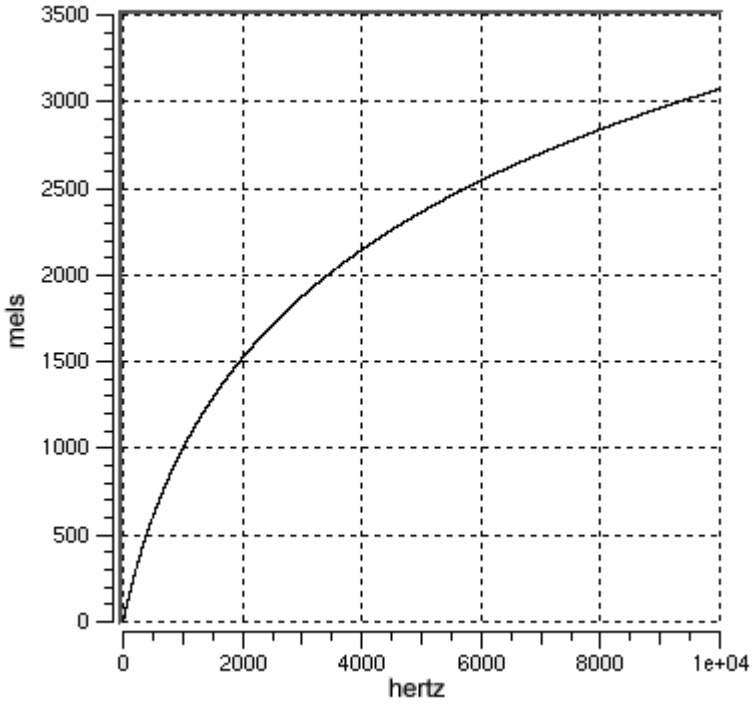


Figure 2 Mel-frequency curve

As shown in figure 2, mel and frequency relation is a logarithmic nature. Human hearing responses have almost a logarithmic nature relation with the frequency. We do not have to handle all frequency bands in the same way as it is not already handled in such way in human hearing mechanism. This is the point; we can try to select analysis bands that coincide as much as possible with equivalent mel bands. In other words, we will try to capture information in a way that is implemented in human hearing mechanism.

Entropy is chosen as a feature in each selected band. The entropy is a randomness measure. As long as the signal is random, it is rich of information. One cannot predict the next sample from the previous one, so the signal is highly informative. Each sample conveys new information. In this case we will have high

entropy. In contrast, in case of periodic signals, one can predict the future samples from the history. The entropy is measured by [3]:

$$I = -\sum_{i=1}^{n} P_i \log(P_i) \qquad \textbf{(2)}$$

where P_i is the probability of sample X_i in the complete set of samples X. Equation 2 indicates that samples with high probability will share by a small term in the summation. Samples with high probability are more predictable and indicate a periodical feature in data samples. Entropy will be positive as long as P is always positive and less than 1.

2. Features extraction process

Wavelet packets analysis (WPA) is a very powerful tool in spectrum analysis. The power of this tool comes from resolution controllability. The signal may be analyzed in controlled bandwidth packets. In addition, the bands of analysis are selective.

In speech signal, WPA is utilized in this paper in order to get a maximum coincidence with human hearing mechanism. For a signal sampled at 32 kHz the optimal level of wavelet packets decomposition is 7 to obtain a bandwidth resolution of about 100 (mel) as illustrated in table 1.

Table 1 Part of Level 7 decomposition of wavelet packets analysis applied on a speech signal sampled in 32KHz. Filter bandwidth is 125 Hz. The table shows 11 filters out of 128 total filters available at this level.

Filter index	Corner frequencies (HZ)	Corner frequencies (mel)	Band Width (mels)
1	125	169.925	169.925
2	250	321.9281	152.0031
3	375	459.4316	137.5035
4	500	584.9625	125.5309
5	625	700.4397	115.4772
6	750	807.3549	106.9152
7	875	906.8906	99.53567
8	1000	1000	93.1094
9	1125	1087.463	87.46284
10	1250	1169.925	82.46216
11	1375	1247.928	78.00251

As shown in table 1, bandwidths in (mels) are not equal along the whole filter banks as those of frequency bandwidths. If we try to figure out the relation between l bandwidth in (mels) and progressive filter indices we will find a logarithmic nature. Figure 3 emphasis this relationship.

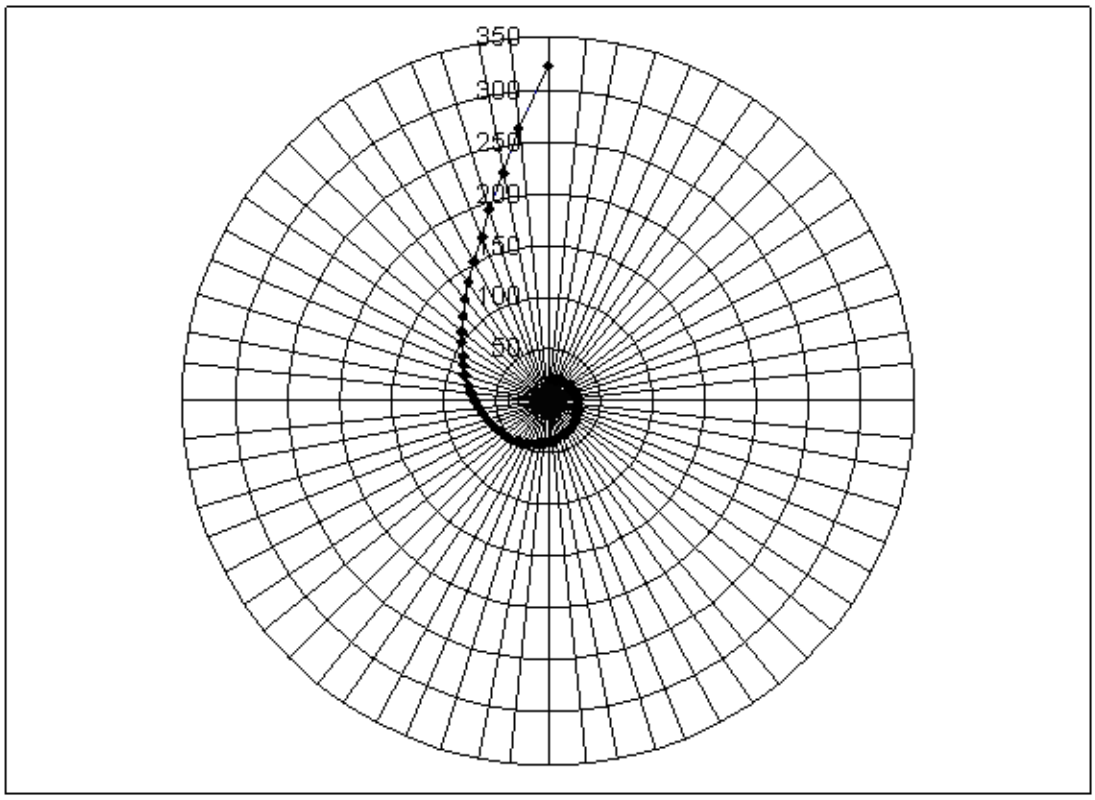


Figure 3 Ray graph representation of level 7 decomposition. Progressive ray lines indicate band index. The value along the ray indicates mel bandwidth.

The key point for feature extraction is to get information from available bandwidths in mel scale for a suitable number of bandwidths not less than 90 (mels). Returning back to table 1, filters from 1 to 8 may be considered as information sources. This will not cover the suitable information frequency band of 5000 Hz. So we have to get the rest of information from another level. Table 2 indicates a part of level 6 decomposition.

Table 2 Part of Level 6 decomposition of wavelet packets analysis applied on a speech signal sampled in 32KHz. Filter bandwidth is 250 Hz. The table shows 12 filters out of 64 total filters available at this level

Filter index	Corner frequencies (HZ)	Corner frequencies (Mel)	Mel BW (mels)
1	250	321.9281	321.9281
2	500	584.9625	263.0344
3	750	807.3549	222.3924
4	1000	1000	192.6451
5	1250	1169.925	169.925
6	1500	1321.928	152.0031
7	1750	1459.432	137.5035
8	2000	1584.963	125.5309
9	2250	1700.44	115.4772
10	2500	1807.355	106.9152
11	2750	1906.891	99.53567
12	3000	2000	93.1094

As shown in table 2 the bands from 5 to 12 may be included. Bands below 5 are already considered in level 7 decomposition. Still we need more extra information to cover the whole 5000(Hz), so we may continue to level 5.

Table 3 Part of Level 5 decomposition of wavelet packets analysis applied on a speech signal sampled in 32KHz. Filter bandwidth is 500 Hz. The table shows 12 filters out of 32 total filters available at this level

Filter index	Corner frequencies (HZ)	Corner frequencies (Mel)	Mel BW (mels)
1	500	584.9625	584.9625
2	1000	1000	415.0375
3	1500	1321.928	321.9281
4	2000	1584.963	263.0344
5	2500	1807.355	222.3924
6	3000	2000	192.6451
7	3500	2169.925	169.925
8	4000	2321.928	152.0031
9	4500	2459.432	137.5035
10	5000	2584.963	125.5309
11	5500	2700.44	115.4772
12	6000	2807.355	106.9152
13	6500	2906.891	99.53567
14	7000	3000	93.1094

Table 3 indicates that we may consider filters 7, 8, 9 and 10 to complete the needed coverage. Now, we have 8 filters in level 7, 8 filters in level 6 and 4 filters in level 5 covering all frequency band needed to make a good analysis for human speech signal. Let us put all together in one table to see what we have.

Table 4 indicates the good choice of the selected bands. We can easily notice that although the selected bands are different in band widths in Hertz scale they are almost fluctuating around 100 (mels). Drawing progressive Feature index – mel bandwidth relationship may give us more illustration of what we have. Figure 4 indicates the logarithmic nature in mel bandwidths over the progressive feature index within the feature vector of 20 components as indicated in table 4.

Now we are ready to extract some information from the selected bands. Entropy is selected as information measurer. It is a measure of randomness in the signal. As was indicated in the introduction, high entropy indicates high randomness in the signal. Randomness indicates that the signal has not stationary property along the analysis period of time. In case of non speech periods, we may expect a very low power signal with almost a stable low energy. This may lead to very low entropy. Also we may except a high power white noise, which indicate high entropy. In both cases the entropy is accepted to be homogeneous in all bands. In other words, all features inside a single vector that represent a non speech are almost fluctuating around a very narrow value. This is the key of this widget. In contrast of this case, the speech period is assumed to have verities of entropy along the available bands. In the next section this will be illustrated by a descriptive graph.

Table 4 Selected bands properties.

Index	Frequency band (Hz)	Bandwidth (Hz)	Bandwidth (mel)
1	0 – 125	125	169.925
2	125 -250	125	152.0031
3	250 – 375	125	137.5035
4	375 – 500	125	125.5309
5	500 – 625	125	115.4772
6	625 – 750	125	106.9152
7	750 – 875	125	99.53567
8	875 – 1000	125	93.1094
9	1000-1250	250	169.925
10	1250-1500	250	152.0031
11	1500-1750	250	137.5035
12	1750-2000	250	125.5309
13	2000-2250	250	115.4772
14	2250-2500	250	106.9152
15	2500-2750	250	99.53567
16	2750-3000	250	93.1094
17	3000-3500	500	169.925
18	3500-4000	500	152.0031
19	4000-4500	500	137.5035
20	4500-5000	500	125.5309

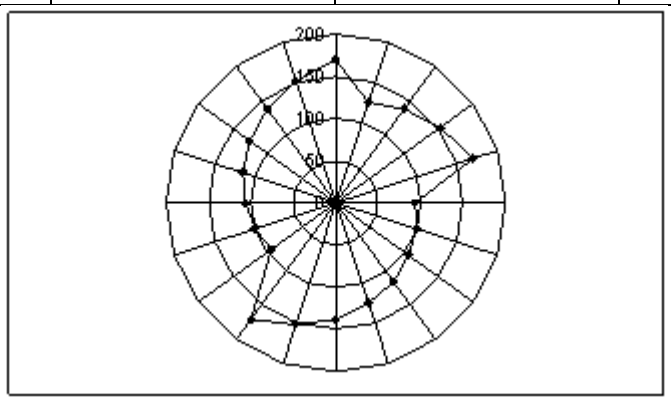


Figure 4 Progressive selected bands with associated mel bandwidth are illustrated in this figure. Ray lines represent the progressive index of selected bands. Value along the ray line represents mel band width associated with the line.

3. Non speech modeling

The way to model non speech signal using the supposed features is simply getting common features in such phenomena. A descriptive speech signal with long period of non speech is utilized in this process. Features are extracted from the non speech frames. A probability model is constructed as shown in figure 5.

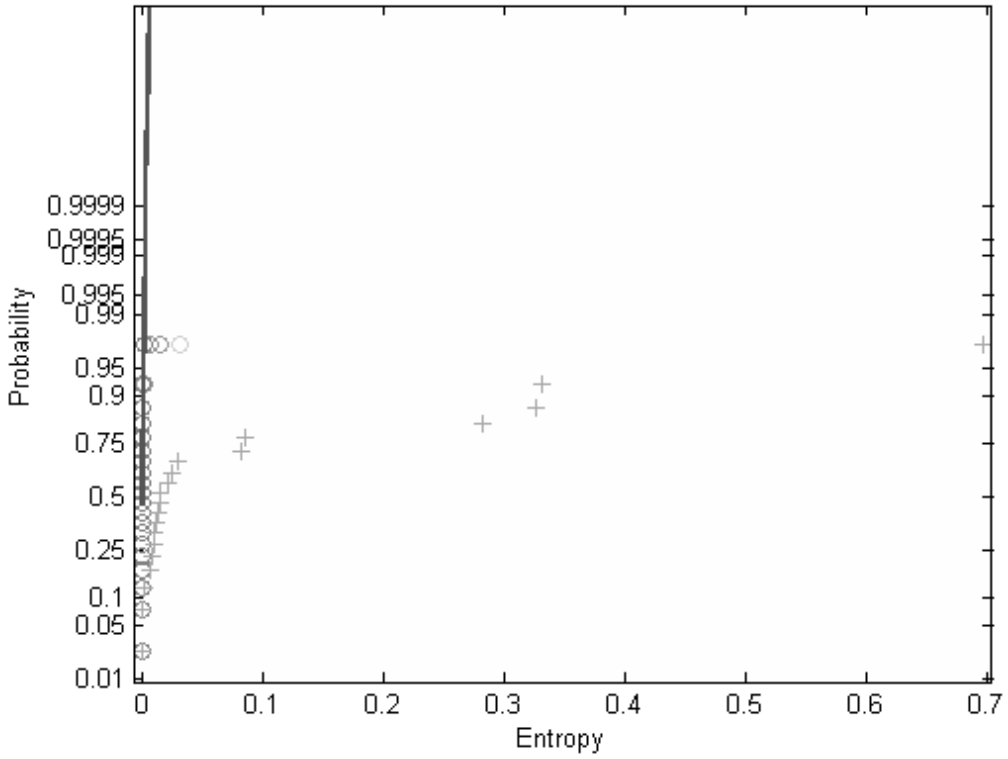


Figure 5 Exponential cumulative probability function that fit the probability points of the non speech features. Points in cross represents speech feature's.

Many feature vectors that represent the non speech periods are collected to calculate the cumulative probability. Figure 5 provides a fitting function of the non speech feature vectors (Circle points). For a comparison a speech features vector (Cross points) is inserted in the same graph. As shown in the figure, the cross points are widely spread over the entropy values. This is very logical result. The speech data will have some features with low information and others with high information. This is the key point to model the non speech data which have no such Entropy variations. To identify unknown frame, the associated features vector will be introduced to the fitting function. The summation of the emission probabilities will be highly discriminating between speech frame and non speech frame as will be explained in the next section. The successive non speech frames will give always almost a constant value fluctuating in a very narrow range. The successive speech frames will give very random values that construct a curve that may have rapid variations. The fitting function is chosen to be exponential fitting function because all entropy values are positive.

$$F = 1 - e^{\frac{x}{\mu}} \qquad \textbf{(3)}$$

Equation 3 represents the exponential cumulative distribution function utilized to represent the non speech feature vector's cumulative probability. Applying some training data gives a value for $\mu = 0.0016$.

The word widget is widely utilized now to express something do an integrated simple process. This term is widely used in the field of information technology. The large information technology manufacturers like Microsoft, yahoo and opera are using this term to express simple products that do an integrated function. For more details you may reference to the following references.
The term widget is utilized here to express a black box fast operation that senses single physical phenomena. The physical phenomenon here is End Points Detection EPD of speech signal. There are many other physical phenomena that may be expressed using other widgets.

4. End Points Detection (EPD) Widget.

As shown in figure 5, there is a great discrimination between speech and non speech feature vectors. The idea is, if all Feature vectors are applied to the EPD widget which is trained with non speech feature vectors, hence the speech vectors will be highly scored.

$$F = \sum_{i=1}^{20} 1 - e^{\frac{V_i}{0.0016}} \qquad \textbf{(4)}$$

F is the indicator function. All frames are applied to this function to construct the indicator curve. V_i is the feature component number i; where i ranges from 1 to 20 which representing the 20 feature components. Equation 4 represents the widget core function. This EPD widget responds to the applied speech frames according to equation 4. The widget has only one parameter which is μ. $\mu = 0.0016$ in our example according to the training process. Connecting the output points will construct a curve as shown in figure 6. Figure 6 indicates the indicator curve and the original speech signal. As shown in figure 6, speech segments have a large discrimination over the non speech segments.

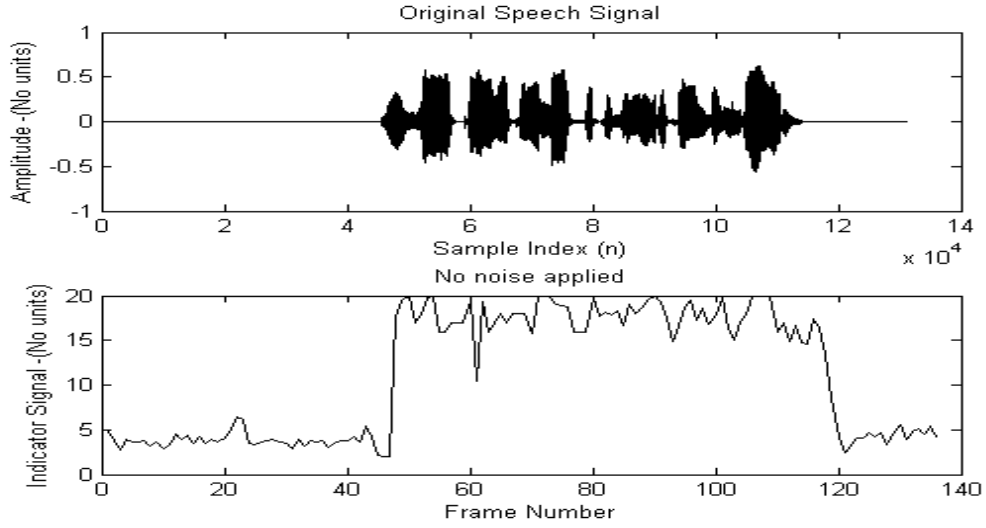
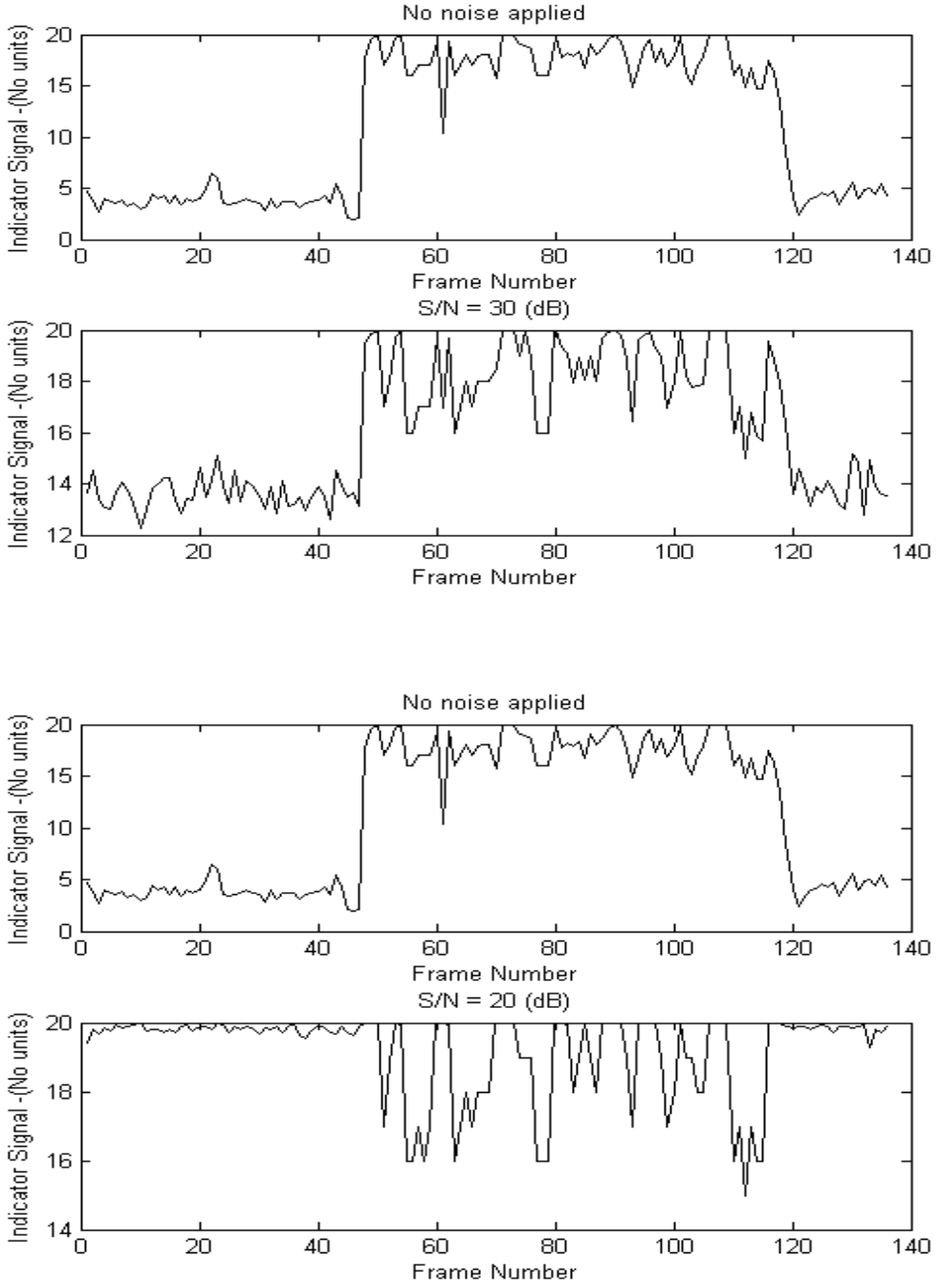


Figure 6 ndicator function that represents the discrimination between the speech segments and the non- speech segments.

5. Signal to Noise ratio Effect

To make a test of robustness of the indicator function, the signal is exposed to white Gaussian noise with different signal to noise ratios. The following figures represent the effects.



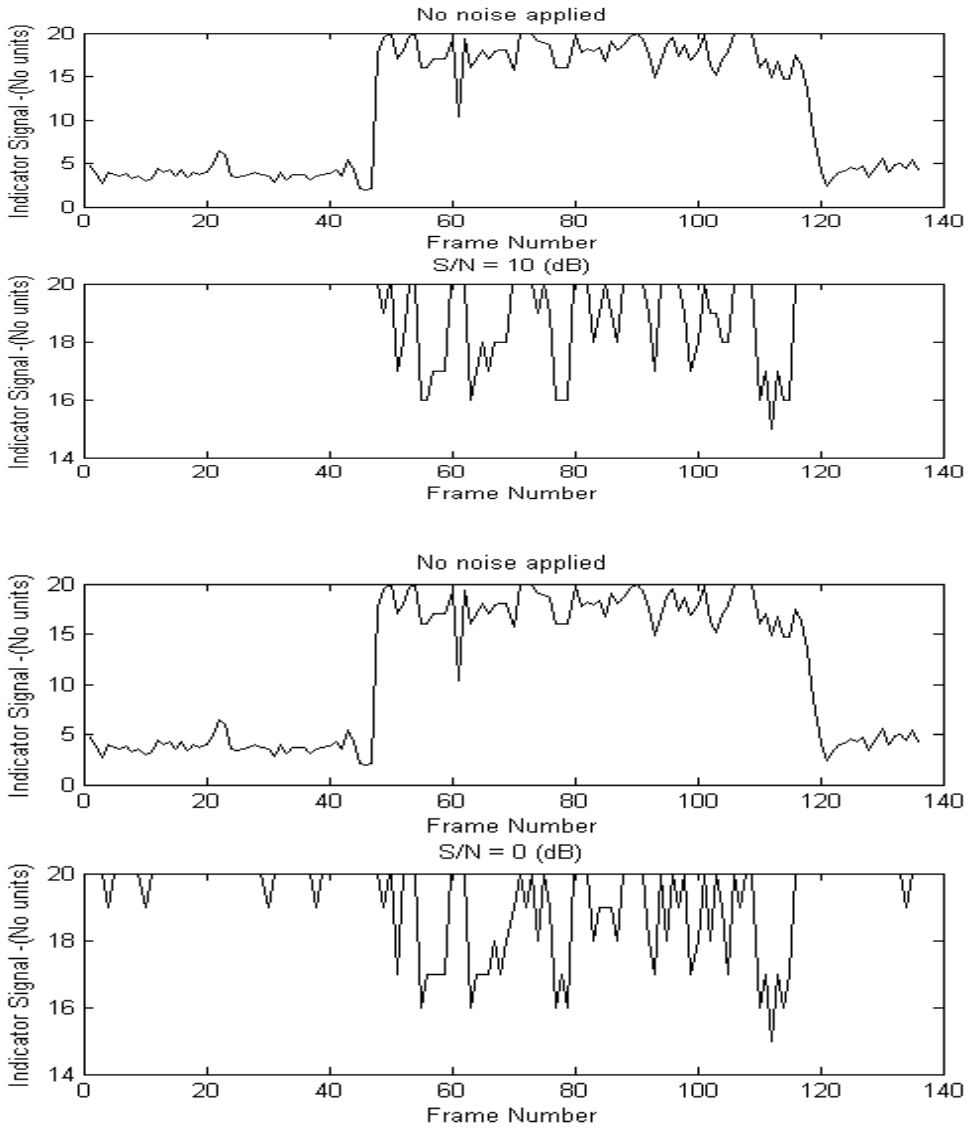


Figure 7 Speech Signal Indicator function. Comparison for different Signal to noise ratios.

Figure 7 indicates a good discrimination even in the very low signal to noise ratio. There is a discrimination that we can distinguish by eyes. As explained in section 4, the entropy variation is almost minimal in the non speech successive periods. In all diagrams in figure 7, it is highly appeared that the curve is almost stable in the non speech periods even in the very low signal to noise ratio. This is the key point that may be utilized to formulate a mathematical model to figure out the not stationary successive periods in the curve. The non stationary successive period's reprints entropy variations which indicate information streaming of speech data.

6. Conclusions

Wavelet packets analysis is utilized in this research to extract features that are much similar to human hearing mechanism. The 2o features vector is used to figure out some physical phenomena that have common characteristics. The term widget is utilized here to represent a fast algorithm that deal with identifying a single physical phenomena. The non speech periods are highly discriminate using End Point Detection EPD widget based on the proposed features. The widget itself is a

probabilistic model that is trained to discriminate the non speech periods. This widget indicates robustness in a very bad signal to noise ratio that may reach to 0 (dB).

7. References

[1] Hai Jiang, Meng Joo Er and Yang Gao ," Feature Extraction Using Wavelet Packets Strategy", Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, Hawaii USA, December 2003

[2] http://en.wikipedia.org/wiki/Mel_scale .

[3] http://en.wikipedia.org/wiki/Information_entropy.

[4] Amr M. Gody, “Natural Hearing Model Based On Dyadic Wavelet”, The Third Conference on Language Engineering CLE’2002, Page(s): 37-43,October 2002