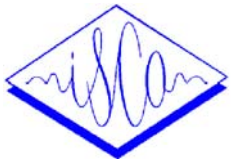

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

End points detection for noisy speech using a wavelet based algorithm

Amin M. Nassar[1], Nemat S. Abdel Kader[2], Amr M. Refat[3]

1. Professor in Electronics and Communication Dept., Faculty of Engineering, Cairo University.
2. Lecturer in Electronics and Communication Dept., Faculty of Engineering, Cairo University.
3. T.A. in Electrical Engineering Dept., Faculty of Engineering, Cairo University, Fayoum Branch. Email: agody@ieee.org

10.21437/Eurospeech.1999-220

ABSTRACT

This paper represents a way for the detection of start-end boundaries of a speech segment in the presence of noise using wavelet transform. The technique is based on generating a certain mathematical function derived from the wavelet parameters that can keep track with the energy changes along the speech duration. The problem in end points detection always appears in words that begin or end in low-energy phonemes. This problem can be entirely eliminated by this technique. Examples are given to show how the algorithm performs under different signal-to-noise ratios.

1. INTRODUCTION

The problem of locating the beginning and end of a speech utterance in background of noise is of importance in many areas of speech processing. In particular, in automatic recognition of isolated words, it is essential to locate the regions of a speech signal that correspond to each word. A scheme for locating the beginning and end of a speech signal can be used to eliminate significant computation in non real-time systems by making it possible to process only the parts of the input that correspond to speech.

For high signal-to-noise ratio environments, the energy of the lowest level speech sounds (e.g. , weak fricatives) exceeds the background noise energy , and thus a simple energy measurement suffices . However such ideal recording conditions are not practical for most applications. The wavelet transform is one of the powerful transforms that are used in the signal processing fields [1]-[4]. The wavelet transform extracts the frequency contents of the signal similar to the Fourier transform but it relates the frequency domain with the time domain [5]. This link between the time and the frequency gives this transform powerful characteristic for the determination of the boundaries of a frequency-band-defined signals such as the speech signals. The wavelet parameters indicate an appropriate mapping for the power distribution of the speech signal along the analysis time period. In this case a radical change in the waveform energy between the background noise and the speech is the cue to locate the boundaries of the segment. A mathematical form derived from the wavelet parameters is used to track the energy changes along the speech duration.

2. THE WAVELET TRANSFORM

The wavelet is a small wave from which many other waves are derived from it by translation and dilation of the wavelet wave.

It can be defined as:

$$W_{ij} = W(2^{i} \cdot t - j \cdot t) \qquad (1)$$

Where:

W_{ij} is the wavelet function obtained by shifting the main wavelet function by j samples and c
by a factor of 2^{i}. The compression in time gives expansion in frequencey. From the previous point of view the index I indecates the frequencey level of the wavelet function.

The time waveform can be expresed in termes of wavelet functions and wavelet coefecients according to the following equation.

$$f(t) = \sum_{i=0}^{m} \sum_{j=0}^{2^{i}} b_{ij} \cdot W(2^{i} \cdot t - j \cdot t) \qquad (2)$$

b_{ij}: The wavelet coefficient at frequency level i and time index j. It is given by:

$$b_{ij} = \int_{0}^{T} f(t) \cdot W_{ij} dt \qquad (3)$$

T: The frame duration.

All wavelets must be orthogonal. The first index makes a dilation of the original wavelet. It gives the indication of the period of the wavelet function so that it conveys information about certain frequency band of the signal. As an example, if the duration of a signal is reduced in the time domain by half then it will expand in the frequency domain by a factor of 2.

Equation 2 can be rearranged as [1]:

$$f(t) = \sum_{j=0}^{1} b_{0j} \cdot W_{0j} + \cdots + \sum_{j=0}^{2^{m}} b_{mj} \cdot W_{mj} \qquad (4)$$

Each summation represents the signal over the whole period in time domain but in different frequency band.

Figure 1 represents a PCM[1] time speech signal. As shown in this figure there is a transition between silence and speech signal. This transition is very hard to be detected using an ordinary way of energy and zero crossing rate because this transition is between a silence and a weak fricative sound which is recorded in a normal environment contains a background noise. Figure 2 and figure 3 indicate the plotting of the wavelet coefficients in a certain band verses the time. A simple interpolation is made for concatenation of the wavelet parameters. Figure 4 represents the crosscorrelation between the wavelet parameters in two different frequency bands.

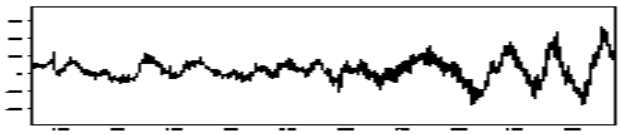

Figure 1 Speech sample contains a silence to consonant transition. The segment is taken from the beginning of the word همس in Arabic. This word is pronounced Hamss. /h//a/m//s/

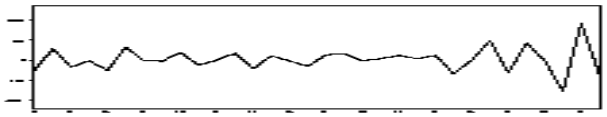

Figure 2 the wavelet transform of figure 1. This figure represents the frequency band 172-344 Hz in case of 1024 sample/Frame and 11025 Hz sampling rate

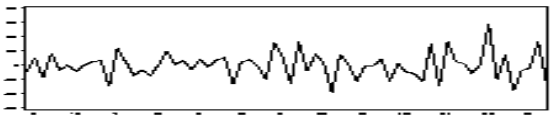

Figure 3 the wavelet transform of figure 1. This figure represents the frequency band 344-689 Hz in case of 1024 sample/Frame and 1025 Sample/sec.

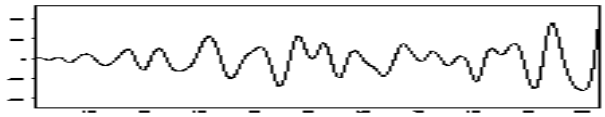

Figure 4 the first 1024 coefficients of the crosscorrelation between figures 2 and 3.

Figure 4 gives indication about how far the wavelet parameters are correlated in the two frequency levels. The two selected levels are chosen to cover the frequency range under the 1000 Hz that contains the most power of the human speech [8]. Any transition between a silence and the speech signal can be detected by the correlation between the different frequency bands. The signal is highly correlated in case of speech and is not correlated in case of silence. This fact makes it possible to track the uncorrelated points and extract the pure speech signal. This way is still effective to detect the end points of the actual speech signal even if the speech is highly noised. The multi-resolution nature of the wavelet transform makes it possible to monitor the signal into many bands each has a portion of the noise power which is much less than the total noise power distributed in all bands (in case of normal distribution of noise which is the natural case).

Table 1 relates the wavelet coefficients to the accordingly frequency bands in case of sampling rate 11025 samples/sec and frame length of 1024 samples.

Window #	Frequency Range in Hz	Number of wavelet parameters
9	2756 - 5512	512
8	1378 - 2756	256
7	689 - 1378	128
6	344- 689	64
5	172 - 344	32
4	86 - 172	16
3	43 - 86	8
2	21 - 43	4
1	10 - 21	2
0	0 - 10	1

Table 1 The wavelet parameters distribution over the whole frequency band in case of 11025 samples/sec and 1024 samples /frame.

3. END POINTS DETECTION

3.1 Selection of Wavelet windows

One of the wavelet windows can be shown in Figure 2. The window represents the energy distribution of the speech signal in a certain frequency band. Table 1 indicates the number of wavelet parameters for each frequency band in case of 1024 frame length and sampling rate of 11025 Hz. A simple interpolation technique is used to insert points between the wavelet parameters to expand them in each frequency band to 1024 points. Windows 5 and 6 are selected. Window 5 covers the range of (172-344) Hz and window 6 covers the range of (344-689) Hz. The selection based on the criteria of speech that indicates that the most power of the speech signal is below the 1 kHz.

3.2 Correlation model

The correlation model is obtained from the correlation between wavelet windows. The following definitions will help us in the rest of this section:

- ***End points*** the points of time at which the speech signal have no energy in all wavelet windows.
- ***Win(n)*** The wavelet window which have 2^n parameters according to table 1.
- ***R(n)*** The crosscorrelation parameter that indicates the correlation between win(5) and win(6) in this work. The crosscorrelation between the prepared win(5) (interpolated so that it contains 1024 points) and the prepared win(6) gives 2047 points of R(n).

The two windows are selected adjacent to insure that the power curves will be alike as much as possible. This is important to link between the correlation parameter and the

[1] PCM: Pulse code modulation. The speech in a normal digital form.

time index. Moreover, the crosscorrelation is used rather than the autocorrelation of one window to get the highest immunization to noise. If the speech unit is weak in one window it may be strong in the adjacent window. For the above two reason the crosscorrelation can give the maximum reliable correlation representation between the two windows win(5) and win(6).

The algorithm begins by dividing the speech signal into smaller windows of 1024 samples each. The wavelet parameters are extracted for each window. The crosscorrelation is performed on win(5) and win(6). The frames of R parameters are concatenated then the absolute value of the points is taken and smoothed using moving average of 1024 points (figure 6). Figures 5 and 6 show how far the energy correlation model tracks the boundaries of the speech signal.

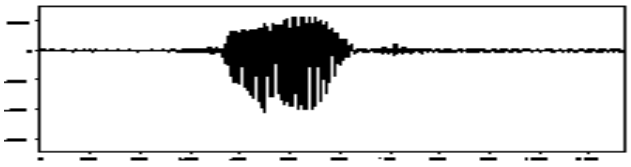

Figure 5 Speech signal contains a whisper consonant /h/ at start and unvoiced fricative /s/ at the end. There are a silence periods before and after the signal. The word is همس in Arabic. This word is pronounced Hamss /h//a//m//s/

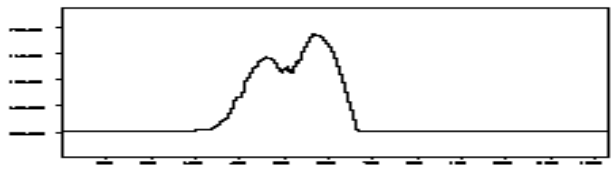

Figure 6 the correlation model. (The crosscorrelation parameters are concatenated).

3.3 End points

As shown in figure 6, the energy changes can easily be detected. Correlating the energy contents of the same signal in two different frequency levels generates the curve shown. The algorithm of detecting the end points from this curve is as follows:

- ***Noise analysis***: The first 20 ms (~220 samples in case of 11025 samples/sec) of the correlation model are used to extract the noise statistics. The moving standard deviation is calculated to each 10 ms (110 samples). The maximum of the first 220 points of the moving standard deviation is multiplied by 4 and taken as a threshold for discriminating the noise from the speech.

Logical series: The standard deviation points are compared with the noise threshold generated in the first step. The logical series is series contains 1 and 0 only. The size of the series is the same as the size of the speech signal. The element in the series can take a value of 1 if the threshold of noise is less than the standard deviation at this point. Else the value of the element is 0.

After this loop the SERIES contains ones 1 at speech duration only and zeroes 0 at the noise or silence periods. So the end points markers are obtained as shown in figure 7.

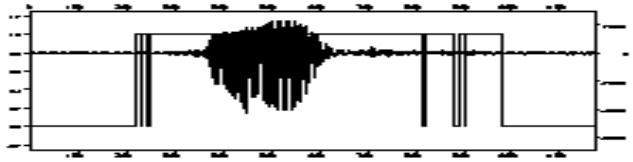

Figure 7 the speech signal and the logic series markers. The first and the last markers represent the speech boundaries.

4. SYSTEM EVALUATION

4.1 End points in high noise environment

To study how far the previous algorithm is valid the normal distribution noise is generated to superimposed on speech signal. The noise is multiplied with factors to control the signal to Noise ratio.

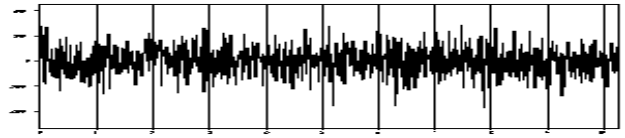

Figure 8 Noise in time domain

After applying the previous algorithm on the noisy speech the following results are obtained as shown in figures 9 and 10.

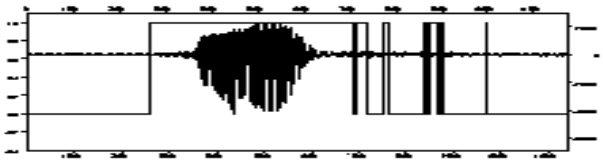

Figure 9 the speech signal and logic series markers in case of 48 dB signal to noise ratio.

As shown in figure 9 the markers still detect the boundaries of speech signal.

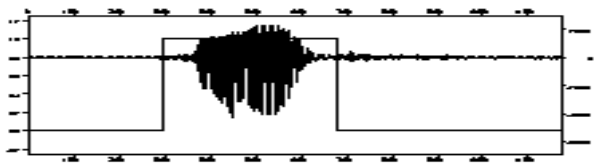

Figure 10 the speech signal and logic series markers in case of 16 dB signal to noise ratio.

Here in figure 10 the last point is shifted left and the starting point still acceptable.

4.3 End points in case of some hard cases

There are some hard cases at which the end points can not be detected accurately in normal noise conditions. These cases are:

1. Weak fricatives at the beginning or end.
2. Weak plosive at the beginning or end.
3. Nasals at the end.
4. Voiced fricatives that become devoiced at the end of words.

These cases are studied and the following results are obtained as shown in figures 11 and 12.

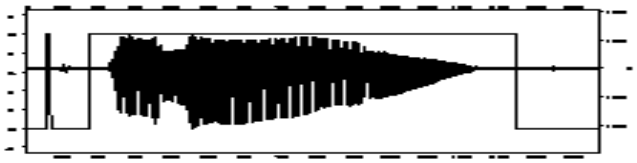

Figure 11 The word contains a weak plosive at the beginning /k/ and a nasal at the end /n/. The word is كمان in Arabic and it is pronounced kaman /k//a//m//a//n/

In figure 11 the weak plosive at the beginning and the nasals at the end are detected accurately. This speech signal is taken in a normal noise condition not in laboratory conditions.

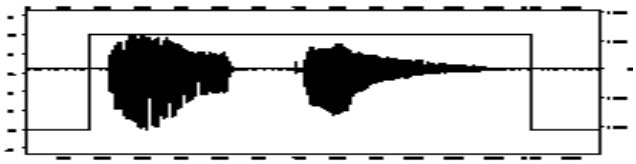

Figure 12 The word contains a voiced fricative at the end of utterance /z/. The word is منتزه in Arabic and it is pronounced Montazh /m//o//n//t//a//z//h/

The first case is previously studied.

5. SUMMARY

The end points of the speech signal can be detected using a wavelet transform with very accurate results. The correlation model is generated from the correlation of wavelet coefficients. This model is used to generate a logical series to extract the speech duration from the whole sample and reject the noise duration at the boundaries of words. An evaluation of the system is made by superimpose a normal distributed noise to a speech signal with different signal to noise ratios. Moreover, the system is tested in the hard cases of end points such as weak fricatives at the beginning or end ...etc. The system gives a high accuracy for end point detection in normal case or in a highly noise cases.

6. REFERENCES

[1]Mark . hensa, he iscrete avelet ransform: Wedding the A Trous and Mallat Algorithms, IEEE Transactions on Signal Processing, VOL. 40, NO. 10, October 1992.

[2]Xiang-Gen Xia and Zhen Zhang, On Sampling Theorem, Wavelets, and Wavelet Transforms, IEEE Transactions on Signal Processing, VOL. 41, NO. 12, December 1993.

[3]Ali N. Akansu, The Binomial QMF-Wavelet Transform for Multiresolution Signal Decomposition, IEEE Transactions on Signal Processing, VOL. 41, NO. 1, January 1993.

[4]Ahmed H. Tewfik, On the Optimal Choice of a Wavelet for Signal Representation, IEEE Transactions on Information theory, VOL. 38, NO. 2, March 1992.

[5]Gilbert Strang, Wavelets and Filter Banks, Wellesley-Cambridge Press, pp 1 - 34, pp 53-60, pp 155-172.

[6]J. D. Markel and A. H. Gray, Linear Prediction of Speech, pp. 1-63.

[7]Thomas W. Parsons, Voice and speech processing, pp. 57-98, 136-192, 291-317.

[8]Lawrence R. Rabiner, Digital Processing of Speech Signals, pp. 43-55, 130-135.