Speech Processing Using Wavelet Based Algorithms

By
Amr Mohamed Refat Mohamed Gody

A thesis submitted to the communtication and electronics Department, Cairo University , Faculty of Engineering in Partial Fulfillment for the Ph.D. Degree
In
Electronics and communication Engineering

Approved by the Examining committee

Prof. Dr. Amin Nassar	**, Thesis Main Advisor**
Prof.Dr. Salwa Hussien El-Ramly	**, Member**
Prof.Dr. Magdy Fekry M. Ragaey	**, Member**

Faculty of Engineering, Cairo University
Giza, EGYPT
1999

Speech Processing Using Wavelet Based Algorithms

By
Amr Mohamed Refat Mohamed Gody

A thesis submitted to the communtication and electronics Department, Cairo University , Faculty of Engineering in Partial Fulfillment for the Ph.D. Degree
In
Electronics and communication Engineering

Under the supervision of

Dr. Nemat Sayed Abdel Kader
Lecturer
Electronics and Communication Dept. , Faculty of Engineering , Cairo University

Prof. Dr. Amin Mohamed Nassar
Professor
Electronics and Communication Dept. , Faculty of Engineering , Cairo University

Faculty of Engineering, Cairo University
Giza, EGYPT
1999

Speech Processing Using Wavelet Based Algorithms

By
Amr Mohamed Refat Mohamed Gody

A thesis submitted to the communtication and electronics Department, Cairo University , Faculty of Engineering in Partial Fulfillment for the Ph.D. Degree
In
Electronics and communication Engineering

Faculty of Engineering, Cairo University
Giza, EGYPT
1999

ACKNOWLEDGMENT

First thank god, I would like to thank every one helped me to complete this work.

Specially, I would like to thank Prof. Amin Nassar for giving me this chance to do my Ph.D. under his supervision. It is my pleasure to thank him for his support and all the help he gave.

It is also my pleasure to express my thanks and appreciation to Dr. Nemat Abd-ElKader for her effort, support and encouragement during the Ph.D. period.

Furthermore, I am indebted to Prof. Mohsen Rashwan for his invaluable technical assistance.

I would like to extend my appreciation to Prof. Salwa El-ramly for her effort in reviewing this thesis.

I want to express my thanks to Prof. Magdy Fekry.

Finally, I would like to express my deep thanks to my family and my wife for their support, help, and encouragement.

List of publications

[1]. Nemat Sayed Abdel Kader, Amr M. Refat ," Voiced/Unvoiced Classification using Wavelet based algorithm", ICSPAT98. (International conference of signal processing and technology)

[2]. Nemat Sayed Abdel Kader, Amr M. Refat, "Voiced/Unvoiced classification using wavelet correlation model", ICSPAT'99

[3]. Nemat Sayed Abdel Kader, Amr M. Refat , "End points detection using wavelet based algorithm", Eurospeech'99

Abstract

The aim of this research is to design an Arabic recognition system based on wavelet transform that is highly reliable even in the presence of noise. **There are many achievements in this research:**

1. New techniques based on wavelet transform are implemented to classify the speech signal into voiced sounds and unvoiced sounds. The system indicates high sensitivity to voice changes even in case of low signal to noise ratio.
2. New technique for end points detection based on wavelet transform is achieved. The system can work in a poor signal to noise ratio (S/N) with a good accuracy of determining the speech boundaries. At about 9dB it gives about 91% of accuracy.
3. New technique for pitch estimation based on wavelet transform is achieved.
4. New technique for vowel/consonant classification is achieved using wavelet transform. The system gives a probability of success more than 95% at 9 dB S/N.
5. New technique for vowel recognition based on wavelet transform is achieved. The gives a recognition rate of about 90%.

The research is divided into smaller objectives. Each one is totally studied as a separate research point.

- **Speech classification into voiced and unvoiced segments**. Speech signal is classified into voiced speech or unvoiced speech using wavelet transform. The effect of noise is taken into

consideration and a good classification accuracy is achieved even in case of very low signal to noise ratio.

- **End points detection**. In this part, the problem is treated using wavelet transform as features of the speech signal. The problem is handled using different methods which are:

1- Correlation between wavelet bands.

2- Using mathematical classifier.

3- Using neural network.

All methods are tested in the presence of noise.

- **Pitch estimation** is one of the fundamental properties that is very important in speech recognition. The problem is handled here using a new algorithm based on wavelet transform. The correlation between wavelet bands gives indication about pitch pulses. The system is also tested in case of low signal to noise ratio.

- **Recognition of Arabic phonemes**. In this part the problem is divided into two parts. The first one is to classify the vowels and the consonants inside the utterance. This problem is manipulated using the wavelet transform and the mathematical classification of wavelet features.

The second part is to discriminate the vowels itself. In this part the wavelet transform and mathematical classification is used to recognize the Arabic vowels. Vowels are monitored in six frequency bands using wavelet features.

The first three objectives have been implemented by different ways in many languages including the Arabic language. The practical

constrains are taken into consideration. New methods are introduced. Practical results have been achieved in the first three objectives. Recognition of Arabic vowels gives best results while consonant phonemes will be considered in future work. The results, which are obtained here, give a promise that the realistic-unlimited-real time speech dictation machine is in the way.

Chapter 1

Speech signal and wavelet transform

1.1 Introduction

Much of our thinking about spoken language has been focused on its use as an interface in human-machine interactions mostly for information access and extraction. With increases in cellular phone use and dependence on networked information resources, and as rapid access to information becomes an increasingly important economic factor, telephone access to data and telephone transactions will no doubt rise dramatically. There is a growing interest, however, in viewing spoken language not just as a means to access information, but as, itself, a source of information. Important attributes that would make spoken language more useful in this respect include: random access, sorting (e.g., by speaker, by topic, by urgency), scanning, and editing. How could such tools change our lives? Enabling such a vision challenges our systems still further in noise robustness and in spontaneous speech effects. Further, the resulting increased accessibility to information from conversational speech will likely also raise increased concern for privacy and security, some of which may be addressed by controlling access by speech: speaker identification and verification. While such near-term application possibilities are exciting, we can envision an even greater information revolution with the development of writing systems if we can successfully meet the challenges of spoken language both as a medium for information access and as itself a source of information. Spoken language is still the means of communication used first and foremost by humans, and only a small percentage of human communication is written. Automatic-spoken-language understanding can add many of the advantages normally associated only with text (random access, sorting, and access at different times and places) to the

many benefits of spoken language. Making this vision a reality will require significant advances.

Speech-understanding research was non-existent 50 years ago[1]. The dramatic changes in speech recognition and in language understanding during the past 50 years, combined with political changes and changes in the computing infrastructure, led to the state of the art that we observe today. Challenges remain in several areas:

Integration. There is much evidence that human speech understanding involves the integration of a great variety of knowledge sources, including knowledge of the word or context, knowledge of the speaker and/or topic, lexical frequency, previous uses of a word or a semantically related topic, facial expressions (in face-to-face communication), prosody, in addition to the acoustic attributes of the words. Our systems could do much better by integrating these knowledge sources.

Prosody. Prosody can be defined as information in speech that is not localised to a specific sound segment, or information that does not change the identity of speech segments. Such information includes the pitch, duration, energy, stress, and other supra-segmental attributes. The segmentation (or grouping) function of prosody may be related more to syntax (with some relation to semantics), while the saliency or prominence function may play a larger role in semantics and pragmatics than in syntax. To make maximum use of the potential of prosody will likely require a well-integrated system, since prosody is related to linguistic units not just at and below the word level, but also to abstract units in syntax, semantics, discourse, and pragmatics.

Spontaneous Speech. The same acoustic attributes that indicate much of the prosodic structure (e.g., pitch, stress, and duration patterns) are also very

common in aspects of spontaneous speech that seem to be more related to the speech planning process than to the structure of the utterance. For example, a long syllable followed by a pause can indicate either an important syntactic boundary or that the speaker is planning the rest of the utterance. Similarly, a prominent syllable may mark new or important information, or a restart intended to replace something said in error. Although spontaneous speech effects are quite common in human communication and may be expected to increase in human machine discourse, as people become more comfortable conversing with machines, modelling of speech disfluencies is only just beginning.

1.2 Speech signal

1.2.1 Speech production:

The study of the nature of speech generation is required as a background of speech modelling and analysis. The understanding of speech generation in human is needed for modelling the organs of speech and controlling of speech model. The organs of speech are discussed first to explain how speech signal is produced and recognised in nature.[2-5].

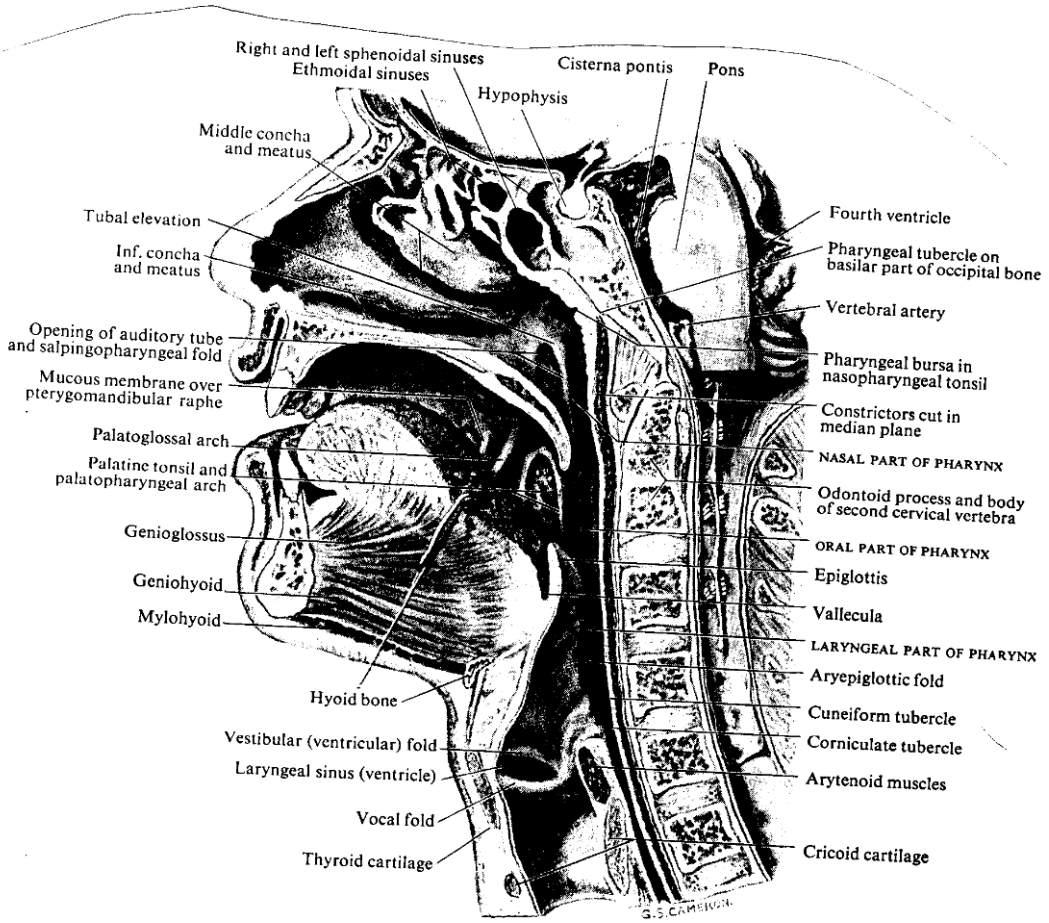


Figure 1. 1 Speech production Organs

Figure 1.1 lists all organs that are responsible for speech formation[4]. The acoustical speech waveform is simply an acoustic pressure wave, which start from intentional physiological movements of the structures shown in Figure 1.1. Air is released from the lungs into the trachea and then forced between the vocal cords. The lungs and trachea also control the intensity of the resulting speech, but they rarely make an audible contribution to speech. The vocal tract plays a very important task in speech signal. It acts as a filter that its input comes from the lungs and trachea through the larynx. It consists of Epiglottis, Lower jaw, Tongue, Velum, Palate, Teeth and Lips.

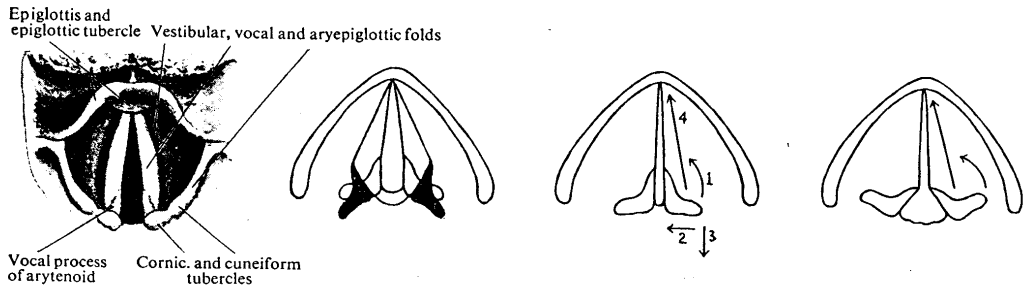


Figure 1. 2 the larynx structure

The diaphragm is a dome shaped muscle attached to the bottom of the rib cage, when this muscle contracts the dome becomes flatter; the volume of pleura increases and air rushes into the lungs. When the diaphragm relaxes, it resumes its dome shape and process reverses[3]. We speak while breathing and must mange to reconcile linguistic and physiological requirements. We learn to do this as children. The vocal cords is included in the larynx. Figure 1.2 describes the structure of the larynx. The larynx consists of four basic elements Cricoid cartilage, Thyroid cartilage, Arytenoid cartilage and Vocal cords. The first two elements are mostly framework. The Cricoid cartilage is essentially another one of the rings making up the trachea, but much higher at the rear in order to support the ends of the vocal cords. The domed shape of the Cricoid cartilage is the Adam's apple.

Comprising phonation, whispering, friction, compression and vibration does excitation. The phonation is the most important excitation source. It is the oscillation of the vocal cords. The opening and closing of the cords break the air stream up into pulses as shown in figure 1.3. The repetition rate of the pulses is termed Pitch.

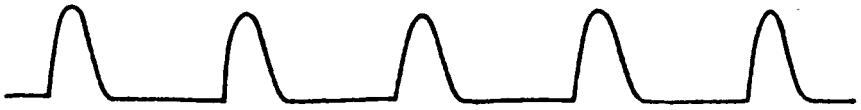


Figure 1. 3 The glottal pulse train

Research in speech processing and communication, for the most part, was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. The earliest attempt of this type was a mechanical mimic of the human vocal apparatus by Wolf-gang von Kempelen, described in his book published in 1791 [1]. Charles wheatstone, some 40 years later, constructed a machine based on Kempelen's specification using a bellow to represent the lung in providing a reservoir of compressed air. The vocal cords were replaced by a vibrating reed that was placed at one end of a flexible leather tube-“the vocal tract”-whose cross-sectional area could be varied to produce various voiced sounds. Other sounds could be produced by the machine as well, e.g., nasals by opening a side branch tube (the "nostrils"), fricatives by shutting off the reed and introducing turbulence at appropriate places in the vocal tract, and stops by closing the tube and opening it abruptly. It appears that Wheatstone was able to produce a fairly large repertoire of vowels and consonants and even some short sentences using this simple mechanical device.

Interest in mechanical analogous of the human vocal apparatus continued into the 20th century. While several notable people (Faber, Bell, Paget, and Riesz) followed Kempelen and Wheatstone's speech-production models, Helmholz, Miller, Koenig, and others pursued a different design principle. They synthesized vowel sounds by superimposing harmonically related

sinusoids with appropriately adjusted amplitudes. These two fundamentally different approaches, source-tract modelling (motivated by physics) and sinusoidal modelling (motivated by mathematics), have dominated the speech signal-processing field for more than 100 years.

Research interest in speech processing today has gone well beyond the simple notion of mimicking the human vocal apparatus (which still intrigues many researchers). The scope (both breadth and depth) of speech research today has become much larger due to advances in mathematical tools (algorithms), computers, and the almost limitless potential applications of speech processing in modern communication systems and networking Conversely, speech research has been viewed as an important driving force behind many of the advances in computing and software engineering, including digital signal processors (DSPs). Such a synergetic relationship will continue for years to come.

With the collaboration of Riesz and Watkins, Dudlev implemented two highly acclaimed devices, the VODER (VOice DEmonstration Recorder) and the VOCODER, based on this principle. The VODER (a schematic diagram of which is shown in Figure 1.4 was a system in which an operator manipulated a keyboard with 14 keys, a wrist bar, and a foot pedal to generate the control parameters required to control the sound source and the filter bank. This system was displayed with great success at the New York World's Fair in 1939. According to Dudley, it took a few weeks of training to be able to operate a VODER and produce intelligible speech on demand.

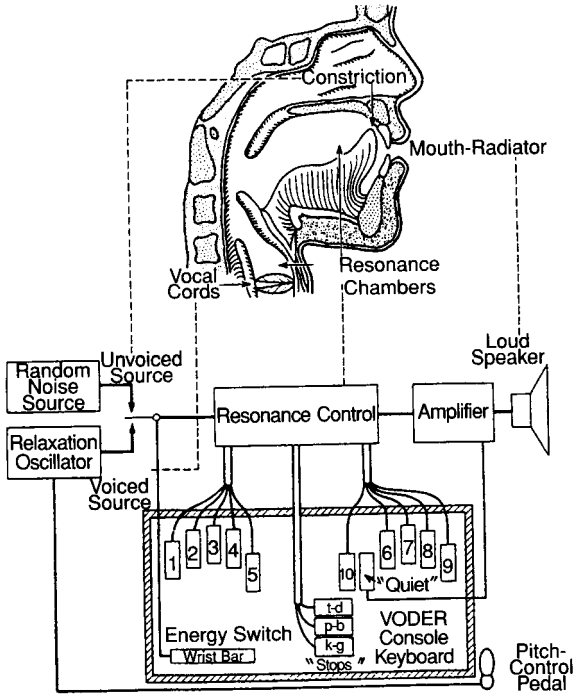


Figure 1. 4 VODER synthesiser model[1]

1.2.2 Linear prediction model

Representation of the vocal-tract frequency response, independent of the source parameters (e.g., voicing and fundamental frequency), captured researchers' interest in the 1960s. One approach to this problem was to analyze the speech signal using a transmission line analog of the wave-propagation equation. This method allows use of a time-varying source signal as excitation to the "linear" system of the vocal tract.

To make analysis of the vocal-tract response tractable, one often assumes that the vocal tract is an acoustic system consisting of a concatenation of uniform cylindrical sections of different areas with planar waves propagating through the system. Each section can be modelled with an equivalent circuit with wave reflections occurring at the junctions between sections. Such a model allows analysis of the system from its input-output characteristics.

In the late 1960s, Atal and Itakura independently developed a spectral analysis method, now known as linear prediction. While the motivations were different, they made an identical assumption; namely, that the speech signal at time t could be approximately predicted by a linear combination of its past values. In a discrete time implementation of the method, this concept is expressed as:

$$S_i \sim \hat{S}_i = \sum_{j=1}^{p} a_j S_{i-j} \qquad \textbf{(1. 1)}$$

S_i : Actual speech value at time index i.

$\hat{S}_i$: predicted speech value at time index i.

Where p is called the order of the predictor. The task is to find the coefficients $\{a_j\}$ that minimize some measure of the difference between S_i and $\hat{S}_i$ over a short-time analysis window. To retain the time-varying characteristics of the speech signal, the analysis procedure updates the

coefficients estimation process progressively over time.

The linear prediction analysis method has several interesting interpretations. In the frequency domain, the computed coefficients $\{a_j\}$ define an all-pole spectrum $\sigma/A(e^{j\omega})$ *where*

$$A(Z) = 1 - \sum_{j=1}^{p} a_j Z^{-1} \qquad \textbf{(1. 2)}$$

with $z = e^{j\omega}$

Such a spectrum is essentially a short-term estimate of the spectral envelope of the speech signal, at a given time [1]. The "envelope" models the frequency response of the vocal tract while the fine structure in the Fourier spectrum is a manifestation of the source excitation or driving function. This spectral envelope estimate can be used for many purposes; e.g., as the spectral magnitude control in a speech synthesizer or as features for speech recognition.

Another interesting result of the linear prediction technique is that it provides an estimate of the reflection coefficients as well as the area functions of a cylindrical tube of the type mentioned above [3]. Linear prediction thus could be viewed as a spectral estimation technique as well as a method for vocal-tract modelling (through the cylindrical tube model approximation).

The all-pole spectrum that resulted from linear prediction is a very efficient representation of the speech short time spectrum and is widely used in a range of speech-coding systems.

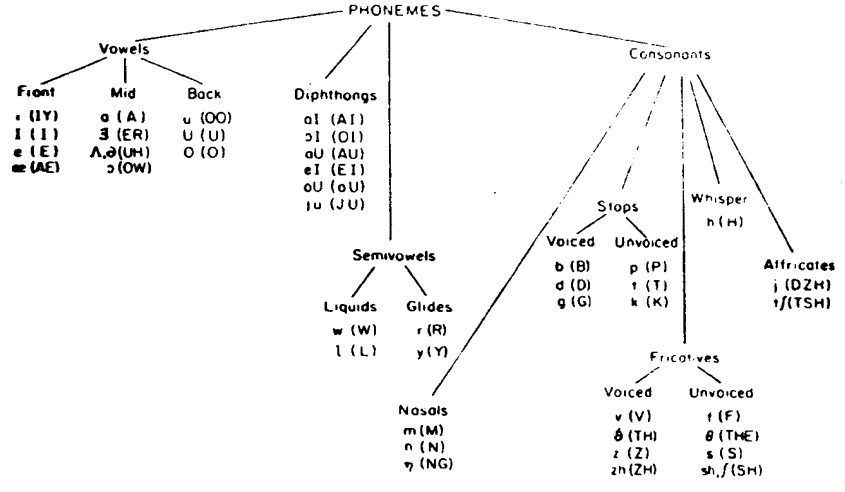
1.2.3 Acoustical parameters

Most 1anguages, including Arabic, can be described in terms of a set of distinctive sounds, or phonemes. In particular, for American English, there are about 42 phonemes including[4] vowels, diphthongs, semivowels and consonants. There are a variety of ways of studying phonetics; e.g., linguists

study the distinctive features or characteristics of the phonemes. For our purposes it is sufficient to consider an acoustic characterisation of the various sounds including the place and manner of articulation, waveforms, and spectrographic characterisations of these sounds.

Figure 1.5 shows how the sounds of American English are broken into phoneme classes.' The four broad classes of sounds are vowels, diphthongs, semivowels, and consonants. Each of these classes may be further broken down into subclasses that are related to the manner, and place of articulation of the sound within the vocal tract.

Each of the phonemes in Figure 1.5 (a) can be classified as either a continuant, or a noncontinuant sound. Continuant sounds are produced by fixed (non-time-varying) vocal tract configuration excited by the appropriate source. The class of continuant sounds includes the vowels, the fricatives (both unvoiced and voiced), and the nasals. The remaining sounds (diphthongs, semivowels, stops and aifricates) are produced by a changing vocal tract configuration. These are therefore classed as noncontinuants.



(a)

Arabic ymbol	IPA Symbol	Phoneme ymbol	Arabic Symbol	IPA Symbol	Phoneme Symbol
ء	ʔ	ʔ	ط	tˤ	tˤ
ب	b	b	ظ	ðˤ	Z
ت	t	t	ع	ʕ	ʕ
ث	θ	θ	غ	ɣ	G
ج	ʤ	g	ف	f	f
ح	ħ	H	ق	q	q
خ	x	x	ك	k	k
د	d	d	ل	l	l
ذ	ɕ	c	م	m	m
ر	r	r	ن	n	n
ز	z	z	هـ	h	h
س	s	s	و	w	w
ش	ʃ	ʃ	ى	j	j
ص	sˤ	sˤ			
ض	dˤ	dˤ			

Vowel symbol	IPA Symbol	Phoneme symbol	Vowel Symbol	IPA Symbol	Phoneme Symbol
ـَ	a	a	ـا	aː	aː
ـِ	i	i	ـي	iː	iː
ـُ	u	u	ـو	uː	uː

(b)

Figure 1. 5 (a)Phonemes in American English[5],(b) Arabic phonemes[72].

The Arabic language has basically 34 phonemes , 28 consonants and six vowels (see fig 1.5 b).

1.2.4 Human ear and speech perception

According to the source-filter model of speech production, the speech signal can he considered to be the output of a linear system. Depending on the type of input excitation (source), two classes of speech sounds are produced: voiced and unvoiced. If the input excitation is noise, then unvoiced sounds

such as /S/, /t/, etc., are produced, and if the input excitation is periodic then voiced sounds such as /a/, /i/, etc., are produced. In the unvoiced case, noise is generated either by forcing air through a narrow constriction (e.g., production of /f/) or by building air pressure behind an obstruction and then suddenly releasing that pressure (e.g., production of /t/). In contrast, the excitation used to produce voiced sounds is periodic and is generated by the vibrating vocal cords. The frequency of the voiced excitation is commonly referred to as the fundamental frequency (F0) or the pitch[2].

The vocal tract shape defined in terms of tongue, velum, lip and jaw position, acts like a "filter" that filters the excitation to produce the speech signal. The frequency response of the filter has different spectral characteristics depending on the shape of the vocal tract. The broad spectral peaks in the spectrum are the resonances of the vocal tract and are commonly referred to as formants. Figure 1.6 shows, for example, the formants of the vowel /eh/ (as in "head"). The frequencies of the first three formants (denoted as Fl, F2, and F3) contain sufficient information for the recognition of vowels as well as other voiced sounds. Formant movements have also been found to be extremely important for the perception of unvoiced sounds. In summary, the formants carry some information about the speech signal.

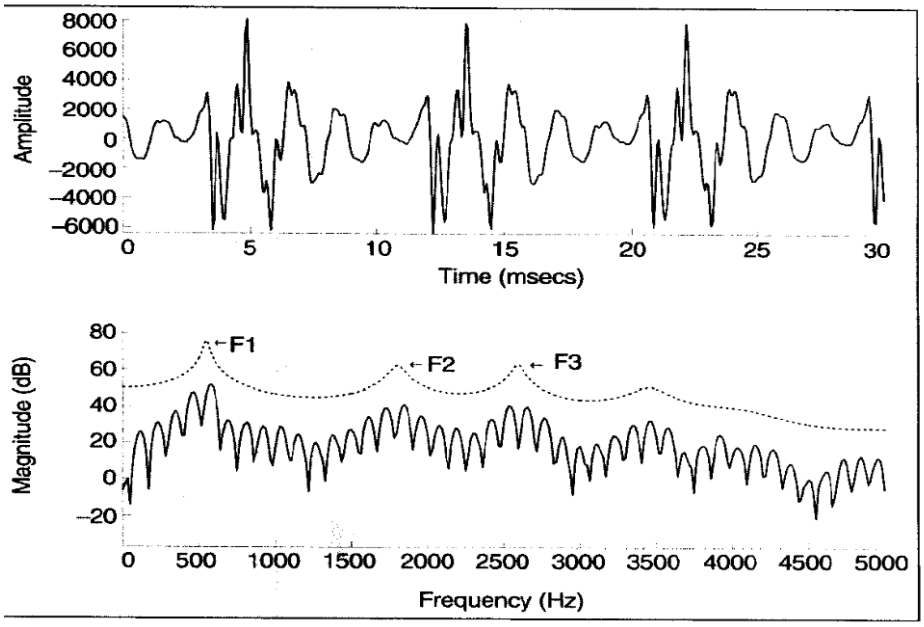


Figure 1. 6 A 30 ms segment of the vowel /eh/ and its spectrum

This leads to the question “How does the auditory system encode frequencies?" The pioneering work of Georg von Bekesy in the 1950s showed that the basilar membrane in the inner ear is responsible for analyzing the input signal into different frequencies. Different frequencies cause maximum vibration amplitude at different points along the basilar membrane (see Figure 1.7).

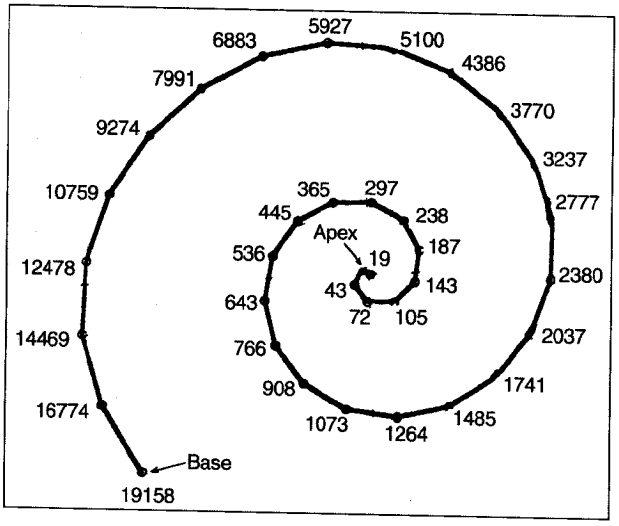


Figure 1. 7 Frequency response distribution in the basiliar membrane [2]

Low-frequency sounds create travelling waves in the fluids of the cochlea that cause the basilar membrane to vibrate with largest amplitude of displacement at the apex (see Figure 1.3) of the basilar membrane. On the other hand, high-frequency sounds create travelling waves with largest amplitude of displacement at the base (near the stapes) of the basilar membrane. If the signal is composed of multiple frequencies, then the resulting travelling wave will create maximum displacement at different points along the basilar membrane. The cochlea therefore acts like a spectrum analyzer. It decomposes complex sounds into their frequency components.

The cochlea is one of the mechanisms used by our auditory system for encoding frequencies. The travelling wave of the basilar membrane in the cochlea vibrates with maximum amplitude at a place along the cochlea that is dependent on the frequency of stimulation. The corresponding hair cells bent by the displacement in the membrane stimulate adjacent nerve fibres, which are organized according to the frequency at which they are most sensitive. Each place or location in the cochlea is therefore responding "best" to a particular frequency. This mechanism for determining frequency is referred to as *place theory.* The place mechanism for coding frequencies has motivated multichannel cochlear implants. Another theory, called *volley theory*, suggests that frequency be determined by the rate at which the neurons are fired. According to the volley theory, the auditory nerve fibres fire at rates proportional to the period of the input signal up to frequencies of 5,000 Hz. At low frequencies, individual nerve fibres fire at each cycle of the stimulus; i.e., they are "phase locked" with the stimulus. At high frequencies, the organized firing of groups of nerve fibers indicates frequency.

1.2.5 Speech processing in time and frequency domains

Speech processing techniques are based on either time analysis methods or frequency analysis methods. The time based methods are those that manipulate the speech signal in time domain such as autocorrelation methods for finding pitch , voiced/unvoiced,.. etc. Frequency methods handle the speech signal via spectral parameters such as cepstrum based pitch determination.

In time-based methods we take the advantage of handling the speech signal as it is which means more faster algorithms. The disadvantage of this manipulation is that we can not eliminate the noise effect [5]. The time-based

techniques are useful in case of high signal to noise ratio's environments. The frequency based methods overcome the last disadvantage. But in general we lose information in the transition from time to frequency domain. The intermediate frequency information is not available rather the information is about a package of time (frame). Any variation within frame can not be predicted. The last statement raise the problem of what is the appropriate frame length that gives a minimum error. In the problem of pitch estimation, selecting frame length affects the whole process as shown in figure 1.8.

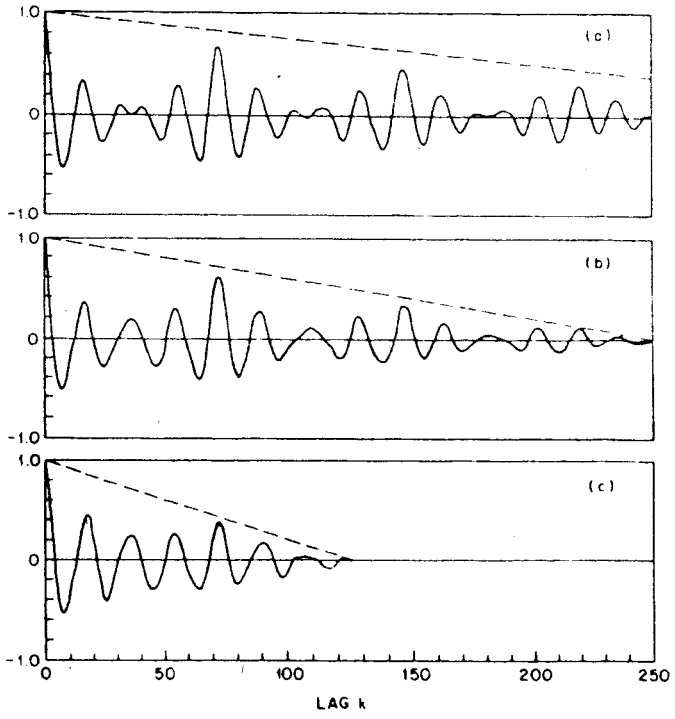


Figure 1. 8 Autocorrelation function for voiced speech with frame length N (a) N=401 samples, (b) N=251 samples, (c) N= 125 samples.

Figure 1.8c corresponds to a window length of 125 samples. Since the period for this example is about 72 samples, not even two complete pitch periods are included in the window. This is clearly a situation to be avoided, but avoiding it is difficult because of the wide range of pitch periods that may be encountered. One approach is to simply make the window long enough to

accommodate the longest pitch period, but this leads to undesirable averaging of many periods when the pitch period is short. Another approach is to allow the window length to adapt to match the expected pitch period.

The joint-time-frequency is the best representation of the speech signal. We can take the advantages of both simplicity of time based methods and powerful of frequency based methods in noise cancellation and signal compression. The joint-time-frequency is what we called wavelet transform.

1.3 Wavelet transform

Strictly speaking, wavelets transform is a topic of pure mathematics, however in only a few years of existence as a theory of its own, it have shown great potential and applicability in many fields.

There are several excellent monographs and articles talking about wavelets[8-15].

1.3.1 What are wavelets?

Wavelets are functions that satisfy certain requirements. The name wavelet comes from the requirement that they should integrate to zero[61], ``waving'' above and below the x-axis. The diminutive connotation of wavelet suggests the function has to be well localized. Other requirements are technical and needed mostly to insure quick and easy calculation of the direct and inverse wavelet transform.

There are many kinds of wavelets. One can choose between smooth wavelets, compactly supported wavelets, wavelets with simple mathematical expressions, wavelets with simple associated filters, etc. The most simple is the Haar[6]. Examples of some wavelets (from the family of Daubechies

wavelets) are given in Figure 1.9. Like sines and cosines in Fourier analysis, wavelets are used as basis functions in representing other functions. Once the wavelet (sometimes called the mother wavelet) Ψ(x) is fixed, one can form translations and dilations of the mother wavelet $\{ \psi(\frac{x-b}{a}) \rightarrow (a,b) \in R^{+} \times R \}$. It is convenient to take special values for a and b in defining the wavelet basis: $a = 2^{-j}, b = k2^{-j}$ where k and j are integers. This choice of a and b is called a critical sampling and will give a sparse basis. In addition this choice naturally connects the multiresolution analysis in signal processing with the world of wavelet.

Wavelet novices often ask, why not use the traditional Fourier methods? There are some important differences between Fourier analysis and wavelets. Fourier basis functions are localized in frequency but not in time. Small frequency changes in the Fourier transform will produce changes everywhere in the time domain. Wavelets are local in both frequency/scale (via dilations) and in time (via translations). This localization is an advantage in many cases.

Many classes of functions can be represented by wavelets in a more compact way. For example, functions with discontinuities and functions with sharp spikes usually take substantially fewer wavelet basis functions than sine-cosine basis functions to achieve a comparable approximation.

This sparse coding makes wavelets excellent tools in data compression. For example, the FBI has standardized the use of wavelets in digital fingerprint

image compression[6]. The compression ratios are on the order of 20:1, and the difference between the original image and the decompressed one can be told only by an expert. There are many more applications of wavelets, some of them very pleasing. Coifman and his Yale team used wavelets to clean noisy sound recordings, including old recordings of Brahms playing his First Hungarian Dance on the piano.

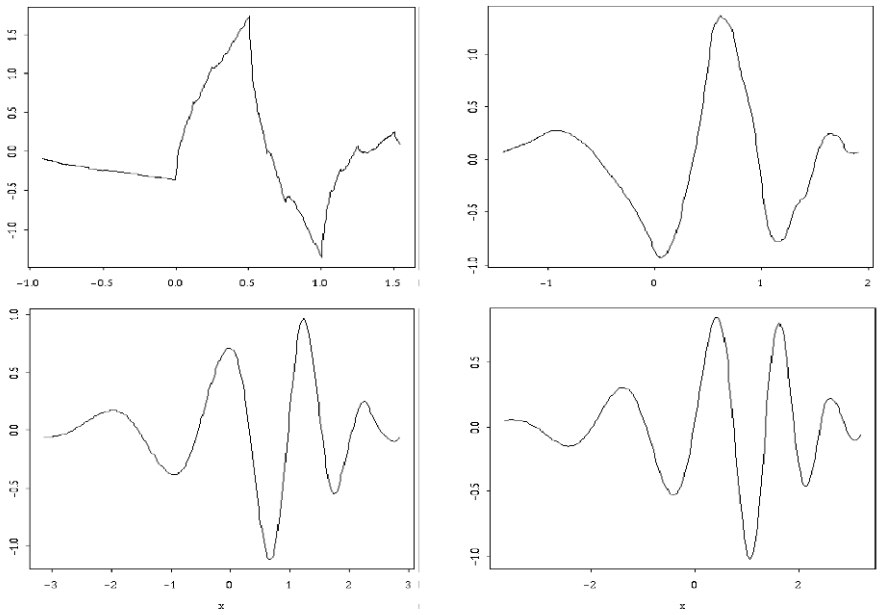


Figure 1. 9 Wavelets from the Daubechies family

1.3.2 Wavelets and filter bank

The wavelet is a small wave from which many other waves are derived by translation and dilation of the wavelet wave. It can be defined as:

$$W_{ij} = w(2^i t - jt) \qquad \textbf{(1. 3)}$$

Where:

W_{ij} is the wavelet function obtained by shifting the main wavelet base function by j samples and compressing the base function's duration by a factor of 2^i. The compression in time gives expansion in frequency. From the previous point of view the index i indicates the frequency level of the wavelet function.

Any function of time can be expressed in terms of wavelet functions and wavelet coefficients according to the following synthesizing equation.

$$f(t) = \sum_{i=0}^{m} \sum_{j=0}^{2^i} b_{ij} W_{ij} \qquad \textbf{(1. 4)}$$

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

$$b_{ij} = \int_0^T f(t) \cdot W_{ij} \, dt \qquad \textbf{(1. 5)}$$

T: The frame duration.

Equation 1.5 is valid if and only if the wavelets are orthogonal i.e.

$$\int_T W_{ij} \, W_{mu} \, dt = 0 \quad \textbf{ij} \neq \textbf{mu} \qquad \textbf{(1. 6)}$$

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 1.4 can be rearranged as:

$$f(t) = \sum_{j=0}^{1} b_{0j} W_{0j} + \cdots + \sum_{j=0}^{2^m} b_{mj} W_{mj} \qquad \textbf{(1. 7)}$$

Each summation represents the signal over the whole period in time domain but in different frequency bands. Table 1.1 represents each summation of equation 1.7. Each one gives a projection of the speech signal in a certain frequency band. As shown in table 1.1 column 3, the signal is represented with different number of parameters in each frequency band. The different number of parameters that represents the speech signal in the different frequency bands is called the multiresolution nature of Dyadic wavelet transform. In this research the dyadic wavelet is used for simplicity.

Table 1.1 The wavelet parameters distribution over the whole frequency band in case of 11025 samples/sec and 1024 samples /frame.

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**

1.3.3 Speech processing using wavelet transform

The application of the wavelet transform in speech gives a powerful tool to manipulate many speech-processing needs. It can be used to detect the pitch period or to classify the speech into voiced or unvoiced.

The speech processing has many fields that gain from wavelet representation of the speech signal. As an announcement not integration of the following areas are briefly discussed.

- **Speech compression[6]**

Speech compression is important in mobile communications, to reduce transmission time. Digital answering machines also depend on compression. The bit-rates are low, typically 2.4 kbits per second to 9.6 kbits/seconds. The best algorithms use either linear predictive models or sinusoidal models.

Speech is classified into *voiced* and *unvoiced* sounds. Voiced sounds are mainly low frequency. In CELP (code excitation linear predictor) the voiced sound is modelled as the output of an all-pole IIR filter with white noise as input. The filter coefficients are found by linear prediction. This filter represents the transfer function of the vocal tract. In a sinusoidal transform, the voiced sounds use a sinusoidal basis. Unvoiced sounds (like sss) have components in all frequency bands and resemble white noise. Model-based techniques achieve reasonable performance at low rates.

At more than 16 kbits/second, subband coding is effective and compatible with the models. Psychoacoustics has associated human hearing to nonuniform critical bands. These bands can be realized roughly as a four-level dyadic tree (Figure 1.10). For sampling at 8kHz, the frequency bands of the dyadic tree are: 0-250 Hz, 250-500 Hz, 500-1000 Hz, l000-2000 Hz and 2000-4000 Hz. These bands can be quantized and coded depending on subband energy; the average signal to noise ratio is maximized. And the *noise masking property* is used.

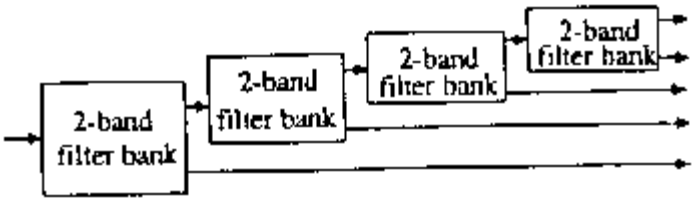


Figure 1. 10 Tree-structured filter banks used to approximate the critical bands[6].

- **Denoising:**

The piecewise constant signal below (Figure 1.11) [6] is corrupted by Gaussian white noise. The corrupted signal is decomposed using the Daubechies wavelet *D6.* The coefficients at level 4 are thresholded using Stein's Unbiased Risk Estimate. Notice that the reconstruction consists of the original signal and *some* of the noise.

In both wavelet shrinkage and denoising, the output is a cleaned-up version of the input. This works only when one knows the signal characteristics in advance. The algorithm will distort the desired signals when thresholding is applied.

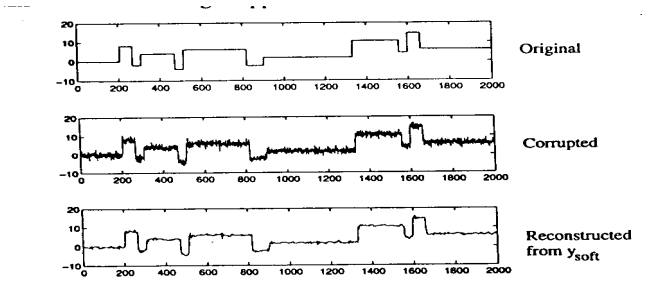


Figure 1. 11 Denoising using wavelet transform[6].

- **Speech classification**

Using wavelet transform the frequency variation of speech signal along the time of utterance can be monitored so that the variation from voiced to unvoiced or vice versa can be detected easily. Furthermore, The speech signal can be manipulated in a very narrow frequency band that corresponds to the maximum frequency of the voiced sounds this makes the effect of noise on the signal negligible.

- **Pitch detection**

Fundamental frequency estimation is one of the difficult problems in speech processing. It is handled using time-based methods and frequency based methods. The wavelet representation of the signal makes it possible to correlate the signal projections in different frequency bands to get the actual fundamental peaks, which is the pitch peaks.

- **End points detection**

End points detection of the speech utterance is one of the major problems in speech processing specially in case of low signal to noise ratios. The importance of this problem comes from the fact that the total efficiency of any speech-based machine is dramatically degraded if the speech boundaries are not accurate. Wavelet transform gives a frequency-time representation of the

speech signal. This makes it possible to find a certain threshold to detect the speech from the background noise as will be illustrated in chapter 2.

1.4 Artificial Neural Network for pattern classification

Many researchers believe that neural networks offer the most promising unified approach to building truly intelligent computer systems.

Artificial neural networks (ANNs) are simplified models of the central nervous system and are networks of highly interconnected neural computing elements that have the ability to respond to input stimuli and to learn to adapt to their environment. Neural networks employ parallel distributed processing (PDP) architectures. Hammerstrom clearly describes the three major advantages of neural networks [36-44].

Fig. 1.12 illustrates the basic neural network. As shown in fig. 1.12 there are 3 different layers: input layer, hidden layer and output layer. Each layer consists of nodes called neurones. The input layer actually not a neurones, it is just a buffer layer that illustrate the inputs to the next layer. As shown in fig. 1.12 there are small bubbles on the end of each arrow. Those bubbles represent the weights. It means that the input is multiplied with weight before introducing it to the neurone. Each neurone makes two fundamental functions. The first is the summation of all inputs from the previous layer after multiplying them with the corresponding weights. The second function is the firing function or comparing the sum with certain threshold. If the sum is higher than the threshold the neurone gives a one, else it gives a zero.

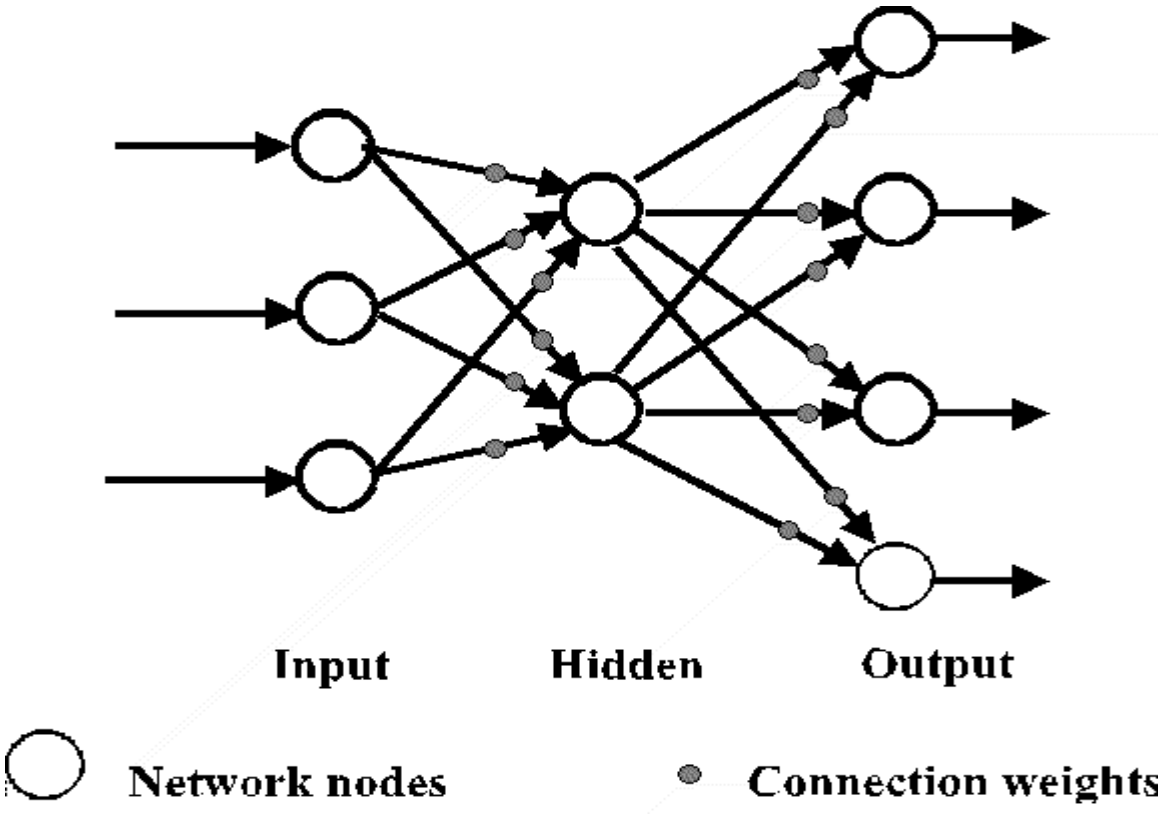


Figure 1. 12 Artificial Neural Network.

The basic anatomical unit in the nervous system is a specialised cell called the neurone. Fig. 1.13 is a view of a typical neurone [36][37].

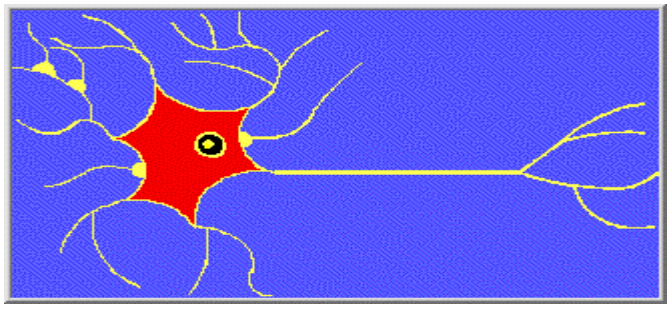


Figure 1. 13 Typical neurone in the nervous system.

Many extensions of the single cell are long and filamentary; these structures are called processes. Every neurone plays several functional roles in a neural system:

Metabolic machinery within the cell provides a power source for information-processing functions. In addition, the cell enforces a certain unity for biochemical mechanisms throughout its extent [36].

A tree of processes called dendrites is covered with special structures called synapses, where junctions are formed with other neurones. These synaptic contacts are the primary information-processing elements in neural systems.

Processes act as wires, conveying information over a finite spatial extent. The resistance of fine dendrites allows the potential at their tips to be computed with only partial coupling to other computations in the tree.

Temporal integration of signals occurs over the short term through charge storage on the capacitance of the cell membrane, and over the longer term by means of internal second messengers and complex biochemical mechanisms.

Certain neurones are equipped with a long, specialised proces called an axon. The axon is used for "digitising" data for local transmission, and for transmitting data over long distances.

The classical neurone is equipped with a tree of filamentary dendrites that aggregate synaptic *inputs* from other neurones. The input currents are integrated by the capacitance of the cell until a critical threshold potential is reached, at which point an *output* is generated in the form of a nerve pulse. This output pulse propagates down the axon, which ends in a tree of synaptic contacts to the dendrites of other neurones.

The resistance of a nerve's cytoplasm is sufficiently high that signals can not be transmitted more than about 1 millimetre before they are hopelessly spread out in time, and their information largely lost. For this reason, axons are equipped with an active amplification mechanism that restores the nerve

pulse as it propagates. In lower animals, such as the squid, this restoration is done continuously along the length of the axon. In higher animals many axons are wrapped with a special insulating material called myelin, which reduces the capacitance between the cytoplasm and the extracellular fluid, and thereby increases the velocity at which signals propagate. The sheaths of these myelinated axons have gaps called nodes of Ranvier every few millimetres. These nodes act as repeater sites, where the signal is periodically restored [39]. A single myelinated fibre can carry signals over a distance of 1 meter or more.

1.4.1 Features of Artificial Neural Network ANN [36][38]

- They are adaptive; they can take data and learn from it. This ability differs radically from standard software because it does not depend upon the prior knowledge of rules. In addition, neural networks can reduce development time by learning underlying relationships even when they are difficult to find and describe. They can also solve problems that lack existing solutions.
- Neural networks can generalise; they can correctly process information that only broadly resembles the original training data set. Similarly, they can handle imperfect or incomplete data, providing a measure of fault tolerance. Generalisation is useful in practical applications, because in the real world data is often noisy.
- Neural networks are non-linear; they can capture complex interactions among the input variables in a system.

1.4.2 Limitations of Neural Network

A limitation of neural networks is that

- they can consume vast amounts of computer time - two months, for example - particularly during training.
- The output from a neural net is usually difficult to directly interpret without the assistance of an expert system.
- Not adaptive, If the environment is changed the training must be repeated.

1.5 Mathematical modelling using multiple regression

In this section, focus will be on how the experimental results can be used to formulate a system model. System model is a mathematical function that can relate output to input. This is the case in pattern recognition methods. Database is collected for independent input variables and the corresponding dependent outputs in the training phase. After that we try to get a relation between inputs and output to model the system. Then, test data are introduced to system model for evaluation of its efficiency. This problem in mathematics is called Regression.

In Matrix notation this can be written as:

$$\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & X_{03} & X_{04} & X_{05} & X_{06} \\ \vdots & & \vdots & & \vdots & \\ X_{m1} & X_{m2} & X_{m3} & X_{m4} & X_{m5} & X_{m6} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \quad \textbf{(1. 8)}$$

[Y] matrix is the outputs that correspond to input vectors ([X] matrix). [B] matrix is the linear statistical model of the system. The term linear comes from the linear relation between [Y] and [X]. [X] is called the design matrix.

In order to find least-squares estimators of the b's, we consider the sums of squares of errors in predicting Y_i by

$$b_0 + b_1 X_{1i} + \cdots + b_k X_{ki} \qquad \textbf{(1. 9)}$$

The demand is to find $\hat{b}_0, \hat{b}_1, \cdots, \hat{b}_k$ that minimize

$$Q(b) \equiv S(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}))^2 \qquad \textbf{(1. 10)}$$

By differentiate Q(b) with respect to b then equating by 0,

$$\hat{b} = (XX')^{-1} XY \qquad \textbf{(1. 11)}$$

For complete details see [47].

In speech processing area this mathematical tool can be useful in classification. The problem of voiced/unvoiced as an example is a classification problem. There are two categories in this case voiced speech or unvoiced speech. The data is collected for each category and aligned into

matrices [X] and [Y] as indicated above. The matrix [B] is obtained which is a system model for voiced/unvoiced classification system.

1.6 Conclusion

As indicated in this chapter, the wavelet transform can give a good representation of the speech signal in multiple frequency bands. The property of joint time frequency of the wavelet transform gives it the facility to keep track with frequency-change-events, such that voiced sounds and unvoiced sounds or pitch or vowels and consonants, along the duration of utterance.

The classification tools such as neural network or mathematical linear regression, can be used in the classification problems of speech using the wavelet parameters as inputs. The use of wavelet parameters that describe the frequency changes of the speech in many bands along the utterance duration makes the classifier to make a good decision. Neural network is used in end points detection problem and the mathematical linear regression is used in all classification problems in this research.

CHAPTER 1 ..... **1**

SPEECH SIGNAL AND WAVELET TRANSFORM 1

1.1 INTRODUCTION **2**

1.2 SPEECH SIGNAL **4**

1.2.1 SPEECH PRODUCTION: 4

1.2.2 LINEAR PREDICTION MODEL 9

1.2.3 ACOUSTICAL PARAMETERS 11

1.2.4 HUMAN EAR AND SPEECH PERCEPTION 13

1.2.5 SPEECH PROCESSING IN TIME AND FREQUENCY DOMAINS 17

1.3 WAVELET TRANSFORM **19**

1.3.1 WHAT ARE WAVELETS? 19

1.3.2 WAVELETS AND FILTER BANK 21

1.3.3 SPEECH PROCESSING USING WAVELET TRANSFORM .. 24

1.4 ARTIFICIAL NEURAL NETWORK FOR PATTERN CLASSIFICATION **27**

1.4.1 FEATURES OF ARTIFICIAL NEURAL NETWORK ANN[38][36] 30

1.4.2 LIMITATIONS OF NEURAL NETWORK 30

1.5 MATHEMATICAL MODELLING USING MULTIPLE REGRESSION **31**

1.6 CONCLUSION 33

Chapter 2

End points Detection

2.1 Introduction

The problem of extracting the speech from the background noise is one of the major problems in speech applications. This is always the first step in any speech-based application. The performance of the application may be degraded dramatically if this point is not handled carefully. The problem of locating the beginning and end of a speech utterance in a 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 in speech transmission.

The problem of discriminating speech from background noise is not trivial, except in the case of extremely high signal-to-noise ratio acoustic environments - e.g., high fidelity recordings made in an isolated chamber or a soundproof room. For such 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 algorithm to be discussed in this section is based on wavelet transform. The wavelet transform as discussed before makes the link between time and frequency domains in one step by splitting the signal into many frequency channels. A new method will be introduced by using the wavelet transform for detecting the speech from the background noise. The algorithm gives

highly accurate results even though in case of very low signal to noise ratio and low energy phonemes at the beginning or end of utterance.

The chapter begins by introducing the old method of end points detection and its advantages and disadvantages. Then the problem will be handled with new algorithm based on wavelet transform.

2.2 Energy and zero crossing rate method

The problem of end points of speech is usually handled in almost all speech applications by two simple time-domain measurements - energy, and zero-crossing rate. Several simple examples will illustrate some difficulties encountered in locating the beginning and end of a speech utterance[5]. Figure 2.1 shows an example (the beginning of the word eight) for which the background noise is easily distinguished from the speech, as denoted in the figure. In this case a radical change in the waveform energy between the background noise and the speech is the cue to the beginning of the utterance. Figure 2.2 shows another example (the beginning of the word /six/) for which it is easy to locate the beginning of the speech. In this case, the frequency content of the speech is radically different from the background noise, as seen by the sharp increase in zero crossing rate of the waveform. It should be noted that, in this case, the speech energy at the beginning of the utterance is comparable to the background noise energy.

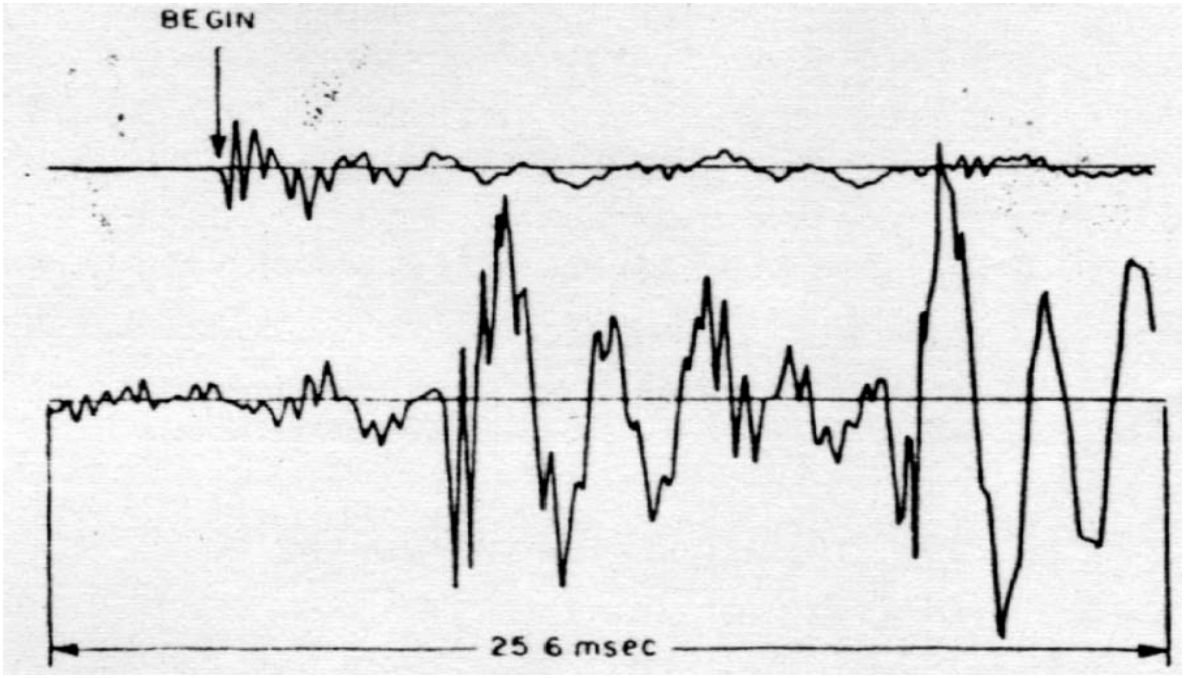


Figure 2.1 Waveform of the beginning of utterance /eight/[5]

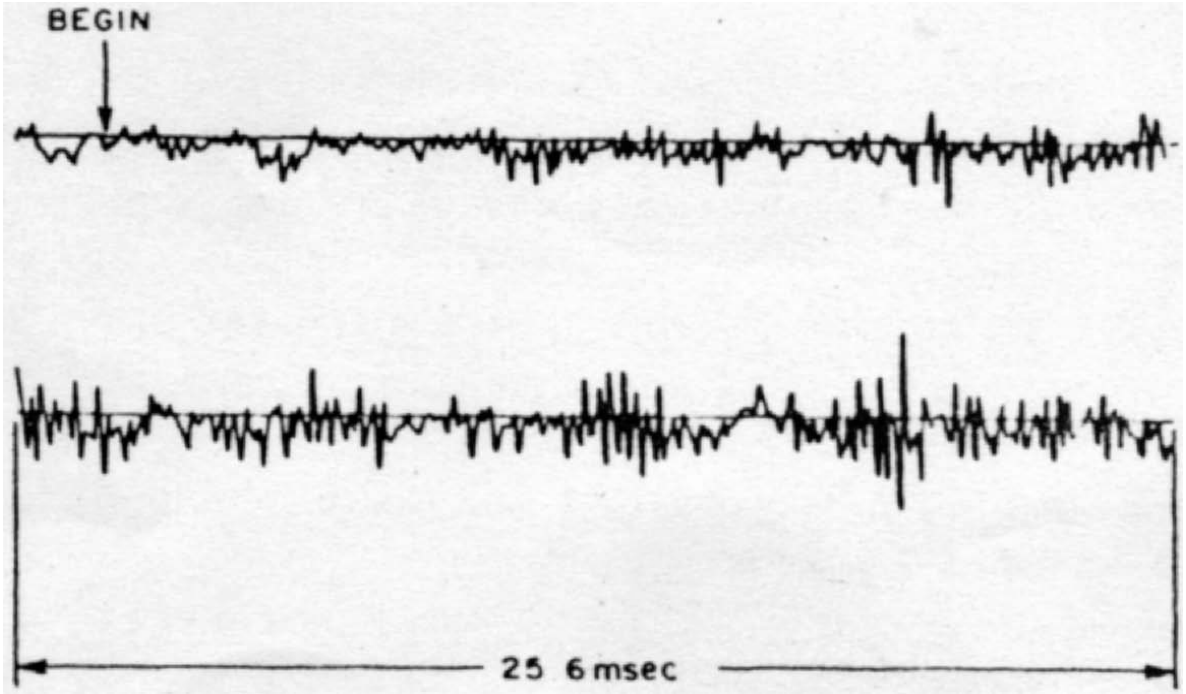


Figure 2.2 The beginning of word /six/[5]

Figure 2.3 gives an example of a case in which it is extremely difficult to locate the beginning of the speech signal. This figure shows the waveform for the beginning of the utterance /four/. Since /four/ begins with the weak fricative /f/ (low energy), it is very difficult to precisely identify the beginning point. Although the point marked B in this figure is a good candidate for the beginning, point A is actually the beginning. In general it is difficult to locate the beginning and end of an utterance if there are:

1. Weak fricatives *(/f/, /th/, /h/)* at the beginning or end.
2. Weak plosive bursts *(/p/,* /t/, /*k/)* at the beginning or end.
3. Nasals at the end.
4. Voiced fricatives which become devoiced at the end of words.
5. Trailing off of vowel sounds at the end of an utterance.

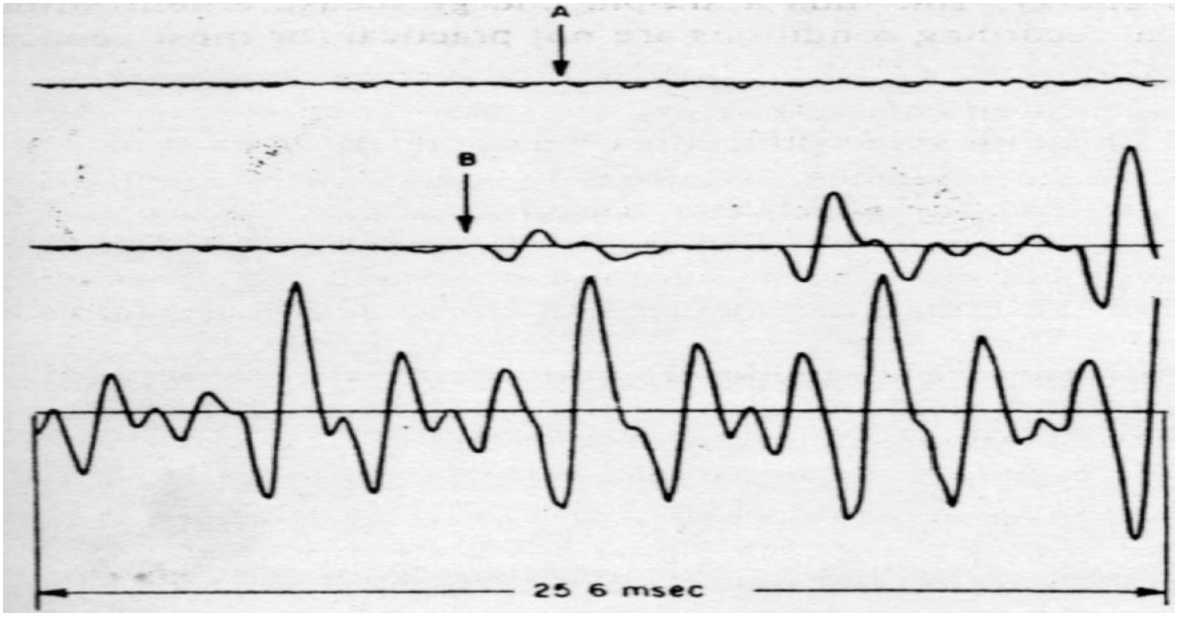


Figure 2. 3 word / four/[5]

In spite of the difficulties posed by the above situations, energy and zero crossing rate representations can be combined to serve as the basis of a useful

algorithm for locating the beginning and end of a speech signal. One such algorithm was studied by Rabiner and Sambur in the context of an isolated-word speech recognition system [5]. In this system a speaker utters a word during a prescribed recording interval, and the entire interval is sampled and stored for processing. The purpose of the algorithm is to find the beginning and end of the word so that subsequent processing and pattern matching can ignore the surrounding background noise.

The algorithm can be described by reference to figure 2.4. The basic representations used are the number of zero-crossings per 10 msec frame and the average magnitude computed with a 10 msec window**.** As follows:

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|$$

$$Z_n = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}(x(m)) - \mathrm{sgn}(x(m-1))| \qquad \textbf{(2. 1)}$$

Where M_n is the short time average magnitude at time index n.

Z_n is the short time average zero crossing rate at time index n.

Both functions are computed for the entire recording interval at a rate of 100 times/sec, It is assumed that the first 100 msec of the interval contains no speech. The mean and standard deviation of the average magnitude and zero crossing rate are computed for this interval to give a statistical characterization of the background noise. Using this statistical characterization and the maximum average magnitude in the interval, zero-crossing rate and energy thresholds are computed. (Details are given in [5].)

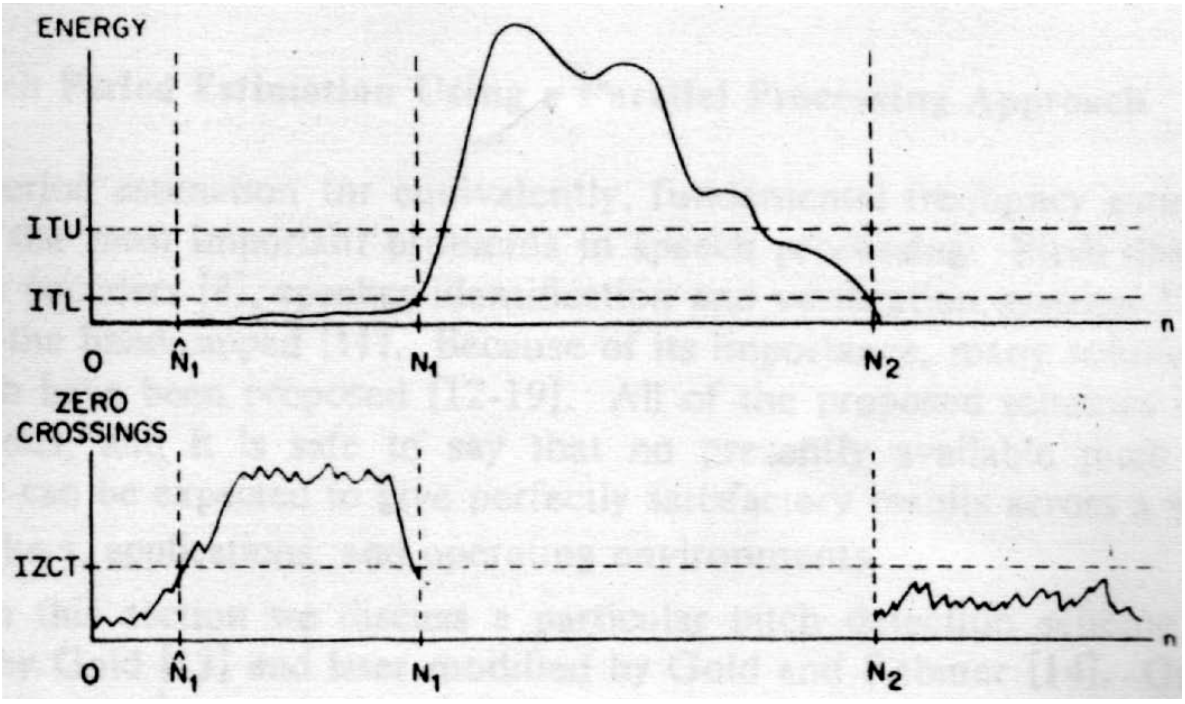


Figure 2. 4 Energy and zero crossing rate algorithm for end points detection [5].

The average magnitude profile is searched to find the interval in which it always exceeds a very conservative\e threshold (ITU in Figure 2.4). It is assumed that the beginning and ending points lie outside this interval. Then working backwards from the point at which energy magnitude first exceeded the threshold ITU, the point (labeled N1 in figure 2.4) where energy first falls below a lower threshold ITL is tentatively selected as the beginning point. A similar procedure is followed to find the tentative endpoint N2. This double threshold procedure ensures that dips in the average magnitude function do not falsely signal the endpoint. At this stage it is reasonably safe to assume that the beginning and ending points are not within the interval N1 to N2. The next step is to move backward from N1 (forward from N2) comparing the zero-crossing rate to a threshold (IZCT in figure 2.4) determined from the statistics of the zero-crossing rate for the background noise. This is limited to

the 25 frames preceding N1 (following N2). If the zero-crossing rate exceeds the threshold 3 or more times, the beginning point N1 is moved back to the first point at which the zero-crossing threshold was exceeded. Otherwise N1 is defined as the beginning. A similar procedure is followed at the end.

The above discussion is briefly introducing the famous method for end points detection. There are many advantages using this method. As it can be seen, it depends on a very simple mathematical basis so that this method is widely used in most of speech applications. It gives good results especially in case of high signal to noise ratio or medium signal to noise ratio.

The disadvantages of this method are that it degrades dramatically in case of highly noise environment (S/N < 16 dB.). It also needs to apply at the whole speech sample before any further processing so that it is not suitable for real time applications. The computation of the end points pass through two mathematical phases one from energy and the other from zero-crossing rate.

The wavelet transform gives an alternative method that combines the energy and zero-crossing rate in one step. Although it is more complex in understanding and calculation but it is fast and can be implemented using fast algorithms such that of fast Fourier transform. The wavelet transform as seen before is splitting the speech sample using quad filters into many frequency channels. We can see the frequency contribution for the signal in different bands along the period of time that contains the speech.

2.3 End points detection using wavelet correlation of wavelet features [69]

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 tools that are used in the signal processing field [10-22]. The wavelet transform extracts the frequency contents of the signal as similar to the Fourier transform do, but it links the frequency domain with the time domain [6]. This link between the time and the frequency gives this transform its powerful characteristic for the determination of the boundaries of frequency-band-defined signals such as the speech signal. 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.

2.3.1 The proposed algorithm

Figure 2.5 is a speech signal of Arabic word “ همس”. The word contains a whisper consonant /h/ at the start and unvoiced fricative /s/ at the end. There are a silence periods before and after the signal. The start and end of this sample is hard-to-detect in case of low signal to noise ratio. Figure 2.6 is the wavelet-based energy function of figure 2.5. As shown in figure 2.6, the energy changes can easily be detected. Correlating the energy contents of the same signal in two different frequency bands generates the curve as shown in figure 2.6.

The algorithm of detecting the end points from this curve is divided into Three parts:

1- Correlation model: The correlation model is obtained from the correlation between wavelet windows. Win(5) and win(6) are selected for correlation. As shown in table 1.1, win(5) covers frequency band 172-344Hz with resolution of 32 parameters and win(6) covers the frequency band 344-689 Hz with resolution of 64 parameters. Most of speech power is concentrated below 1000 Hz[5]. So the above two bands are selected because they have the minimum number of parameters beside they are in the middle of the range of frequencies below 1000 Hz. The two windows are selected adjacent to insure that the power curves will be alike as much as possible. This is important to get the correlation information. Moreover, the crosscorrelation is used rather than the autocorrelation of one window to get the highest immunity to noise. To illustrate this point, if the speech sample is weak in one window (due to noise strike) it may be strong in the adjacent window. For the above two reasons 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 (~ 92 ms in case of 11025 Hz sampling rate). 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 2.6). Figures 2.5 and 2.6 show how far the energy correlation model tracks the boundaries of the speech signal

2-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) to monitor the rate of change. The maximum of the first 220 points of the moving standard deviation is multiplied by 4 and taken as a threshold to discriminate the noise from the speech. This threshold is obtained after many trials. In the noise there is no correlation between windows so the rate of change is very small.

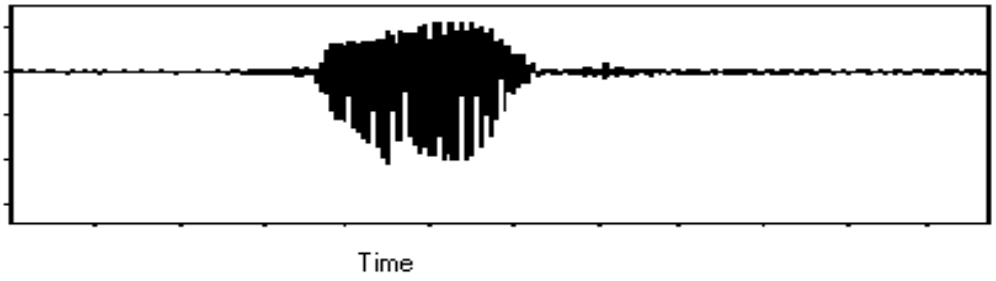


Figure 2.5 Speech signal contains a whisper consonant /h/ at the start and unvoiced fricative /s/ at the end. There are a silence periods before and after the signal. The word is هس in Arabic. This word is pronounced /h//Θ//M//σ/

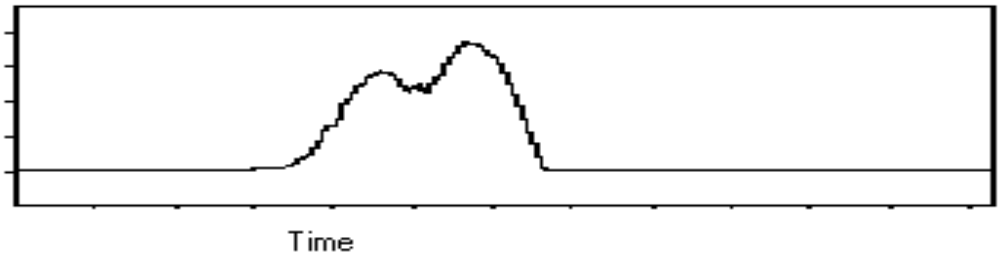


Figure 2. 6 The correlation model. (The crosscorrelation parameters are concatenated)

3-Logical series: The moving standard deviation is applied over the whole speech utterance. The standard deviation points are compared with the noise threshold generated in the first step. The logical series, a series which, contains 1's and 0's only where the number of ones and zeroes equals to speech samples. The element in the series takes 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. Figure 2.7 gives an example of some speech utterances and the markers of the logical series after the application of the proposed algorithm.

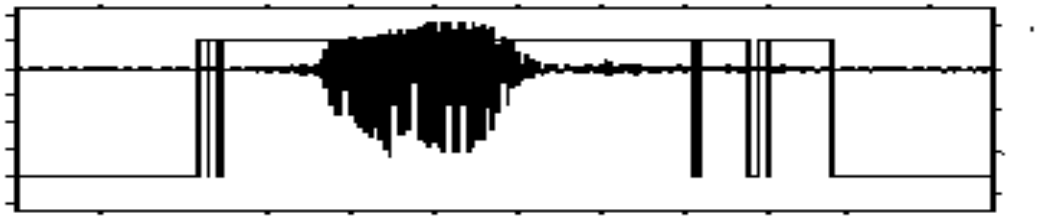


Figure 2.7 The speech signal and the logical series markers. The first and the last markers represent the speech boundaries.

Consider the following definitions:

- ***Win (n)*** The wavelet window which has 2^n parameters according to table1.1.
- ***R (n)*** The crosscorrelation's parameters number n that indicates the correlation between win (5) and win (6) in table 1.1. It indicates the correlation at t±n. 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). R(0) is the energy of speech frame.

2.3.2 System performance in case of noise

To study how far the previous algorithm is valid in case of noise, the normal distribution noise is generated to superimpose it on speech signal. The noise is multiplied with different values to control the signal to Noise ratio.

After applying the previous algorithm on the noisy speech the following results are obtained as shown in figures 2.8 and 2.9.

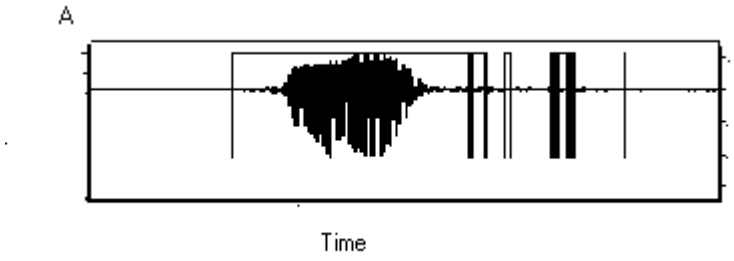


Figure 2. 8 The speech signal and logical series markers in case of 48 dB signal to noise ratio.

The markers still detect the boundaries of speech signal.

Figure 2.9 indicates that in case of 16 dB S/N the last point is shifted left and the starting point is still acceptable.

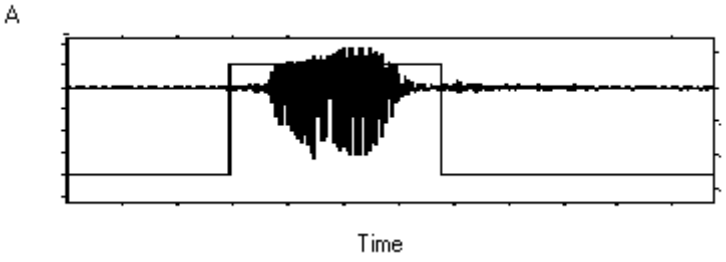


Figure 2.9 the speech signal and logic series markers in case of 16 dB signal to noise ratio.

In figure 2.10 the weak plosive /k/ at the beginning and the nasal /n/ at the end is detected accurately. This speech signal is acquired in normal noise condition not in laboratory conditions.

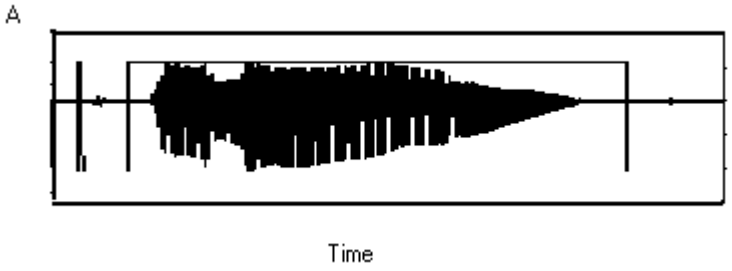


Figure 2.10 The word contains a weak plosive at the beginning /k/ and a nasal at the end /n/. The word is كمان in Arabic and it is pronounced /κ//Θ//Μ//Θ//ν/

The case of weak fricatives at the beginning or end is previously illustrated in the previous section.

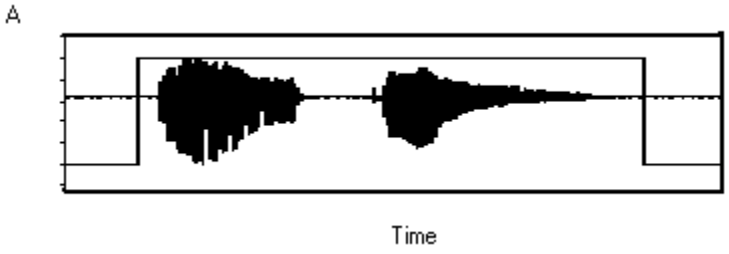


Figure 2. 11 The word contains a voiced fricative at the end of utterance /z/.The word is منتزه in Arabic and it is pronounced /Μ//ο//ν//τ//Θ//ζ//αΗ/

Figure 2.12 summarizes the overall system performance in case of noise.

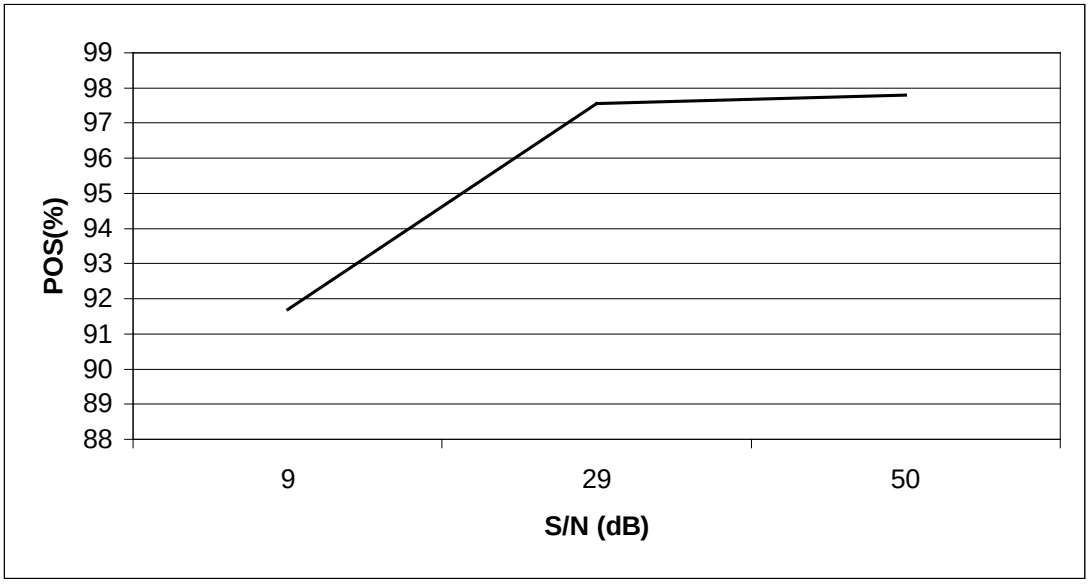


Figure 2.12 EPD'S System performance with correlation of win(5) and win(6).

POS is Probability Of Success.

POS is calculated by measuring how far the logical markers matches the actual pre-calculated markers. A tolerance of 5 ms is taken into consideration.

As shown in figure 2.12, the system indicates a good noise immunity in case of low signal to noise ratio. The system performance is the same for signal to noise ratios from 29dB up to large values and degrades slowly when we go lower than 29 dB. The performance at 9 dB is about 91 %.

2.4 End points Detection using wavelet transform and Neural Network as a classifier

The previous section illustrates how far the wavelet transform can success for extracting the speech signal from the background noise. In the previous

section, all information about speech signal is extracted from two wavelet windows only. The other windows are omitted. In this section all available windows are included. The decision of speech boundaries will be taken via neural network. The neural network takes the input from all windows and gives the decision of speech or nonspeech.

*2.*4.1 Neural Network design

Neural network of our concern will take its information from six different wavelet channels and the decision will be either speech or nonspeech. So NN will have a six nodes in the input layer and only one node in the output layer. The hidden layer is assumed to be 20 nodes.

2.4.2 Training data preparation

In this phase, data is collected from speech and prepared into input output vectors to train NN. A speech of about 20sec contains many words and silence is captured. The speech signal is segmented into smaller windows each of 1024 samples. Wavelet transform is applied on all windows. Wavelet parameters are interpolated into all wavelet channels so that each wavelet channel contains 1024 wavelet parameters. To trace energy changes in each wavelet channel, each channel is prepared such that the following equation.

$$B_N = \frac{\sum_m \left| W_N(m) \right|}{200} \qquad \textbf{(2. 2)}$$

m : Moving index. It takes a sequence of 200 samples starting at the first sample in Wavelet channel and ending at the last point in wavelet channel.

N: Wavelet channel index. It takes values from 1 to 6.

B: Moving-average-wavelet-series name of band index N.

W: Wavelet-series name of raw wavelet parameters in band index N.

The input vector is constructed from the following equation:

$$V_i = \left(B_1^i, B_2^i, B_3^i, B_4^i, B_5^i, B_6^i, D \right) \downarrow_{50} \qquad \textbf{(2. 3)}$$

i: Index of sample within B. Note size of B is the same as size of speech sample.

V: Training vector.

$\downarrow_{50}$: Decimate sequence by a factor of 50 samples ~ 5ms(sampling rate is 11025 Hz). i.e. a training vector is assembled every 5ms of training speech.

D: Desired output value which in this case either 0 for nonspeech and 1 for speech. The decision is made according to spectrogram and listening (see figure 2.13).

Training vectors are introduced to NN. The network is sensitively trained to avoid overtraining i.e. a test is made every about 5 traces over a complete training set.

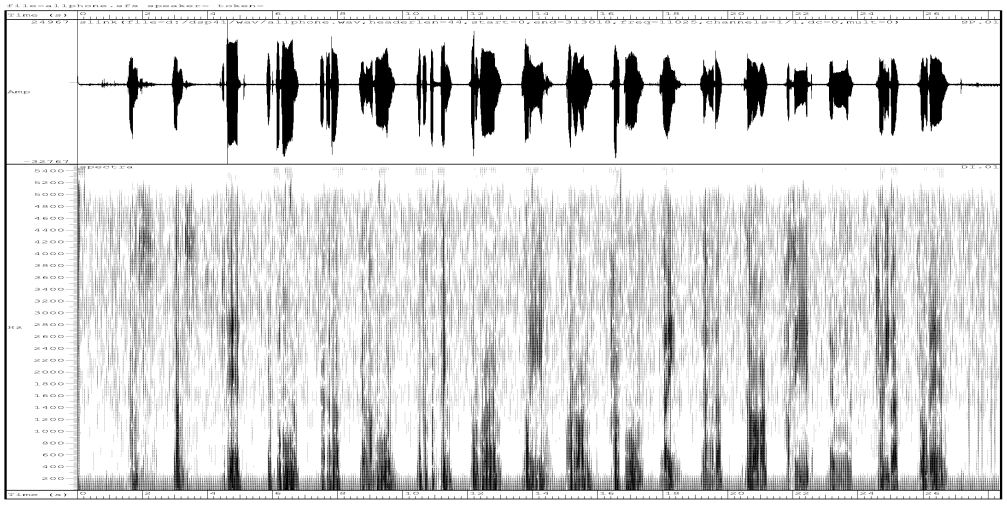


Figure 2.13 Speech data and spectrogram.

2.4.3 Testing NN with prepared test data.

In this phase, many speech samples are used to test the system (about 3 minutes). Vectors such that of the previous section vectors are constructed from a different speech file contains words and silence periods. The vectors are delivered to the input of the trained NN

Results are summarized in figure 2.14. System performance degraded dramatically in case of low signal to noise ratio.

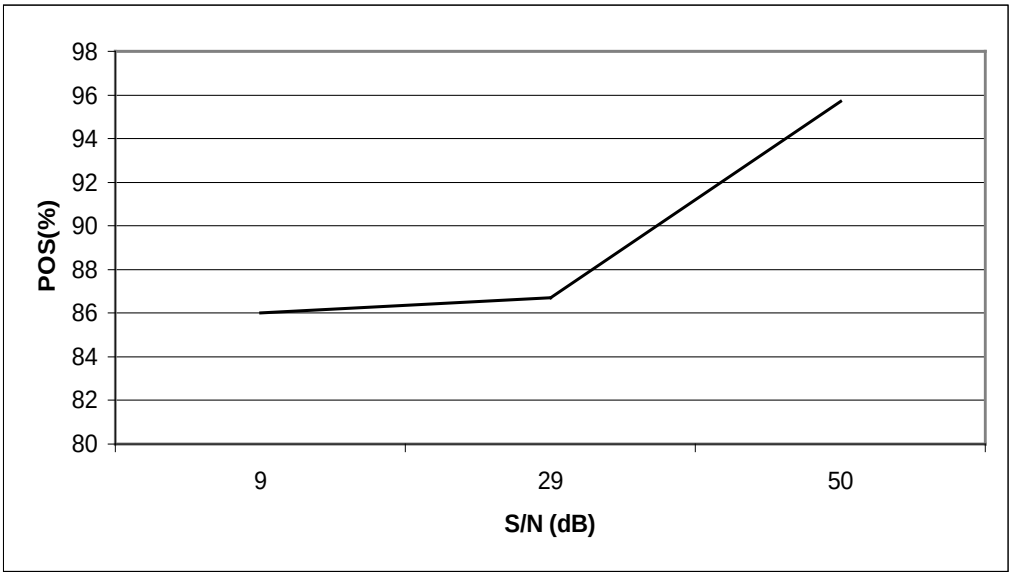


Figure 2.14 EPD's system performance in case of noise. Neural Network is used as a classifier.

As shown in figure 2.14, system performance is degraded slightly in case of S/N ratio less than 50 dB. In case of high S/N ratio it gives a good system performance. This indicates that if the neural network is chosen as a classifier, the training phase must cover the noisy environments. This makes it more complex to learn a single neural network to make the same decision for widely variant process (High S/N and low S/N). So it is decided to design different neural networks for different S/N ratios.

Figure 2.15 introduce an example of applying NN in EPD. In figure (2.15 a) the speech signal represents the Arabic word كتاب is captured. Markers that indicate the speech regions in the original speech signal are indicated in figure (2.15 c). Figure (2.15 b) is the speech data extracted from the original speech according to EPD markers.

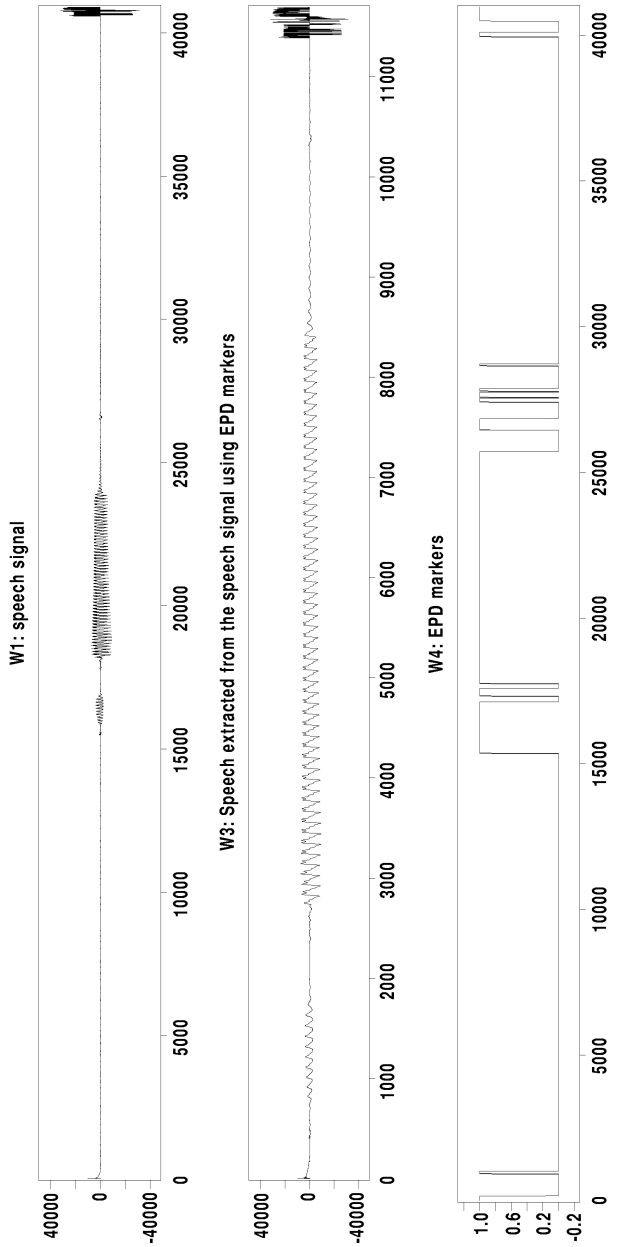


Figure 2. 15 EPD using neural network for Arabic word كتاب. a-Speech signal, b-Raw speech from the speech signal in a, c- EPD markers (high in speech regions).

2.5 Mathematical classifier

A training data is prepared for regression process. A training period of 20 sec of speech and silences is used to prepare the training data set.

Wavelet parameters are extracted, interpolated and smoothed as previous method. The bands under study are six bands so that a single piece of information is taken from each band. X-vector of 6 elements, each is a smoothed-interpolated-wavelet parameter from a single band, is constructed. The corresponding Y-output of X-vector is 0 in case of silence or 1 in case of speech. The following table is constructed from subsequence of X-vectors and corresponding Y-outputs.

X						Y
B0	B1	B2	B3	B4	B5	
54000	30200	2230	1000	650	120	1 or 0

Y is regressed on X to find the mathematical model of the system.

Equation (2.4) represents the system equation. [B] Matrix is the system model that is obtained from training as discussed above. [X] Matrix is the input speech signal after preparation (Smoothed-interpolated-wavelet parameters from the six bands). [Y] Matrix is the output decision.

$$\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & X_{03} & X_{04} & X_{05} & X_{06} \\ \vdots & & \vdots & & \vdots & \\ X_{m1} & X_{m2} & X_{m3} & X_{m4} & X_{m5} & X_{m6} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \qquad \textbf{(2. 4)}$$

After training the system Matrix is:

$$[B] = \begin{bmatrix} 0.0031 \\ 0.0012 \\ 0.0036 \\ -0.0253 \\ 0.0332 \\ 0.0033 \end{bmatrix}$$

To Evaluate the efficiency of this method a test data from the database is applied on the system matrix according to equation (2.4) with different signal to noise ratios. Figure 2.16 summarizes the output results.

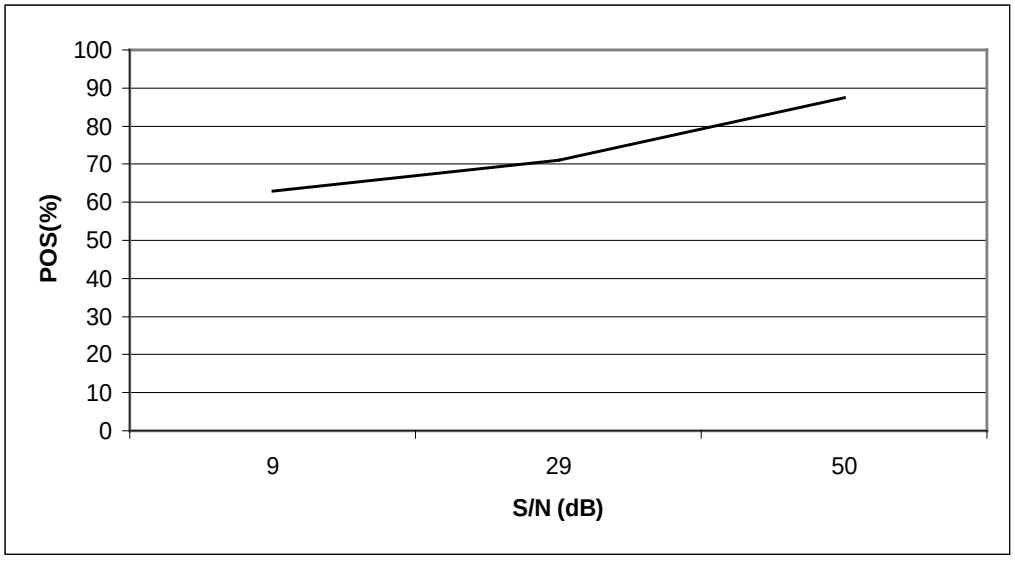


Figure 2. 16 EPD'S system performance. Mathematical regression is used as classifier.

The system of figure 2.16 behaves like neural network-based classifier system. Actually this result is expected because this system depends on training data. The system gives a good results in case of high S/N and the performance degrades for lower S/N ratios (< 50 dB).

Figure 2.17 a and b, illustrate EPD markers using the mathematical classifier. In fig. 2.17 a, the word begin with plosive /k/ and end with plosive /b/. In fig. 2.17 b, the word begin with whispering /h/ and end with fricative /s/.

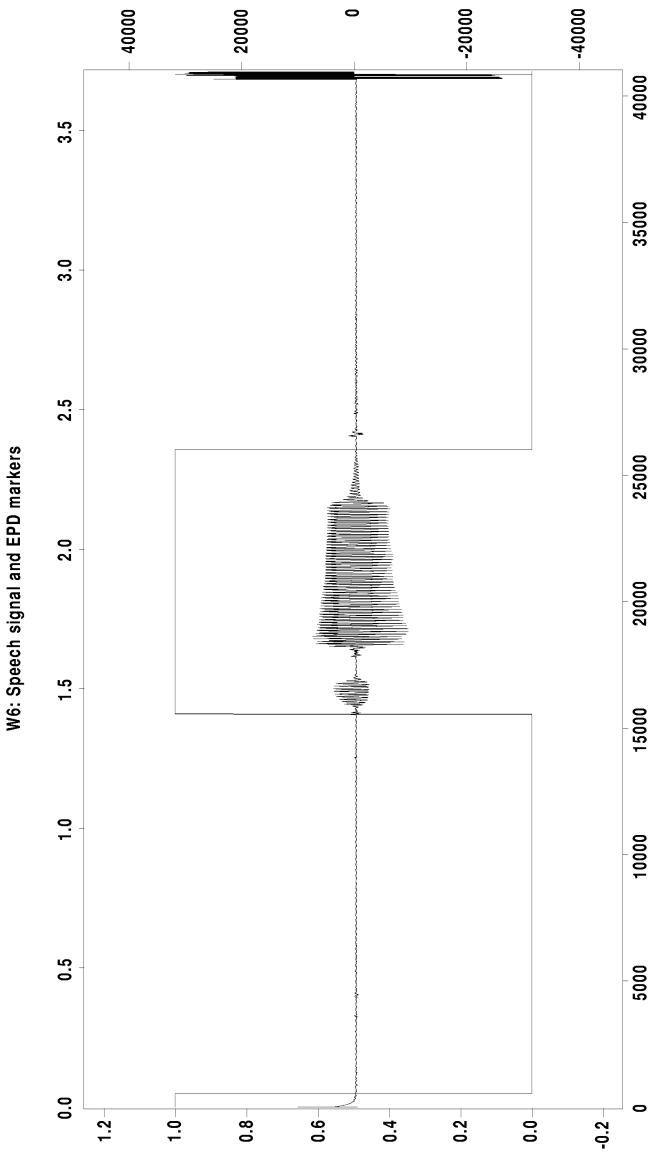


Figure 2. 17 a- EPD markers of speech signal كتاب.
Mathematical classifier is used.

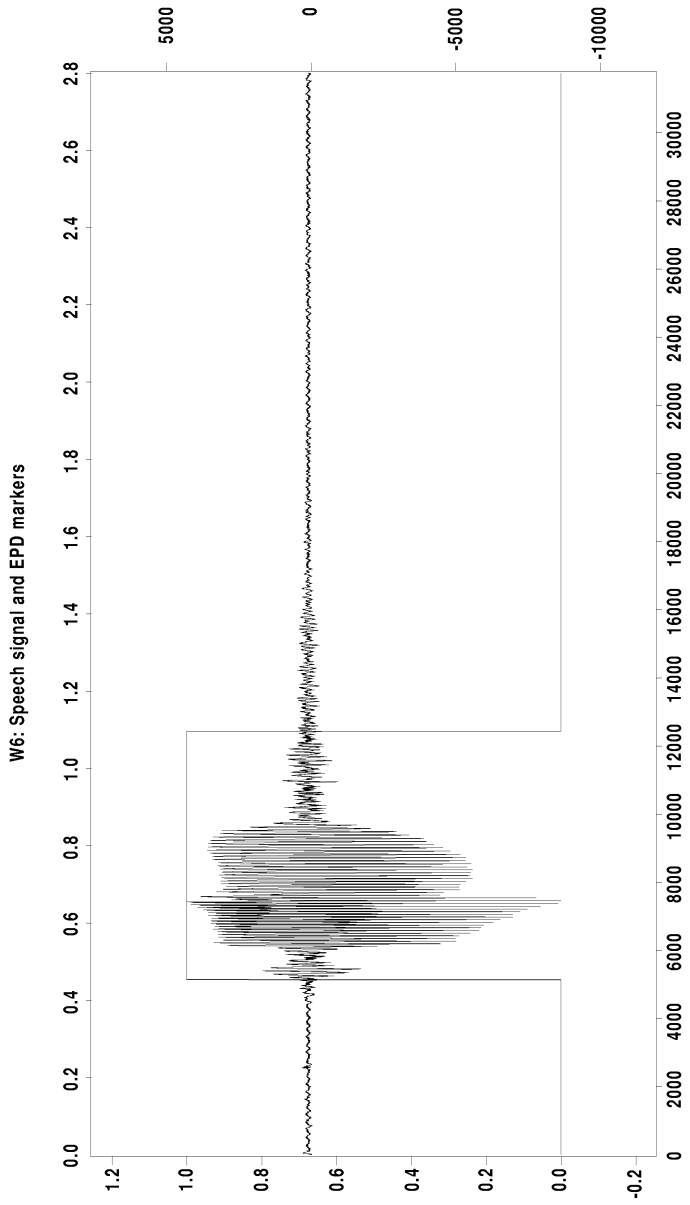


Figure 2. 17 cont. b- EPD markers of speech signal شمس.
Mathematical classifier is used.

2.6 Conclusion

It is clear that from the above discussion the wavelet transform can be used efficiently in EPD problem. The problem is treated with several methods. The first one based on manipulating the speech signal itself to find a threshold for

EPD calculations. This method gives a good performance over a wide range of S/N ratios. But it needs some pre analysis for noise threshold calculation. It can be used in the applications where the speed is not a critical factor.

The last two methods of classifier are extremely alike. They are based on training the system then finding a model. They can be used in systems that have relatively stable environment (approximately constant S/N ratio). They are faster than the first method because no extra calculations are needed.

The mathematical-based method is faster than the neural network-based method because the mathematical operations needed to find the output are less than those of the neural network.

2.1 INTRODUCTION 36

2.2 ENERGY AND ZERO CROSSING RATE METHOD 37

2.3 END POINTS DETECTION USING WAVELET CORRELATION OF WAVELET FEATURES 42

2.3.1 THE PROPOSED ALGORITHM 43

2.3.2 SYSTEM PERFORMANCE IN CASE OF NOISE 46

2.4 END POINTS DETECTION USING WAVELET TRANSFORM AND NEURAL NETWORK AS A CLASSIFIER 49

2.4.1 NEURAL NETWORK DESIGN 50

2.4.2 TRAINING DATA PREPARATION 50

2.4.3 TESTING NN WITH PREPARED TEST DATA. 52

2.5 ..... MATHEMATICAL CLASSIFIER 54

2.6 ..... CONCLUSION 59

Chapter 3

Classification of voiced/unvoiced utterances and pitch period estimation

3.1 Introduction

Speech classification is one of the basic points in speech processing. Speech signals are composed of a sequence of sounds. These sounds and the transitions between them serve as a symbolic representation of information. The arrangement of these sounds (symbols) is governed by the rules of language. The study of these rules and their implications in human communication is the domain of *linguistics* and the study and classification of the sounds of speech is called *phonetics.* A detailed discussion of phonetics and linguistics would take us too far afield. However, in processing speech signals to enhance or extract information, it is helpful to have as much knowledge as possible about the structure of the signal; i.e., about the way in which information is encoded in the signal.

The following section deals with the problem of classifying the speech signal into voiced or unvoiced sound. This problem is handled by different methods.

3.2 Voiced / unvoiced classification

3.2.1 Voiced sound versus unvoiced sound

The Speech sounds can be classified into 3 distinct classes according to their mode of excitation. **Voiced sounds** are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract[5].

Forming a constriction at some point in the vocal tract (usually toward the mouth end), and forcing air through the constriction at a high enough velocity to produce turbulence generates fricatives or **unvoiced sounds**. This creates a broad-spectrum noise source to excite the vocal tract.

Plosive sounds result from making a complete closure (again, usually toward the front of the vocal tract), building up pressure behind the closure, and abruptly releasing it.

The vocal tract and nasal tract are shown in Figure 3.1 as tubes of non-uniform cross-sectional area. As sound is generated, it propagates down these tubes, the frequency spectrum is shaped by the frequency selectivity of the tube. This effect is very similar to the resonance effects observed with organ pipes or wind instruments. In the context of speech production, the resonance frequencies of the vocal tract tube are called formants. The formant frequencies depend upon the shape and dimensions of the vocal tract; each shape is characterized by a set of formant frequencies. Varying the shape of the vocal tract forms different sounds. Thus, the spectral properties of the speech signal vary with time as the vocal tract shape varies.

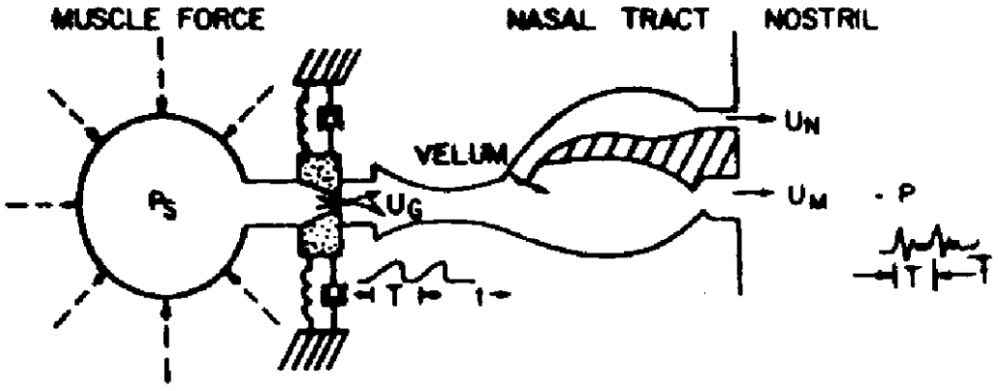


Figure 3.1 Schematics of vocal tract system

As shown in figure 3.1, T is the pitch period in case of voiced sound. U_G is the generated velocity function (excitation of the vocal tract tube). U_M and U_N are the output speech velocity function from mouth and nose respectively.

The following section will discuss the differences between voiced and unvoiced sounds in terms of energy and frequency contents.

3.2.2 Signal characteristics of voiced and unvoiced sounds

The underlying assumption in most speech processing schemes is that the properties of the speech signal change relatively slowly with time. This assumption leads to a variety of "short-time" processing methods in which short segments of the speech signal are isolated and processed as if they were short segments from a sustained sound with fixed properties. This is repeated (usually periodically) as often as desired. Often these short segments which are some-times called analysis frames, overlap one another. The result of the processing on each frame may be either a single number, or a set of numbers.

We have observed that the amplitude of the speech signal varies appreciably with time. In particular, the amplitude of **unvoiced** segments is generally much lower than the amplitude of **voiced** segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations. In general, we can define the short-time energy as

$$En = \sum_{m=n-N+1}^{n} x(m)^2 \qquad \textbf{(3. 1)}$$

Where N is the window length or frame length. If N is too small, i.e., on the order of pitch period or less, E_n will fluctuate very rapidly depending on the exact details of the waveform. If N is too large, i.e., on the order of several pitch periods. E_n will change very slowly and thus will not adequately reflect the changing properties of the speech signal. Unfortunately this implies that no single value of N is entirely satisfactory because the duration of a pitch period varies from about 20 samples (at a 10 kHz sampling rate) for a high pitch female or a child, up to 250 samples for a very low pitch male. With these conditions in mind, a suitable practical choice for N is on the order of 100-200 for a 10 kHz sampling rate (i.e., 10-20 msec duration).

Figures 3.2 and 3.3 show the effects of varying the duration of the window (for the rectangular and Hamming windows, respectively) on the energy computation for the utterance /What, she said/ spoken by a male speaker. It is readily seen that as N increases, the energy becomes smoother for both windows.

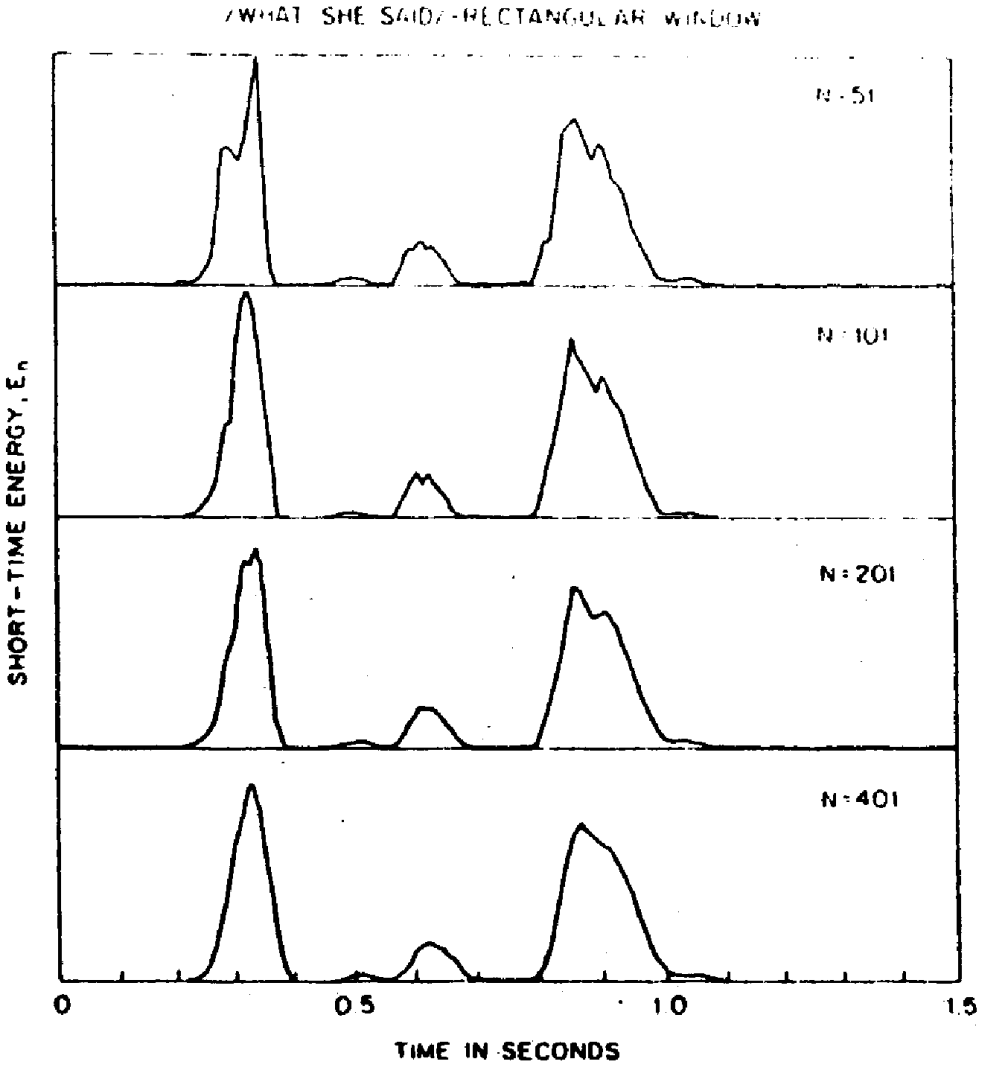


Figure 3. 2 Energy distribution for Rectangular weighted frames for different frame lengths[5].

The major significance of E_n is that it provides a basis for distinguishing voiced speech segments from unvoiced speech segments. As can be_seen in Figures 3.2 and 3.3, the values Of E_n for the unvoiced segments are significantly smaller than for voiced segments. The energy function can also be used to locate approximately the time at which voiced speech becomes

unvoiced, and vice versa, and for very high quality speech (high signal-to-noise ratio), the energy can be used to distinguish speech from Silence [5] as was shown in chapter 2.

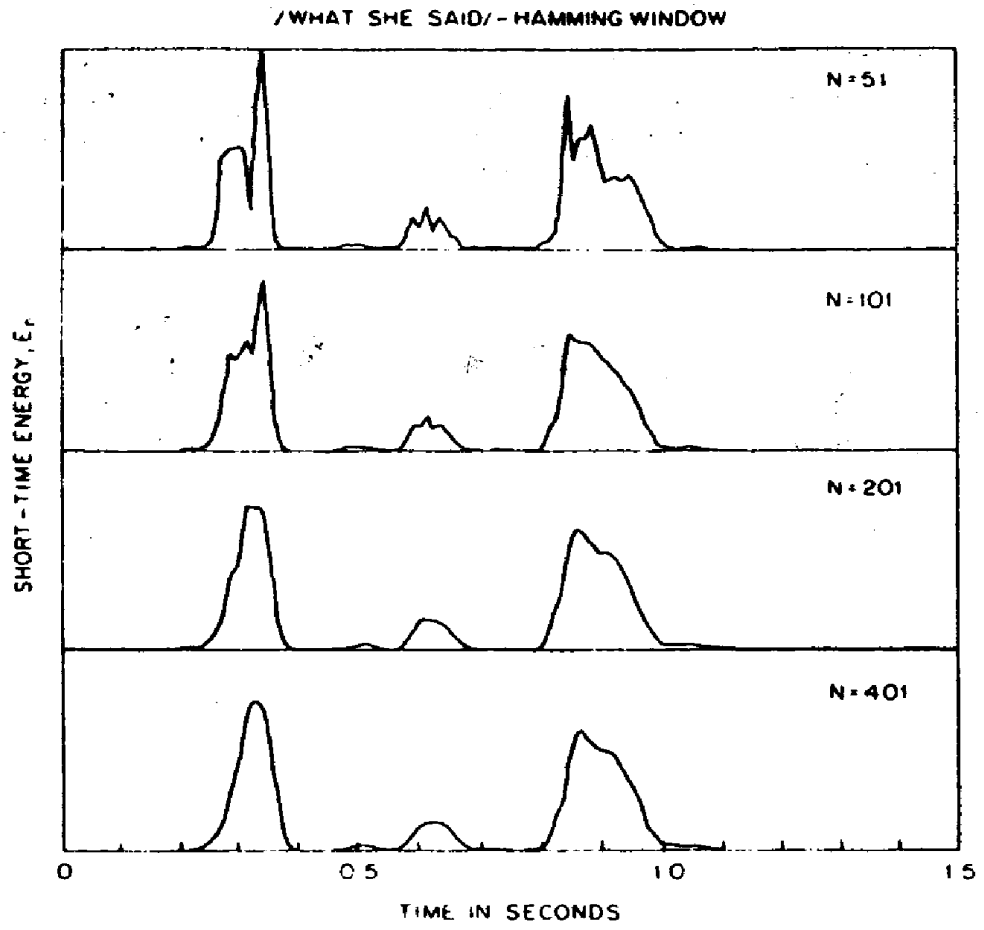


Figure 3. 3 Energy distribution for Hamming weighted frames for different frame lengths[5].

The above discussion illustrates the speech signal properties from the energy point of view.

Now let us see how the short-time average zero-crossing rate applies to speech signals. The model for speech production suggests that the energy of

voiced speech is concentrated below about 3kHz because of the spectrum fall-off introduced by the glottal waveform whereas for unvoiced speech, most of the energy is found at higher frequencies. Since high frequencies imply high zero-Crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero-crossing rate and energy distribution with frequency. A reasonable generalization is that if the zero-crossing rate is high the speech signal is unvoiced, while if zero-crossing rate is low, the speech signal is voiced. This, however, is a very imprecise statement because we have not said what is high and what is low, and, of course, it is really not possible to be precise. Figure 3.4 shows a histogram of average zero-crossing rates (averaged over 10 msec) for both voiced and unvoiced speech. Note that Gaussian curve provides a reasonably good fit to each distribution. The mean short-time average Zero-crossing rate is 49 per 10 msec for unvoiced and 14 per 10 msec for voiced. Clearly the two distributions overlap so that an unequivocal voiced/unvoiced decision is not possible based on short-time average zero crossing rate alone.

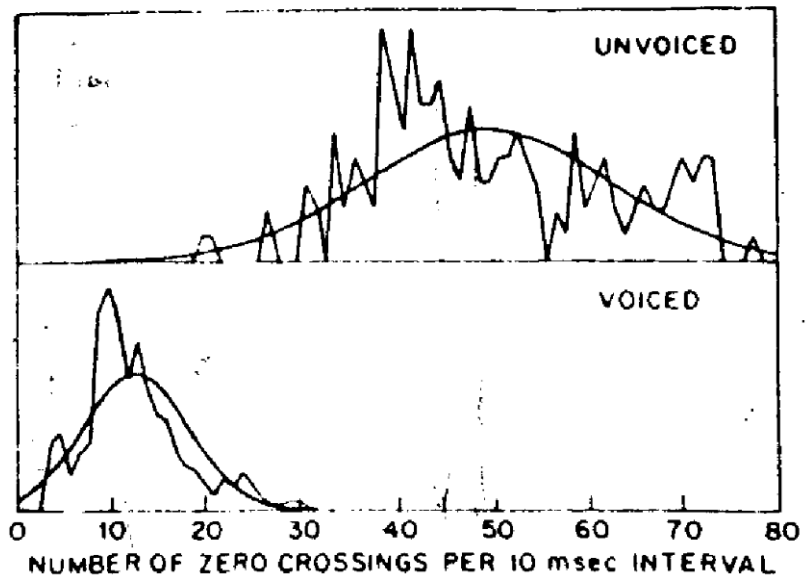


Figure 3.4 **Histogram of average zero-crossing rates over 10 mec for both voiced and unvoiced speech[5].**

3.3 Voiced / unvoiced classification using Dyadic wavelet

3.3.1 Dyadic wavelet

Corresponding to the GCI (glottis closure), the glottal pulse exhibits a peak that can be regarded as a transient phenomenon, a singularity carrying information about the vibration of the vocal folds. Until recently, the Fourier transform was the main mathematical tool for analyzing signal singularities. Unfortunately, the Fourier transform is global and provides only an overall description of the regularity of the signal, not being well adapted to finding the time location and distribution of singularities. This was a major motivation for studying the wavelet transform in mathematics and in applied science domains. The wavelet transform is reforming a decomposition of signals into

elementary building blocks that are well localized both in time and in frequency. The wavelet transform is suitable for characterizing the local regularity of signals [6].

In dyadic form, The wavelet transform of a signal x(t) is defined by the relation:

$$DW(\tau, j) = \frac{1}{2^j} \int_{-\infty}^{\infty} x(t)\Psi^*(\frac{t-\tau}{2^j})dt = x(t) * \Psi^*(t) \qquad \textbf{(3. 2)}$$

where:

τ : The time delay.

j: The scale parameter.

$\Psi^*(t)$: The complex conjugate wavelet function for which:

$$\int_{-\infty}^{\infty} \Psi(u)du = 0 \qquad \textbf{(3. 3)}$$

From a signal processing point of view the Dyadic Wavelet can be considered as the output of a bank of constant Q, octave band, band-pass filters whose impulse response is $\frac{1}{2^j}\Psi(\frac{t}{2^j})$ for each scale 2^j .

Mallat has shown in [7] that if a signal x(t) or its derivatives have discontinuities, then the modulus of the DW of x(t) , | DW (τ, a)| exhibits local maxima around the point of discontinuity at $t=t_0$. So, if we choose a

wavelet function Y (t) that is the first derivative of a smoothing function f (t), then the local maxima of the |DW | will indicate the sharp variations of the signal. This property is used in estimating the instantaneous pitch period, by noting that at the instant of the glottis closure, the speech signal has a discontinuous behaviour, and hence., the |DW| will have maxima. The important difference from other functions that have maxima at the GCI is that these maxima can be detected across several dyadic scales. This fact ensures a better reliability of the method, a multichannel (multiscale) decision being possible. The wavelet transform may be calculated in discrete form with the pyramidal algorithm proposed by Mallat in [7]. The band-pass filter for each scale is made up of a pair of low-pass and high-pass quadrature mirror filters with impulse responses h(k) and g(k). For one scale the processing chain is depicted in figure 3.5: the entire algorithm is represented in figure 3.6. The number of coefficients of the transform decreases for each scale yielding a multiresolution representation.

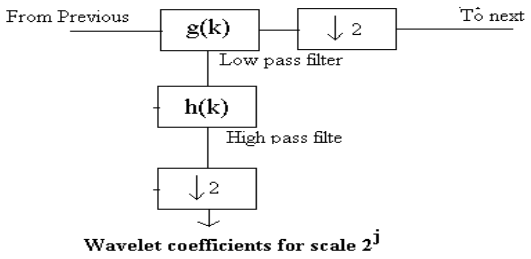


Figure 3. 5 **The basic unit of wavelet transform mechanism (DWT Block in figure** 3**.6).**

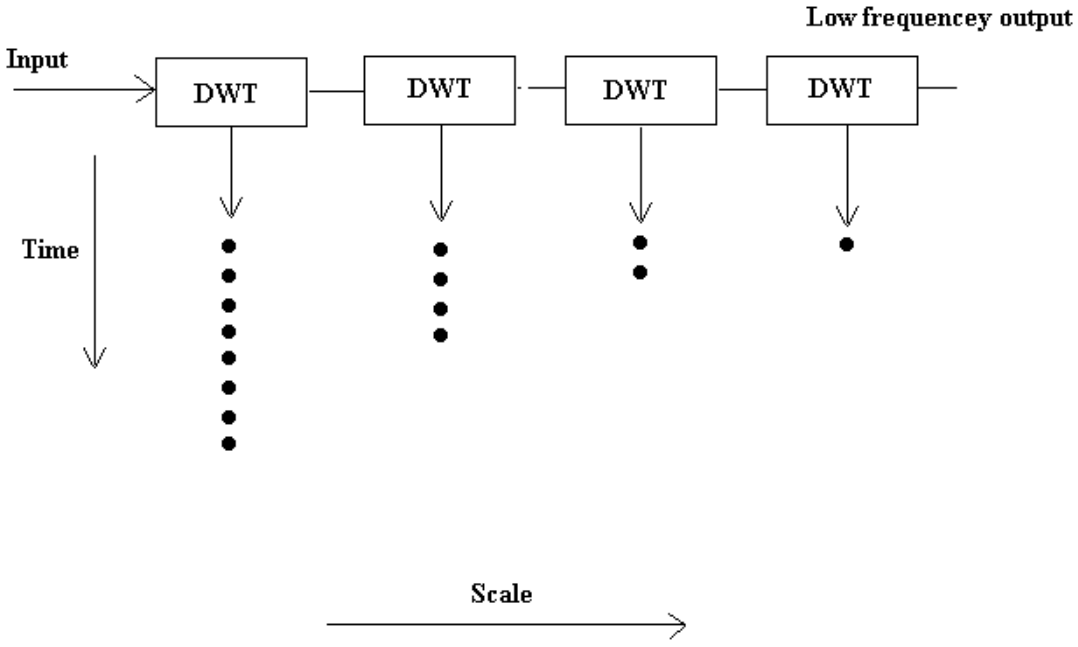


Figure 3.6 Pyramidal algorithm for the processing of Dyadic wavelet transform.

3.3.2 Classification using single band

A band of which the vowels or voiced sounds are dominant in the speech signal is selected for the analysis[48]. Our work is oriented to the Arabic language so the selected words are all in Arabic. The speech samples are digitized with a 16 bit sound card. The sampling rate was 11025 samples per second. The Mathcad[1] software package is used as a platform of all mathematics such as wavelet transform, interpolation ... etc. Window of 1024 samples is used in the analysis. Table 3.1 relates the wavelet coefficients to the according frequency bands.

[1] © 1986-1994 Mathsoft Inc. Version 5.0. © 1993 by Houghton Mifflin Company.

Table 3. 1 Wavelet parameters distribution over the whole frequency band in case of 1024 samples window length and 11025 kHz sampling rate.

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

The frequency band of 172-344 Hz is chosen here for the tracking method. This band is represented by 32 wavelet parameters as shown in Table 1. Daubechies four-coefficient wavelet filter[6] was used in the wavelet analysis.

The unvoiced sound is modeled in speech as a white noise distributed in all ranges and the voiced is modeled by the vocal tract filter excited with a pulse train having a frequency equals to the pitch [4], [3]. The voiced sound is a limited band sound because both the excitation and the vocal tract filter are band limited. The wavelet transform of a given signal may be interpreted as a decomposition of the signal into a set of frequency channels of equal bandwidth on a logarithmic scale.

Most of the speech signal power is contained around the first formants. The statistical results for many vowels of adult, males, and females indicate that the first formant frequency doesn't exceed 1000 Hz and isn’t below 100Hz [4]

approximately. The 172-344 Hz level is chosen for analysis because it has the minimum number of wavelet parameters than the other two levels as shown in table3.1, beside that, it contains most of the speech energy.

The algorithm generates a mathematical function that depends on the wavelet transform and reflects the energy changes along the speech utterance. The first step toward generating this function is introduced in the previous paragraph. In this step the wavelet parameters are extracted. The magnitude of the 32 wavelet parameters in the 172-344Hz band are used to make the appropriate mapping for the power distribution of the speech samples along the analysis time period in this frequency band. The entire analysis period is distributed over those 32 parameters. Each parameter concerns of one window length divided by 32. Time is given by the following formula:

$$tn = \frac{n}{F_s} + m\frac{w}{2} \qquad \textbf{(3. 4)}$$

where:

F_s: Sampling rate.

m: Frame number.

w: Window length in samples.

n: Time index.

The frame number is the number of the analysis window. A 50% overlapping between the analysis windows is implemented. This overlapping is needed to eliminate the error produced from the frame discontinuity. The suffix "n" is the index of the wavelet parameter within the selected band. Each wavelet parameter represents a point in the time-power domain. The x-axis

represents the time and the Y-axis (log scaled) represents the power in figure 3.7-a. A simple interpolation is made to smooth these points by using low pass filter. The generated smoothing tracking function is shown in figure 3.7-b. The characteristics of the low pass filter are:

1- Very narrow bandwidth.

2- Critical edge transition.

3- No ripple in the stop band and flat response in the pass band.

4- Small order as much as possible to insure a good speed in a real-time application.

Figure 3.8 indicates the designed and the implemented digital filter.

The narrow band width is to smooth the curve of figure 3.7-b. The abrupt change in the filter is to eliminate the sudden variations totally. The different manipulations of the pass band components makes reshaping of the slow variations which may give harmful results so that the filter is flat in the pass band.

The tracking function is a level sensitive function, i.e. thresholds will be extracted from it in the training phase. Those thresholds give the information about the unvoiced level.

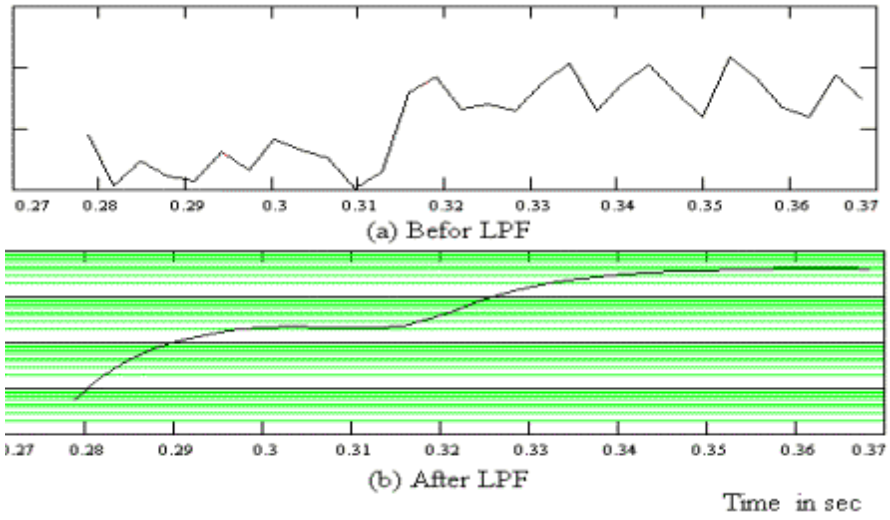


Figure 3. 7 **a-The wavelet parameters before applying to the smoothing filter. b- The wavelet parameters after applying the smoothing filter.**

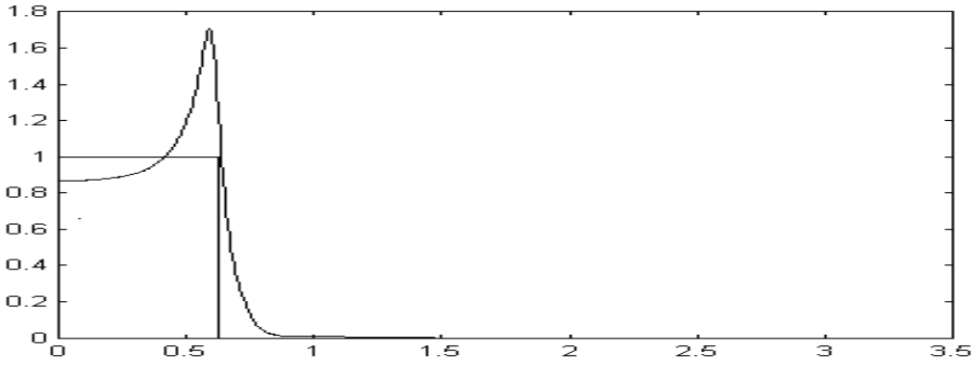


Figure 3. 8 **The proposed digital filter for smoothing process.**

3.3.2.1 Training phase

The process of finding appropriate thresholds is based on statistical data collection. The data is extracted from linking between the time waveform

curve, the listening and the tracking function curve as shown in figure 3.9. This phase of the process is called the training phase.

Assume the following definitions:

- L: the minimum limit that represents the starting of the unvoiced segment.
- U: the upper limit, which can not be exceeded by the tracking function during the unvoiced sound duration.
- Y: tracking function.
- Y_{max}: Maximum statistical value of tracking function
- .$Y_{n:}$ Normalized tracking function.
- Y_{i_max} : Maximum statistical value of tracking function of frame i.
- Y_{i_mean} : average value of tracking function of frame i.
- Min_U_Duration: minimum unvoiced duration.

In the training phase, many speech samples are taken from many speakers (males and females). The tracking function is a power-related function. It depends on the signal level so that the tracking function must be normalized to be a signal level independent function. If the curve goes above the U limit it can not represent unvoiced sound.

The tracking function Y will be normalized with respect to the statistical maximum value Y_{MAX} rather than the absolute maximum. This is because a fatal error can occur if there is a value which is very large with respect to all others due to any error in the process (hazard). If the tracking function (Y_n) is

normalized with respect to this unexpected value, it will give a false information about the signal phonetic levels.

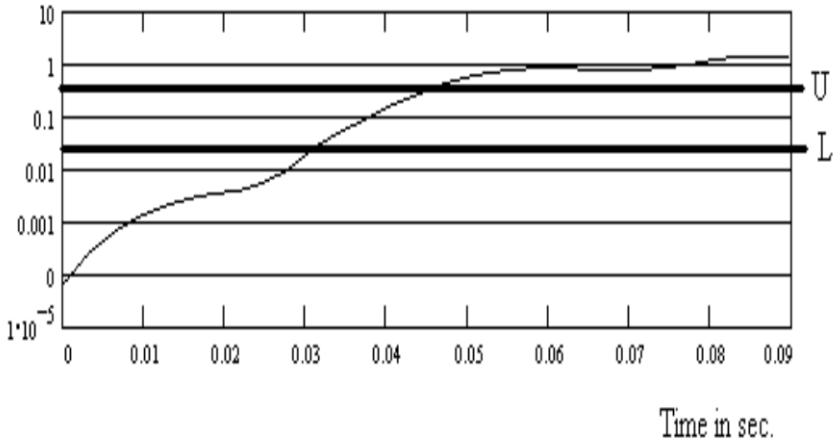


Figure 3. 9 **Upper and lower threshold**

To overcome this error, the statistical maximum value is taken instead of the absolute maximum value. In each frame (analysis window), the mean value and the standard deviation are computed. It is found statistically that the maximum value is:

$$Y_{i\_max} = Y_{i\_mean} + \sigma_i \qquad \textbf{(3. 5)}$$

"i" refers to the frame number.

σ_i : Standard deviation of frame i.

Y_{MAX} is the maximum value of all Yi_MAX.

The above algorithm is applied over the training set to extract "L","U" and Min_U_Duration (Minimum Unvoiced time Duration). According to the experiment results L= 0.1 , U= 0.4 units of the log scale as shown in figure 3.9.

3.3.2.2 Test phase

The automatic tracking algorithm is introduced in figure 3.10. The sampling process is applied to the speech, then the speech samples are divided into frames. The wavelet transform is applied on each window, the wavelet parameters for the tracking function are extracted and applied to the previous low pass filter. Y_n is generated for all frames then it is normalized as described before. Now Y_n can be used for extracting the unvoiced boundaries.

Figure 3.11 a illustrates a comparison between the actual boundaries, which are marked by using the time waveform drawing and listening test, of the unvoiced sounds and the boundaries which are generated from the above tracking algorithm. A rate of **98.7%** of accurate recognition is achieved.

Figure 3.11 b,c and d, illustrate the method in work. As shown in figure (3.11 b) , Arabic word كتاب , W4 contains a curve that represents the regions where Y exceed U limit and W5 contains the curve that indicate the regions where Y exceed L limit. If Y exceeds L then U within certain time as illustrated before then the marker indicating the beginning of voiced segment is generated. If the curve of Y goes below L then unvoiced sound is started.

W2 and W3 in figure 3.11 b,c and d are wavelet parameters in 3D plot and the tracking function Y respectively.

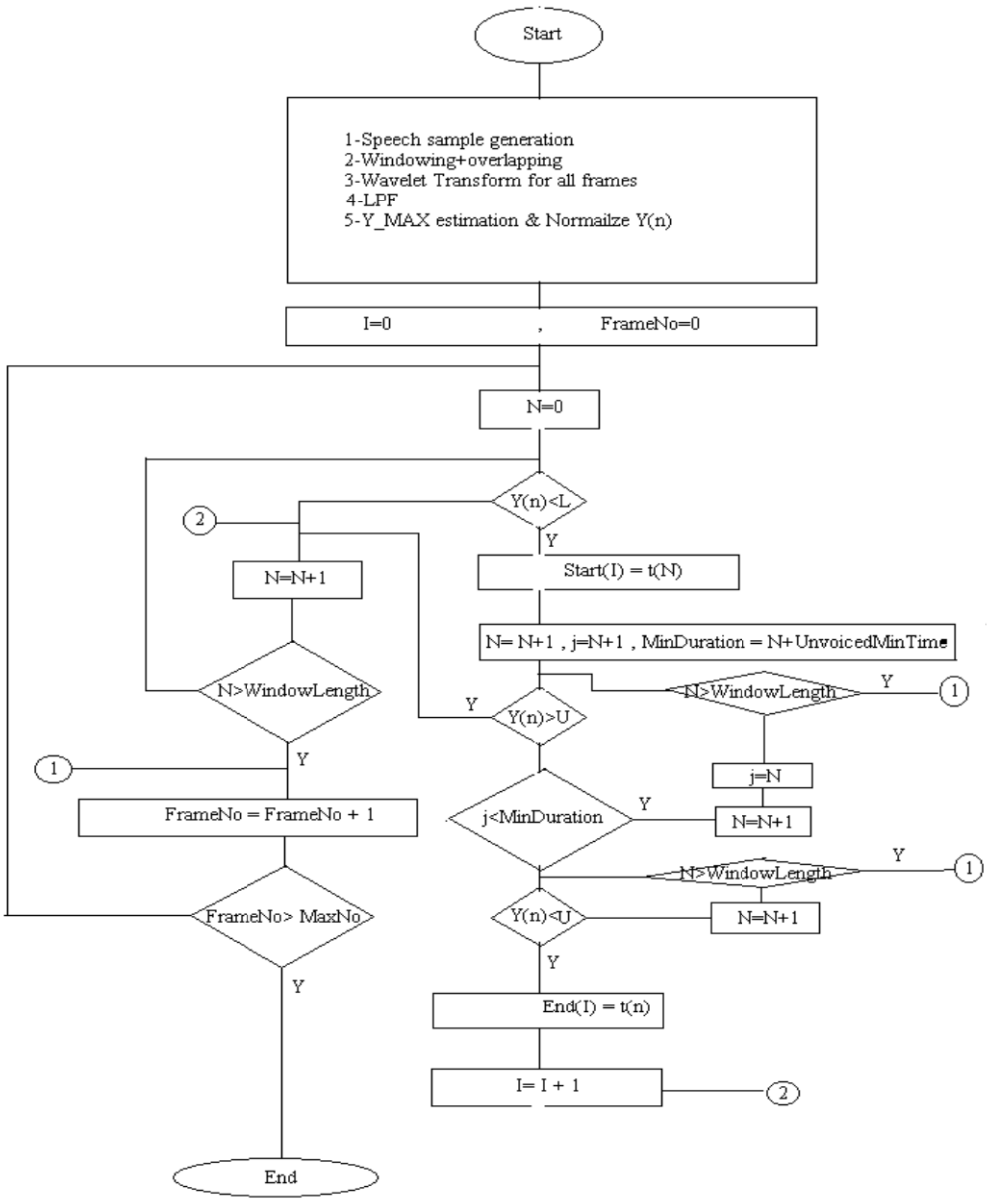


Figure 3. 10 **Flow chart of the automatic tracking algorithm.**

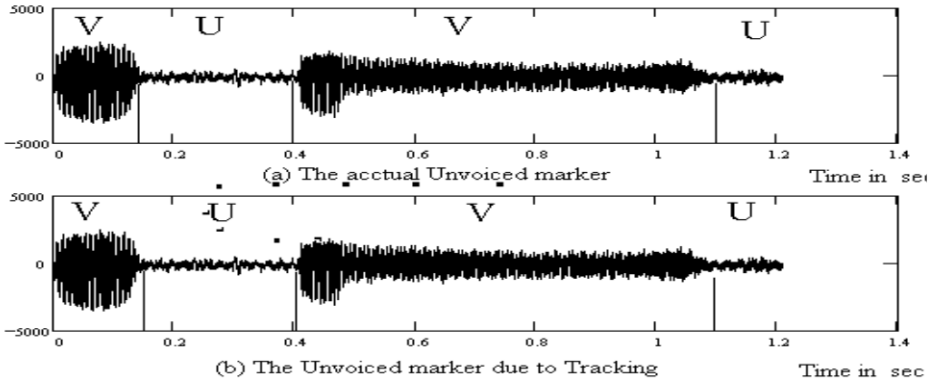


Figure 3. 11 **a -Comparison between the ordinary method and automatic tracking algorithm.**

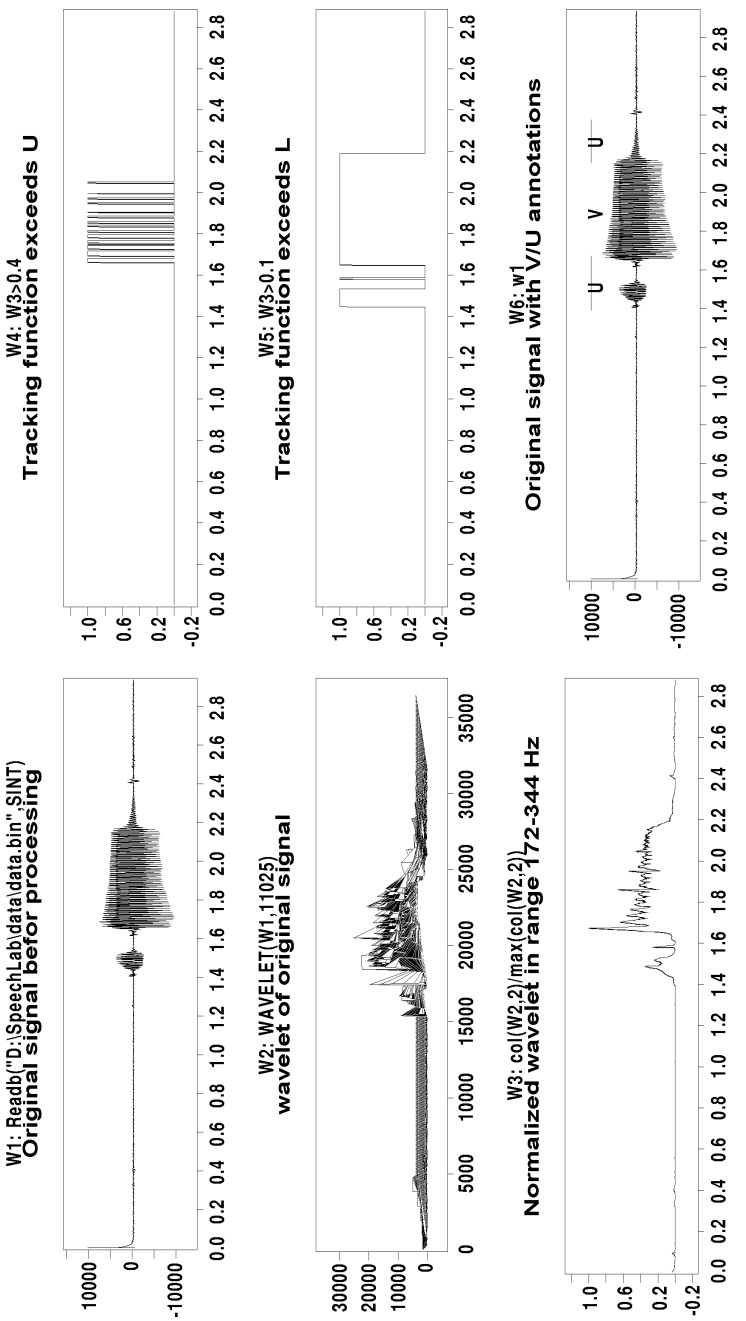


Figure 3. 11**-cont. b- Work sheet represents the traking function method**

The Arabic word is كتاب

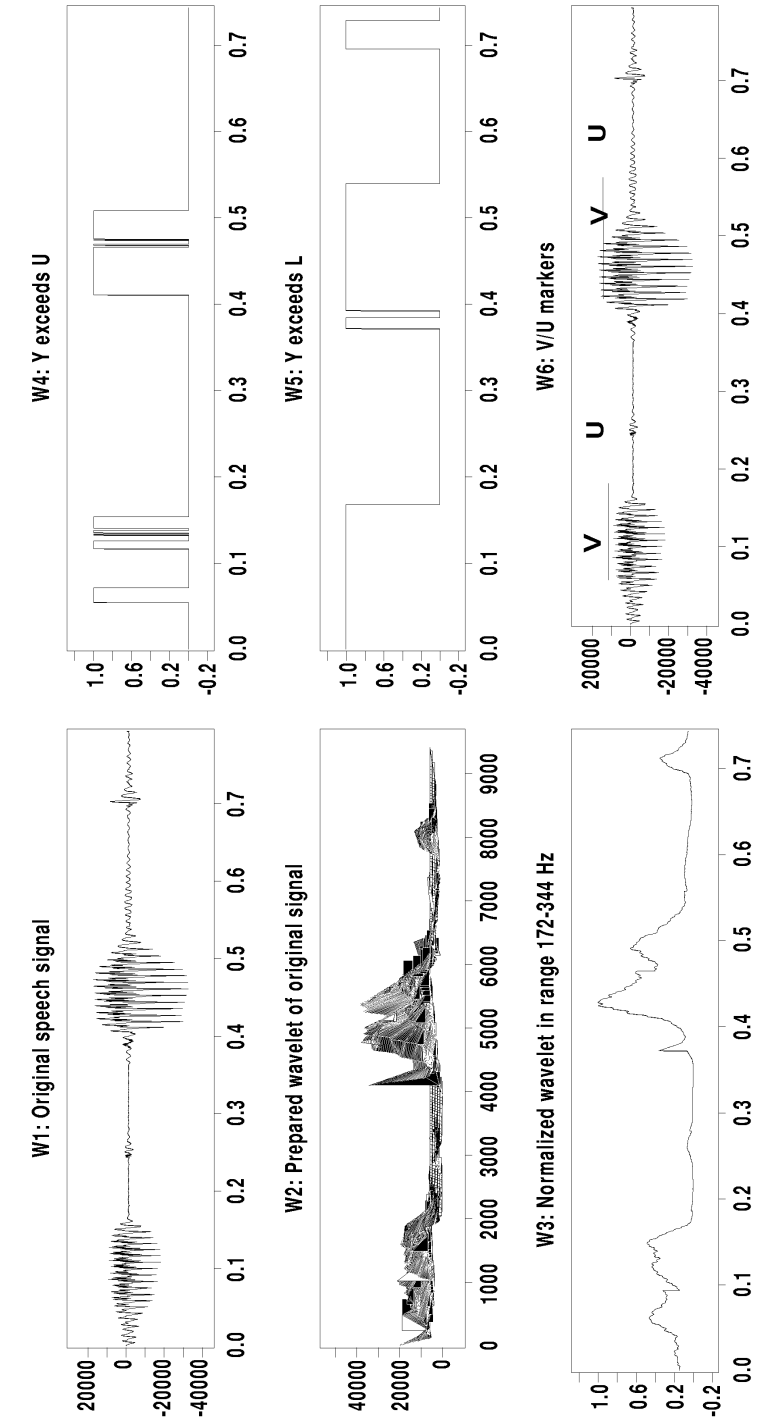


Figure 3. 11**-cont. c- Work sheet represents the traking function method The Arabic word is** يكتب

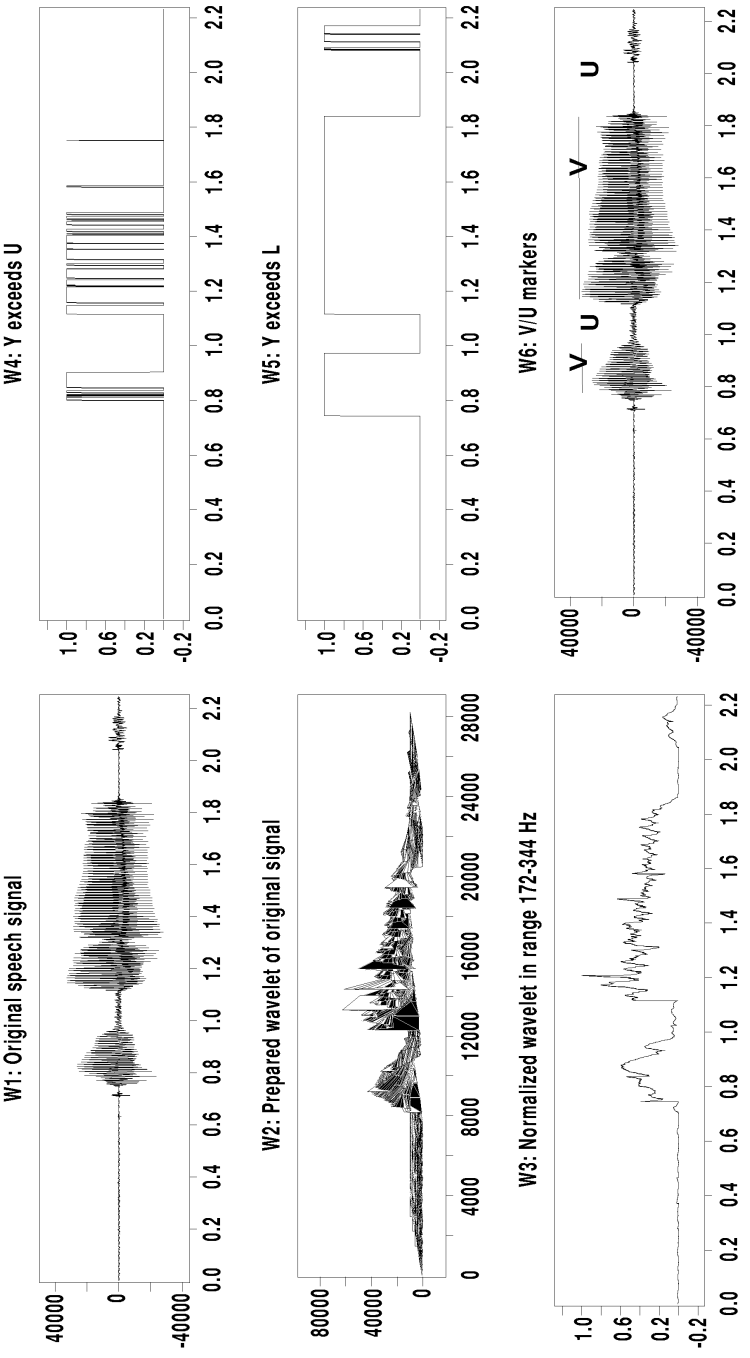


Figure 3. 11**-cont. d- Work sheet represents the traking function method The Arabic word is** إذدراء

3.3.3 Correlation based method

A new method for classifying the speech signal into voiced and unvoiced sounds using the discrete wavelet transform is introduced [49]. The technique is a modified version of the tracking function that is presented in section 3.3.2.

A correlation model that is generated from the wavelet transform of the speech signal is used to make the classification. This way is highly immune to noise. It works with a good accuracy for signals with low signal to noise ratio (less than 9 dB). This way is fast and can be implemented in real time applications.

Figure 3.12 gives a view about how the wavelet transform is powerful in representing the variations of the speech sounds from voiced to unvoiced or from unvoiced to voiced. The figures are constructed by interpolating the wavelet parameters in each frequency level.

The relation between the energy and the frequency clearly appears in figure 3.12. The energy of voiced speech is approximately vanishing in the higher ranges of the frequency, so the low frequency bands are chosen (172-344 Hz and 344-689 Hz).

3.3.3.1 Algorithm

Assume the following definitions:

- R : Crosscorrelation parameters
- UTR : unvoiced threshold.
- MUT : maximum unvoiced threshold.
- MVT : maximum of the moving standard deviation.

Figure 3. 12 Speech segment and the corresponding wavelet transform that is distributed over the whole frequency band as indicated in table 3.1.

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 the wavelet parameters of ranges [172-344Hz] and [344-689Hz](win(5) and win(6) in Table 1.1). To generate the correlation function, the frames of "R" parameters (The crosscorrelation parameters) are concatenated, then the absolute values of the points are taken and smoothed using moving average of 1024 points (about 90ms of speech in case of 11025 Hz sampling rate). The moving standard deviation is applied on the correlation function to reflect the variation in the correlation parameters along 100ms that is sufficient to detect any phonetic changes. The unvoiced threshold UTR is calculated as follows.

The first 100 ms of speech is assumed to be unvoiced or silence. Maximum unvoiced threshold is obtained from the first 100 ms (about 1024 samples) of the moving standard deviation.

The maximum voiced threshold is obtained from the whole speech duration. (MVT= the maximum of the moving standard deviation along the speech signal which only occurs in case of transition from unvoiced to voiced or vice virsa).

Let UTR=0.01*(MVT-MUT) the constant (0.01) is obtained by many trials of speech samples in the training phase.

R> UTR gives 1 that indicates a voiced. R< UTR gives 0 that indicates unvoiced.

Figures 3.13, 3.14, and 3.15 give some examples of the proposed algorithm applied on the words (سياره ، سياسة، شمسا) indicate the results that are obtained by use the above technique. It is clearly shown that the markers indicate

accurately the voiced segments. In figure 3.13 the word begins by the unvoiced sound /s/ followed by three different voiced sounds the short vowel /Θ/, the long consonant /y/ and the long vowel /Θ/. The markers track the voiced sounds along the duration of the three different sounds. A small duration drop in markers indicating unvoiced sound occurred in the transition between /Θ/ and /y/ ,which may include whispering, then in the transition between /y/ and /Θ/. A small duration drop in markers also occurred before the end indicating the location of unvoiced /r/. These false markers can be neglected by the software.

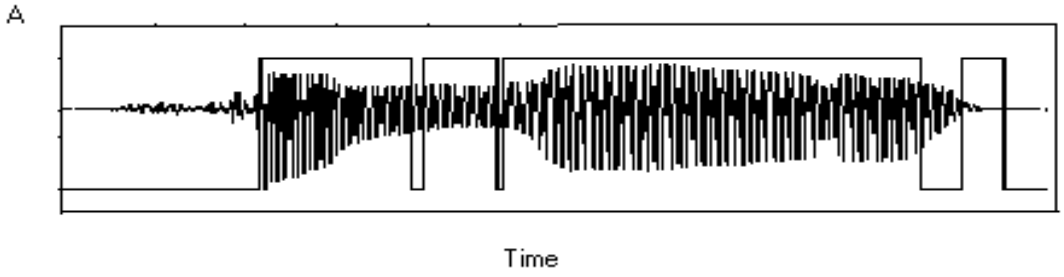


Figure 3. 13 Speech signal and logic markers. The markers are high in case of voiced sound. The word is /s//Θ//y//Θ//r//αH/سیاره

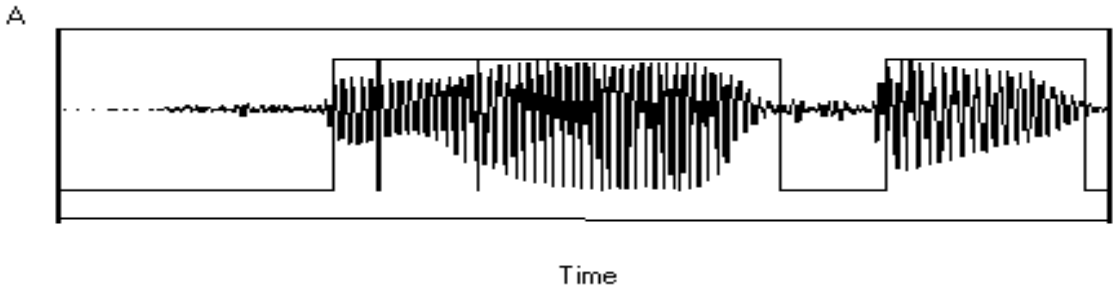


Figure 3.**14 Speech signal and logic markers. The markers are high in case of voiced sound. The word is /s//y//a//s//aH/**سیاسه

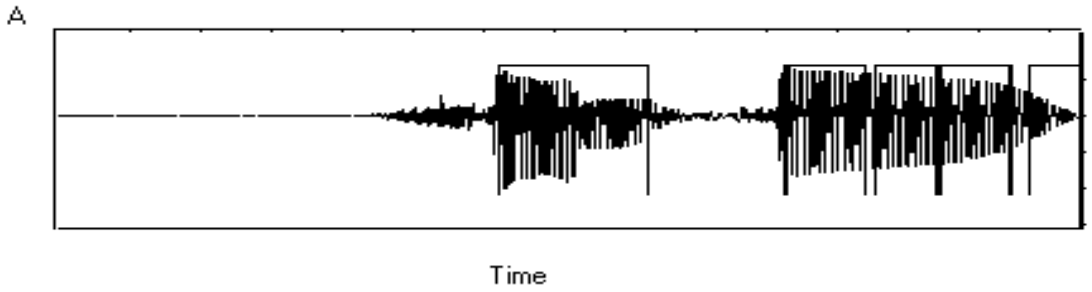


Figure 3.15 Speech signal and logic markers. The markers are high in case of voiced sound. The word is /Σ//Θ//M//s//aH/شمسا.

In figure 3.14, the word starts by unvoiced /s/ followed by voiced consonant /y/ then long vowel /Θ/. The markers are still high during the two different voiced sounds. The markers are dropped in the transition duration between /y/ /Θ/. The markers are dropped again in the duration of the internal unvoiced sound /s/ then become high again at the beginning of the end long vowel /Θ/.

Figure 3.15 is a speech signal which begins with a consonant unvoiced sound /Σ/. The markers start to be high at the beginning of vowel /Θ/ and continue in high position along the voiced consonant /M/. It goes down at the beginning of consonant /s/ then it goes back high at the beginning of vowel /Θ/.

This method gives a classification accuracy of about 98.4% for a test of 2 minutes of speech. This method is much immune to noise than the tracking function.

3.3.4 Voiced/Unvoiced classification using mathematical model based on wavelet features.

A trial to build a mathematical model to classify the speech into voiced/unvoiced is done. This model has the advantages of:

1- Once the model is found there is no need to make pre-estimation for unvoiced threshold.

2- It is easier to implement as hardware or software.

but there are many drawbacks:

1- It needs in the training phase a Database which must be handled carefully for best classification accuracy.

2- Efficiency of the system is environment-sensitive. In other words, training database must be collected in environment similar to the practical environment in which the system will be installed.

As introduced before in chapter 1, database are collected and aligned into the following table.

X						Y
B0	B1	B2	B3	B4	B5	
54000	30200	2230	1000	650	120	1 or 0

Wavelet parameters are extracted, interpolated and smoothed as in the previous method. The first six bands (B0, B1, B2, B3, B4, B5) that cover the frequencey range 86-5512 Hz are chosen. The algorithm is as follows:

1. A training period of 4 minutes of speech is used to prepare the training data set.

2. Wavelet parameters are extracted, interpolated and smoothed.

3. Training matrix is prepared. It contains rows called X-vectors. Each row represents the power distribution of the signal at certain time in the different six bands.

4. X-vector contains 6 elements as follows:

   X[i] = { B0,B1,B2,B3,B4,B5}

   Where each element in vector X represents the wavelet function (smoothed interpolated wavelet parameters) at time index i in the frequency bands 86-172Hz, 172-344 Hz,344-689 Hz, 689-1378 Hz, 1378-2756 Hz, 2756-5512 Hz respectively.

5. A pre-estimation of the state of X[i] vector into Voiced or unvoiced is made manually. The decision is put into vector Y. The i^{th} element of Y is a decision of x[i] vector as indicated below:

$$\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & X_{03} & X_{04} & X_{05} & X_{06} \\ \vdots & & \vdots & & \vdots & \\ X_{m1} & X_{m2} & X_{m3} & X_{m4} & X_{m5} & X_{m6} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \qquad \textbf{(3. 6)}$$

Where [B] is calculated :

$$[B]=\begin{bmatrix} 0.0115 \\ 0.0034 \\ 0.0095 \\ -0.1058 \\ 0.1374 \\ 0.0200 \end{bmatrix} \qquad \textbf{(3. 7)}$$

Now the [B] matrix is the system model for V/U classification. Many speech signals are tested. The system gives around **90.7**% classification rate which is less than the previous correlation method but is much faster as it does not need pre calculations as the past two methods. Figures 3.16 and 3.17 show two examples of Arabic words (كتاب، سياسة). The markers indicate the classification of V/U regions using the proposed algorithm.

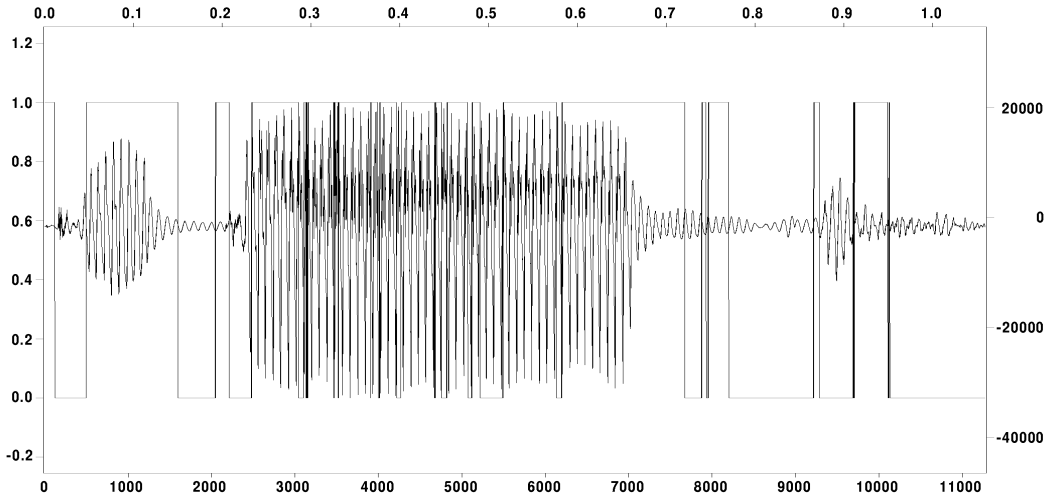


Figure 3.**16 V/U Markers for a speech utterance /k//i//t//Θ//b/. Markers are generated using the mathematical regression model. The word is** كتاب

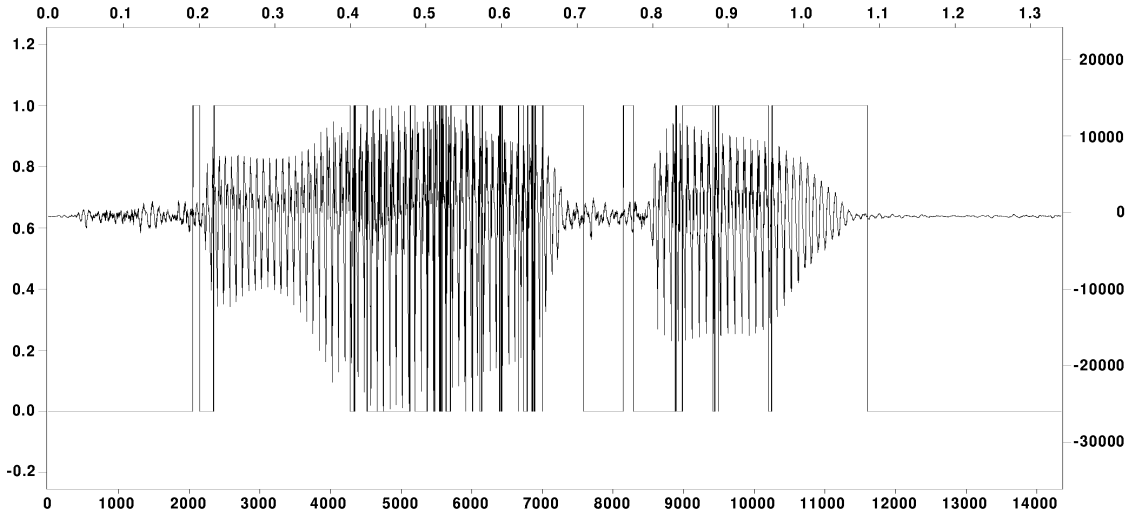


Figure 3.**17 V/U Markers for a speech utterance /s//y//Θ//s//aH/. Markers are generated using mathematical regression model** .**The word is** سیاسه

As shown in figures 3.16 and 3.17, the small drops in markers occur frequently within the voiced or unvoiced period. That is because the system here is highly sensitive to environmental changes. Practically the drop's duration is very small and can be corrected by software.

3.4 Pitch period estimation

Pitch period estimation (or equivalently, fundamental frequency estimation) is one of the most important problems in speech processing. Pitch detectors are used in vocoders, speaker identification , verification systems and many other applications[5]. Because of its importance, many solutions to this problem have been proposed [52-68]. All of the proposed schemes have their limitations, and it is safe to say that no presently available pitch detection scheme can be expected to give perfectly satisfactory results across a wide range of speakers, applications, and operating environments[5].

The time domain methods give good results for pitch estimation especially for low noise environment. The frequency or spectral methods, such that LPC-

based pitch detector, give good results in some cases but it gives a poor results in case of high pitch speakers.

In this section a general review of some of the pitch detection methods is given.

3.4.1 The parallel processing method

The scheme was first proposed by Gold [5] and later modified by Gold and Rabiner [5]. Our reasons for discussing this particular pitch detector in this chapter are:

(1) It has been used successfully in a wide variety of applications.

(2) It is based on purely time domain processing as this point of research.

(3) It can be implemented to operate very quickly on a general-purpose computer or it can be easily constructed in digital hardware.

(4) It illustrates the use of the basic principle of parallel processing in speech processing.

The basic principles of this scheme are as follows;

1. The speech signal is processed so as to create a number of impulse trains that retain the periodicity of the original signal and discard features. which are irrelevant to the pitch detection process.

2. This processing permits very simple pitch detectors to be used to estimate the period of each impulse train.

3. The estimates of several of these simple pitch detectors are logically combined to infer the period of the speech waveform.

The particular scheme proposed by Gold and Rabiner [5] is depicted in Figure 3.18. The speech waveform is sampled at a rate sufficient to give adequate time resolution; e.g., sampling at 10 kHz allows the period to be determined to within $T = 10^{-4}$ sec. The speech is lowpass filtered with a cutoff of about 900 Hz to produce a relatively smooth waveform. A bandpass filter passing frequencies between 100 Hz and 900 Hz may be necessary to remove 60 Hz noise in some applications. (This filtering can be done either with an analog filter before sampling or with a digital filter after sampling.)

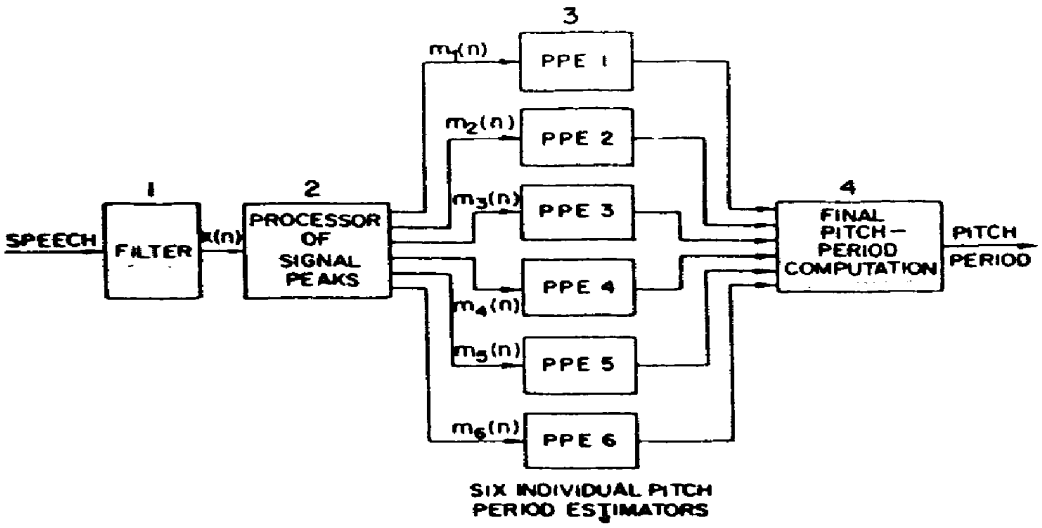


Figure .3 18 Block diagram of a parallel processing time domain pitch detector.

Following the filtering the peaks and valleys (local maxima and minima) are located, and from their locations and amplitudes, several impulse trains (6 in figure 3.18) are derived from the filtered signal. Each impulse train consists of positive impulses occurring at the location of either the peaks or the valleys. The 6 cases used by Gold and Rabiner [5] are:

1. m1(n): An impulse equal to the peak amplitude occurs at the location of each peak.

2 m2(n): An impulse equal to the difference between the peak amplitude and the preceding valley amplitude occurs at each peak.

3. m3(n): An impulse equal to the difference between the peak amplitude and the preceding peak amplitude occurs at each peak. (if this difference is negative the impulse is set to zero.)

4. m4(n): An impulse equal to the negative of the amplitude at a valley occurs at each valley.

5. $m_5(n)$: An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding peak occurs at each valley.

6. $m_6(n)$*:* An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding local minimum occurs at each valley. (If this difference is negative the impulse is set to zero.)

Figures 3.19 and 3.20 show two examples - a pure sine wave and a weak fundamental plus a strong second harmonic - together with the resulting impulse trains as defined above. Clearly the impulse trains have the same fundamental period as the original input signals, although m5(n) of Fig. 3.20 is close to being periodic with half the fundamental period. The purpose of generating these impulse trains is to make it simple to estimate the period on a short-time basis. The operation of the simple pitch period estimators is depicted in Figure 3.21. Each impulse train is processed by a time varying nonlinear system (called a peak detecting exponential window circuit in [5]).

When an impulse of sufficient amplitude is detected in the input, the output is reset to the value of that impulse and then held for a blanking interval, $\tau(n)$ - during which no pulse can be detected. At the end of the blanking interval, the output begins to decay exponentially. When an impulse exceeds the level' of the exponentially decaying output, the process is repeated. The rate of decay and the blanking interval are dependent upon the most recent estimates of pitch period. The result is a kind of smoothing of the impulse train, producing a quasi-periodic sequence of pulses as shown in Fig. 3.21. The length of each pulse is an estimate of the pitch period. The pitch period is estimated periodically (e.g., 100 times/sec) by measuring the length of the pulse spanning the sampling interval.

This technique is applied to each of the six impulse trains thereby obtain-ins six estimates of the pitch period. These six estimates are combined with two of the most recent estimates for each of the six pitch detectors. These estimates are then compared and the value with the most occurrences (within some tolerance) is declared the pitch period at that time. This procedure produces very good estimates of the period of voiced speech. For unvoiced speech there is a distinct lack of consistency among the estimates. When this lack of consistency is detected the speech is classified as unvoiced The entire process is repeated periodically to produce an estimate of the pitch period and voiced/unvoiced classification as a function of time.

Although the above description may appear very involved, this scheme for pitch detection can be efficiently implemented either in special purpose hardware or on a general-purpose computer. Indeed, near real-time operation (within a factor of 2 times real-time) is possible on present computers.

Furthermore it has been observed that at the initiation of voicing (i.e., the first 10-30 msec of voicing) the speech is often classified as unvoiced. This result is due to the decision algorithm that requires about 3 pitch periods before a reliable pitch decision can be made - thus a delay of about 2 pitch periods is inherently built into the method

In summary, the details of this particular method are not so important as the basic principles that are introduced. First, note that the speech signal was processed to obtain a set of impulse trains which retain only the essential feature of periodicity (or lack of periodicity). Because of this simplification in the structure of the signal, a very simple pitch estimator suffices to produce good estimates of the pitch period. Finally, several estimates are combined to increase the overall reliability of the estimate. Thus, signal processing simplicity is achieved at the expense of increased logical complexity in estimating the desired feature of the speech signal. Because the logical operations are carried out at a much lower rate (e.g., 100 times/sec) than the signal processing, this results in an overall speed-up in processing. A similar approach was used by Barnwell et al. [5] in designing a pitch detector in which the outputs of four simple zero-crossing pitch detectors were combined to produce a reliable estimate of pitch.

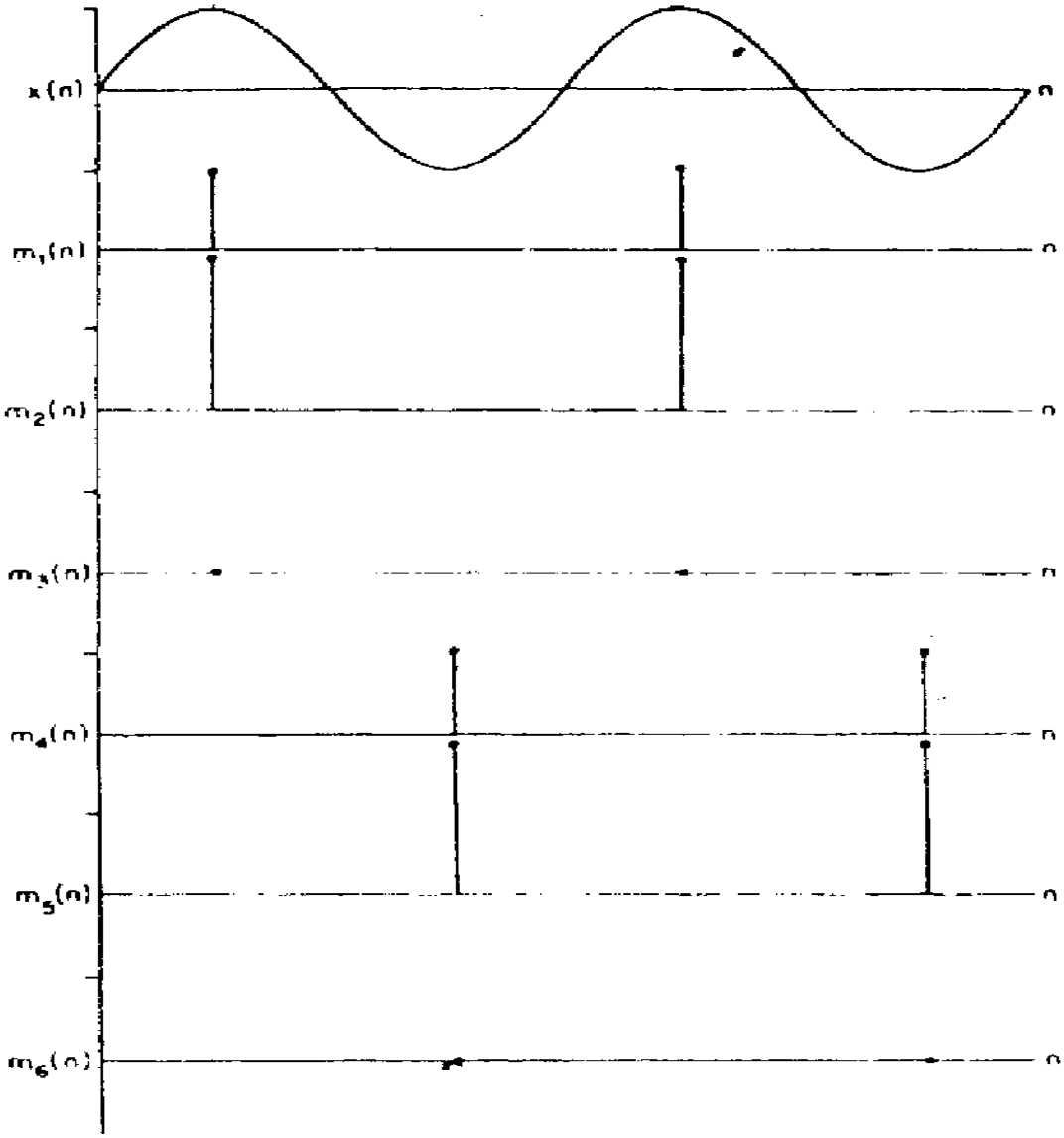


Figure 3. 19 Impulse trains generated from peaks and valleys of a pure sin wave[5].

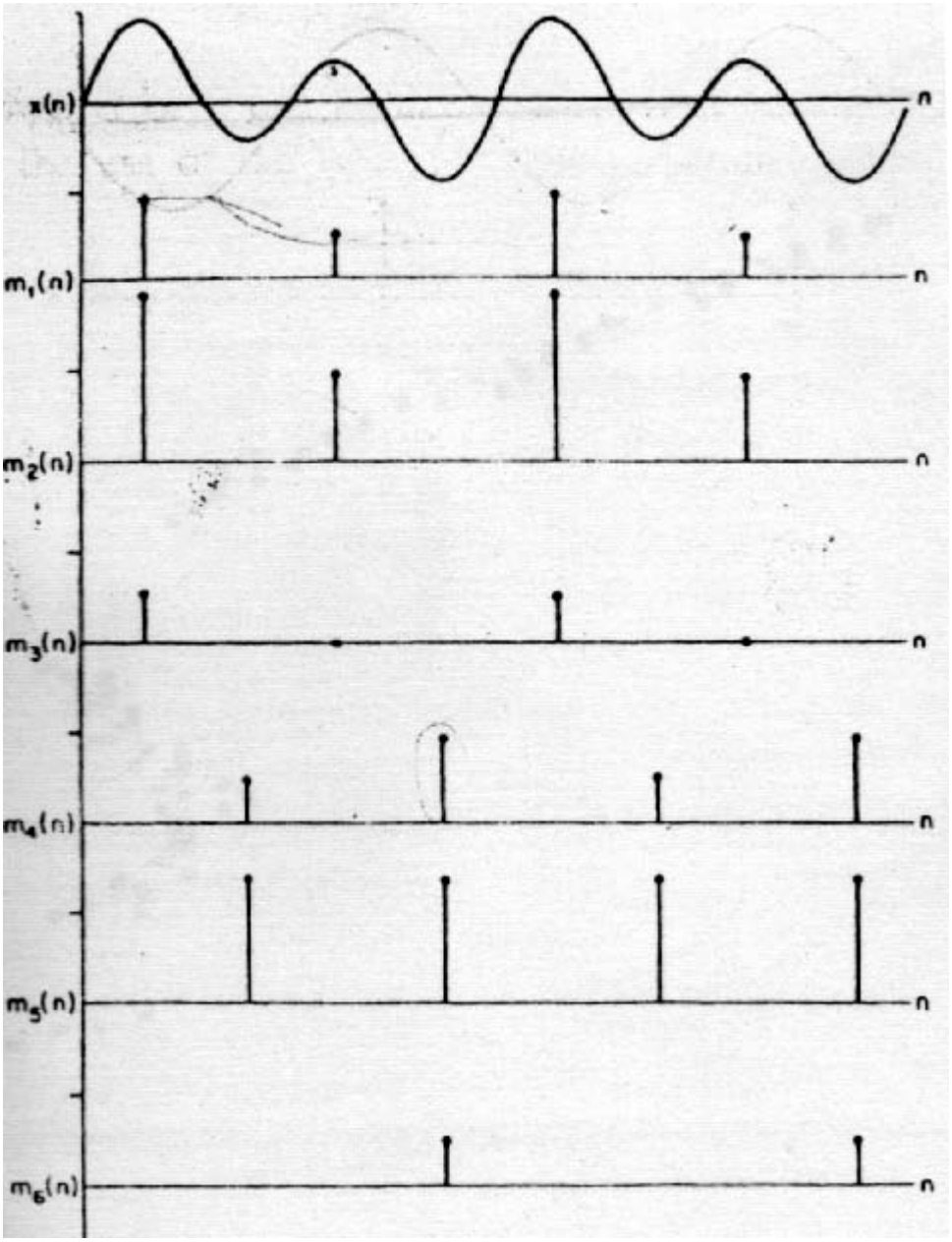


Figure 3. 20 Impulse trains generated from peaks and valleys of a weak fundamental and second harmonic[5].

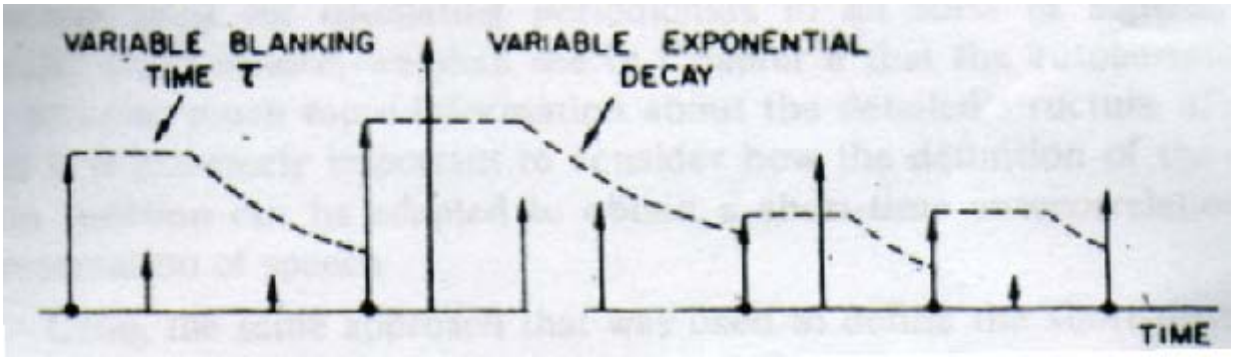


Figure 3.**21 Basic operation of pitch estimator[5].**

3.4.2 The simplified inverse filter tracking SIFT method

An efficient and accurate pitch extraction method based upon linear prediction principles for the range 50-250 Hz is the simplified inverse filter tracking (SIFT) algorithm [Markel, 1972c]. A down-sampling procedure is used so that the effective sampling frequency for *F0* analysis is about 2 kHz. Therefore, only the most reliable frequency range up to about 1 kHz is processed and in addition, the necessary number of operations is substantially reduced. A block diagram of the SIFT algorithm, represented in two steps, is shown in Figure 3.22. Efficient preprocessing to reduce formant and fundamental frequency interaction is performed in step 1. A sequence of speech samples corresponding to frame *k* is pre-filtered with a cutoff close to f_s/I=2kHz. where *I* is the integer down-sampling factor.

Down-sampling is performed to reduce the effective sampling rate to f_s/I. The samples are differenced to accentuate the region of the second formant, and multiplied by a Hamming window. A fourth-order inverse filter *A* (z) is then designed using the autocorrelation method. Due to the fact that at most two formants can reside in the range (0, 1 kHz), four coefficients have been demonstrated to be sufficient.

Although the inverse filter was designed on the basis of differenced windowed data the output is obtained by applying the unwindowed non-differenced data. The effect of this operation is to produce a low-pass filtered error signal without low-frequency bias. This signal is then multiplied by a second Hamming window.

In step 2, an autocorrelation sequence is obtained and then the peak within the minimum-to-maximum desired pitch range is obtained. Parabolic interpolation is applied to provide greater pitch period resolution. (Without interpolation, the maximum resolution would be *1/* f_s*).* A variable threshold has been found to be of significant utility with a filtered error signal. The threshold is defined by two linear segments intersecting at some quiescent threshold location. As the peak location becomes smaller, the threshold is raised. Since proportionally more pitch periods will be obtained per analysis interval. As the peak location increases, the threshold is lowered. If a peak crosses the variable threshold, its location becomes the pitch period candidate for that frame. Otherwise the frame is defined as unvoiced *(P=0).* An attempt at error detection and correction is made by storing several pitch period candidates. After this operation, the pitch period estimate with maximum delay is output.

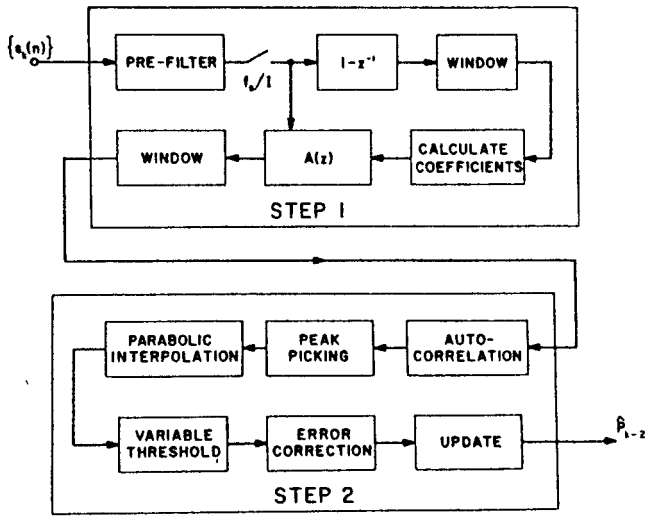


Figure 3. 22 Block diagram of the SIFT algorithm[5].

3.4.3 Pitch estimation using Cepstrum

Figures 3.23 suggest a powerful means for pitch estimation based on cepstrum. It is observed that for the voiced speech, there is a peak in the cepstrum at the fundamental period of the input speech segment. No such peak appears: in the cepstrum of the unvoiced speech segment.

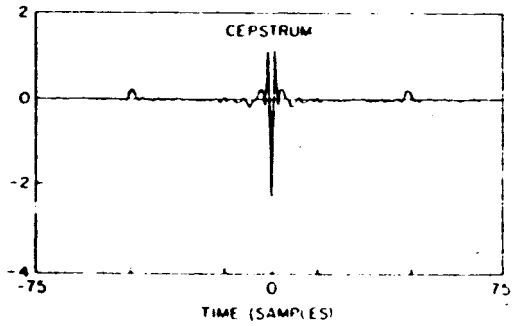


Figure 3. 23 Cepstrum of a voiced speech segment[5].

These properties of the cepstrum can be used as a basis for determining whether a speech segment is voiced or unvoiced and for estimating the fundamental period of voiced speech.

The outline of the pitch estimation procedure based on the cepstrum is rather simple. The cepstrum is searched for a peak in the vicinity of the expected pitch period. If the cepstrum peak is above a pre-set threshold, the input speech segment is likely to be voiced, and the position of the peak is a good estimate of the pitch period. If the peak does not exceed the threshold, it is likely that the input speech segment is unvoiced. The time variation of the mode of excitation and the pitch period can be estimated by computing a time-dependent cepstrum based upon a time dependent Fourier transform. Typically, the cepstrum is computed once.

Figures 3.24 shows example due to A. M. Noll [5], who first described a procedure for estimating pitch using the cepstrum. Figure 3.24 shows a series of log spectra and corresponding cepstra for a male speaker. The cepstra plotted in this example are the square of cepstrum. In this example, the sampling rate of the input was 10 kHz. A 40 msec (400 samples) Hamming window was moved in jumps of 10 msec; i.e., log spectra on the left and corresponding cepstra on the right are computed at 10 msec intervals. It can be seen from Figure 3.24 that the first seven 40 msec intervals correspond to unvoiced speech, while the remaining cepstra indicate that the pitch period increases with time (i.e., fundamental frequency decreases).

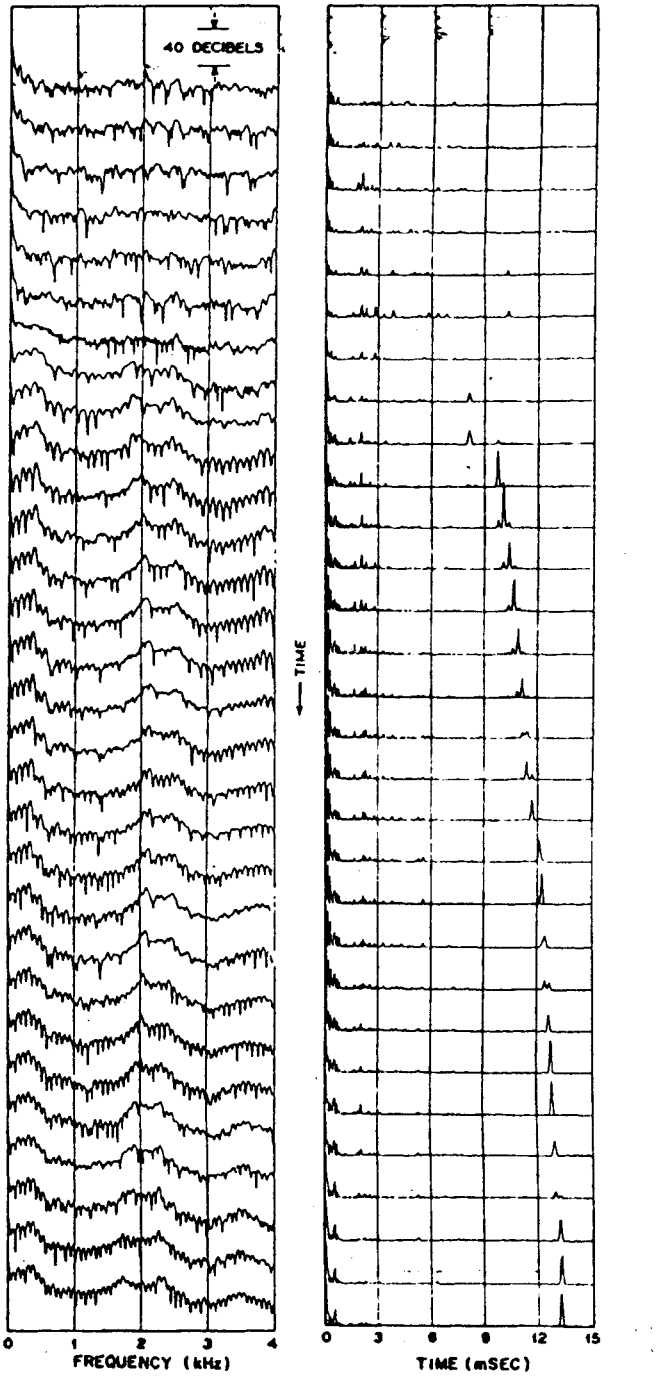


Figure 3. 24 Series of log spectra and cepstrum for a male speaker[5].

Unfortunately, as is usually the case in speech analysis, there are numerous special cases and trade-offs that must be considered in designing a cepstrum pitch detection algorithm.

- First, the presence of a strong peak in the cepstrum in the range 3-20 msec is a very strong indication that the input speech segment is voiced. However, the absence of a peak or the existence of a low-level peak is not necessarily a strong indication that the input speech segment is unvoiced. That is, the strength of or even the existence of a cepstrum peak for voiced speech is dependent on a variety of factors, including the length of the window applied to the input signal and the formant structure of the input signal. It is easily shown that the maximum height of the "pitch peak" is unity[5]. This can be achieved only in the case of absolutely identical pitch periods. This is, of course, highly unlikely in natural speech, even in the case of a rectangular window that encloses exactly an integer number of periods. Rectangular windows are rarely used due to the inferior spectrum estimates that result, and in the case of, for example, a Hamming window, it is clear that both window length and the relative positions of the window and the speech signal will have considerable effect upon the height of the cepstrum peak. As an extreme example, suppose that the window is less than two pitch periods long. Clearly it is not reasonable to expect any strong indication of periodicity in the spectrum or the cepstrum in this case. Thus, the window duration is usually set so that, taking account of the tapering of the data window, at least two clearly defined periods remain in the windowed speech segment. For low pitched male speech, this requires a window on the order of 40 msec in duration. For higher pitched voices, proportionately

shorter windows can be used. It is, of course, desirable to maintain the window as short as possible so as to minimize the variation of speech parameters across the analysis interval. The longer the window, the greater the variation from beginning to end and the greater will be the deviation from the model upon which the analysis is based. One approach to maintaining a window that is neither too short nor too long is to adapt the window length based upon the previous (or possibly average) pitch estimates

- Second, if the signal is band-limited, it will deviate from the model, In this case there is only one peak in the log spectrum. If there is no periodic oscillation in the log spectrum, there will be no peak in the cepstrum. In speech, voiced stops are generally extremely band-limited, with no clearly defined harmonic structure at frequencies above a few hundred Hertz. In such cases there is essentially no peak in the cepstrum. Fortunately, for all but the shortest pitch periods, the pitch peak occurs in a region where the other cepstrum components have died out appreciably. Therefore, a rather low threshold can be used in searching for the pitch peak (e.g., on the order of 0.1).

3.4.4 Pitch estimation using wavelet

A new method for pitch estimation of the speech signal is introduced. The technique is based on the discrete wavelet transform. The algorithm is highly immunized to noise. A fair comparison between the ordinary methods and this new one is presented.

The wavelet transform creates a link between the time domain and the frequency domain. So, the methods that are based on the wavelet transform can take the advantages of both time domain and frequency domain.

3.4.4.1 Detection of pitch using two band correlation of wavelet features.

Table 3.1 indicates the number of wavelet parameters for each frequency band in case of 1024 samples 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 is based on the criteria which indicates that most of the power in the voiced speech is below the 900 hz [4]. A Crosscorrelation algorithm is applied between Window#5 and Window# 6 (Table 3.1) rather than the autocorrelation of one window to get the highest immunity to noise. That is because if the speech features are weak in one window it may be strong in the adjacent window. For the above two reasons the crosscorrelation can give the maximum reliable correlation representation between the two windows.

The procedure can be arranged as follows:

1) The speech signal is low pass filtered at 900 Hz.
2) The speech signal is classified into voiced and unvoiced speech.
3) The algorithm is applied on the voiced section only by dividing them into smaller windows of 1024 samples each.
4) The wavelet parameters are extracted for each window.

5) The crosscorrelation is performed on win(5) and win(6) To generate the correlation function.

6) The frames of "R" parameters (The crosscorrelation parameters) are concatenated to compose a continuos correlation function along the voiced segment of speech signal.

7) A peak detection algorithm is applied on the generated function.

The duration between the fundamental peaks correspond to the pitch period. The pitch contour will be established by using frames of speech signal of 100 ms.

The above procedure is applied on the speech signal in figure 3.25 (Arabic word ذهب).

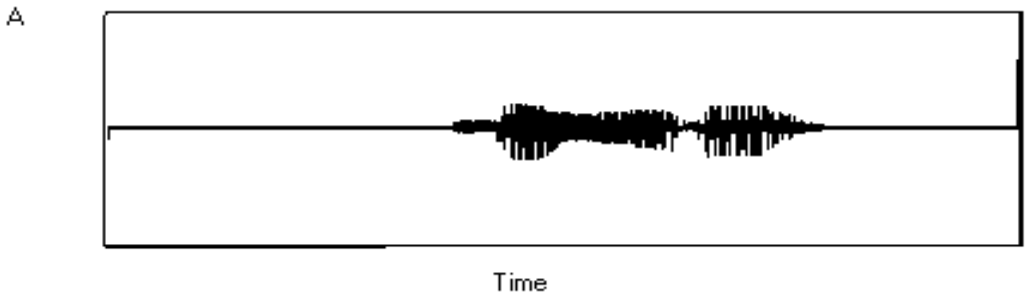


Figure 3. 25 Speech sample of the word "ذهب" in Arabic. It is pronounced /ζ//Θ//η//Θ//βH/

Figure 3.26 indicates the impulse train after applying the algorithm over the utterance of figure 3.25.

Figure 3.27 focus on part of the voiced segment.

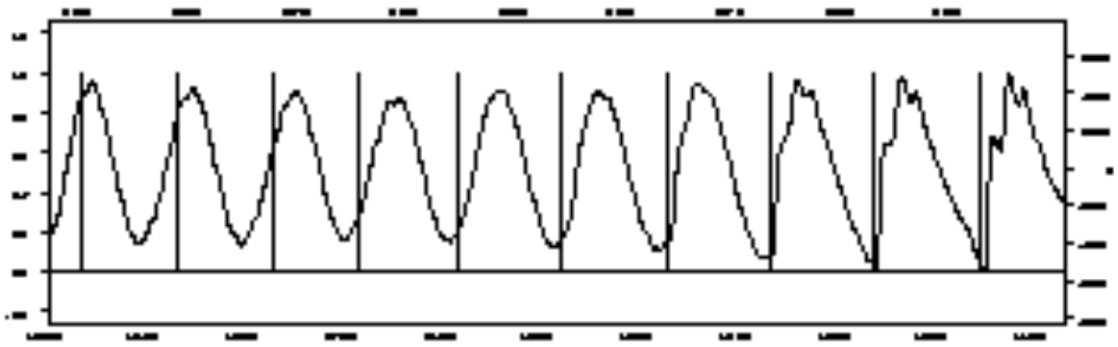


Figure 3.27 The impulses train of part of voice segment in figure 3.26

The power of this technique appears clearly in case of noisy environments. This technique is highly robust in any noisy environment even in case of very low signal to noise ratios as will be shown.

The old time based techniques are highly affected with the environmental condition. This problem is partly solved in the case of this algorithm.

The peak detector extracts all peaks of the correlation function to generate an impulse train. To achieve this point the first 200 ms of utterance is processed to extract the noise level. The correlation parameters of this period are calculated. The maximum parameter is taken as the noise threshold.

The whole correlation parameters of the whole utterance are compared with the noise threshold. The logical function (impulse train) is generated by this comparison. If the correlation parameter is bigger than the noise threshold an impulse is generated.

The above technique is applied for the speech sample in figure 3.28 to test how far this algorithm is robust in the presence of noise. The signal to noise ratio of 7dB is achieved and the results are the same. The following figures summaries the results.

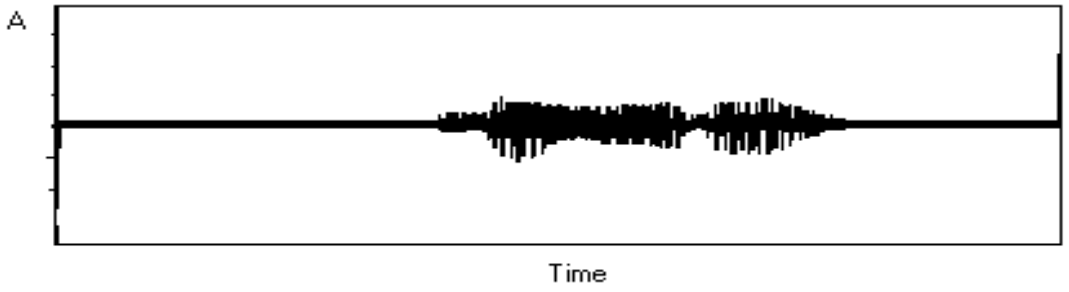


Figure 3.2**8 Arabic word "ذهب" /z//Θ//h//Θ//bH/. S/N = 15 dB.**

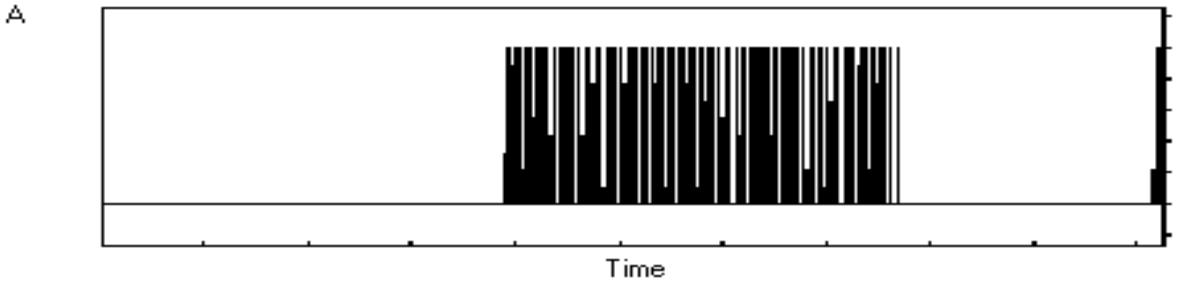


Figure 3. 2**9 Pitch markers of figure 3.28.**

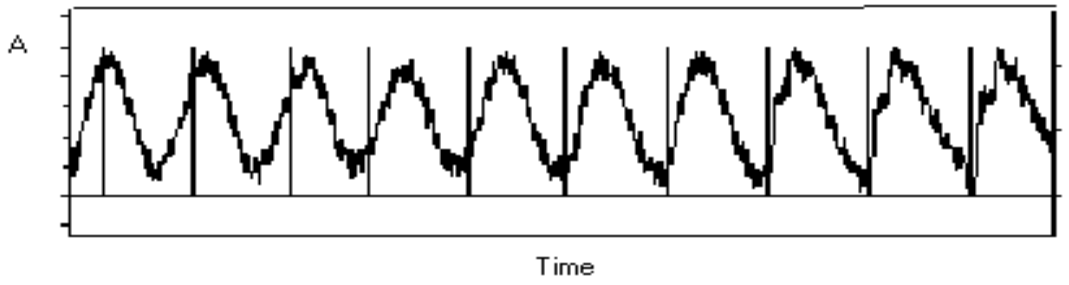


Figure 3.**30 Focus on apart of voiced segment /a/ of figure** 3.**29. The speech segment corresponds to this part is overlaid on the impulse train.**

As shown in figure 3.30, the pitch impulse train still keep track with the speech energy in case of S/N= 16 dB.

Figures 3.31,3.32 and 3.33 indicate the result in a very poor noise environment. The speech signal is superimposed to a uniform noise to reach a signal to noise ratio of 7 dB only.

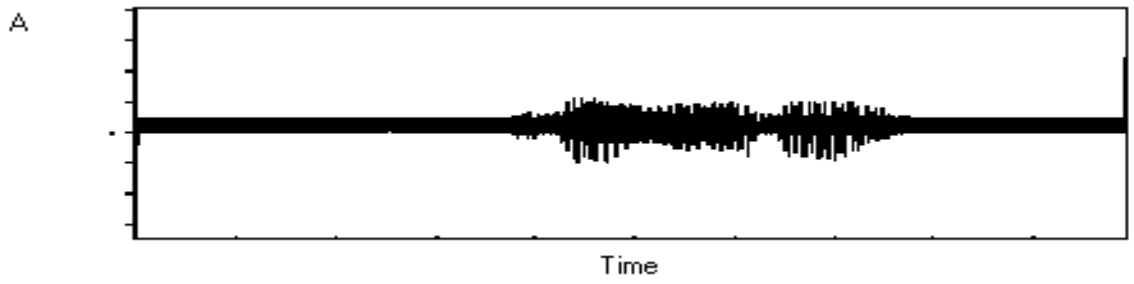


Figure 3. 31 Arabic word "ذهب". S/N = 7 dB.

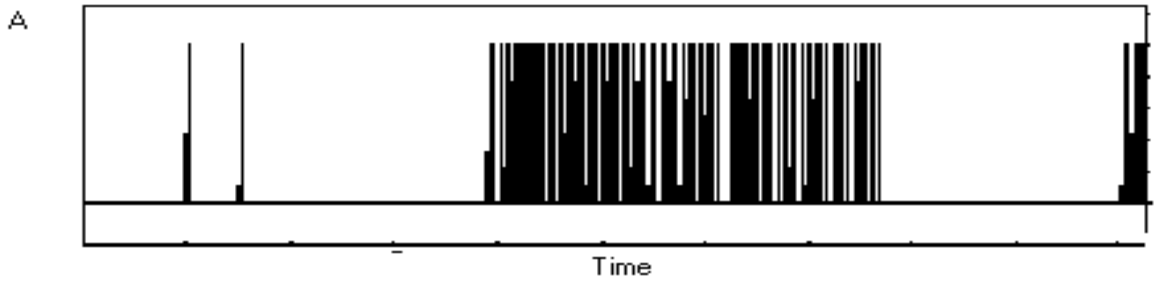


Figure 3. 32 Pitch markers of figure 3.31.

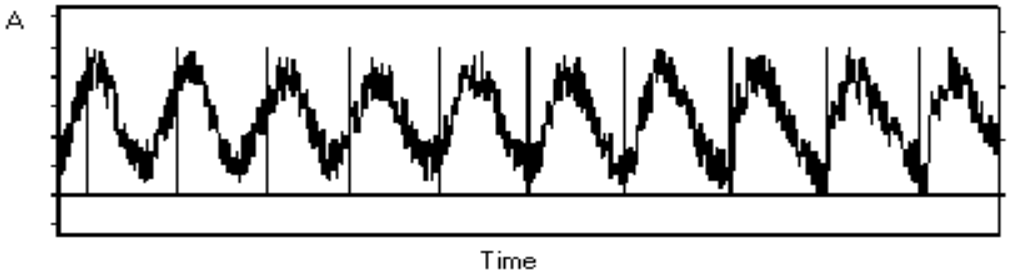


Figure 3.33 **Focus on a part of voiced segment /a/ of figure** 3.32**. The speech segment corresponds to this part is overlaid on the impulse train.**

3.4.4.2 Pitch detection using two wavelet based estimators in parrallel

The above discussion illustrate how far wavelet succeed to track the fundamental frequency of speech utterance. In this section the algorithm is compared with well-known pitch estimators (Autocorrelation and Cepstrum).

Figure 3.34 is a flow chart which represents the algorithm. Following is a discussion of each block in flow chart.

Framing and overlapping: Speech signal is segmented into frames. Each frame contains 1024 samples. The frames are overlapped by 975 samples. This overlapping makes the steps of unoverlapped period is 50 samples (about 5 ms in case of 11025 Hz sampling rate).

- **Wavelet:** Performs the wavelet transform on a frame which contains 1024 samples. The wavelet filter is Doubchi filter. The output of this block are six series each contains 1024 samples representing the utterance in different frequency bands. The bands are summarized below:

  B0 86-172 Hz.

  B1 172-344 Hz.

  B2 344-689 Hz.

  B3 689-1378 Hz.

  B4 1378-2756 Hz.

  B5 2756-5512 Hz.

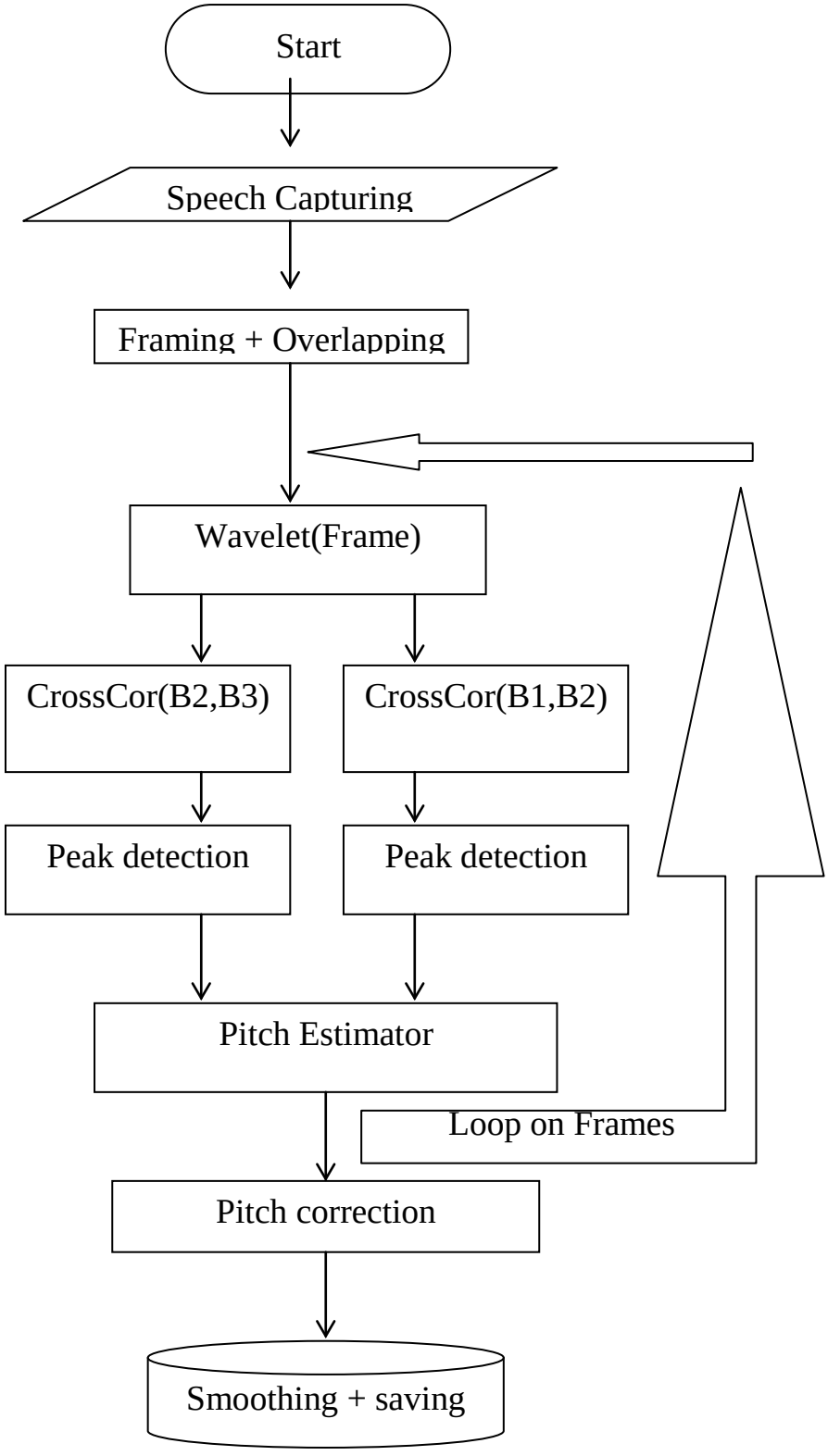


Figure 3.34 Flow chart of two parallel pitch estimator

- **CrossCor(B_m, B_n):** Performs the crosscorrelation between $\mathbf{B_m}$ and $\mathbf{B_n}$**.** The crosscorrelation give indication about the dependencies of signal components in the two selected bands.
- **Peak Detection:** Peaks of crosscorrelation function are detected. Peaks occur at distance representing the repetition of the speech signal fundamental frequency.
- **Pitch Estimator:** Actually the first harmonics can interfere the process. So the information of peak detector is correlated in different bands. The fundamental frequency appears in the two bands so correlation between them eliminates harmonics.
- **Pitch verification:** Pitch contour is verified to eliminate unexpected values or variation. Moving standard deviation of 5 points is applied. Parts of speech contains deviation more than 15 Hz are eliminated and assumed to be unvoiced.
- **Pitch smoothing:** A 5-points smoothing filter is applied on the pitch contour.

***Results and comparison**

The above algorithm is applied on 40 sec speech utterances. The technique is compared with the familiar pitch estimators such as Autocorrelation pitch estimator and Cepstrum pitch estimator. The technique is applied on normal speech utterances as well as synthetic speech utterances for both male speaker and female speaker. The following figures summarize the results of comparison.

Figures 3.35,3.36,3.37 and 3.38 indicate how far the system performs with respect to a well-known systems (Autocorrelation and Cepstrum). Figure 3.35 is a comparison of pitch contours calculated using three different pitch methods for a speech signal on the top of the figure. The word is " كتاب" in Arabic and it is pronounced /κ//ε//τH//Θ//βH/. The word contains two vowels /ε/ αnδ /Θ/. The first starts at 0.1 ms and ends at 0.2 ms. The second vowel starts at 0.3 ms and ends at 0.8 ms as shown in figure 3.35. Female speaker pronounces the word. The second graph from the top of figure 3.35 is a pitch contour calculated with Cepestrum method. The third graph from the top of figure 3.35 is a pitch contour calculated with the wavelet-based method. It is clearly apparent that, the two curves give approximately the same results but the wavelet-based method is more stable in the transition regions. The last graph on the bottom of figure 3.35 is the pitch contour calculated using the autocorrelation method. It is clear that it gives unstable graph compared with the other two methods.

Figure 3.36 is the same as figure 3.35 except that the word under test is a synthesized female word. The top graph is the original pitch of the speech signal. The second curve from the top is the synthesized speech signal. The third graph is the pitch of the synthesized utterances calculated with cepstrum method. The fourth graph is the pitch contour calculated with the autocorrelation method. The bottom graph is the pitch contour of the wavelet-based method. It is clear that the wavelet-based pith contour is the best one approximation of the original pitch contour on the top.

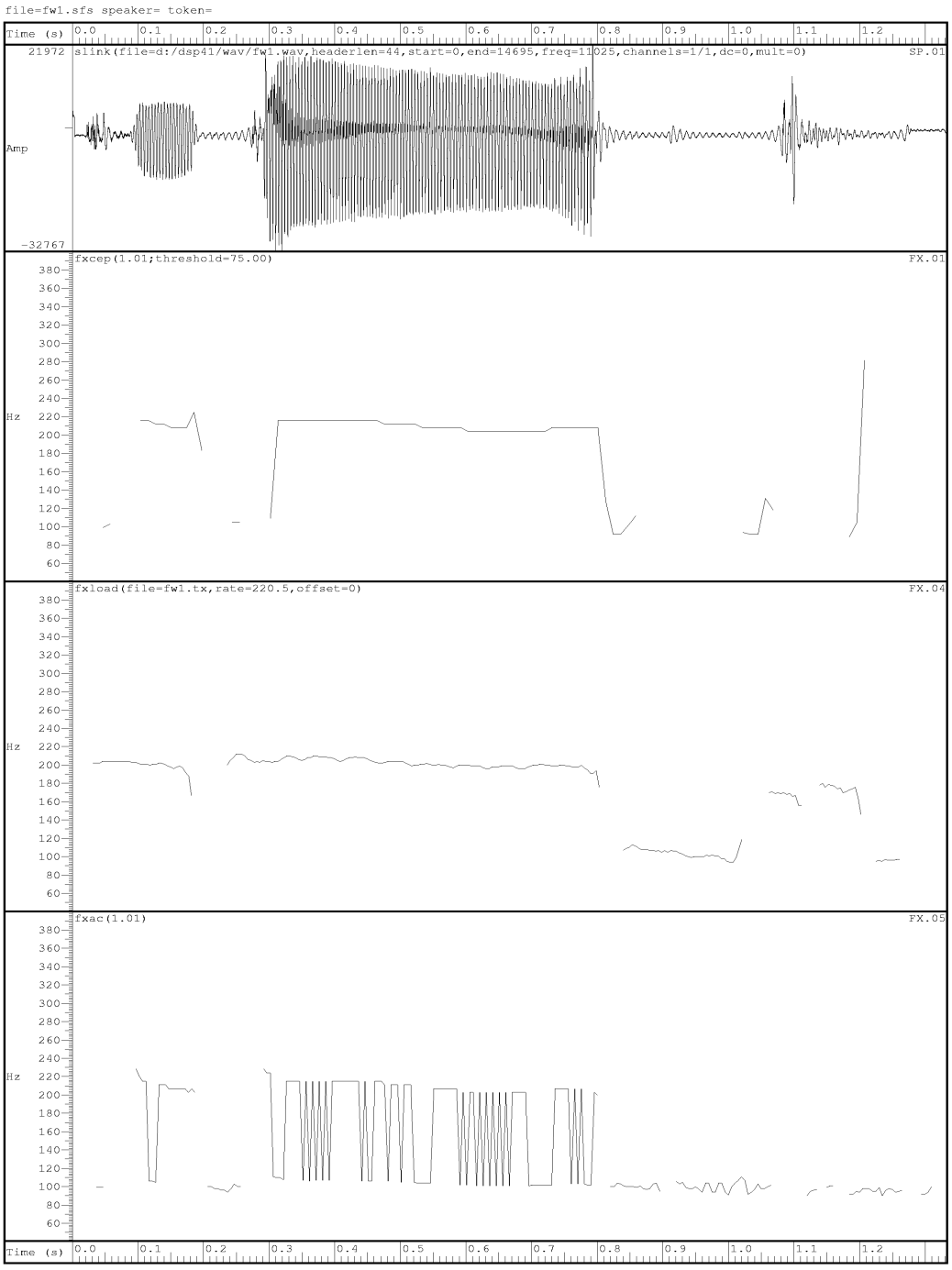


Figure 3. 35 Female pitch comparison. The top graph is aligned as speech sample, Cepstrum-based pitch, wavelet-based pitch and Autocorrelation-based pitch.

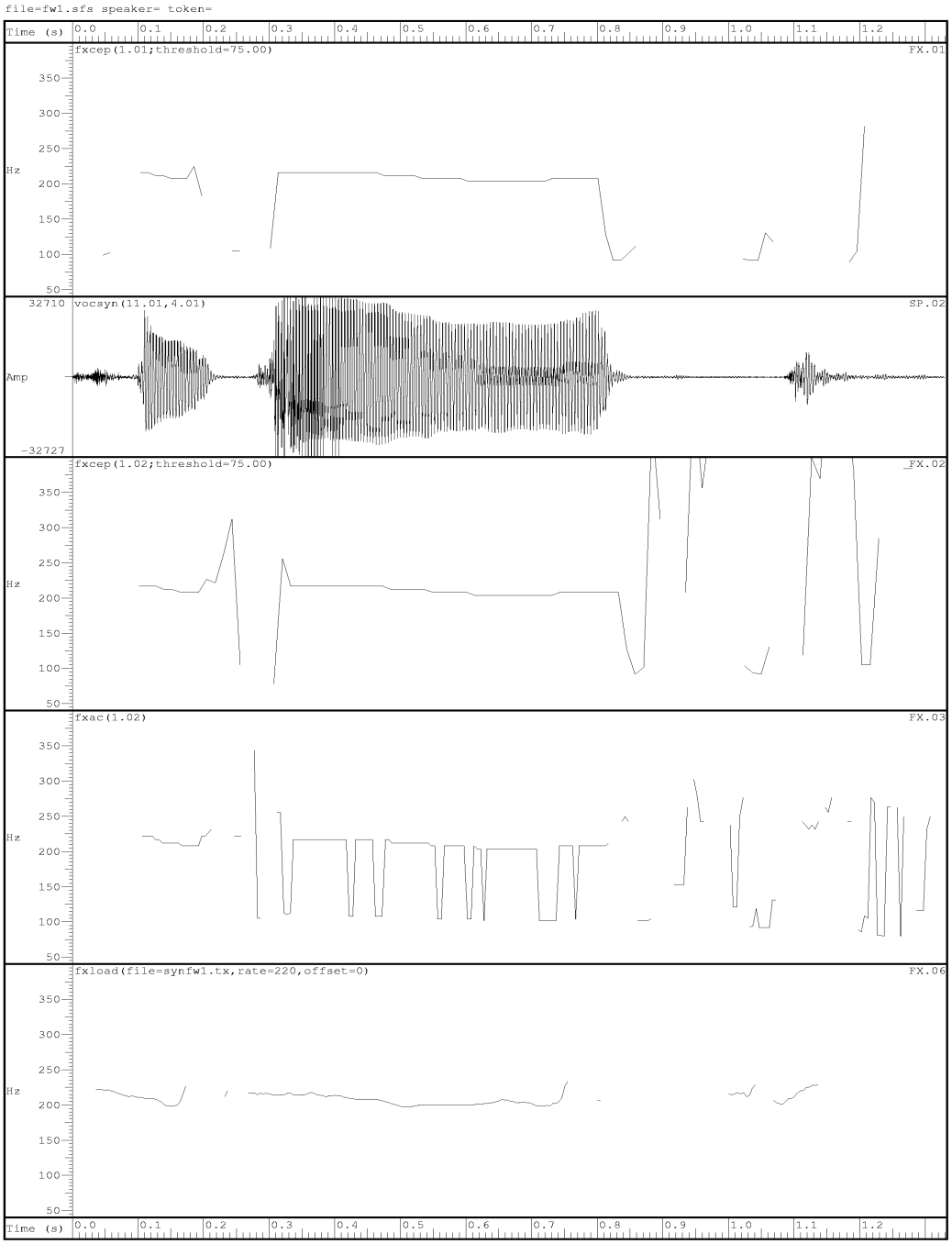


Figure 3. 36 Female synthetic speech pitch comparison. The graph is aligned as Cepestrum-based pitch for normal speech, Synthesized speech, Cepstrum-based pitch for synthesized speech, Autocorrelation-based for synthesized speech. Wavelet-based pitch for synthesized.

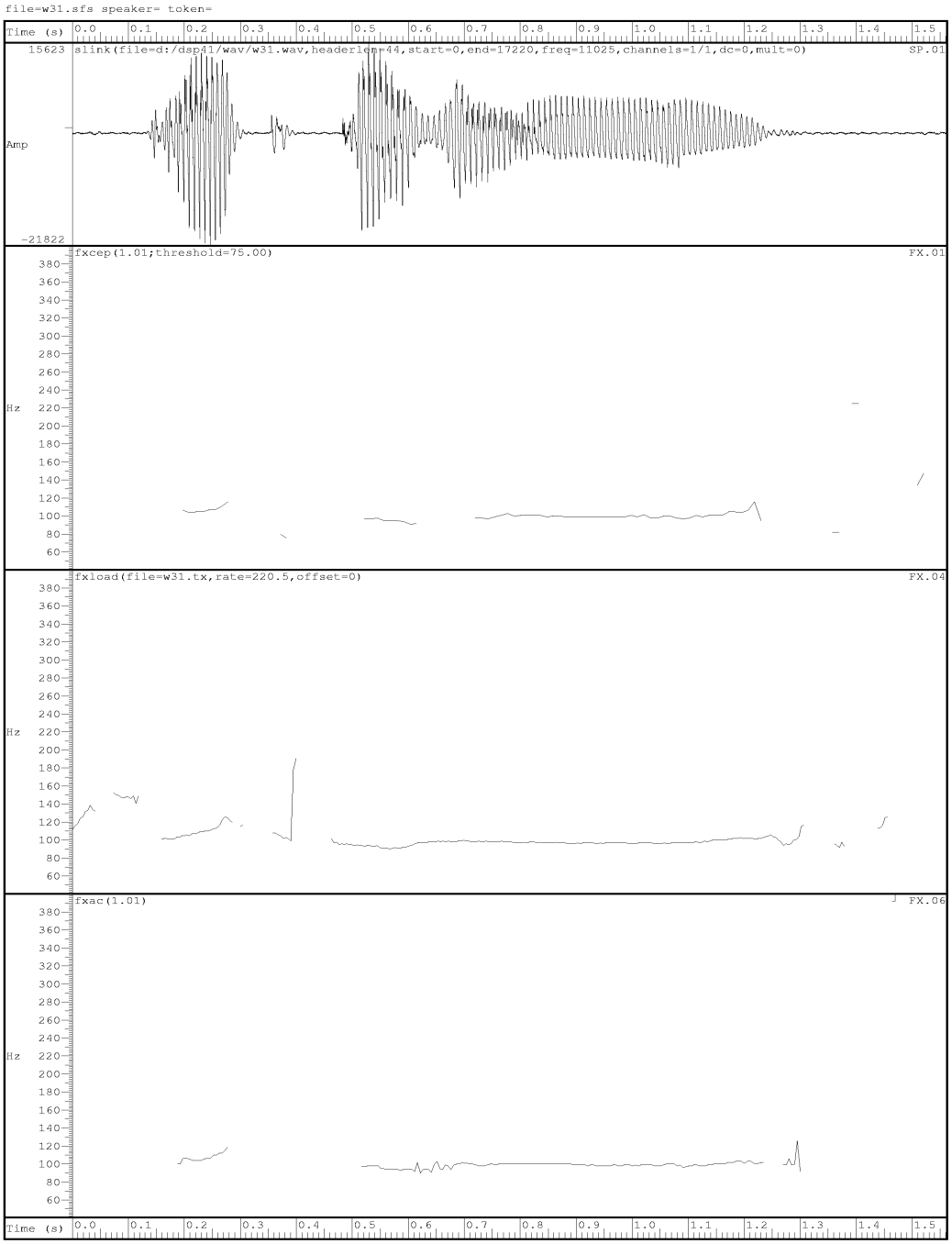


Figure 3. 37 Male pitch comparison. The graph is aligned as speech sample, Cepstrum-based pitch, wavelet-based pitch and Autocorrelation-based pitch.

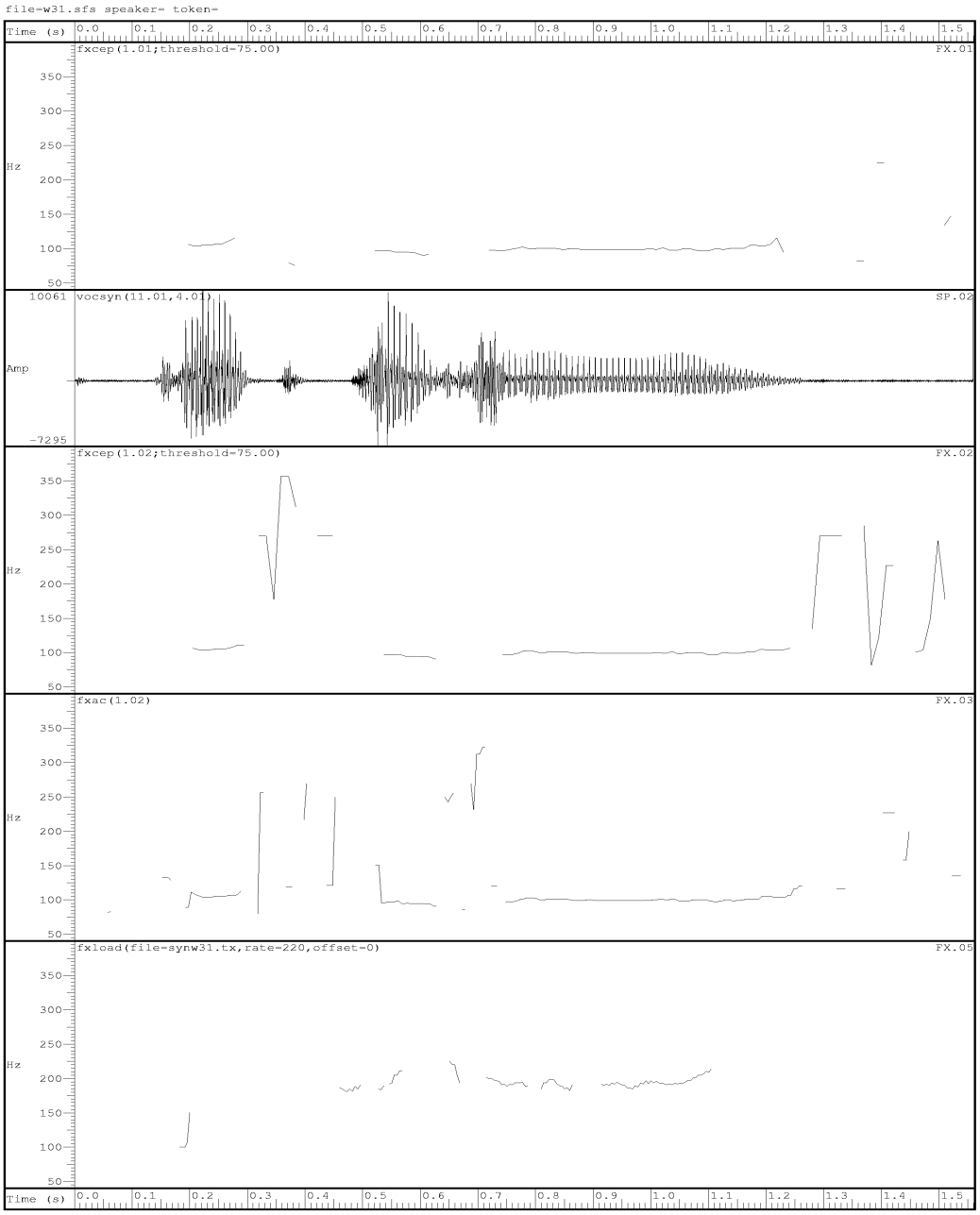


Figure 3.**38 Male synthetic speech pitch comparison. The graph is aligned as Cepestrum-based pitch for normal speech, Synthesizd speech, Cepstrum-based pitch for synthesized speech, Autocorrelation-based for syntheszed speech. Wavelet-based pitch for synthesized**.

Figure 3.37 is a repetition of figure 3.35 but for a male speaker. The second graph from the top is the cepstrum-based pitch contour. The third one is the wavelet-based pitch contour and the bottom one is the autocorrelation-based pitch contour.

Figure 3.38 is a graph for the synthesized word of figure 3.37. The top graph is the original pitch contour before synthesizing. The second one is the synthesized word. The third one is the cepstrum-based pitch contour for the synthesized word. The fourth one is the autocorrelation-based pith contour and the last one is the wavelet-based pitch contour. As indicated in the figure the best one fits the original curve is the cepstrum-based method. That is because the synthesized male utterance is approximately distorted with the synthesizer. The synthesizer is a vocoder synthesizer which generates the synthesized speech using the filter bank outputs and pitch contour.

3.5 Conclusion

In this chapter the problem of classifying the speech into voiced or unvoiced sounds is handled. Wavelt transform can represent the phonetic variation along the utterance duration. This property is used in two algorithms to find V/U boundaries. The correlation of wavelt parameters gives robust decision.

The pitch period estimation problem is handled using wavelet transform. It is clearly apparent that wavelet can keep track with pitch variation even in case of poor signal to noise ratio. Pitch contour that is generated using wavelet algorithm is more stable and smoothed than those generated using autocorrelation method or cepstrum.

CHAPTER 3 ..... **62**

CLASSIFICATION OF VOICED/UNVOICED UTTERANCES AND PITCH PERIOD ESTIMATION 62

3.1 INTRODUCTION **63**

3.2 VOICED / UNVOICED CLASSIFICATION **63**

3.2.1 VOICED SOUND VERSUS UNVOICED SOUND 63

..... 64

3.2.2 SIGNAL CHARACTERISTICS OF VOICED AND UNVOICED SOUNDS 65

3.3 VOICED / UNVOICED CLASSIFICATION USING DYADIC WAVELET **70**

3.3.1 DYADIC WAVELET 70

3.3.2 CLASSIFICATION USING SINGLE BAND 73

3.3.2.1 Training phase *77*

..... 79

3.3.2.2 Test phase *80*

3.3.3 CORRELATION BASED METHOD 86

3.3.3.1 Algorithm *86*

3.3.4 VOICED/UNVOICED CLASSIFICATION USING MATHEMATICAL MODEL BASED ON WAVELET FEATURES 90

3.4 PITCH PERIOD ESTIMATION **94**

3.4.1 THE PARALLEL PROCESSING METHOD 95

3.4.2 THE SIMPLIFIED INVERSE FILTER TRACKING SIFT METHOD 102

..... 104

3.4.3 PITCH ESTIMATION USING CEPSTRUM 104

3.4.4 PITCH ESTIMATION USING WAVELET 108

3.4.4.1 Detection of pitch using two band correlation of wavelet features *109*

3.4.4.2 Pitch detection using two wavelet based estimators in parrallel *113*

* RESULTS AND COMPARISON 116

3.5 CONCLUSION 122

Chapter 4

Speech segmentation and vowel recognition

4.1 Introduction

In this chapter, one of the most complicated areas of speech processing is considered. The speech segmentation into basic units is a very hard and complicated process. It is expected that the speech recognition systems will be enhanced if it is based on a reliable data bank. The best data bank is the basic speech units because each unit represents pure sound rather than a complex combination of sounds. This is the dream but if we try to get it into reality we must face a large number of problems. The first and basic one is no standard way to detect the phonemes until now. Many trials are made to find a model for phonemes [70] and in general it succeed only in a simple phonetic classification problem.

The phonemes in general are divided into two categories (vowels and consonants). Actually there is some extra categories such that Diphthongs and semivowels in some languages such as English. The study here is concentrated on the vowels and consonants only. The vowels and consonants have many differences in the acoustical characteristics. Also, there are many rules that control the existence of them into the context. These rules differ from one language to another. The vowels themselves behave differently according to their position in the utterance, also the consonants do the same.

For the above difficulties the problem is divided into three parts.

1- Determination of vowels and consonants boundaries.

2- Collecting database for each vowel and each consonant in the studied language.

3- Differentiating between different vowels and different consonants in the studied language.

The first step is the base of the two further steps. The efficiency of the next two steps will be affected dramatically if the first step is not handled with extra care. This work tries to solve the first step. To discriminate the different kinds of phonemes, a large database must be built, so that only vowels are taken as target and the whole recognition system can be left for future work.

The next section illustrates the acoustic phonetics in brief. Next the problem is handled using the wavelet transform.

4.2 Acoustic phonetics

Most languages can be described in terms of a set of distinctive sounds, or phonemes. A phoneme is the smallest unit of speech. It does not typically have meaning but is used to distinguish meanings between words. Number of phonemes can range between 30 and 40 depending on the language.

The brain decides what phonemes to be said. It then takes this sequence and translates it into neural commands that actually move the tongue, jaw, and lips into target positions. However, other commands may be issued and executed before these targets are reached, and this accounts for articulation effects.

Because we deal her with Arabic language, it has basically 34 phonemes containing 28 consonants and six vowels[72]. A List of Arabic language phonemes is introduced in Appendix A.

4.3 Method of segmentation

Segmentation of speech into vowels and consonants is manipulated in different ways. In this work, the following two methods are introduced:

1- Band selection method.

2- Math classifier method.

In addition to the above two methods, neural network is considered as a classifier. It gives poor results in case of using the same training data set supplied to the mathematical classifier. For the same test data it gives a recognition rate less than 45%. These poor results let us to exclude neural network from further work and concentrate the work on mathematical regression classifier. But it may be considered again in the future work of all phones recognition.

The above two methods depend on the features of the wavelet transform. Speech signal is captured, wavelet transform is applied then wavelet parameters are handled.

4.3.1 Band Selection Method (BSM)

4.3.1.1 Method description and algorithm

In this way, some wavelet bands are chosen for information extraction. Figure 4.1 outlines the worksheet for segmentation. The algorithm will be as follows:

- Figure (4.1 a) indicates the speech signal under test. The speech signal is captured using a 16-bit sound card.
- Speech signal is processed. It is framed into smaller frames with 1024 samples each. The wavelet transform is applied into all frames.
- The wavelet parameters in each frequency band are interpolated to achieve 1024 wavelet parameters in each frequency level. That is because each frequency level has a different number of parameters to describe the signal (see Table 1.1).

- Wavelet parameters are smoothed to eliminate unpredictable peaks. Smoothing is made by using the moving average of 200 samples (~20 ms in case of 11025 sampling rate). Wavelet features of the six bands are created. Table of figure (4.1 b) show two bands of the six bands.

- Figures (4.1 c),(d), (e),(f),(g),(h) are the graphical representation of the smoothed interpolated wavelet parameters. Figure (4.1 c) represents the frequency band 86-172 Hz, (d) represents frequency band 172-344 Hz, (e) represents frequency band 344-689 Hz, (f) represents frequency band 689-1378 Hz, (g) represents frequency band 1378-2756 Hz and (h) represents frequency band 2756-5512 Hz.

- The first 4 bands are taken into consideration because most of speech power is concentrated below 1000 Hz [5]. The idea is how to get the points of the large variation in the first four bands. At those points a transition from vowel to consonant occurs. In Arabic language there is no transition between vowel to vowel [72] rather there are always one of six patterns CV, CV:,CVC,CVCC,CV:C,CV:CC where C denotes to consonant and V denotes vowel and V: denotes to long vowel. So any transition in wavelet curves will occur at the boundaries of V or C. The only source of error is the pattern CC that can be overcome by choosing a reasonable threshold of variation.

- Figures 4.1 c, d, e and f are normalized and summed to construct a single curve that reflects any variation in any frequency band as shown in figure (4.1 i).

- To measure the stability of the curve of figure (4.1 i) and to find the points of large variations, the moving standard deviation of 550 points (~50 ms) is applied. 50 ms is a reasonable duration of phonemes to be stable [5] then figure (4.1 j) is created.
- Figure (4.1 k) is compared with a reasonable threshold (obtained from many trials of different cases) to get markers at the large transition boundaries. The word in figure (4.1 k) is كتاب in Arabic. It is pronounced /k//i//t//Θ/β/. The pattern of it is CVCV: C. It is clear in figure (4.1 k) that the markers surround the vowel periods.

Figure 4.2 illustrates many other examples of V/C classification. Figure 4.2 a is an Arabic word يكتب. It contains V/C pattern as CVCCVC. The problem of CC is appearing here. As shown in window –k- in figure 4.2 a , markers bound all vowels.

The last marker of figure (4.2 a) represents sudden change of the stop plosive /b/. It is a false marker that can easily be removed by the software where no period is detected for a vowel.

Figure (4.2 b) represent the word يكتبون. Window k shows wrong markers within the duration of the first vowel. It can be removed by checking on the duration between the two markers is very low with respect to a vowel duration. The second check is that no adjacent vowels can be found in Arabic language.

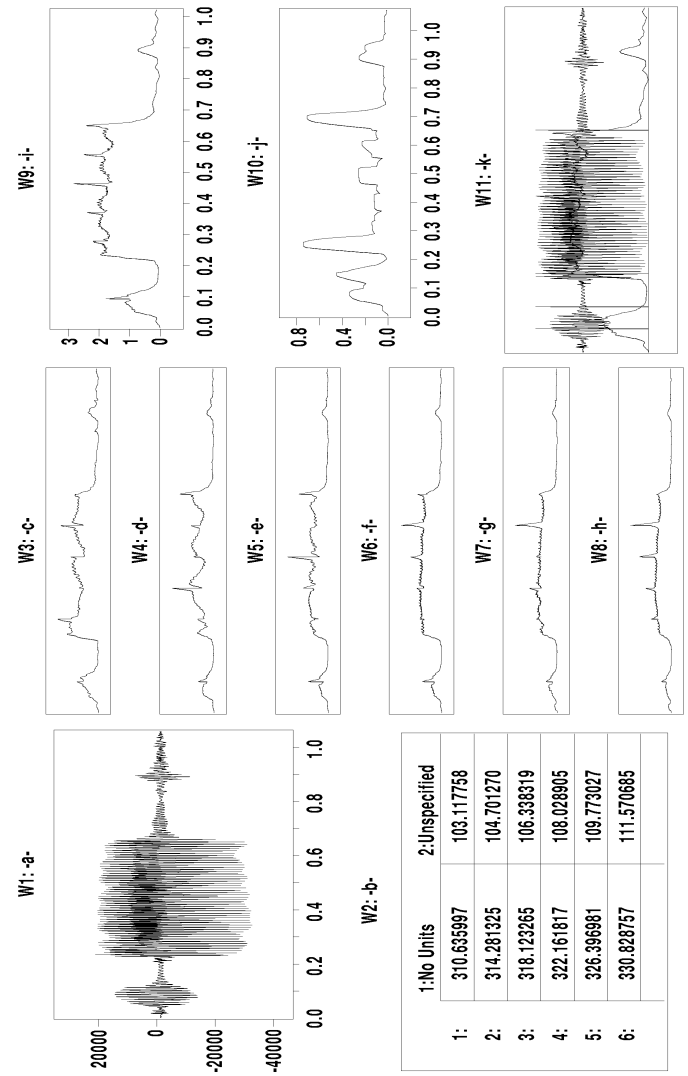


Figure 4. 1 Work Sheet for V/C using wavelet. a-speech signal, b-wavelet table, c through h graphical representation of each column in the wavelet table, i- Summation of normalized curves c through h. j- moving standard deviation of –i-, k- Speech signal overlaid with V/C markers.

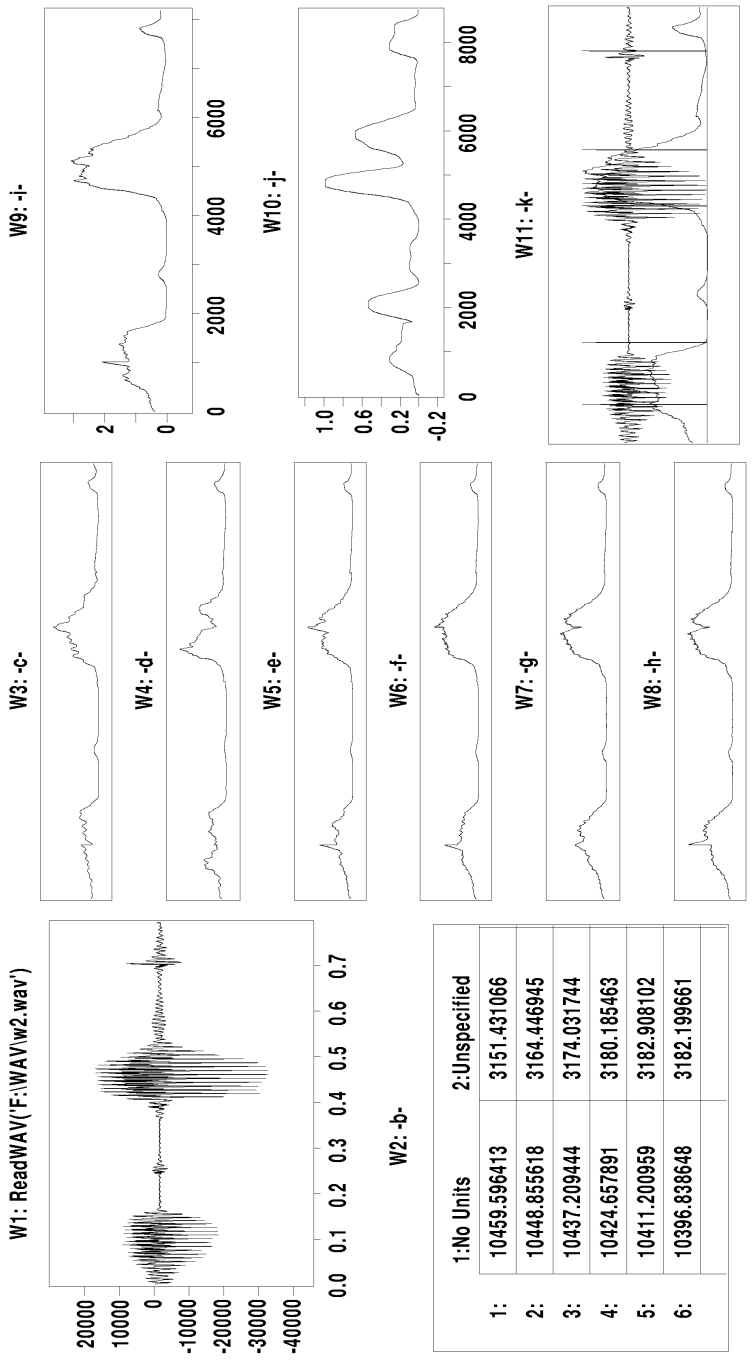


Figure 4.2 a- Work sheet for V/C classification by the band selection method. Word يكتب in Arabic. It contains CVCCVC.

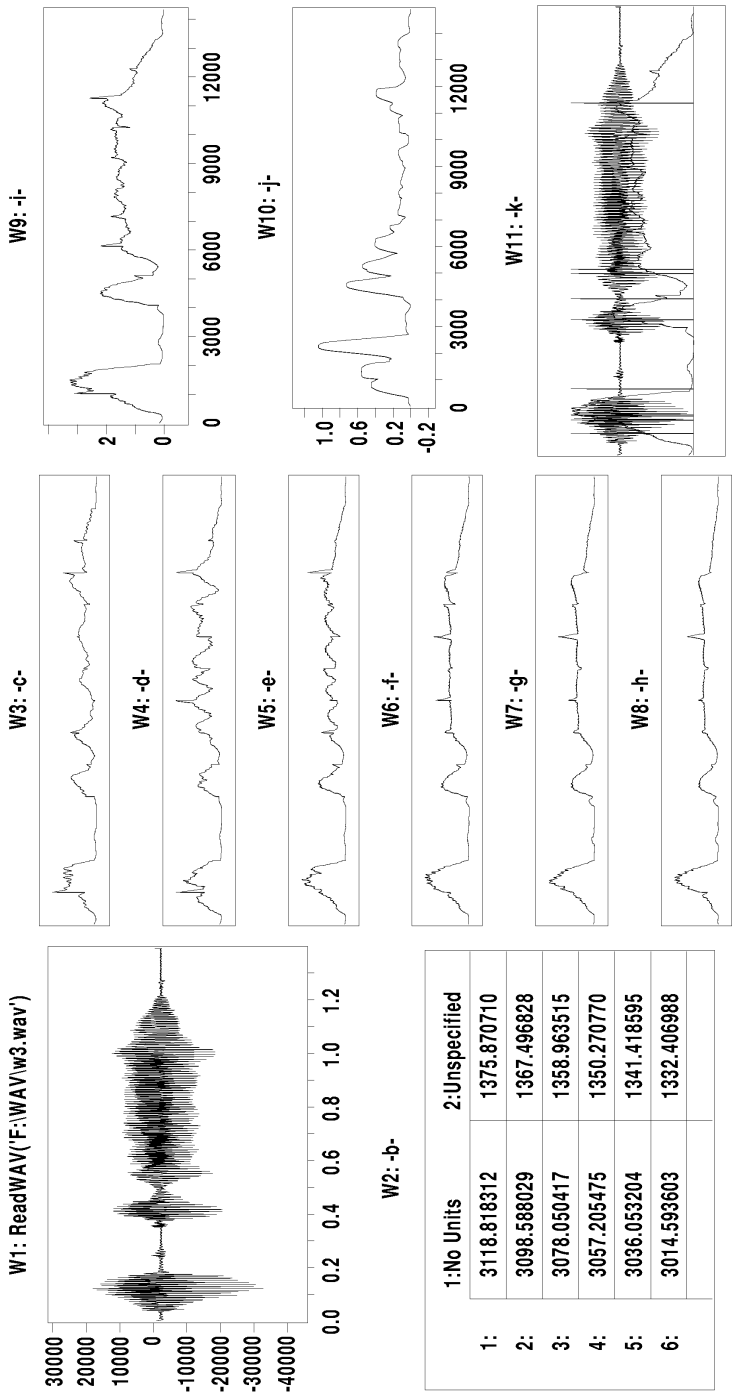


Figure 4.2-cont. b- Work sheet for V/C classification by the band selection method. Word يكسرن in Arabic. It contains CVCCVCV:C.

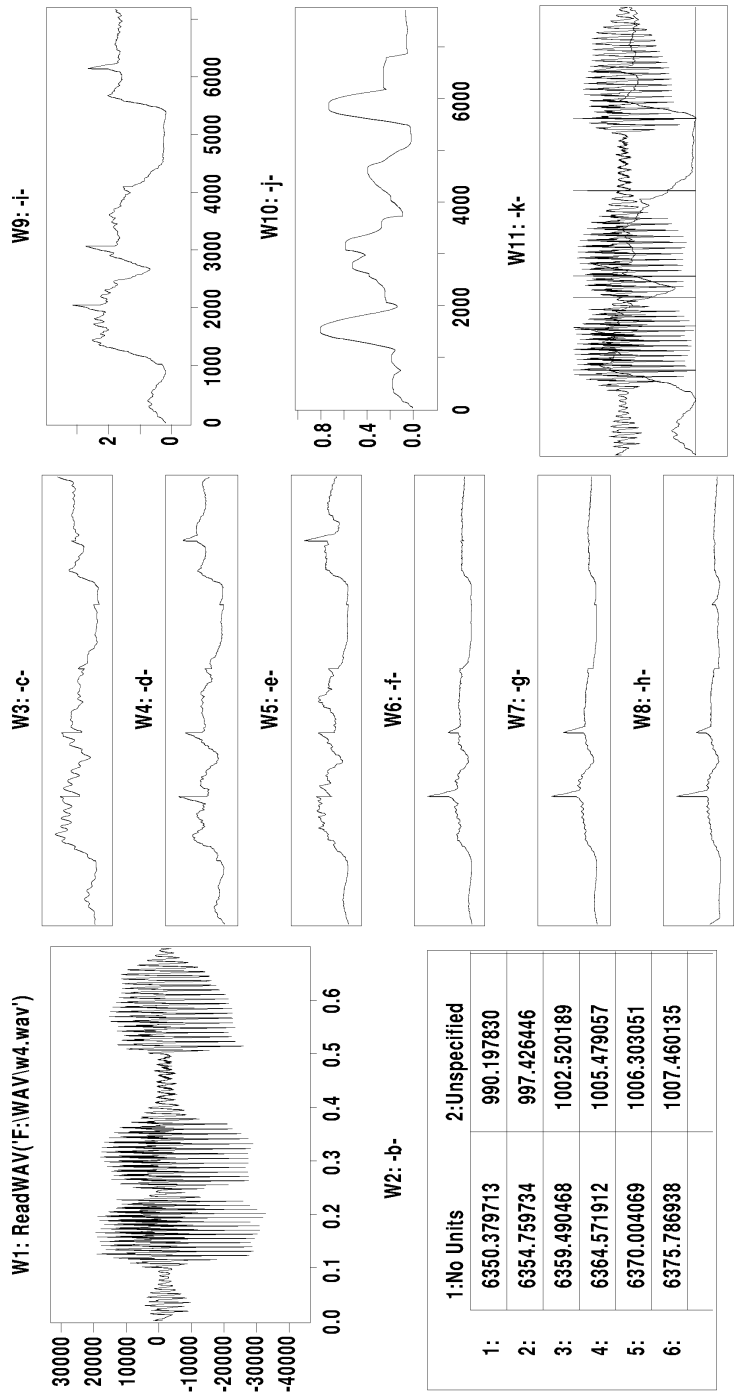


Figure 4.2-cont. c- Work sheet for V/C classification by the band selection method. Word دَرَسَ in Arabic. It contains CVCVCV.

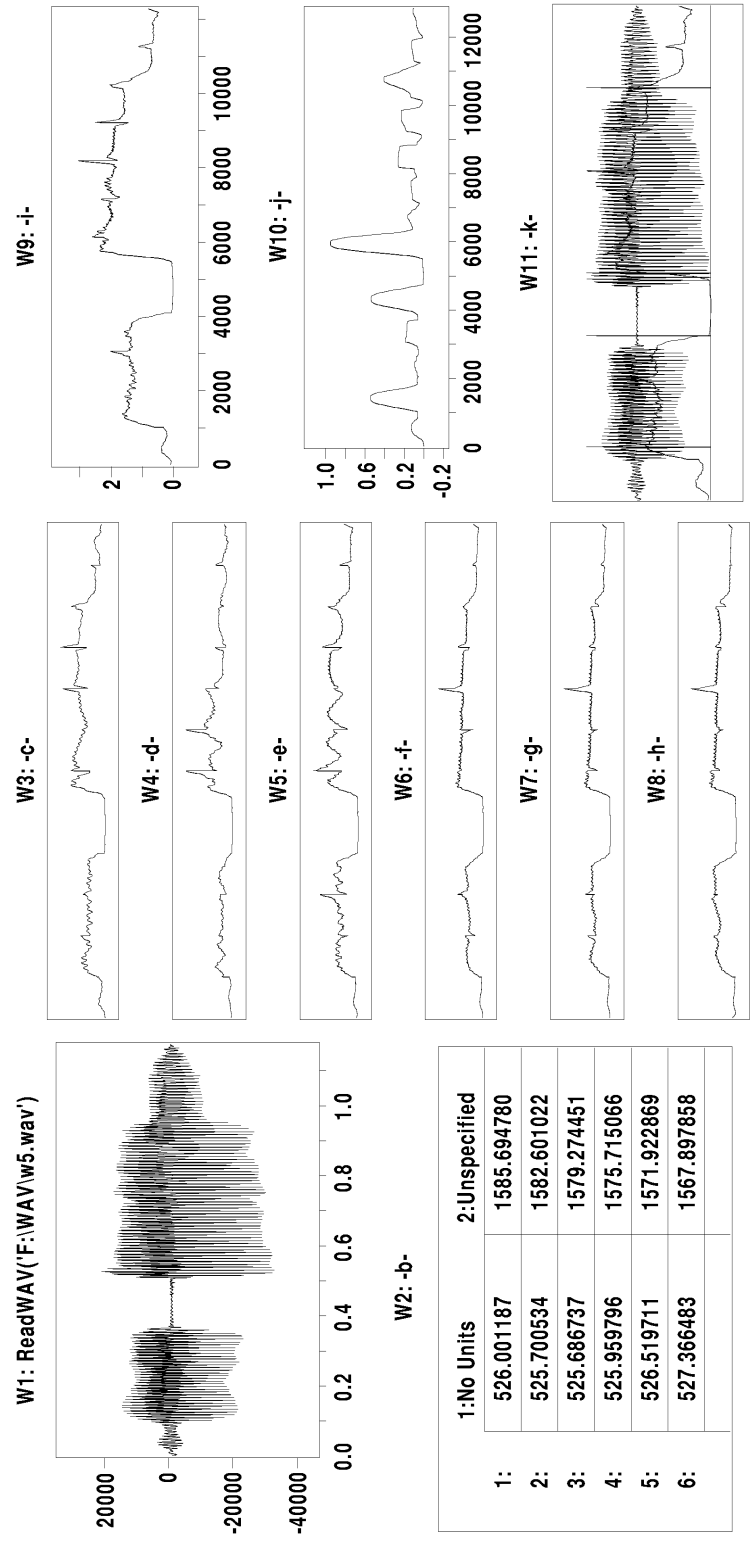


Figure 4.2-cont. d- Work sheet for V/C classification by the band selection method. Word باقون in Arabic. It contains CVCV:C.

4.3.1.2 Test and evaluation

The system is evaluated in the presence of noise. White noise is superimposed on speech signals to achieve different signal to noise ratios. Steps of efficiency measure are:

1. Vowel periods are those speech periods between vowel markers. Vowel periods are marked high “1” and consonant periods are marked low “0”. False markers will be rejected by the software check.
2. Pre-calculation of V/C periods is made for about 14 minutes of speech under test. (Actual classification from manual test)
3. White noise is superimposed on speech under test to control signal to noise ratio.
4. The algorithm of V/C using band-selected method is applied on speech of step 3.
5. V/C periods are obtained using BSM method.
6. Error signal is calculated by subtracting curve of step 5 from curve of step 2 and taking the absolute value as shown in figure 4.3.
7. To make a tolerance, the curve of step 6 is shifted 5 ms to generate a tolerance curve see figure 4.3.
8. Curve of step 6 is multiplied with curve of step 7 to remove the tolerance periods from the error signal, figure 4.3 c.
9. The total error is the summation of error periods of the curve in step 8.

10. $POS = (1 - \frac{T_{err}}{T}) * 100\%$. Where T_{err} is total duration of error and T is total period and η is the efficiency.

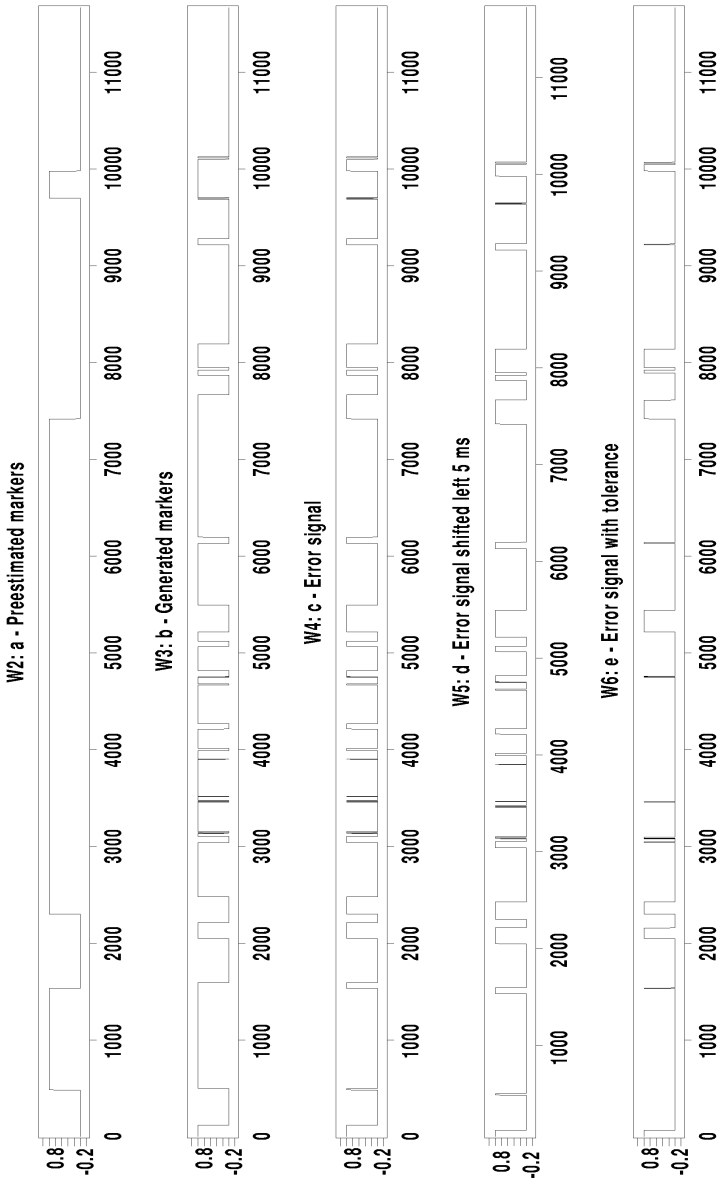


Figure 4. 3 Error calculation.

Figure 4.4 illustrate the performance of the system for different signal to noise ratios.

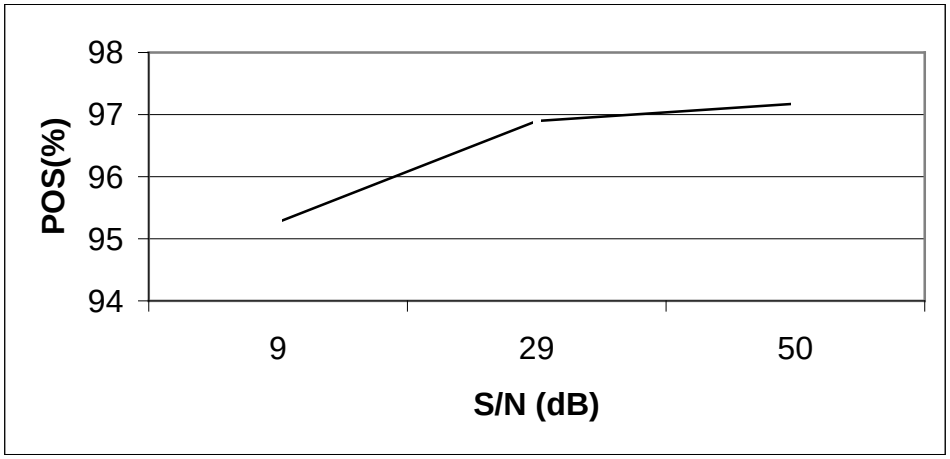


Figure 4. 4 System performance in the presence of noise for vowel/consonant classification using BSM.

4.3.2 Math classification method (MCM)

It is obviously clear that BSM (Band Selection Method) gets its information from selected bands and ignore the other bands. In this section, all bands are taken into consideration and a mathematical way is used to get the combined information from all bands. A mathematical linear regression is used here to handle vowel consonant classification[47].

4.3.2.1 Training phase

A training data is prepared for regression process.

1. A training period of 27 sec of phonemes is used to prepare the training data set.
2. Wavelet parameters are extracted, interpolated and smoothed as previous method. The bands under study are six bands.

3. Training matrix is prepared. It contains rows called X-vectors. Each row represents the power distribution of the signal at certain time in the different six bands.

4. X-vector contains 6 elements as follows:

   X[i] = { B0,B1,B2,B3,B4,B5}

   Where each element in vector X represents the wavelet function (smoothed interpolated wavelet parameters) at time index i in the frequency bands 86-172Hz, 172-344 Hz,344-689 Hz, 689-1378 Hz, 1378-2756 Hz, 2756-5512 Hz respectively.

5. A pre-estimation of the state of X[i] vector into Vowel or consonant is made manually. The decision is put into vector Y. The i^{th} element of Y is a decision of x[i] vector as indicated below:

X						Y
B0	B1	B2	B3	B4	B5	
54000	30200	2230	1000	650	120	1 or 0
⋮		⋮		⋮		⋮
23223	20345	5428	300	250	70	1 or 0

6. Y is regressed on X to find the mathematical model of the system as equation 4.1.

$$\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & X_{03} & X_{04} & X_{05} & X_{06} \\ \vdots & & \vdots & & \vdots & \\ X_{m1} & X_{m2} & X_{m3} & X_{m4} & X_{m5} & X_{m6} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \qquad \textbf{(4. 1)}$$

Equation (4.1) represent the system equation. [B] Matrix is the system model that is obtained from training as discussed above..

$$[B] = \begin{bmatrix} 0.0009 \\ -0.0001 \\ 0.0037 \\ 0.0024 \\ 0.0278 \\ -0.0225 \end{bmatrix}$$

4.3.2.2 Test phase

To Evaluate the efficiency of this method a test data from the database is applied on the system matrix according to equation (4.1) with different signal to noise ratios. The steps are as follows:

1. Vowel periods are those speech periods between vowel markers. Vowel periods are marked high "1" and consonant periods are marked low "0".

2. Pre-calculation of V/C periods is made for about 14 minutes of speech under test.

3. White noise is superimposed on speech under test to control signal to noise ratio.

Wavelet transform is applied on the speech under test. The wavelet parameters are prepared as shown in figure (4.1 c) through (h) of word كتاب.

4. X-vector is created each 2 ms.

5. [X] Matrix is multiplied with [B] vector that is obtained in the training phase. [Y] vector is obtained from the previous multiplication.

6. [Y] vector contains high at the vowel periods and low at the consonant periods. It can be represented graphically as per-calculated periods of step 2.

7. Error Calculated by the same way as previous method BSM..

Figure 4.5 a and b illustrate examples of V/C markers generated using this method.

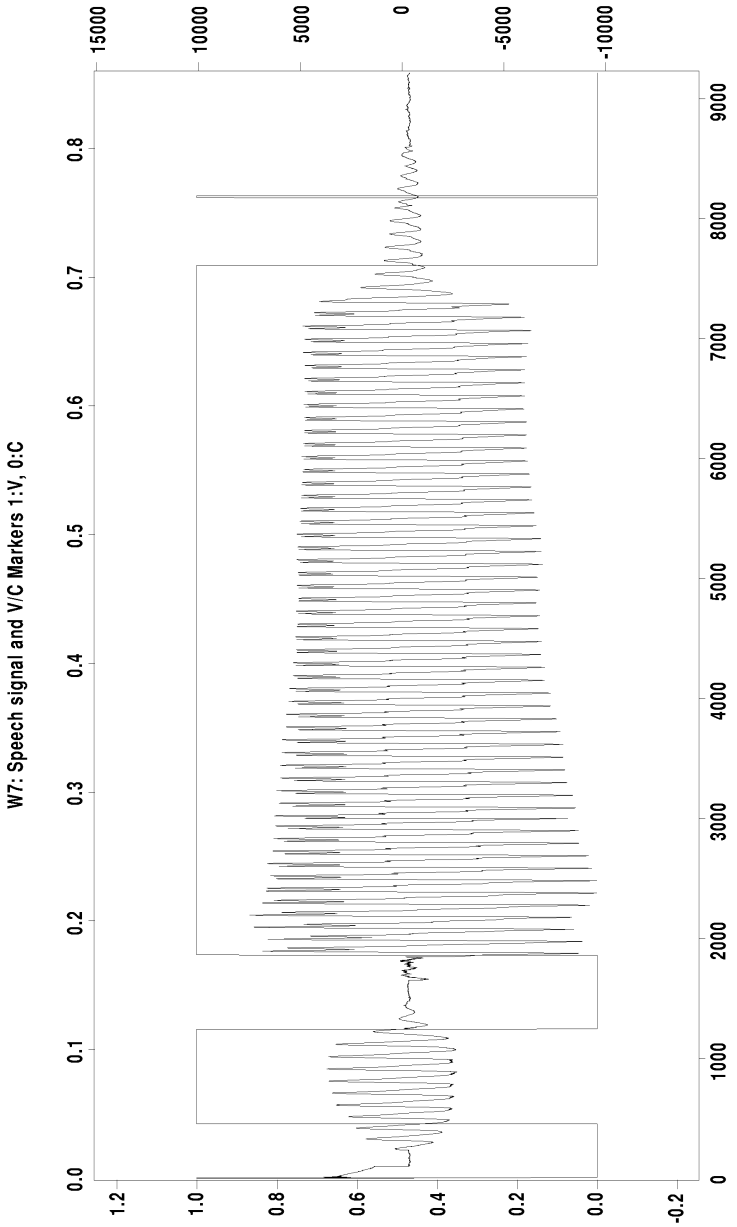


Figure 4.5 a- V/C classification by MCM method. Word كتاب in Arabic. It contains CVCV:C.

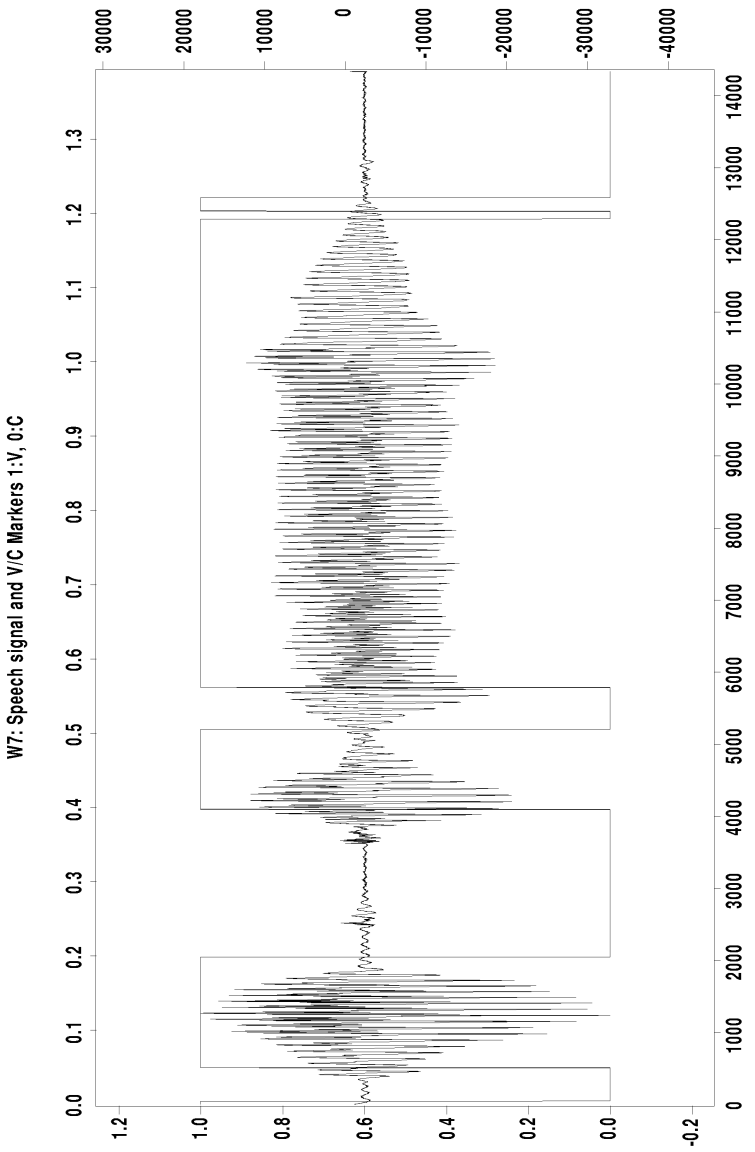


Figure 4.5-cont. b- V/C classification by MCM method. Word يكتبون in Arabic. It contains CVCCVCV:C.

Figure 4.6 summarizes the output results.

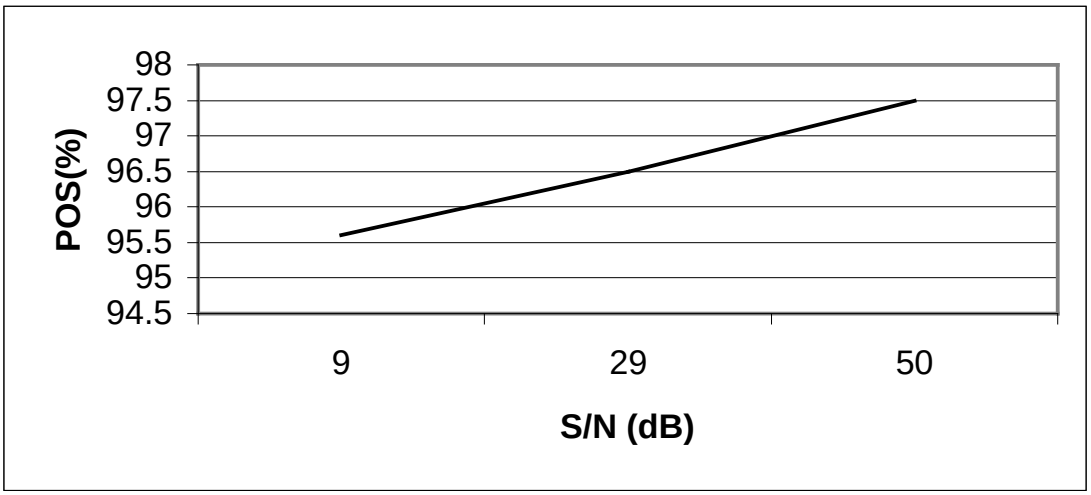


Figure 4. 6 V/C system performance in case of MCM.

4.4 Vowels recognition

In Arabic language there are only 6 main vowels. Three short vowels and three long vowels. The short vowels are / ◌َ / فتحه / □ / ضمه / ◌ِ / كسره.

The long vowels are /Θ/, /o/ and /i/. In this part, discrimination of vowels is the target. It is very difficult to discriminate between them by using the wavelet features of single band because all of them have approximately the same characteristics of high-energy distribution over a low frequency range [5].

Expressing speech signal with wavelet parameters makes a joint time-frequency representation of the signal. That makes it possible to trace the variation of energy with time in different frequency bands. Vowels are closely alike in frequency and time domains but their characteristics in wavelet bands are different a little along the time. So, it is expected that all bands are important for differentiating between vowels.

In section 4.3.2, the problem of V/C (Vowels and consonant) classification was handled using information supplied from all wavelet bands. The problem here is treated by the same approach of mathematical-based classification.

4.4.1 Vowel classifications using a single math classifier

4.4.1.1 Training phase

In this case the training data set is aligned as section 4.3.2. In this problem of classification there are three different values in the decision vector Y. Steps of training are:

1. Apply V/C algorithm on the training data set to obtain the vowel boundaries.
2. Verify of boundaries manually to insure that error-free data set in the training phase is obtained.
3. Calculate the wavelet features of the training data set as those of figure 4.3 –c- through –h-.
4. Construct X-vectors every 2 ms as illustrated before in section 4.3.1 and 4.3.2.
5. There are three possible decisions of each X-vector as follows:

   <1> in case of vowel /Θ/ or / ◌́ /.

   <2> in case of vowel /i/ or / ◌ /.

   <3> in case of vowel /o/ or / □ /.

X	Y

B0	B1	B2	B3	B4	B5	
54000	30200	2230	1000	650	120	1 or 2 or 3
	⋮			⋮		⋮
35155	24254	2341	2134	432	432	1 or 2 or 3
56234	31435	1223	1236	643	21`	1 or 2 or 3

6. Y vector is supplied with the proper decision value of each X-vector.
7. Y is regressed on X to obtain the system model [B]

$$[B] = \begin{bmatrix} 0.0050 \\ 0.0009 \\ 0.0162 \\ -0.0062 \\ -0.0849 \\ 0.0592 \end{bmatrix}$$

4.4.1.2 Test and evaluation

In this part the system is tested on 12 minutes of speech data containing different vowels. Steps of testing are:

1. Test speech data are prepared to extract the X-vectors every 2 ms.

2. V/C algorithm is applied to extract the vowel periods.
3. A pre-calculation of vowels is made to get the reference markers.
4. X-vectors that correspond to vowel periods are collected into [X] matrices. Each matrix of the [X] matrices contains a collection of X-vectors within a vowel period.
5. Each [X] Matrix is multiplied with [B] vector to get the decision vector [Y].
6. Efficiency is made by comparing the decision vector [Y] of each vowel with the pre-calculated one at step 3.

Figure 4.7 illustrates one sample of the test process. Figure (4.7 a) is speech signal that contains the Arabic word كِتاب. This word contains two vowels. The first one is /i/ and the last one is /Θ/. Figure (4.7 a) indicates that in each vowel period there are more than one decision for the vowel. Note that vowel markers has a level of 4 in figure (4.7 a). /Θ/ has a level 1 , /i/ has a level 2 and /o/ has a level 3. Calculating the maximum stable period of [Y] within the vowel period makes the final decision.

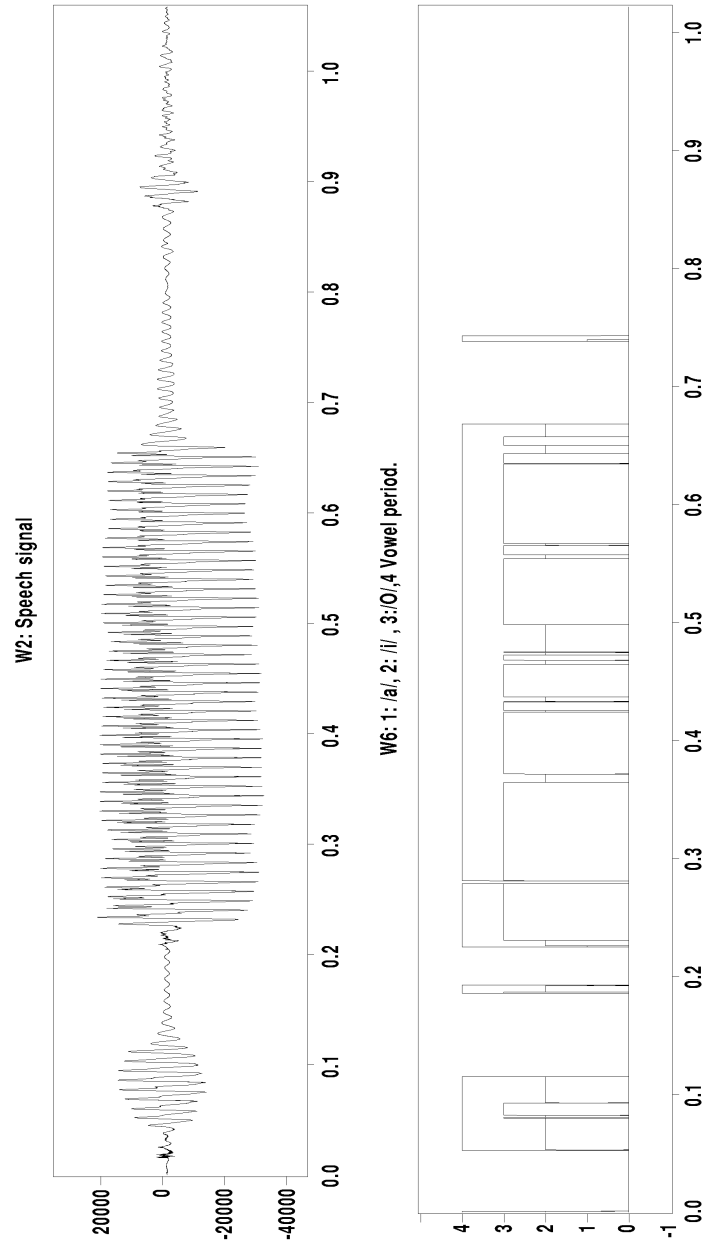


Figure 4. 7 a- Vowel recognition using single math classifier.

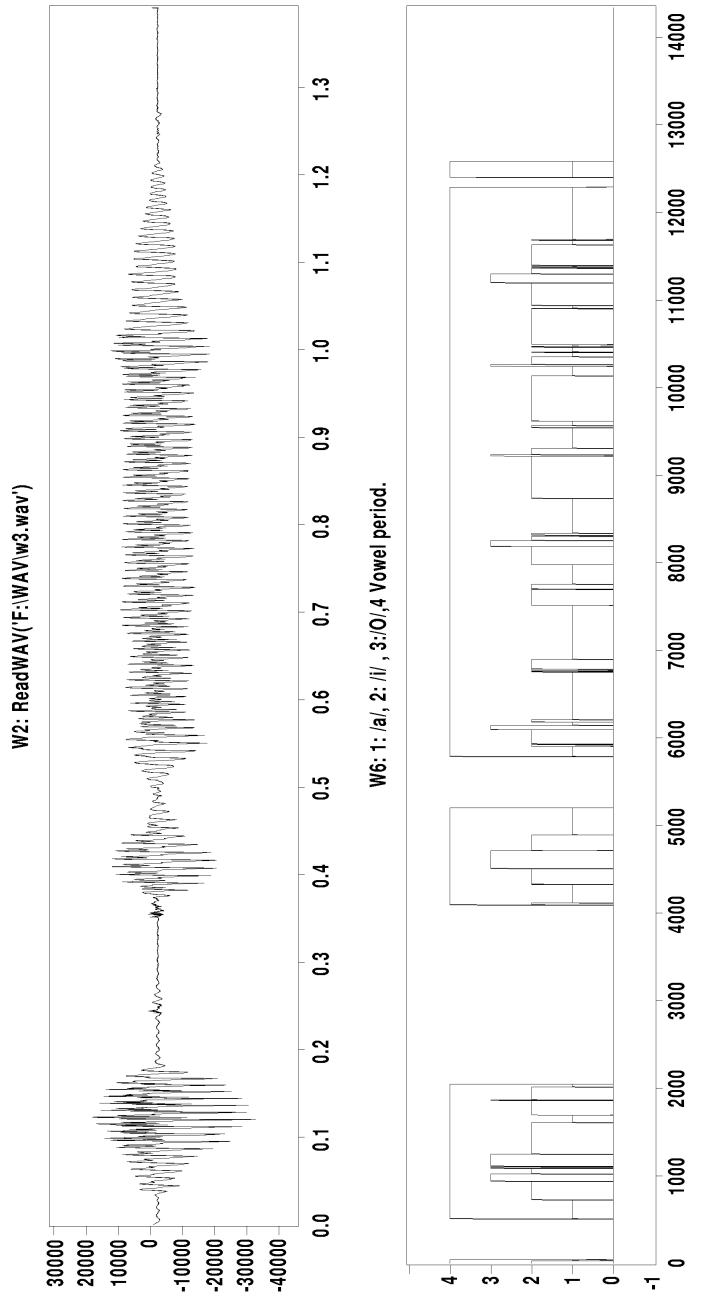


Figure 4. 7 cont. b- Vowel recognition using single math classifier.

Figure (4.7 b) is another example that indicates that one linear mathematical classifier is not sufficient to distinguish between vowels.

This technique failed to give acceptable results. It can not differentiate /o/ and /i/ but it gives good results in case of /a/. Recognition accuracy less than 53% is obtained.

The linear regression process can not find a suitable single system model that can distinguish the three kinds of vowels. That leads to the idea of parallel processing of the vowel. In other words, What will happen when treat the problem using three different Systems model working in parallel. Each system is responsible to find one of the three basic vowels.

4.4.2 Vowel classification using multiple math classifiers

The problem of low recognition rate in case of handling all vowels with a single classifier directs the work to a parallel classification. In this case each vowel is handled with a separate classifier.

4.4.2.1 Training phase

In this section database is collected and prepared to design three system models for the three different vowels. The process is as follows:

1. Training data set is prepared as in section 4.4.2.1.
2. [X] matrix is created by collecting X-vectors each 2 ms.
3. Three different [Y] vectors are created. Each one gives two decisions "1" in case of the focused vowel and "0" in case of other vowels as shown in the following tables.

X						Y1
B0	B1	B2	B3	B4	B5	

54000	30200	2230	1000	650	120	1
⋮			⋮			⋮
21342	12113	1233	6541	341	121	1 or 0
21412	76542	1243	3532	321	464	1 or 0

Training set of vowel /Θ/

X						Y2
B0	B1	B2	B3	B4	B5	
54000	30200	2230	1000	650	120	1 or 0
⋮			⋮			⋮
21342	12113	1233	6541	341	121	1 or 0
21412	76542	1243	3532	321	464	1 or 0

Training set of vowel /ι/

X						Y3
B0	B1	B2	B3	B4	B5	

54000	30200	2230	1000	650	120	1 or 0
⋮			⋮			⋮
21342	12113	1233	6541	341	121	1 or 0
21412	76542	1243	3532	321	464	1 or 0

Training set of vowel /o/

4. Each [Y] matrix is regressed on the same [X] matrix to obtain a system model. So, three different system models are obtained each one corresponds to a different vowel (B1,B2,B3).

$$[B1] = \begin{bmatrix} 0.0003 \\ 0.0007 \\ -0.0068 \\ 0.0030 \\ 0.0859 \\ -0.0460 \end{bmatrix}$$

$$[B2] = \begin{bmatrix} 0.0014 \\ -0.0005 \\ 0.0024 \\ 0.0069 \\ -0.0309 \\ 0.0021 \end{bmatrix}$$

$$[B3] = \begin{bmatrix} 0.0014 \\ -0.0005 \\ 0.0024 \\ 0.0069 \\ -0.0309 \\ 0.0021 \end{bmatrix}$$

[B1] is the system model of vowel /Θ/ , [B2] is the system model of vowel /o/ and [B3] is the system model of vowel /i/.

4.4.2.2. Test and evaluation

In this phase the system is tested using the same speech data as in section 4.4.2.1. Steps of the test are as follows:

1 X-vectors are obtained as in 4.4.2.1

2 V/C algorithm is applied on the speech under test to extract the vowel periods.

3 X-vectors of each vowel are collected into different [X] matrices.

4 Pre-calculation of vowels is made to construct reference markers.

5 Each [X] matrix is multiplied with the three [B] matrices. That generates three different [Y] matrices each one gives focus on one vowel corresponding to the system matrix which generate it..

6 [Y] matrix that gives a maximum area under its curve within the vowel period which indicates that the corresponding vowel is the decision.

Testing the above system on the test database of 4 minutes gives a correct recognition rate of 80.6%.

Figures 4.8 and 4.9 indicate the process. Figure 4.8 is Arabic word كتاب. It contains two vowels /i/ and /Θ/. V/C algorithm is applied on it to get the boundaries of vowels as it is shown in figure 4.8. Figure 4.9 represents the integration of [Y] vectors in regions of each vowel to get the area under their curves. [Y] Vector that represents the maximum area gives the decision. In other words, the system that generates [Y] vector of the maximum area is the system of the target vowel. If that system is for /i/ detection then the decision is /i/ and so on.

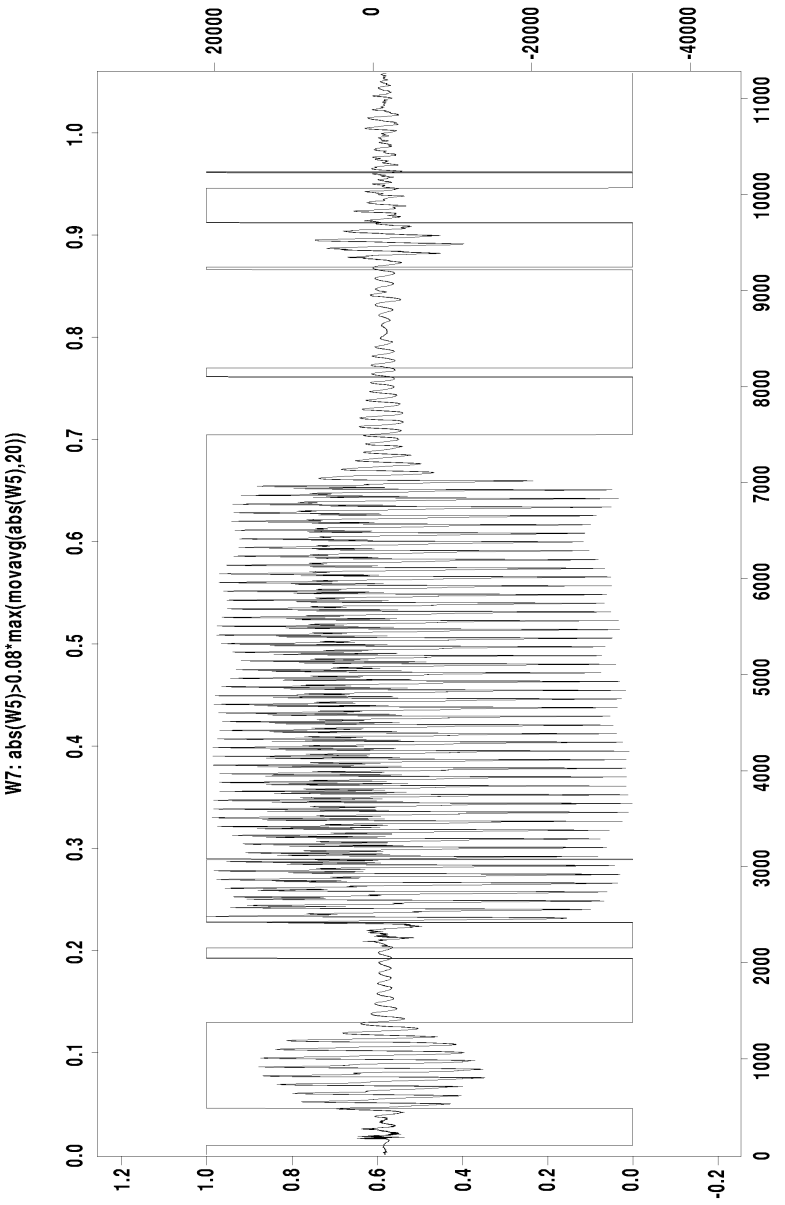


Figure 4. 8 V/C Markers. The word is /κ//ι//τ//Θ//βH/. High in markers represents Vowel and low or zero is a consonant.

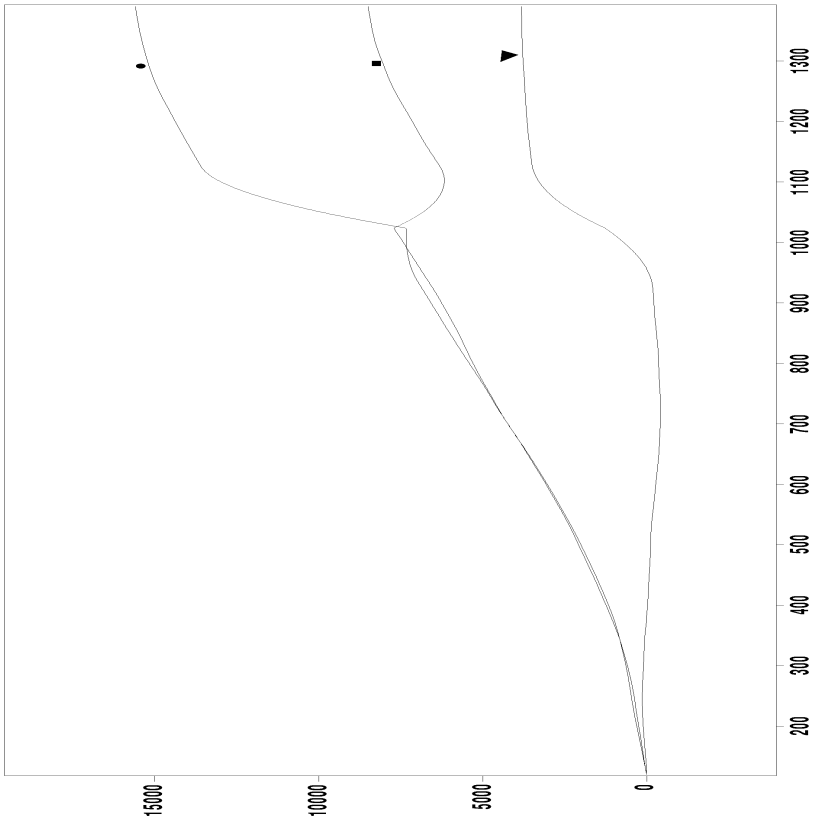


Figure 4. 9 Bound integration of the three [Y] vectors. ○ for /i/ and □ for /Θ/ and Δ for /o/. This integration is made for the first vowel of word كتاب.

Figures 4.10, 4.11 and 4.12 indicate a complete example of vowel recognition using multiple math classifier. As shown in figure 4.10 c, the bound integration of [Y] vectors give indication that /Θ/ and /o/ having the same probability in this period. Figure 4.11 c and figure 4.12 c give indication that the vowel is /o/.

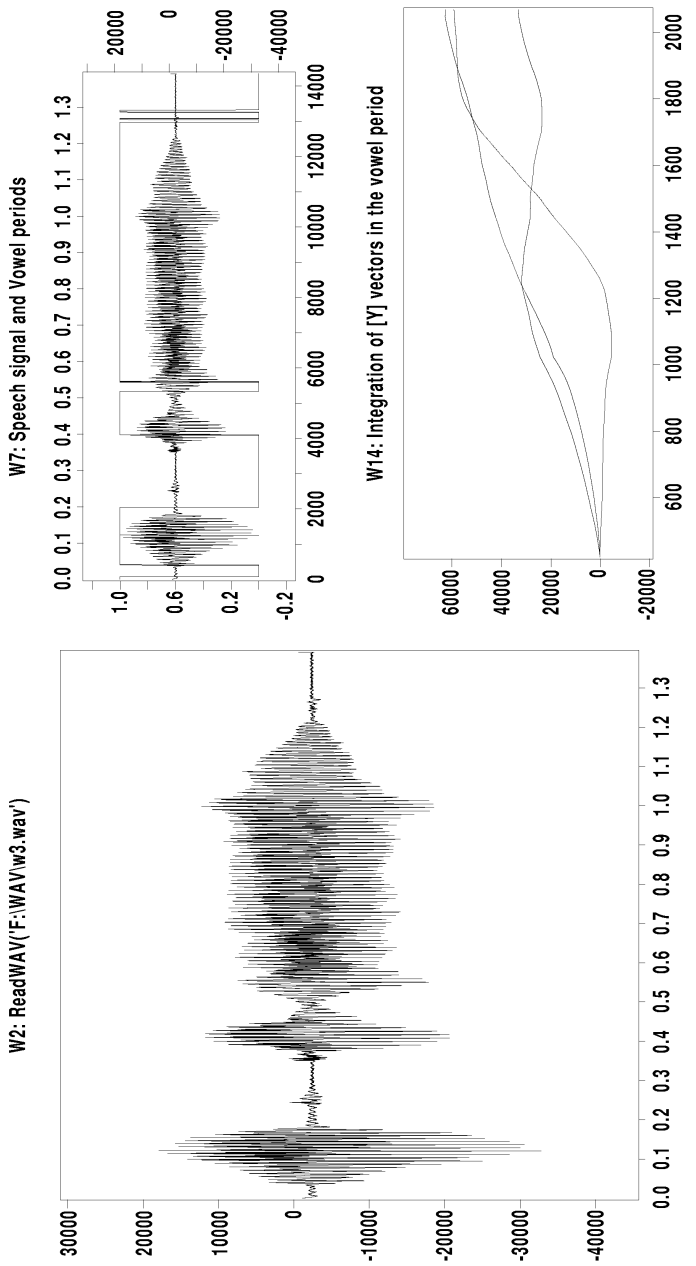


Figure 4.10 Vowel recognition using MMC method. The word is يكتبون in Arabic. Work sheet indicates the integration of [Y] vectors in the first vowel period.

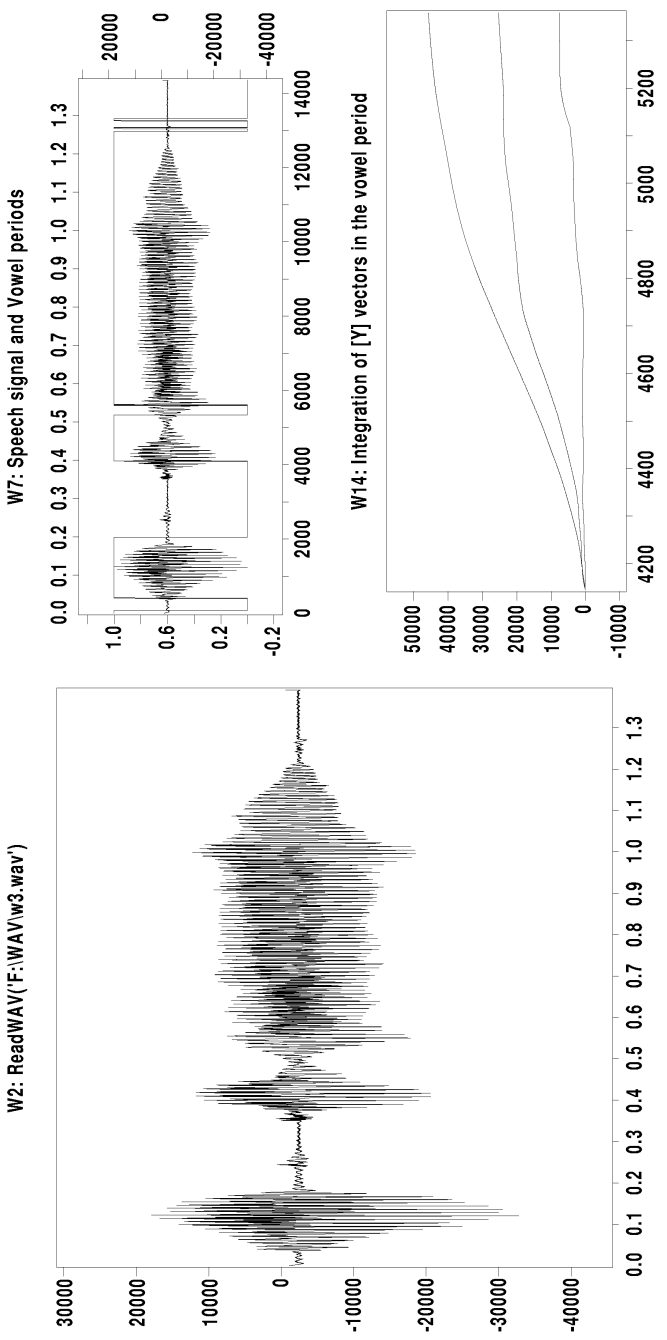


Figure 4.11 Vowel recognition using MMC method. The word is يكتبون in Arabic. Work sheet indicates the integration of [Y] vectors in the second vowel period.

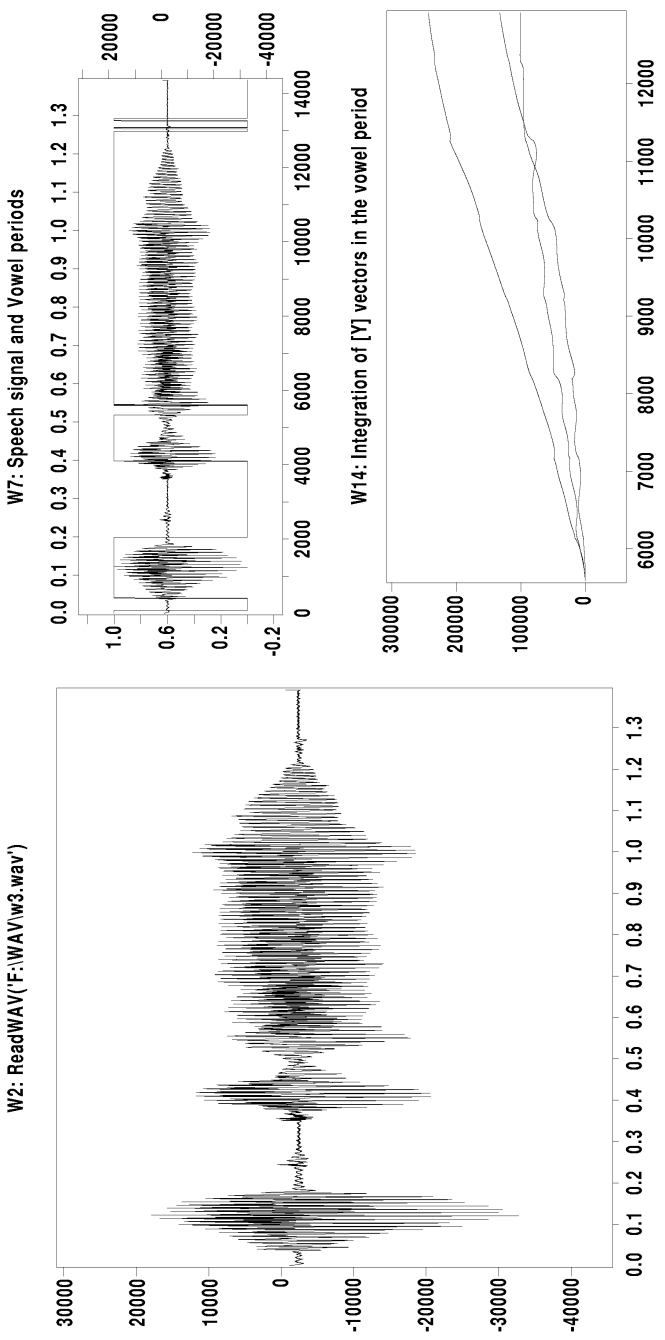


Figure 4.12 Vowel recognition using MMC method. The word is يكتبون in Arabic. Work sheet indicates the integration of [Y] vectors in the third vowel period.

4.5 Conclusion

Wavelet transform can be used in problems that needs joint time frequency analysis. The problem Of V/C classification is solved here using wavelet based algorithm. The technique is highly sensitive to acoustical variation along utterance duration.

With mathematical handling of wavelet parameters that represents the vowels, the problem of Arabic vowel recognition is solved. The recognition of Arabic vowels is high accurate relative to similar methods of English language.

4.1 INTRODUCTION 127

4.2 ACOUSTIC PHONETICS 128

4.3 METHOD OF SEGMENTATION 128

4.3.1 BAND SELECTION METHOD (BSM(..... 129

4.3.1.1 Method description and algorithm 129

..... 136

4.3.1.2 Test and evaluation 137

4.3.2 MATH CLASSIFICATION METHOD (MCM(..... 139

4.3.2.1 Training phase 139

4.3.2.2 Test phase 141

4.4 VOWELS RECOGNITION 145

4.4.1 VOWEL CLASSIFICATIONS USING A SINGLE MATH CLASSIFIER 146

4.4.1.1 Training phase 146

4.4.1.2 Test and evaluation 147

4.4.2 VOWEL CLASSIFICATION USING MULTIPLE MATH CLASSIFIERS 151

4.4.2.1 Training phase 151

.4.4.2.2 Test and evaluation 154

4.5 CONCLUSION 161

Chapter 5

System implementation

5.1 Introduction

This chapter illustrates the methods, which are discussed in the previous chapters, in work as a complete speech analysis system. The system is called SpeechLab.

The system now covers the following topics:

1- Speech acquiring.

2- End points detection using energy &zero crossing , wavelet based method and mathematical classification based on wavelet methods.

3- Voiced/Unvoiced classification using tracking function and mathematical classification methods.

4- Pitch period estimation using wavelets, autocorrelation and cepstrum methods.

5- Vowel/Consonants classification using wavelet transform.

6- Arabic vowels recognition.

It is proposed to extend it, in the future, to cover all Arabic phonemes using wavelet transform.

5.2 Block diagram of the system

Figure 5.1 is a block diagram of the SpeechLab system in a simplified form. The complete one is very complicated because of interconnections between the basic blocks.

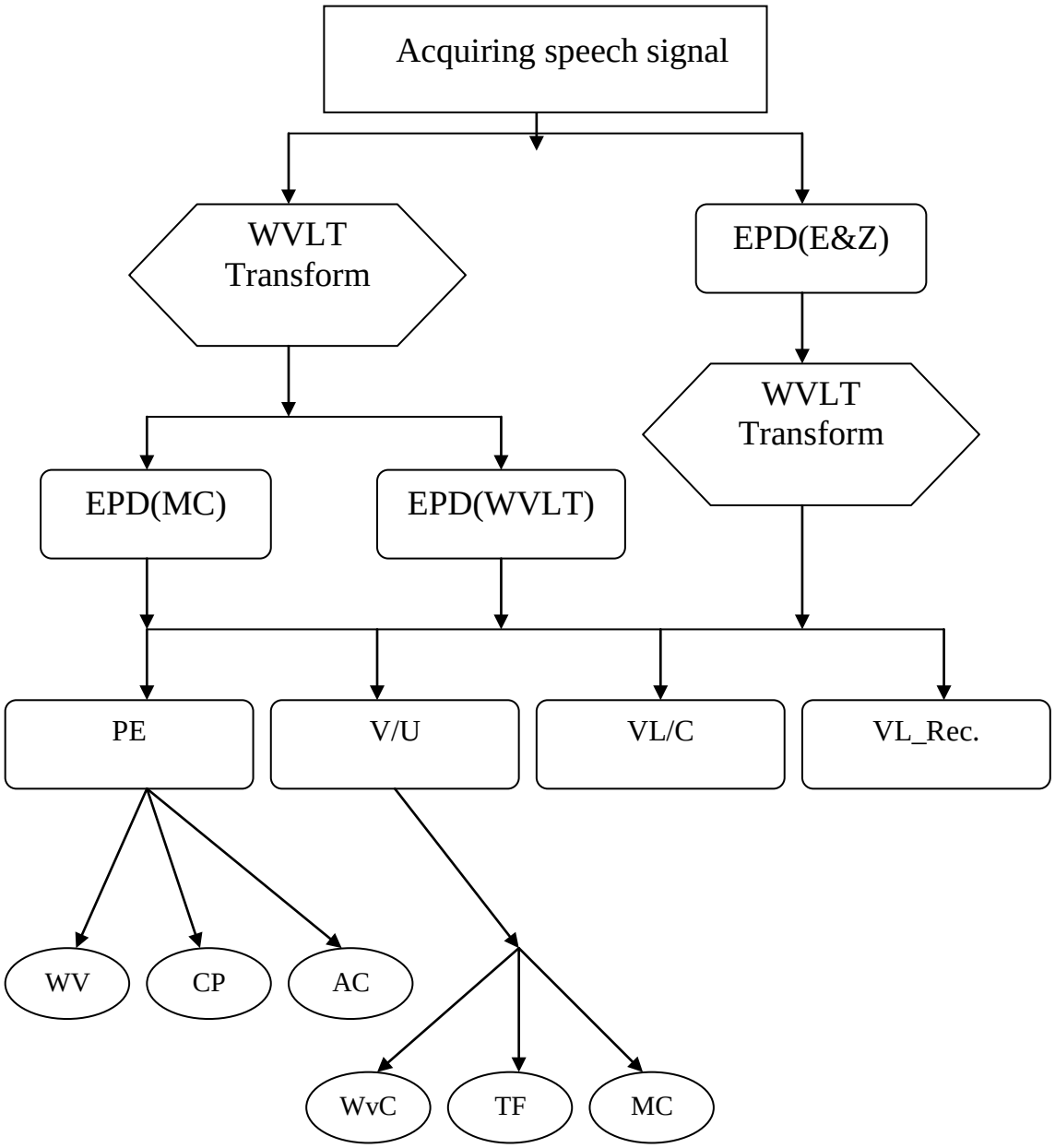


Figure 5.1: Block diagram of the complete system model.

Where:

- EPD: End points detection.
- PE : Pitch Estimation.
- WV : wavelet method.
- CP: Cepstrum Method.
- AC: Autocorrelation method.
- WvC: Wavelet Correlation Method.
- TF: Tracking function method.
- MC: Mathematical classifier of wavelet parameters.
- V/U : Voiced /Unvoiced classification.
- VL/C: Vowel / Consonants classification.
- VL_Rec: Arabic vowel recognition.

As shown in figure 5.1, the first step is to capture speech signal. There are three methods of capturing, by microphone, or from file , or from examples. The first two methods are used in the interactive mode, which allow the user to control the program. The last one is a demo mode, which the user lose the control.

The core of those programs are made using DaDisp 4.1[1]. All algorithms are made using SPL (series processing languages inside DaDisp) and the microsoft visual basic is used for interface.

Figure 5.2 illustrates the interface of the complete system.

[1] Digital signal processing software introduced in appendix.

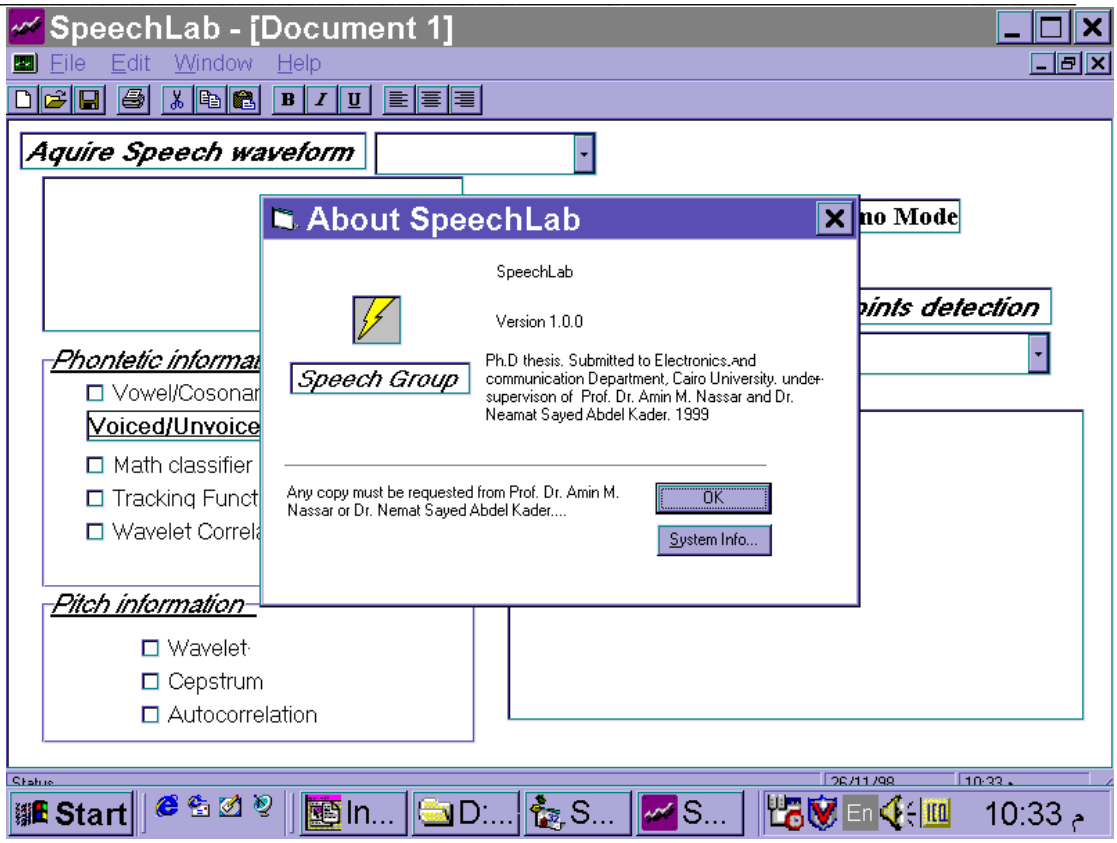


Figure 5. 2 The interface of the complete system.

5.3 The implemented system

In this section, SpeechLab will be illustrated. The system has two modes of operation. The first mode is the interactive mode and the second mode is the demo mode. As shown in figure 5.3, the first step is to capture speech signal. If “From file or from Mic.” options are chosen, then the system will operate in active mode else it will operate in the Demo mode. Some examples are given below to describe the operation of the system.

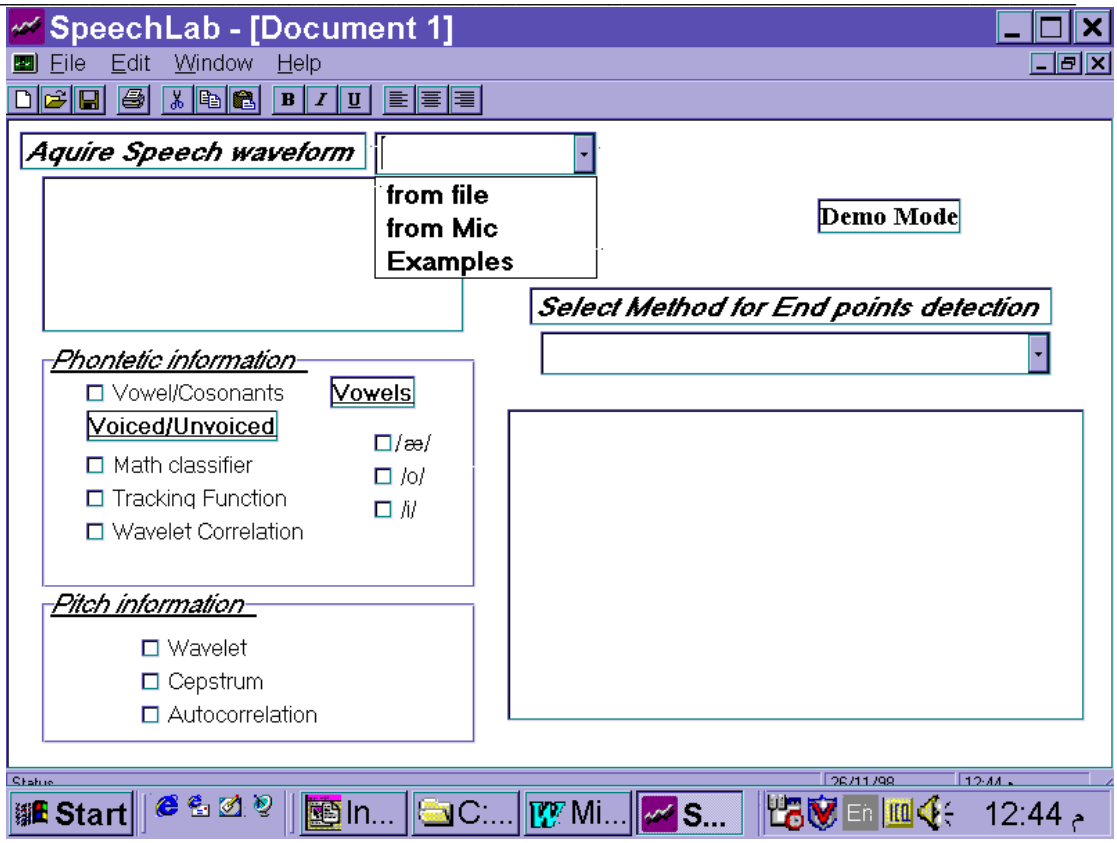


Figure 5. 3 Selecting the way for speech capturing.

Figure 5.4 is an example for choosing an example file. I n this case the example option is chosen and the word كتاب "ketab" is selected. Now, speech sample is ready for further processing. The first process is EPD. This is necessary to eliminate non-speech periods.

Figure 5.5 indicates EPD process. There are three techniques that can be chosen.

1- EPD using mathematical. classifier.

2- EPD using wavelet only.

3- EPD using energy and zero crossing.

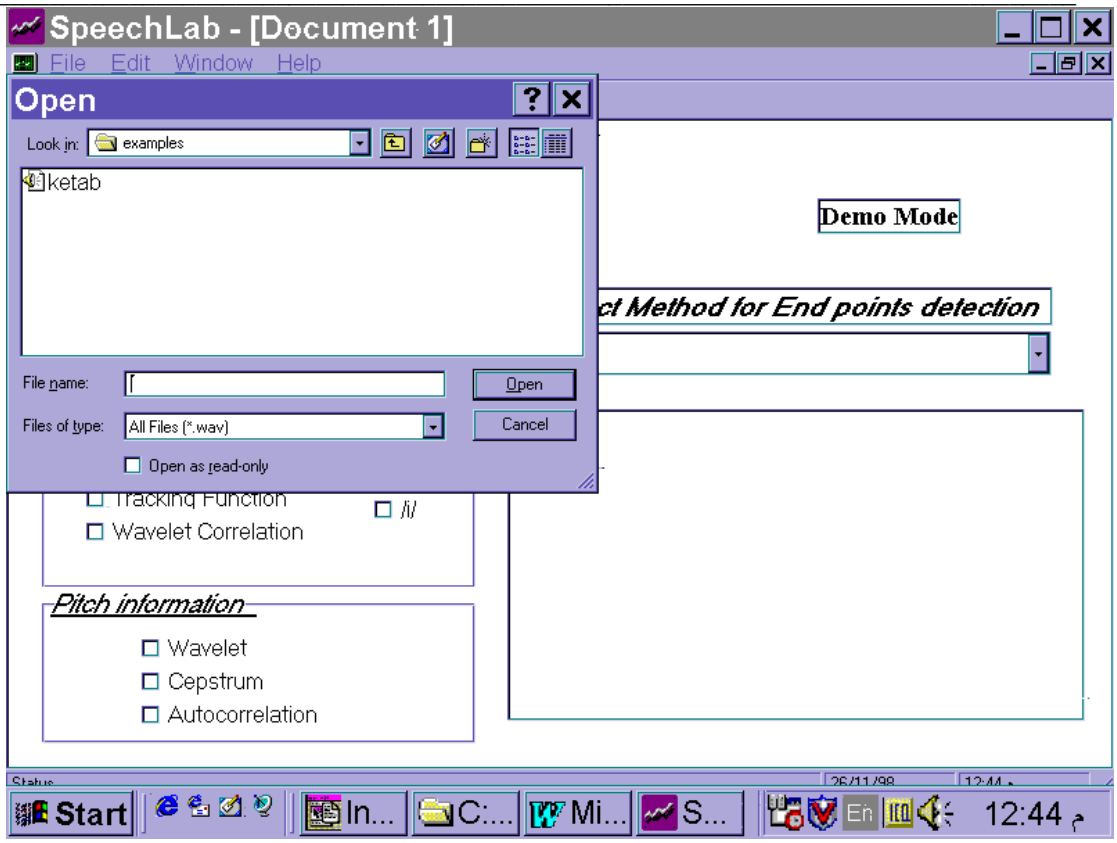


Figure 5. 4 Choice of speech file to be processed.

After EPD step, other processing can be applied.

Figure (5.6 a) indicates V/U process using tracking function. V/U markers are overlaid on the speech to indicate the duration of voiced or unvoiced sounds. Also it is indicated in figure (5.6 b) the pitch contour using wavelet transform method.

Figure (5.7 a) illustrate the process of Vowel/Consonant classification and (5.7 b) illustrates vowels recognition. Markers in figure (5.7 a) are

high in case of vowel and are low in case of consonant.

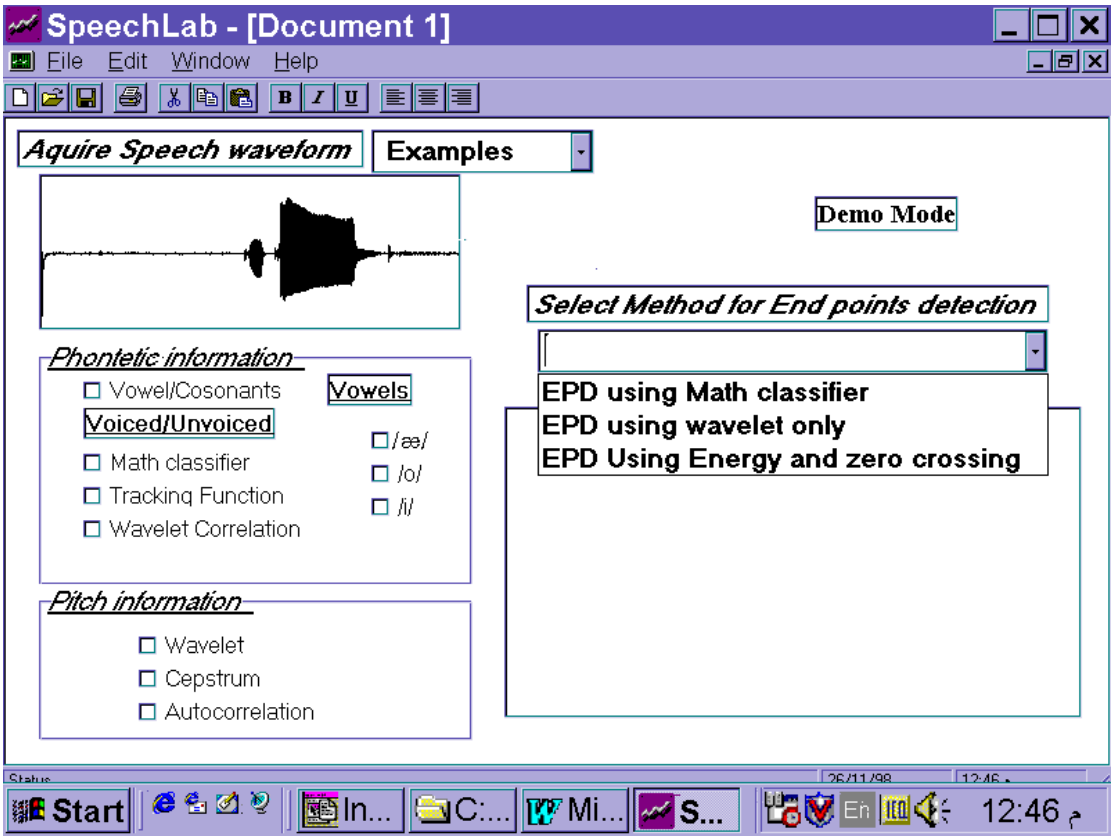


Figure 5. 5 EPD Step.

Figure 5.8 introduces a complicated process for advanced users. The user is allowed to combine different processes at a time to investigate his problem.

The system is highly flexible. Any beginner can use it to handle complicated speech problems. System can be extended to cover all speech areas in a simplified way. This version is just a beginning to the complete system.

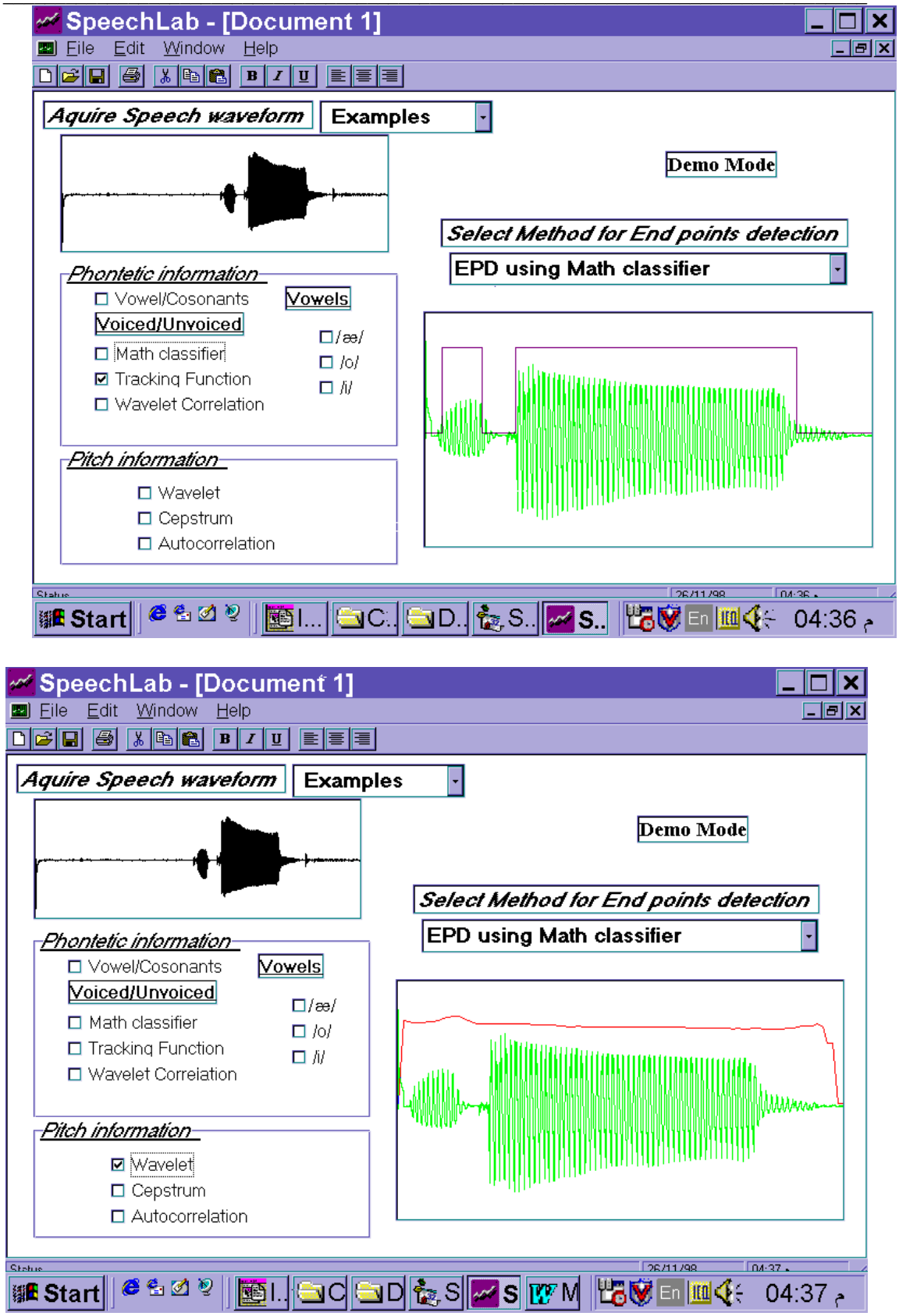


Figure 5.6 V/U using tracking function and pitch contour using wavelet.

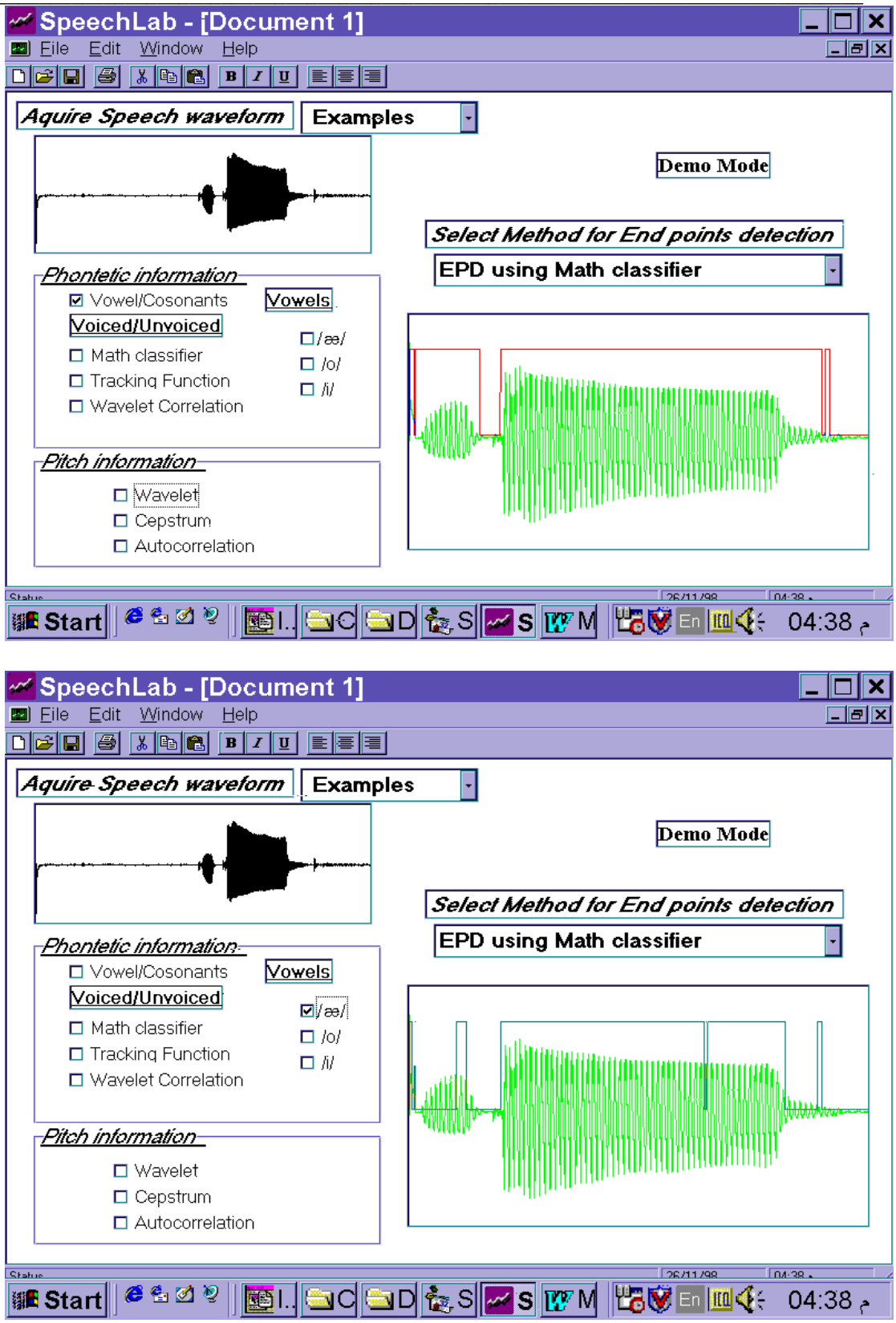


Figure 5. 7 Vowel/Consonant classification and vowels recognition.

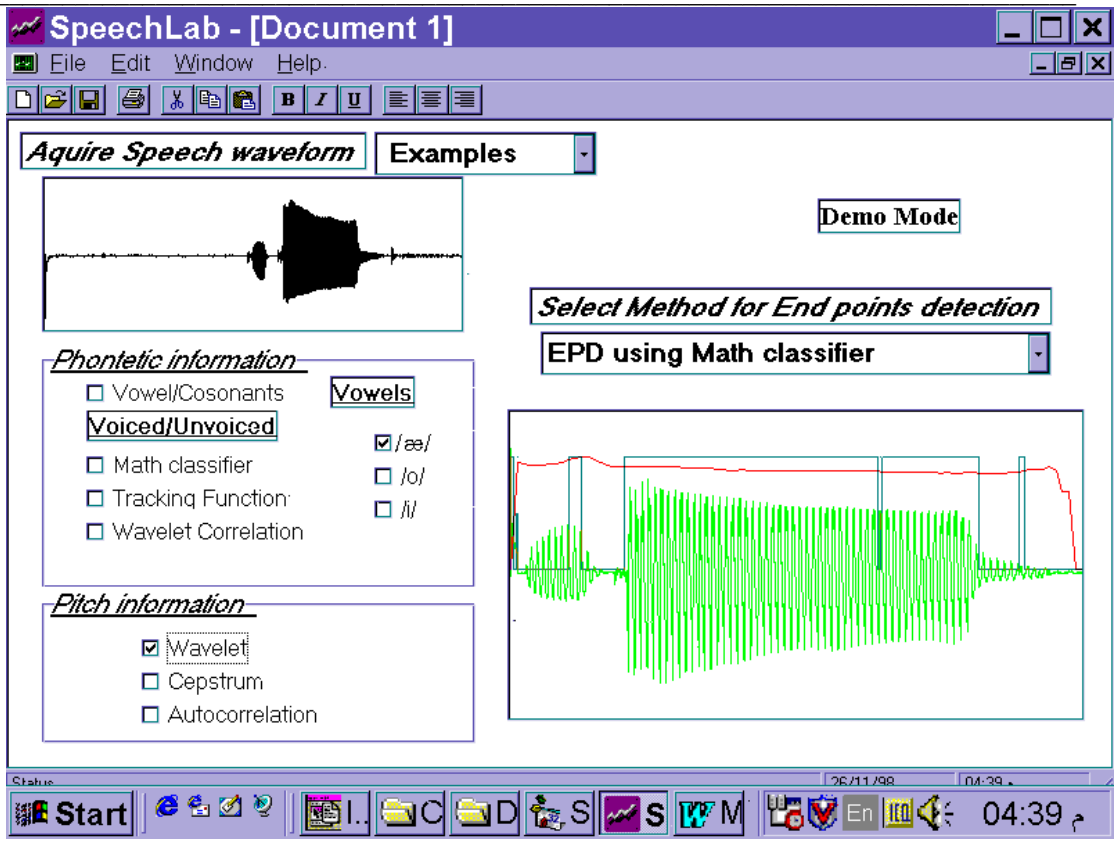


Figure 5. 8 Combination of processes.

5.4 Conclusion

This chapter illustrates a system implementation of all wavelet based algorithms that introduced in the previous chapters. The system is made using visual basic as interface while the core of software is DaDisp, which is introduced in the appendix.

5.1 Introduction 163
5.2 Block diagram of the system 163
5.3 The implemented system 166
5.4 Conclusion 172

Chapter 6

Summary, Conclusion and Future work

6.1 Summary

This work illustrates how far wavelet transform can be used in handling speech-processing problems. Work is divided into four chapters.

- In Chapter 1, speech signal and different classification techniques that are used in the subsequent chapters are introduced.

The study of the nature of speech generation is required as a background of speech modeling and analysis. The understanding of speech generation in human is needed for modeling the organs of speech and controlling of speech model. Representation of the vocal-tract frequency response, independent of the source parameters (e.g., voicing and fundamental frequency), captured researchers' interest in the 1960_S. One approach to this problem was to analyze the speech signal using a transmission line analog of the wave-propagation equation. This method allows use of a time-varying source signal as excitation to the "linear" system of the vocal tract.

To make analysis of the vocal-tract response tractable, one often assumes that the vocal tract is an acoustic system consisting of a concatenation of uniform cylindrical sections of different areas with planar waves propagating through the system. Each section can be modeled with an equivalent circuit with wave reflections occurring at the junctions between sections. Such a model allows analysis of the system from its input-output characteristics.

Most 1arnguages, including English, can be described in terms of a set of distinctive sounds, or phonemes. In particular, for American English, there are about 42 phonemes including vowels, diphthongs, semivowels and consonants. There are a variety of ways of studying phonetics; e.g.,

linguists study the distinctive features or characteristics of the phonemes. For our purposes it is sufficient to consider an acoustic characterization of the various sounds including the place and manner of articulation, waveforms, and spectrographic characterizations of these sounds.

The vocal tract shape defined in terms of tongue, velum, lip and jaw position, acts like a "filter" that filters the excitation to produce the speech signal. The frequency response of the filter has different spectral characteristics depending on the shape of the vocal tract. The broad spectral peaks in the spectrum are the resonance of the vocal tract and are commonly referred to as formants.

Chapter 1 goes to answer the question What are wavelets?.Wavelets are functions that satisfy certain requirements. The very name wavelet comes from the requirement that they should integrate to zero, ``waving" above and below the x-axis. The diminutive connotation of wavelet suggest the function has to be well localized. Other requirements are technical and needed mostly to insure quick and easy calculation of the direct and inverse wavelet transform.

There are many kinds of wavelets. One can choose between smooth wavelets, compactly supported wavelets, wavelets with simple mathematical expressions, wavelets with simple associated filters, etc.

Many researchers believe that neural networks offer the most promising unified approach to building truly intelligent computer systems.

Artificial neural networks (ANNs) are simplified models of the central nervous system and are networks of highly interconnected neural computing elements that have the ability to respond to input stimuli and

to learn to adapt to their environment. Neural networks employ parallel distributed processing (PDP) architectures

- Chapter 2 discuss the problem of end points detection. The problem of extracting the speech from the background noise is one of the major problems in speech applications. This is always the first step in any speech-based application.

Three ways of end points detection are discussed. The first one depends on correlating information of two adjacent wavelet frequency bands then obtain a threshold. The second and third methods get information about speech from all available wavelet frequency bands. The second method uses the Artificial Neural networks as a classifier and the third method uses the mathematical statistical regression for classification. A table comparing the three proposed methods is introduced at the end of the chapter. The table also gives indication of how they perform in different signal to noise ratios.

- Chapter 3 deals with the problem of classifying the speech signal into voiced or unvoiced sound and pitch period estimation.

The problem of V/U is handled by different methods. The differences between voiced sounds and unvoiced sounds are discussed. The wavelet transform is reforming a decomposition of signals into elementary building blocks that are well localized both in time and in frequency. The wavelet transform is suitable for characterizing the local regularity of signals.

From a signal processing point of view the Dyadic Wavelet can be

$$\frac{1}{2^j}\ \Psi\left(\frac{t}{2^j}\right)$$

considered as the output of a bank of constant Q, octave band, band-pass filters whose impulse response is for each scale 2^j .

Three methods for classifying speech into V/U are discussed. The first one is **Single band selection method.** A wavelet frequency band of which the vowels or voiced sounds are dominant in the speech signal is selected for the analysis. Mathcad[1] software package is used as a platform of all mathematics such as wavelet transform, interpolation ... etc. The frequency band of 172-344 Hz is chosen here for the tracking method. Tracking function is obtained. The system indicates high recognition accuracy of about 97.4%.

The second way for classifying speech into V/U is the **Correlation based method.** In this way information about the signal from two-wavelet frequency bands are correlated. This correlation makes the system more immune to noise. A correlation tracking function is formulated. This system appears reliable even in case of low signal to noise ratio (less than 9 dB). The first 100 ms of speech is assumed to be unvoiced. Maximum unvoiced threshold is obtained from the first 100 ms (about 1024 samples) of the moving standard deviation.

The third way for classifying speech into V/U is **Voiced/Unvoiced classification using mathematical model.** In this way all information available about the signal is taken into consideration to formulate a system model. The system model depends on linear statistical regression. The system is robust but it is highly dependent on database collected in the training phase. It does not need pre-estimation of any thresholds as the previous two ways so that it is more practical than the previous two ways. But it gives less recognition accuracy than they do, about 90%.

[1] © 1986-1994 Mathsoft Inc. Version 5.0. © 1993 by Houghton Mifflin Company.

At the end of this chapter the problem of pitch period estimation is considered. Pitch period estimation (or equivalently, fundamental frequency estimation;' is one of the most important problems in speech processing. Pitch estimation using dyadic wavelet is the point that is studied in this work. Pitch detectors are used in vocoders, speaker identification and verification Systems and aids-to-the handicapped. Because of its importance, many solutions to this problem have been proposed. All of the proposed schemes have their limitations, and it is safe to say that no presently available pitch detection scheme can be expected to give perfectly satisfactory results across a wide range of speakers, applications, and operating environments.

Two ways of pitch estimation using wavelet are introduced. The **Two band correlation method ,** which generates a pulse train that have a period between pulses equal to the pitch period. This way correlates the information from two adjacent wavelet frequency bands to formulate a correlation function. Then by peak detection algorithm the pulse train is generated. The method can track the peaks even in case of low signal to noise ratio (less than 10 dB).

The second way is the **Pitch detection using dependencies.** This method is much alike the previous one except that it takes the information from four adjacent bands. Two pitch estimators like the previous one is constructed. Each one estimate pitch period from different two adjacent bands in range of frequencies less than 1000 Hz. Then dependencies between the two systems are measured to eliminate false pulses from the pulse train. This method is highly reliable and more stable than the previous one. A comparison between this method and well-known techniques such as Autocorrelation and Cepestrum is illustrated at the end of this chapter.

- In chapter four the problem of basic unit recognition is discussed. The chapter starts with introduction to vowels and consonants. The problem of segmentation into vowels and consonants are illustrated. The problem is handled using two ways. The first is **Band selection method**. The second is **Math classification method**. As discussed before the first method depends on selected bands and the second one depends on all available bands. The segmentation is studied for different signal to noise ratios.

The problem of vowel recognition is illustrated. In Arabic language there are six different vowels. The problem is handled using mathematical statistical regression.

6.2 Conclusion

Wavelet transform is suitable for handling speech signal. It gives a good representation of many features of speech signal. It can be used for monitoring acoustic phonetics variations in utterance.

Wavelet transform can be used in case of high noise environments. Due to the nature of wavelet transform it handles speech signal approximately with the same manner as human ear does. That makes it highly immune to noise.

Voiced/Unvoiced recognition rate is highly increased using wavelet based algorithms. The system indicates good immunity to noise and can work reliably at low signal to noise ratios.

Pitch period estimation using wavelet based algorithm gives very accurate results compared with familiar algorithms such as autocorrelation and Cepestrum. The fundamental frequency can be tractable even in case of very low signal to noise ratio environments.

Speech detection from the background noise (end points detection) using wavelet based algorithm gives reliable results. It can work in environment of low signal to noise ratio that the ordinary method of energy and zero crossing rates failed.

Speech segmentation into vowels and consonants problem is handled using a technique based on wavelet transform. The system success with a high recognition rate to trace the boundaries of vowels and consonants. In addition to this, the technique is enhanced to distinguish between different vowels.

6.3 Application

All the above techniques to solve common speech problems can be used to make a speech analyzer system that is based on wavelet transform. This system can analyze the speech signal in case of highly noisy environments. This is very suitable in practical world. The system can be used in environments containing heavy machines.

6.3.1 Fault detection of a heavy machine

The area of fault decision of machines based on harmonics requires handling the sounds in a very low signal to noise ratio environments. The machine is tested in the factory environment. Bugs are detected and sounds corresponding to each bug are collected. Sounds are analyzed and saved as a database for bug detector machine. Algorithms of bug detector machine must handle sounds with very low signal to noise ratios which is not reasonable in ordinary techniques. Actually sounds of machines is

totally different from human speech but is simpler. Each bug is a combination of few harmonics that is deterministic and can be calculated mathematically.

*6.*3.2 Speech dictation machine

Speech dictation machine exists now and many researchers and companies introduce a solution of this problem (IBM). This system always depends on training a system with extra large vocabulary to make it. After that a tree of decisions are made to speed up the decision process.

If the system of dictation machine is designed to detect basic speech units it will be simpler, faster and reliable. The problem of phone recognition is the barrier that makes manufacturer leaving this way. Even if this problem is solved the system will be critically stable. It can work in one environment with high signal to noise ratio and fail in another one with low signal to noise ratio. That makes it not suitable for commercial purposes.

The proposed system which is introduced in this work indicates a very high accuracy in determining boundaries of phones even in case of low signal to noise ratio. The system can distinguish between vowels with a very high rate approaching 81%. It needs extra work to verify the same results in case of consonants as well.

This work promise that the dictation machine based on basic speech units can be founded.

6.4 Future work

It is planned to design of Arabic phone recognition system. The system will be extension of this work to complete all phones recognition rather than vowels only in this work.

Huge database will be collected from different speakers. The data will be classified into Arabic phones. It is planed to verify segmentation manually using Spectrogram and listening test. It is acceptable to include English database Such as TIMIT to enhance segmentation process.

Database for each phoneme will be prepared for handling using wavelet based algorithms such as those in this work.

6 1. Summary 174
6 2. Conclusion 179
6 3. Application 180
6 3.1. Fault detection of a heavy machines 180
6 3.2. Speech dictation machine 181
6.4 Future work 181

Appendix A

Database

A.1 Database collection

The database used here contains all Arabic language's phonemes. Database consists of 18 Arabic words with 6 repetitions. The total duration of utterance is 163.08Sec. The following table contains the words and it's phonetic contents.

/η//Θ//μ//σ/	همس
/λ//Θ//μ//σ/	لمس
/κ//ε//τ//α//β/	كتاب
/ψ//Θ//κ//τ//o//β//o//ν/	يكتبون
/κ//Θ//τ//Θ//β//Θ/	كتب
/ψ//Θ//λ//N//α//β//o//ν/	يلعبون
/μ//o//τ//α//θ//Θ//δ//ε//μ/	متقدم
/δ//o//ρ//Φ//Θ//M/	ضرغام
/Λ//Θ//ρ//ι//ϕ/	ظريف
/α//P//N//ν/	أرعن
/ξ//Θ//ρ//τ//o//M/	خرطوم
/Λ//Θ//ι//λ/	ذيل

/Λ//Θ//η//Θ//β//Θ/	ذهب
/Λ//Θ//P//Θ//N//Θ/	زرع
/η#//Θ//σ//ι//δH/	حصيد
/γ//Θ//μ//ι//λ/	جميل
/Σ//Θ//ρ//ι//β//Θ/	شرب
/T//o//N//β//Θ//ν/	ثعبان

All phonetics are written using IPA* language. A table of IPA is illustrated in next section.

* International Phonetic Alphabet

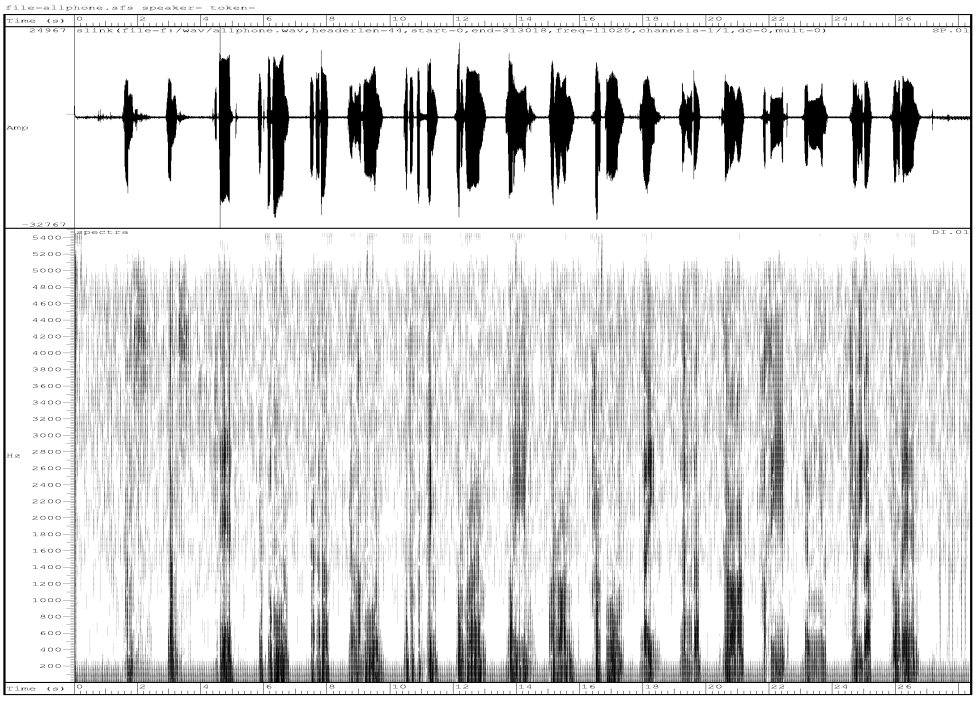


Figure 4. 1 Speech signal of single database file contains Arabic words in the previous table. The lower half's graph is spectrogram of the speech signal.

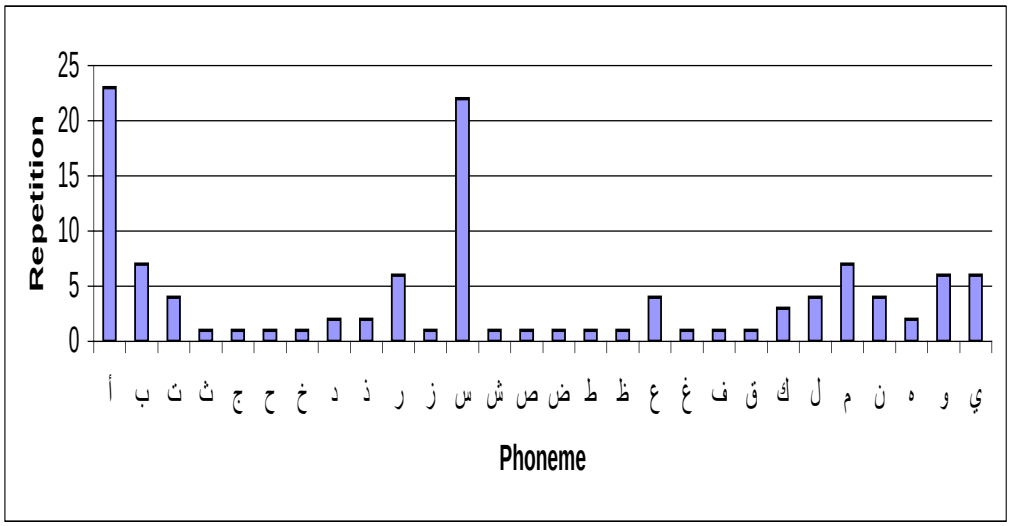


Figure A. 1 Histogram of phonemes in a single database file.

A.2 IPA characters according to articulation

Vowels

	Front		Central		Back	
	Unrounded	Rounded	Unrounded	Rounded	Unrounded	Rounded
Close	i	y	ɨ	ʉ	ɯ	u
Near-close	ɪ	ʏ				ʊ
Close-mid	e	ø	ɘ	ɵ	ɤ	o
Mid			ə			
Open-mid	ɛ	œ	ɜ	ɞ	ʌ	ɔ
Near-open	æ		ɐ			
Open	a	ɶ			ɑ	ɒ

Other symbols

ʍ	Voiceless labial-velar approximant	ʘ	Bilabial click
w	Voiced labial-velar approximant	ǀ	Dental click
ɥ	Voiced labial-palatal approximant	ǃ	(Post-)alveolar click
ɕ	Voiceless alveolopalatal fricative	ǂ	Palatoalveolar click
ʑ	Voiced alveolopalatal fricative	ǁ	Alveolar lateral click
ʃʝ	Simultaneous ʃ and x	ʡ	Voiced epiglottal plosive
ɺ	Voiced alveolar lateral flap	ʜ	Voiceless epiglottal fricative
		ʢ	Voiced epiglottal fricative

Consonants

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar
Plosive	p			t	
	b			d	
Nasal	m	ɱ		n	
Trill	ʙ			r	
Flap				ɾ	
Fricative	ɸ	f	θ	s	ʃ
	β	v	ð	z	ʒ
Lateral fricative				ɬ	
				ɮ	
Approximant		ʋ		ɹ	
Lateral approximant				l	
Implosive	ɓ			ɗ	

	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	ʈ	c	k	q		ʔ
	ɖ	ɟ	g	ɢ		
Nasal	ɳ	ɲ	ŋ	ɴ		
Trill				ʀ		
Flap	ɽ					
Fricative	ʂ	ç	x	χ	ħ	h
	ʐ	ʝ	ɣ	ʁ	ʕ	ɦ
Lateral fricative						
Approximant	ɻ	j	ɰ			
Lateral approximant	ɭ	ʎ	ʟ			
Implosive		ʄ	ɠ	ʛ		

Appendix B

Software (DADISP)

B.1 DADiSP

B.1.1 The Task

Scientists and engineers (S&Es) are in the business of converting data into information. With the incredible increase in processing power of personal computers and data acquisition software, scientists and engineers can now collect streams of data at the push of a button. However, converting that data into useful information often remains a daunting task.

B.1.2 The Scientific Method

Scientific inquiry is rooted in the basic tenets of the scientific method:

- Ask a question.
- Formulate a hypothesis as a possible answer to the question.
- Design an experiment to test the hypothesis.
- Collect data from the experiment.
- Analyze the data.
- Accept or reject the hypothesis based on the results of the analysis.

Thus, data analysis is a fundamental and necessary step in virtually every scientific endeavor. As mentioned, personal computers are rapidly becoming the tool of choice for both scientific data acquisition and data analysis. To understand the necessary components of data analysis software, we must first look at the data analysis user.

B.1.3 Common User Attributes

S&Es who use data analysis software share four common attributes:

1. S&Es are not professional programmers. Although often familiar with the tasks required to write software routines, technical professionals get paid to produce results, not code.
2. S&Es are experts in their application area. The technical professional knows precisely what methods, calculations and graphics are required to produce acceptable results in their particular field.
3. S&Es work in technical application areas that are extremely diverse. Applications run the full gamut of scientific inquiry including signal processing, statistical analysis, test and measurement, noise and vibration, medical research, process monitoring, image processing, communications, quality management and just about anything and everything else.
4. S&Es routinely work with huge volumes of data and rely on graphical representation as an interpretation aid. The raw numbers are overwhelming and must be reduced to application specific graphical form to convey meaningful information. The great diversity of graphs employed by S&Es has lead to the term scientific visualization.

B.1.4 Two Approaches

Because of the numerous target applications, there are at least two avenues of designing data analysis software:

- Create many application specific programs, such as chromatography, modal analysis, filter design, etc. that target specific customers.
- Create a general purpose tool that can be adapted to the many application areas.

Obviously, a general purpose tool is highly preferable from a software development and marketing point of view. In addition, add in modules can be produced to allow the tool to further target specific applications similar to an application specific product.

B.1.5 The Traditional Approach

The traditional approach of creating a technical data analysis tool has been to provide an interactive, high level language. To meet the requirements of S&Es, these languages offer the following features:

- Canned routines such as FFT, INTEGRATE, INVERT, etc. to prevent the customer from needlessly "re-inventing the wheel".
- An interpreted language to avoid the tedious "compile and link" development process of base level programming languages.
- Integrated graphics capability to present results in a meaningful form.
- Products such as Matlab, APL, IDL and a host of other analysis languages fall into this category.

The great benefit of a language based solution is flexibility - almost any application requirement can be programmed. Of course, this flexibility comes at a tremendous price - the S&E must program almost

everything! Programming is a difficult, low productivity chore not in the realm of the S&E's expertise.

B.1.6The Business Spreadsheet

The business spreadsheet is an extremely popular and flexible software tool. The spreadsheet derives its tremendous power from the ability of the user to easily set up relationships between numeric cells in a relatively intuitive manner. When cells are updated with new values, dependent cells automatically recalculate. The user is effectively writing an application specific program without actually programming in the traditional sense. In addition, almost all spreadsheets provide a mechanism to reduce numeric data to graphical form. Thus, the spreadsheet represents a flexible, easy to use tool that provides some degree visualization without the heavy burden of programming. Not surprisingly, surveys consistently show the overwhelming majority of S&Es use business spreadsheets for technical data analysis over every other solution - even though this tool was not designed to handle technical data.

In fact, the business spreadsheet is designed to manipulate a small collection of scalar values. These values are processed and perhaps displayed as a final graph. For example, a user might enter values such as sales, cost of sales, expenses, taxes and more taxes to produce a basic income statement. Several periods of this data could then be appended together to produce a simple trend chart. The business user starts with numbers and perhaps ends up with a graph.

In contrast, in the course of data analysis, the S&E begins with graphs, almost always creates additional graphs, and perhaps produces a

meaningful scalar as a final result. For example, a mechanical engineer would integrate the acceleration data of a vehicle chassis crash test to produce a velocity graph. This graph by itself conveys valuable information. However, the derived velocity data would in turn be converted into the frequency domain to isolate the important natural frequencies. Finally, the most prominent frequency in a certain band would be singled out as the resonant frequency of the chassis.

In this case, the S&E starts with a graph and ends up with a scalar - the exact opposite reduction chain of the business user. In addition, the volume of data routinely processed by the S&E rapidly chokes the business spreadsheet.

B.1.7 DADiSP - the S&E's Spreadsheet

The business spreadsheet is a flexible and powerful tool that S&Es often "shoehorn" to meet their analysis requirements. However, because it was designed for business use, the standard spreadsheet presents many limitations for S&E data analysis applications:

1. Restrictive Data Size
2. Slow Graphics for Large Data
3. Data Must be Saved with Spreadsheet
4. Numeric Focus Inappropriate for S&E Data
5. Lack of S&E Analysis Routines
6. Inability to Handle Complex Numbers
7. Inability to Handle Binary Data
8. Limited Data Import Capabilities

Is there a better solution than the business spreadsheet? Yes there is. It is called **DADiSP**.

DADiSP (pronounced day-disp) is spreadsheet designed specifically for S&Es. DADiSP capitalizes on the power and familiarity of the business spreadsheet while at the same time, overcoming its limitations in S&E applications.

Instead of cells that contain numbers, a DADiSP Worksheet consists of analysis windows that automatically display data as a table or graph. Like a business spreadsheet, when the data in an analysis window changes, all dependent windows automatically update. Specific, custom analysis can be accomplished naturally without the need for traditional programming. DADiSP employs contemporary user interface elements such as pull down menus, dialog boxes, toolbar buttons and on line help to provide a productive, familiar environment. And unlike business spreadsheets, DADiSP is designed to accommodate huge data series and render graphs with optimal speed.

Data import is extremely flexible with support for ASCII and binary file types. Imported data resides in a separate series data base and can be exported to several file formats. Complex numbers are fully supported. DADiSP includes 1000 built-in analysis routines tailored specifically to S&E applications. DADiSP also offers several optional processing modules that target specific application areas.

B.1.8 DADiSP - Language Included

To provide full user customization, DADiSP includes SPL, Series Processing Language. SPL is a full featured, incrementally compiled series processing language based on the omnipresent C language. As a

result, SPL programs have a clean and familiar style about them. SPL also contains useful constructs of languages such as APL and Matlab. Thus, the C programmer is immediately at home with SPL and the Matlab or APL programmer will recognize their favorite programming idioms.

B.1.9 DADiSP - The Best of Both Worlds

By combining the ease of use and familiarity of the business spreadsheet with the power and flexibility of an interpreted analysis language, DADiSP is designed to be the analysis tool of choice for both the "point and click" and "type and enter" S&E user. A few of DADiSP's more popular features include:

1. Graphical Worksheet Windows
2. Unlimited Data Size
3. 1000 built-in analysis functions
4. Tabular, 2D, 3D and Image- optimized graphics
5. Standard GUI Interface
6. Cross Platform Availability
7. SPL - Series Processing Language
8. Inter-Application Communication
9. Line, Legend and Text Annotations
10. Custom Menus, Dialog Boxes and Toolbar Buttons
11. Scrolling Graphs and Cross Hair Cursors
12. Overplot and Overlayed Graphs
13. On Line Help

With DADiSP, "you can have your mouse and program too."

B.2 SPL* Routines

Series **P**rocessing **L**anguage is a special purpose language that concerns with series operations. Almost series operations are included in SPL as simple functions. DADiSP allows using of SPL. The following sections views all routines used in this work which are written using SPL.

B.2.1 End points detection

```
                                                    epd(ser)
                                                          {
                              FrameTable =  RAVEL(ser,1024,1,0);
                          NumberOfFrames =  SERCOUNT(FrameTable);
 NoiseThreshold = MAX ( movstd ( extract ( crosscor ( col
                                                  ( waves ( col
     (FrameTable,1)),1),col(waves(col(FrameTable,1)),2) )
                                                ,1,1024),110) );
Mrkrs = movstd ( extract ( crosscor ( col ( waves ( col (
     FrameTable ,1)),1),col(waves(col(FrameTable,1)),2) )
                                ,1,1024),110)>NoiseThreshold;
                                for(u=2;u<NumberOfFrames;u++)
                                                          {
                                  prcnt=u*100/NumberOfFrames;
                                                 echo(prcnt);
      Temps = movstd ( extract ( crosscor ( col ( waves (
 col ( FrameTable,u)),1),col(waves(col(FrameTable,u)),2))
                               ,1,1024) ,110) > NoiseThreshold;
                                      mrkrs=concat(mrkrs,Temps);
                                                          }
                                           moveleft(mrkrs,2048);
                                                 return(mrkrs);
                                                          }
```

* Series Processing Language

B.2.2 Pitch period estimation

```
pitche(filen,sar)
{
local p;
local XX;
local PP;
datab=READWAV(filen);
FrameTable= RAVEL(datab,1024,1,975);
NumberOFFrames=SERCOUNT(FrameTable);
p=1..NumberOfFrames-1;
for(u=1;u<=NumberOfFrames-1;U++)
{
prcnt=u*100/NumberOfFrames;
echo(prcnt);
p[u]=0;
WvltTable = WAVES(COL(FrameTable,u));
C0 =
Getpeak(extract(Crosscor(COL(WvltTable,1),COL(WvltTable,2
)),1024,1024),.01,1,0);
setdeltax(C0,1/sar);
C1 =
Getpeak(extract(Crosscor(COL(WvltTable,2),COL(WvltTable,3
)),1024,1024),.01,1,0);
setdeltax(C1,1/sar);
C2 =
Getpeak(extract(Crosscor(COL(WvltTable,3),COL(WvltTable,4
)),1024,1024),.01,1,0);
setdeltax(C2,1/sar);
CT = REGION(RAVEL(C0,C1,C2),1,550,1,3)>0;
p[u]=PitchEstimate1(CT,sar);


}
XX = movstd(p,5)<15;
PP = movavg2(p,5);
p = PP * XX;
return(p);
```

```
}
PitchEstimate1(TSer,sr)
{
local dx;
local T;
local Last;
dx=1/sr;
setdeltax(TSer,dx);
C0=col(TSer,1);
C1=col(TSer,2);
C2=col(TSer,3);
P0=GETCONDXS(C0>0);
P1=GETCONDXS(C1>0);
P2=GETCONDXS(C2>0);

Last=sersize(P1);
T=1..Last;
pitch=0;
k=1;
for(u=1;u<=Last;u++)
{
msk= GETCONDXS(ABS(P2-P1[u])<0.005);
y = isnavalue(msk);
if(y[1] != 0)
{
T[k]=P1[u];
k++;
}
}
if(SERSIZE(T)>1)
{
mt = abs(T[2]-T[1]);
pitch=1/mt;
if (pitch<70) pitch=0;
}
```

```
return(pitch);
}
```

B.2.3 Data preparation for neural network and math classifier

```
Inputs:
Co: Count of files to be prepared from names saved
into “TSET.INP”. Full names of files without
extensions are saved into text file called
“TSET.INP”.
Sr: Sampling Rate

Createnna(co,sr)
{
fclose("TSET.INP");
fopen("TSET.INP","r+");
for(u=1;u<=co;u++)
{
prcnt = u * 100 / co ;
echo ( prcnt );
FileName = FGETS("TSET.INP");
FileName = strextract(FileName,1,strlen(FileName)-
1);
infile = STRCAT(FileName,".WAV");
outfile = STRCAT(FileName,".nna");
mrkfile = STRCAT(FileName,".drk");
mrkbuffer = READA(mrkfile);
databuffer = READWAV(infile);
sz = SERSIZE(databuffer);
framenumbers = sz / 1024+1;
dx=1/sr;
setdeltax(databuffer,dx);
mrkrsbuffer = databuffer * 0.0;
echo(rate(mrkrsbuffer));
setdeltax(mrkrsbuffer,dx);
msz = SERSIZE(mrkbuffer);

for(myc=1;myc<=msz;myc++)
{
```

```
                                    startp= mrkbuffer[myc];
                                                  myc++;
                                     endp=mrkbuffer[myc];

        mrkrsbuffer = SUBSTX(mrkrsbuffer,startp,endp,100);
                                 setdeltax(mrkrsbuffer,dx);
                                                  }



                                for(k=1;k<=framenumbers;k++)
                                                  {
                          mrkfram = EXTRACT(mrkrsbuffer,k*1024-
                                                  1023,1024);

              frame = EXTRACT(databuffer,k*1024-1023,1024);
                                  wvltable = WAVES(frame);

                        b0 = movavg2(abs(col(wvltable,1)),200);
                        b1 = movavg2(abs(col(wvltable,2)),200);
                        b2 = movavg2(abs(col(wvltable,3)),200);
                        b3 = movavg2(abs(col(wvltable,4)),200);
                        b4 = movavg2(abs(col(wvltable,5)),200);
                        b5 = movavg2(abs(col(wvltable,6)),200);

             nnatable = RAVEL ( b0,b1,b2,b3,b4,b5,mrkfram);

                           WRITETABLE(outfile,nnatable,2);

                                                  }
                            nnatable = READTABLE(outfile);
                                                  }
                                       fclose("TSET.INP");
                                         return(nnatable);
                                                  }
```

B.2.4 Wavelet routines

```
                                        wavelet(FileName,sr)
                                                          {
                              infile = STRCAT(FileName,".WAV");
                             outfile = STRCAT(FileName,".wvt");
                                  databuffer = READWAV(infile);
                                      sz = SERSIZE(databuffer);
                                      framenumbers = sz / 1024;
                                                      dx=1/sr;
                                      setdeltax(databuffer,dx);

                                frame = EXTRACT(databuffer,1,1024);
                                           wvltable = WAVES(frame);

                           b0 = movavg2(abs(col(wvltable,1)),200);
                           b1 = movavg2(abs(col(wvltable,2)),200);
                           b2 = movavg2(abs(col(wvltable,3)),200);
                           b3 = movavg2(abs(col(wvltable,4)),200);
                           b4 = movavg2(abs(col(wvltable,5)),200);
                           b5 = movavg2(abs(col(wvltable,6)),200);

                            nnatable = RAVEL ( b0,b1,b2,b3,b4,b5);

                                  WRITETABLE(outfile,nnatable,1);

                                    for(k=2;k<=framenumbers;k++)
                                                          {

                       frame = EXTRACT(databuffer,k*1024-1023,1024);
                                           wvltable = WAVES(frame);

                           b0 = movavg2(abs(col(wvltable,1)),200);
                           b1 = movavg2(abs(col(wvltable,2)),200);
                           b2 = movavg2(abs(col(wvltable,3)),200);
```

```
b3 = movavg2(abs(col(wvltable,4)),200);
b4 = movavg2(abs(col(wvltable,5)),200);
b5 = movavg2(abs(col(wvltable,6)),200);

 nnatable = RAVEL ( b0,b1,b2,b3,b4,b5);

        WRITETABLE(outfile,nnatable,2);

                                      }
         nnatable = READTABLE(outfile);
  nnatable = RAVEL ( b0,b1,b2,b3,b4,b5);
               setdeltax(nnatable,dx);
                     return(nnatable);

                                      }

                                  wave(y)
                                       {
                                  local buff;
                 writea("data.dat",y,1);
RUN("wxfrm -Q33 data.dat>datar.dat",-1);
                 buff=reada("datar.dat");
                           return (buff);
                                       }

                                  extractw(y,st)
                                       {
                                    local sz,fn;

                                  sz=sersize(y);
                                   fn=sz/1024+1;
                                for(k=1;k<=fn;k++)
                                       {
                                  echo (k/fn*100);
                                      frr=k-1;
```

```
        writea("data.dat",extract(y,(1024*frr+1),1024),1);
                    RUN("wxfrm -D8 data.dat>datar.dat",-1);
                    n0=sprintf("waveletband0.%d.f%d",st,k);
                    n1=sprintf("waveletband1.%d.f%d",st,k);
                    n2=sprintf("waveletband2.%d.f%d",st,k);
                    n3=sprintf("waveletband3.%d.f%d",st,k);
                    n4=sprintf("waveletband4.%d.f%d",st,k);
                    n5=sprintf("waveletband5.%d.f%d",st,k);
      saveseries(interpr(extract(reada("datar.dat"),17,16)
                                              ,0.01465),n0);
      saveseries(interpr(extract(reada("datar.dat"),33,32)
                                               ,0.0303),n1);
      saveseries(interpr(extract(reada("datar.dat"),65,64)
                                              ,0.06153),n2);
      saveseries(interpr(extract(reada("datar.dat"),129,12
                                              8),0.1241),n3);
      saveseries(interpr(extract(reada("datar.dat"),257,25
                                              6),0.2492),n4);
      saveseries(interpr(extract(reada("datar.dat"),513,51
                                              2),0.4991),n5);
                                                      }
                                                      }


                                                  waves(y)
                                                         {
                                                   local sz,fn;
                                                   local myset;
                                                      local bt;
                                                  sz=sersize(y);
                                                  myset=wave(y);
    b0=interpr(extract(reada("datar.dat"),17,16),0.01465);
     b1=interpr(extract(reada("datar.dat"),33,32),0.0303);
    b2=interpr(extract(reada("datar.dat"),65,64),0.06153);
  b3=interpr(extract(reada("datar.dat"),129,128),0.1241);
  b4=interpr(extract(reada("datar.dat"),257,256),0.2492);
  b5=interpr(extract(reada("datar.dat"),513,512),0.4991);
                               bt=ravel(b0,b1,b2,b3,b4,b5);
```

```
                                                    return(bt);

                                                }
```

Appendix C

Survey on wavelet and speech

C.1 Discrete wavelet transform techniques in speech processing

This Paper Appears in :

TENCON '96. Proceedings., 1996 IEEE TENCON. Digital Signal Processing Applications on Pages: 514 - 519 vol.2 This Conference was Held : 26-29 Nov. 1996 Vol. 2 ISBN: 0-7803-3679-8

Abstract:

The trend towards real-time, low-bit-rate speech coders dictates current research efforts in speech compression. A method being evaluated uses wavelets for speech analysis and synthesis. Distinguishing between voiced and unvoiced speech, determining pitch, and methods for choosing optimum wavelets for speech compression are discussed. It is observed that wavelets concentrate speech energy into bands which differentiate between voiced or unvoiced speech. Optimum wavelets are selected based on energy conservation properties in the approximation part of the wavelet coefficients. It is shown that the Battle-Lemarie wavelet concentrates more than 97.5% of the signal energy into the approximation part of the coefficients followed closely by the Daubechies D20, D12, D10 or D8 wavelets. The Haar wavelets are the worst. Listening tests show that the Daubechies 10 preserves perceptual information better than other Daubechies wavelets and, indeed, a host of other orthogonal wavelets. Pitch periods and evolution can be identified from contour plots of coefficients obtained at several scales.

C.2 Adaptive pitch period decimation and its application in speech compression

This Paper Appears in :

Southeastcon '96. Bringing Together Education, Science and Technology., Proceedings of the IEEE on Pages: 220 – 222 This Conference was Held : 11-14 April 1996 ISBN: 0-7803-3088-9

Abstract:

This paper presents a new method of speech coding that takes advantage of the repetitiveness inherent in voiced speech. Voiced speech is broken into pitch period lengths (wavelets) and these signals are compared with one another to determine If two wavelets differ significantly. If the wavelets are significantly different, then they are encoded and transmitted; otherwise, the current wavelet Is not transmitted, and the next pitch period wavelet is compared. This results in encoding only a representative fraction of the speech signal and significantly lowers the number of bits required to transmit the signal. Pitch period determination is done by using the autocorrelation method and a median smoothing filter. The pitch period wavelets are preprocessed using a time weighted averaging method that allows concatenation of wavelets without sharp transitions at pitch boundaries, therefore reducing high frequency noise. Wavelets are compared using the Itakura distance measure, which is usually employed in speech recognition applications. The transmitted wavelets are encoded using a differential PCM method to further reduce the bit rate of the transmission. Unvoiced speech is encoded using an LPC method on a frame by frame basis. This results in high quality speech transmission at bit rates of approximately 3.8 kb/s.

C.3 New pitch detection algorithm based on wavelet transform

This Paper Appears in :

Time-Frequency and Time-Scale Analysis, 1998. Proceedings of the IEEE-SP International Symposium on Pages: 165 – 168 This Conference was Held : 6-9 Oct. 1998 ISBN: 0-7803-5073-1

Abstract:

A new pitch detection algorithm based on wavelet transform analysis is presented. This algorithm uses a family of modulated Gaussian wavelets adapted to the Bark scale to analyze speech signals decomposing the input signal into different bands. Then, a maxima detector and a new confirmation algorithm are used to extract pitch period information. Evaluation results and comparison tests with standard SIFT algorithm are presented.

C.4 Wavelet algorithm for the estimation of pitch period of speech signal

This Paper Appears in :

Electronics, Circuits, and Systems, 1996. ICECS '96., Proceedings of the Third IEEE International Conference on Pages: 471 - 474 vol.1.This Conference was Held : 13-16 Oct. 1996 Vol. 1 ISBN: 0-7803-3650-X

Abstract:

An algorithm based on dyadic wavelet transform (DyWT) has been developed for detecting pitch period. Pitch period is regarded as an important feature in designing and developing automatic speaker recognition/identification systems. In this paper, we have developed two methods for detecting pitch period of synthetic signals. In the first method, we estimated the pitch period using the original signal. In the second method, pitch period was estimated from the power spectrum of the signal. Several experiments were performed, under noisy and ideal environmental conditions, to evaluate the accuracy and robustness of the

proposed methodology. It was observed from the experiments that the proposed techniques were successful in estimating pitch periods.

C.5 Pitch determination and speech segmentation using the discrete wavelet transform

This Paper Appears in :

Circuits and Systems, 1996. ISCAS '96., Connecting the World., 1996 IEEE International Symposium on Pages: 45 - 48 vol.2 1996 Vol. 2 ISBN: 0-7803-3073-0

Abstract:

Pitch determination and speech segmentation are two important parts of speech recognition and speech processing in general. This paper proposes a time-based event detection method for finding the pitch period of a speech signal. Based on the discrete wavelet transform, it detects voiced speech, which is local in frequency, and determines the pitch period. This method is computationally inexpensive and through simulations and real speech experiments we show that it is both accurate and robust to noise.

C.6 Wavelet based feature extraction for phoneme recognition

This Paper Appears in:

Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on Pages: 264 - 267 vol.1.This Conference was Held : 3-6 Oct. 1996, Vol. 1 ISBN: 0-7803-3555-4

Abstract:

In an effort to provide a more efficient representation of the acoustical speech signal in the pre classification stage of a speech recognition system, we consider the application of the Best-Basis Algorithm of R.R. Coifman and M.L. Wickerhauser (1992). This

combines the advantages of using a smooth, compactly supported wavelet basis with an adaptive time scale analysis, dependent on the problem at hand. We start by briefly reviewing areas within speech recognition where the wavelet transform has been applied with some success. Examples include pitch detection, formant tracking, phoneme classification. Finally, our wavelet based feature extraction system is described and its performance on a simple phonetic classification problem given.

C.7 Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms

This Paper Appears in :

Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on Pages: 1209 - 1212 vol.2. This Conference was Held : 3-6 Oct. 1996, Vol. 2 ISBN: 0-7803-3555-4.

Abstract:

An improvement of an existing pitch detection algorithm is presented. The solution reduces the computational load of its precedent algorithm and introduces a voiced/unvoiced decision step to reduce the number of errors. The efficiency of this improved system is tested with a semi-automatically segmented speech database according to the information delivered by an attached laryngograph signal. The results show its periodicity detection.

C.8 Optimal wavelet representation of signals and the wavelet sampling theorem

This Paper Appears in :

Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on Pages: 262 – 277 April 1994 Vol. 41 Issue: 4 ISSN: 1057-7130.

Abstract:

The wavelet representation using orthonormal wavelet bases has received widespread attention. Recently M-band orthonormal wavelet bases have been constructed and compactly supported M-band wavelets have been parameterized. This paper gives the theory and algorithms for obtaining the optimal wavelet multiresolution analysis for the representation of a given signal at a predetermined scale in a variety of error norms. Moreover, for classes of signals, this paper gives the theory and algorithms for designing the robust wavelet multiresolution analysis that minimizes the worst case approximation error among all signals in the class. All results are derived for the general M-band multiresolution analysis. An efficient numerical scheme is also described for the design of the optimal wavelet multiresolution analysis when the least-squared error criterion is used. Wavelet theory introduces the concept of scale which is analogous to the concept of frequency in Fourier analysis. This paper introduces essentially scale limited signals and shows that band limited signals are essentially scale limited, and gives the wavelet sampling theorem, which states that the scaling function expansion coefficients of a function with respect to an M-band wavelet basis, at a certain scale (and above) completely specify a band limited signal (i.e., behave like Nyquist (or higher) rate samples).

C.9 Robust classification of speech based on the dyadic wavelet transform with application to CELP coding

This Paper Appears in :

Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on Pages: 546 - 549 vol. 1. This Conference was Held : 7-10 May 1996, Vol. 1 ISBN: 0-7803-3192-3.

Abstract:

This paper describes a new algorithm for the classification of telephone-bandwidth speech that is designed for efficient control of bit allocation in low bit-rate speech coders. The algorithm is based on the dyadic wavelet transform (D/sub y/WT) and classifies each unit subframe into one of the three categories background noise/unvoiced, transients/voicing onsets, periodic/voiced. A set of three parameters is derived from the D/sub y/WT coefficients, each giving a decision score that the associated class is active. Taking the history into account, a finite-state model controlled by these parameters computes the classifier's decision. The proposed algorithm is robust to various types of background noise. In comparison with a classifier based on the long-term autocorrelation function, the D/sub y/WT classifier proves to be superior. To evaluate its performance in CELP-type speech coders, a variety of excitation coding schemes with bit rates between 2200 and 4800 bit/s is investigated.

C.10 Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms

This Paper Appears in :

Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on Pages: 1209 - 1212 vol.2. This Conference was Held : 3-6 Oct. 1996, Vol. 2 ISBN: 0-7803-3555-4

Abstract:

An improvement of an existing pitch detection algorithm is presented. The solution reduces the computational load of its precedent algorithm and introduces a voiced/unvoiced decision step to reduce the number of errors. The efficiency of this improved system is tested with a semi-automatically segmented speech database according to the

information delivered by an attached laryngograph signal. The results show its periodicity detection.

Appendix D

Efficiency measure

Efficiency measure is very important in evaluating a technique. In this part, the technique that is used in systems evaluation will be discussed.

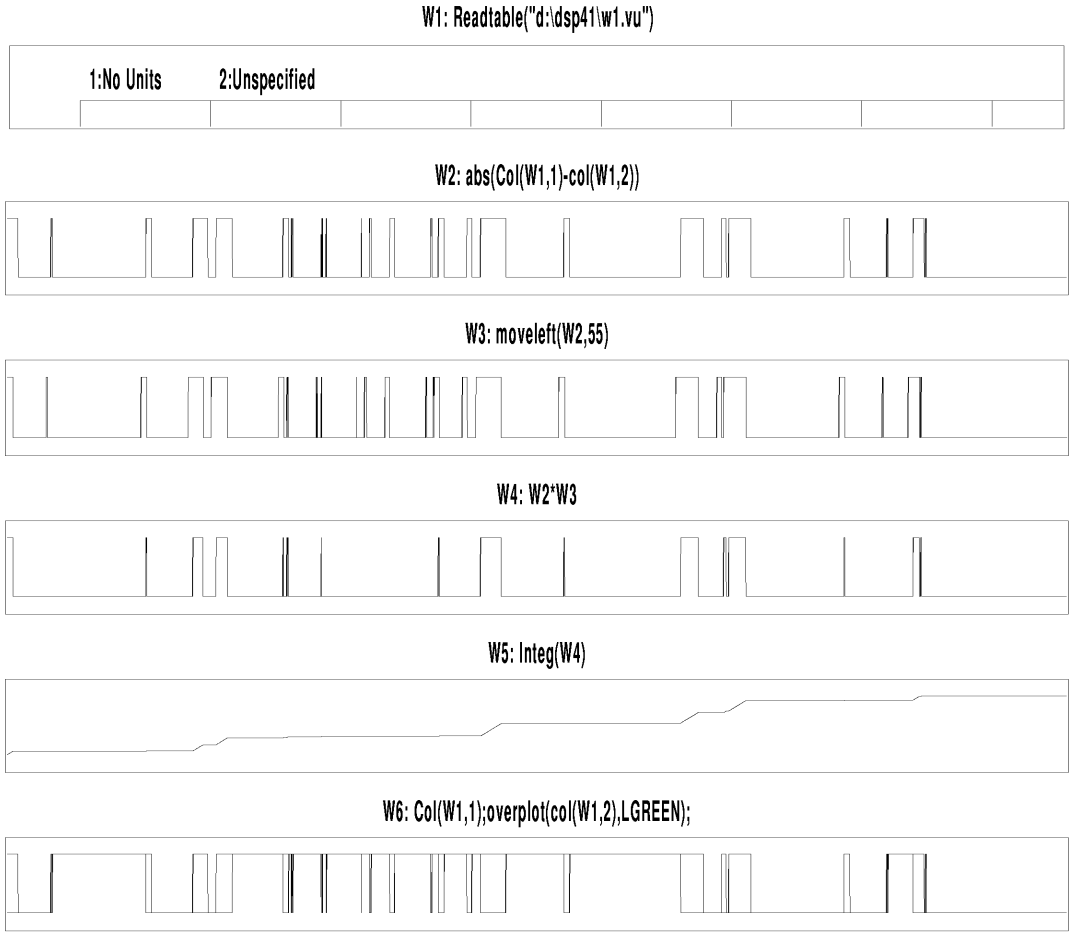


Figure D. 1 System of efficiency measure.

Figure D.1 represents the worksheet for systems evaluation. In the first window ‘W1’ the file contains two columns of data is read. The first column is the markers corresponds to the optimal output (required output from the tested system). The second column contains current output of the system under test.

Window 2 ‘W2’ of figure D.1 contains the absolute difference between Column 2 and column 1 in ‘W1’. In ideal output this must be zero.

Window 3 ‘W3’ of figure D.1 contains the same information in ‘W2’ shifted left by about 55 samples (~5 ms). This window is the allowed tolerance of output.

Window 4 ‘W4’ of Figure D.1 contains results of multiplying ‘W2’ and ‘W3’. This reduces error according to the allowed tolerance.

Area under marker of ‘W4’ is calculated in ‘W5’ by integrating the normalized ‘W4’. This area represents the total error in the system (this area represents the total duration of error markers due to integration of normalized curve).

Dividing it over the total period of markers averages the error area. The efficiency is calculated by the following equation:

$$\eta = (1 - \frac{T_{err}}{T}) * 100\% \qquad \textbf{(D. 1)}$$

T_{err} : Total duration of error.

T : Total period of markers.

Window 5 ‘W5’ contains the required markers overlaid with the current markers.

References

[1]. B. H. Juang,"The past, Present, and Future of speech processing", IEEE Signal Processing magazine, May 1998, vol. 15, No.3.

[2]. Philipos C. Loizou, " Mimicking the Human Ear", IEEE Signal Processing magazine, September 1998, vol. 15, No.5.

[3]. J. D. Markel and A. H. Gray, Linear Prediction of Speech,Springer-Verlag,Berlin Heidelberg,Germany,1976, pp. 1-63.

[4]. Thomas W. Parsons, Voice and speech processing,McGraw-Hill inc.,1987,pp. 57-98, 136-192, 291-317.

[5]. Lawrence R. Rabiner, Digital Processing of Speech Signals,Englewood Cliffs New Jersey: Prentice-Hall inc., 1978, pp. 43-55, 130-135.

[6]. Gilbert Strang , Wavelets and Filter Banks , Wellesley-Cambridge Press ,1996, pp 1 - 34 ,pp 53-60, pp 155-172.

[7]. Mark J. Shensa , " The Discrete Wavelet Transform: Wedding the À Trous and Mallat Algorithms" , IEEE Transactions on Signal Processing, VOL. 40, NO. 10, October 1992.

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

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

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

[11]. Agbinya, J.I.," Discrete wavelet transform techniques in speech processing", TENCON '96. Proceedings., 1996 IEEE TENCON. Digital Signal Processing Applications, pp: 514 - 519 vol.2. 1996 Vol. 2 ISBN: 0-7803-3679-8

[12]. Gopinath, R.A.; Odegard, J.E.; Burrus, C.S.," Optimal wavelet representation of signals and the wavelet sampling theorem", Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions, pp: 262 – 277. April 1994 Vol. 41 Issue: 4 ISSN: 1057-7130

[13]. Rumelhart D.E., and J.L. McCelland, Parallel Distributed Processing PDP: Explorations in the Microstructure of Cognition, Vol. 1,MIT Press, Cambridge MA. 1986.

[14]. Patterson D.W. Artificial Neural Networks, Theory and Application,, Prentice Hall 1996.

[15]. Hammerstrom D. ,"Neural Networks at Work", IEEE Spectrum June 1993.

[16]. Edward J. Dudewicz, Satya N. Mishra, "Modern Mathematical Statistics", John Wiley & Sons, New York, 1988, pp:694-697.

[17]. Nemat Sayed Abdel Kader, Amr M. Refat ," Voiced/Unvoiced Classification using Wavelet based algorithm", ICSPAT98.

[18]. Nemat Sayed Abdel Kader, Amr M. Refat, "Voiced/Unvoiced classification using wavelet correlation model", ICSPAT'99

[19]. Mati Zirra," Pitch Detection of speech by Dyadic Wavelet Transform",ICSPAT97.

[20]. M. S. Obaidat , " Wavelet algorithm for the estimation of pitch period of speech signal" , ICECS 96.

[21]. Wendt, C.; Petropulu, A.P.," Pitch determination and speech segmentation using the discrete wavelet transform" ,Circuits and Systems, 1996. ISCAS '96., Connecting the World., 1996 IEEE International Symposium on Pages: 45-48 vol.2.

[22]. Logan, J.; Gowdy, J.," Adaptive pitch period decimation and its application in speech compression", Southeastcon '96. Bringing Together Education, Science and Technology., Proceedings of the IEEE,pp : 220 – 222, 1996 ISBN: 0-7803-3088-9,

[23]. Janer, L.," New pitch detection algorithm based on wavelet transform", Time-Frequency and Time-Scale Analysis, 1998. Proceedings of the IEEE-SP International Symposium, pp : 165 – 168, 1998 ISBN: 0-7803-5073-1

[24]. Obaidat, M.S.; Lee, T.; Zhang, E.; Khalid, G.; Nelson, D.," Wavelet algorithm for the estimation of pitch period of speech signal", Electronics, Circuits, and Systems, 1996. ICECS '96., Proceedings of the Third IEEE International Conference, pp: 471 - 474 vol.1, 1996 Vol. 1 ISBN: 0-7803-3650-X

[25]. Janer, L.; Bonet, J.J.; Lleida-Solano, E.," Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms",

Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference , pp: 1209 - 1212 vol.2 .

[26]. Nemat Sayed Abdel Kader, Amr M. Refat , “End points detection using wavelet based algorithm”, Eurospeech’99

[27]. C. J. Long and S. Datta , “Wavelet based feature extraction for phoneme recognition” ,

[28]. Long, C.J.; Datta, S.,” Wavelet based feature extraction for phoneme recognition”, Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference, pp : 264 - 267 vol.1., 1996 ISBN: 0-7803-3555-4.

[29]. Stegmann, J.; Schroder, G.; Fischer, K.A.,” Robust classification of speech based on the dyadic wavelet transform with application to CELP coding”, Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference , pp: 546 - 549 vol. 1,1996 Vol. 1 ISBN: 0-7803-3192-3

[30]. Nemat SayedAbdel Kader, “ Arabic Text-to-Speech Synthesis by Rule”, Ph.D. thesis , Cairo Univesity, Faculty of Engineering, Electronics and communication Dept., 1992. PP: 84-90,135-137

References

[1]. B. H. Juang,"The past, Present, and Future of Speech Processing", IEEE Signal Processing magazine, May 1998, vol. 15, No.3.

[2]. Philipos C. Loizou, " Mimicking the Human Ear", IEEE Signal Processing magazine, September 1998, vol. 15, No.5.

[3]. J. D. Markel and A. H. Gray," Linear Prediction of Speech", Springer-Verlag,Berlin Heidelberg,Germany,1976, pp. 1-63.

[4]. Thomas W. Parsons, "Voice and Speech Processing " ,McGraw-Hill inc.,1987,pp. 57-98, 136-192, 291-317.

[5]. Lawrence R. Rabiner, "Digital Processing of Speech Signals",Englewood Cliffs New Jersey: Prentice-Hall inc., 1978, pp. 43-55, 130-135.

[6]. Gilbert Strang , "Wavelets and Filter Banks" , Wellesley-Cambridge Press ,1996, pp 1 - 34 ,pp 53-60, pp 155-172.

[7]. Mark J. Shensa , " The Discrete Wavelet Transform: Wedding the À Trous and Mallat Algorithms" , IEEE Transactions on Signal Processing, VOL. 40, NO. 10, October 1992.

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

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

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

[11]. Madhukumar, A.S.; Premkumar, A.B.; Abut, H Wiener, "Wavelet Quantization of Noisy Speech using Constrained

Filtering", Conference Record of the Asilomar Conference on Signals, Systems & Computers v 1 Nov 2-5 1997 1998 Sponsored by: IEEE IEEE Comp Soc p 39-43.

[12]. Boland, S.; Deriche , "High Quality Audio Coding using Multipulse LPC and Wavelet Decomposition", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 5 May 9-12 1995 1995 Sponsored by: IEEE p 3067-3069.

[13]. Carnero, Benito; Drygajlo, Andrzej, "Perceptual Speech Coding using Time and Frequency Masking Constraints", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings v 2 Apr 21-24 1997 1997 Sponsored by: IEEE IEEE p 1363-1366.

[14]. Heitz, C.; Becker, J.D., "Optimized Time-Frequency Distribution for Speech Analysis" , Speech Communication v 14 n 1 Feb 1994 Publ by Elsevier Science Publishers B.V. p 1-18

[15]. Benedetto, John J.;Teolis, Anthony ," Wavelet Auditory Model and Data Compression", Applied and Computational Harmonic Analysis v 1 n 1 Dec 1993 p 3-28.

[16]. Anderson, David V. ," Speech Analysis and Coding using a Multi-Resolution Sinusoidal Transform", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2 May 7-10 1996 1996 Sponsored by: IEEE p 1037-1040.

[17]. Gidas, Basilis; Murua, Alejandro, "Classification and Clustering of Stop Consonants via Nonparametric Transformations and Wavelets", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 1 May 9-12 1995 Sponsored by: IEEE p 872-875.

[18]. Whitmal, Nathaniel A.; Rutledge, Janet C.; Cohen, Jonathan, "Wavelet-Based Noise Reduction", Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal

Processing 5 May 9-12 1995 1995 Sponsored by: IEEE IEEE p 3003-3006

[19]. Sarikaya, Ruhi; Gowdy, John N., " Wavelet Based Analysis of Speech Under Stress", Conference Proceedings-IEEE SOUTHEASTCON Apr 12-14 1997 1997 Sponsored by: IEEE p 92-96.

[20]. Boland, Simon D.; Deriche, Mohamed, "Hybrid LPC and Discrete Wavelet Transform Audio Coding With a Novel Bit Allocation Algorithm", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings v 6 May 12-15 1998 1998 Sponsored by: IEEE p 3657-3660.

[21]. Wang, Kuansan; Shamma, Shihab A.; Byrne, William J. , "Noise Robustness in the Auditory Representation of Speech Signals", Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing v 2 Apr 27-30 1993 1993 Sponsored by: IEEE; Signal Processing Society Publ by IEEE p II-335-II-338

[22]. Chong, Wonyong; Kim, Jongsoo, "Speech and Image Compressions by DCT", wavelet, and wavelet packet, Proceedings of the International Conference on Information, Communications and Signal Processing, ICICS v 3 Sep 9-12 1997 1997 Sponsored by: IEEE IEEE p 1353-1357.

[23]. Kadambe, Shubha; Bourdeaux-Bartels, G. F. ," A Comparison of a Wavelet Functions for Pitch Detection of Speech Signals", Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing v 1 May 14-17 1991 1991 Sponsored by: IEEE Signal Processing Soc Publ by IEEE p 449-452 .

[24]. Guelzow, T.; Engelsberg, A.; Heute, "Comparison of a Discrete Wavelet Transformation and a Nonuniform Polyphase Filterbank Applied to Spectral-Subtraction Speech Enhancement", Signal Processing v 64 n 1 Jan 1998 Elsevier Sci B.V. p 5-19.

[25]. Ris, Christophe; Fontaine, Vincent; Leich, Henri, "Speech Analysis Based on Malvar Wavelet Transform", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 1 May 9-12 1995 1995 Sponsored by: IEEE IEEE p 389-392.

[26]. Ooi, James; Viswanathan, Vishu, "Computationally Efficient Wavelet Transform CELP Coder", Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing 2 Apr 19-22 1994 1994 Sponsored by: IEEE IEEE p 101-104.

[27]. Souza, M.N.; Caloba, L.P., "Comparison between Fourier and Biological Auditory Based Time-Frequency Distributions, Applied to the Speech Signals", Midwest Symposium on Circuits and Systems v 2 Aug 18-21 1996 1996 Sponsored by: IEEE IEEE p 807-810 .

[28]. Ai, Hongmei; Yang, Xingjun; Lu, Dajin , "Wavelet-Excited Linear Prediction (WELP) - New Method for Lower Bit Rate Speech Coding ", Tien Tzu Hsueh Pao/Acta Electronica Sinica v 25 n 4 Apr 1997 Chinese Institute of Electronics p 120-124.

[29]. Agbinya, J.I., "Discrete Wavelet Transform Techniques in Speech Processing", TENCON '96. Proceedings., 1996 IEEE TENCON. Digital Signal Processing Applications, pp: 514 - 519 vol.2. 1996 Vol. 2 ISBN: 0-7803-3679-8

[30]. Seok, Jong Won; Bae, Keun Sung , "Speech Enhancement with Reduction of Noise Components in the Wavelet Domain", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings v 2 Apr 21-24 1997 1997 Sponsored by: IEEE IEEE p 1323-1326.

[31]. Quirk, Patrick J.; Tseng, Yi-Chyun; Adhami, Reza R., "Efficient Wavelet-Based Voice/Data Discriminator for Telephone Networks", Proceedings of SPIE - The International Society for Optical Engineering v 2750 Apr 10-11 96 1996

Sponsored by: SPIE - Int Soc for Opt Engineering, Bellingham, WA USA p 139-146

[32]. Averbuch, Amir; Bobrovsky, Sheinin, "Speech Compression using Wavelet Packet and Vector Quantizer with 8-msec Delay", Proceedings of SPIE - The International Society for Optical Engineering v 2569/1 Jul 12-14 1995 1995 Sponsored by: SPIE - Int Soc for Opt Engineering, Bellingham, WA USA Society of Photo-Optical Instrumentation Engineers p 320-332 .

[33]. Kadambe, Shubha L.; Srinivasan, Pramila, "Applications of Adaptive Wavelets for Speech", Optical Engineering 33 7 7 1994 Society of Photo-Optical Instrumentation Engineers p 2204-2211.

[34]. Soon, Ing Yann; Koh, Soo Ngee; Yeo, Chai Kiat, "Wavelet for Speech Denoising", IEEE Region 10 Annual International Conference, Proceedings/TENCON v 2 Dec 2-4 1997 1997 Sponsored by: IEEE IEEE p 479-482.

[35]. Gopinath, R.A.; Odegard, J.E.; Burrus, C.S., "Optimal Wavelet Representation of Signals and the Wavelet Sampling Theorem", Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions, pp: 262 – 277. April 1994 Vol. 41 Issue: 4 ISSN: 1057-7130

[36]. Rumelhart D.E., and J.L. McCelland, "Parallel Distributed Processing PDP: Explorations in the Microstructure of Cognition", Vol. 1,MIT Press, Cambridge MA. 1986.

[37]. Agbinya, Johnson Ihyeh, "Discrete Wavelet Transform Techniques in Speech Processing", IEEE Region 10 Annual International Connference, Proceedings/TENCON v 2 Nov 26-29 1996 1996 Sponsored by: IEEE IEEE p 514-519

[38]. Pinter, Istvan ," Perceptual Wavelet-Representation of Speech Signals and its Application to Speech Enhancement", Computer Speech & Language 10 1 Jan 1996 Academic Press Ltd p 1-22 0885-2308.

[39]. Patterson D.W.," Artificial Neural Networks, Theory and Application", Prentice Hall 1996.

[40]. Teuvo Kohonen, " Neural Phonetic Typewriter", IEEE in computer, March 1988, p 12-20.

[41]. Hammerstrom D. ," Neural Networks at Work", IEEE Spectrum June 1993.

[42]. Gavin J. Gibsson," A Combinational Approach to Understanding Perceptron Capabilities", IEEE Transactions on neural networks. Vol. 4, No. 6, November 1993.

[43]. David S. Chen , "A Robust Back Probagation Learning Algorithem for Function Approximation", IEEE Transactions on neural networks. Vol. 5, No. 3, May 1994.

[44]. Rafik Braham," The Design Of Neural Network With A Biologically Motivated Architecture" , IEEE Transactions on neural networks. Vol. 1, No. 3, September 1990.

[45]. Zezhen Huang, "A Combined Self-Organizing Feature Map and Multilayer Perceptron for Isolated Word Recognition", IEEE proc., 1992.

[46]. Davenport, Michael R.; Garudadri, Harinath ," A Neural Net Acoustic Phonetic Feature Extractor Based on Wavelets", Proceedings of the 1991 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing Processing. Conference Proceedings May 9-10 1991 1991 Sponsored by: IEEE Victoria Section; IEEE Region 7; Univ of Victoria Publ by IEEE p 449-452.

[47]. Edward J. Dudewicz, Satya N. Mishra, "Modern Mathematical Statistics", John Wiley & Sons, New York, 1988, pp:694-697.

[48]. Nemat Sayed Abdel Kader, Amr M. Refat, "Voiced/Unvoiced Classification using Wavelet Based Algorithm", ICSPAT98.

[49]. Nemat Sayed Abdel Kader, Amr M. Refat, "Voiced/Unvoiced Classification using Wavelet Correlation Model", ICSPAT'99.

[50]. Tan, Beng T.; Lang, Robert; Schroder, Heiko; Spray, Andrew; Dermody, Phillip, "Applying Wavelet Analysis to Speech Segmentation and Classification", Proceedings of SPIE - The International Society for Optical Engineering v 2242 Apr 5-8 1994 1994 Sponsored by: SPIE - Int Soc for Opt Engineering, Bellingham, WA USA Publ by Society of Photo-Optical Instrumentation Engineers p 750-761.

[51]. Stegmann, Joachim; Schroeder, Gerhard , "Robust Voice-Activity Detection Based on the Wavelet Transform ", IEEE Workshop on Speech Coding for Telecommunications Proceedings Sep 7-10 1997 1997 Sponsored by: IEEE p 99-100.

[52]. Evangelista, Gianpaolo," Pitch-Synchronous Wavelet Representations of Speech and Music Signals", IEEE Transactions on Signal Processing v 41 n 12 Dec 1993 p 3313-3330.

[53]. Obaidat, M.S.; Brodzik, Andy; Sadoun, "Performance Evaluation Study of Four Wavelet Algorithms for the Pitch Period Estimation of Speech Signals", Information Sciences v 112 n 1-4 Dec 1998 Elsevier Science Inc p 213-221

[54]. Nam, Hojung; Kim, Hyoung-soo; Kwon, Y.; Yang, Sung-il, "Speaker Verification System using Hybrid Model with Pitch Detection by Wavelets", Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis Oct 6-9 1998 1998 Sponsored by: IEEE IEEE p 153-156.

[55]. Yang, Haiyun; Qiu, Lunji; Koh, Soo-Ngee , "Application of Instantaneous Frequency Estimation for Fundamental Frequency Detection", Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis Oct 25-28 1994 Sponsored by: IEEE p 616-619.

[56]. Erogul, Osman; Serinken, Nur, "Multiresolutional Analysis/Synthesis Approach For the Pitch Modification of Speech signals", Turkish Journal of Electrical Engineering & Computer Sciences v 5 n 3 1997 Sci Tech Res Counc Turkey p 315-323

[57]. Logan, Joseph; Gowdy, John , "Adaptive Pitch Period Decimation and its Application in Speech Compression", Conference Proceedings - IEEE SOUTHEASTCON Apr 11-14 1996 1996 Sponsored by: IEEE IEEE p 220-222.

[58]. Sasou, Akira;Nakamura, Shogo, "Pitch Extraction Method using the Wavelet Transform", Electronics & Communications in Japan, Part III: Fundamental Electronic Science (English translation of Denshi Tsushin Gakkai Ronbunshi) v 82 n 6 Jun 1999 Scripta Technica Inc p 36-45.

[59]. Du, Limin; Hou, Ziqiang, "Manifestation of Glottal Closure Singularity in Wavelet Transform Domain", Tien Tzu Hsueh Pao/Acta Electronica Sinica v 25 n 8 1997 Chinese Institute of Electronics p 6-13.

[60]. Mati Zirra," Pitch Detection of Speech by Dyadic Wavelet Transform",ICSPAT97.

[61]. M. S. Obaidat , "Wavelet Algorithm for the Estimation of Pitch Period of Speech Signal" , ICECS 96.

[62]. Du, Limin; Hou, Ziqiang , "Determination of the Instants of Glottal Closure from Speech Wave using Wavelet Transform", International Conference on Signal Processing Proceedings, ICSP v 1 Oct 14-18 1996 1996 Sponsored by: IEEE IEEE p 273-275.

[63]. Yip, Wing-kei; Leung, Kwong-sak; Wong, Kin-hong, "Pitch Detection of Speech Signals in Noisy Environment by Wavelet", Proceedings of SPIE - The International Society for Optical Engineering v 2491/1 1995 Sponsored by: SPIE - Int Soc for Opt Engineering, Bellingham, WA USA Society of Photo-Optical Instrumentation Engineers p 604-614.

[64]. Wendt, C.; Petropulu, A.P., "Pitch Determination and Speech Segmentation using the Discrete Wavelet Transform" ,Circuits and Systems, 1996. ISCAS '96., Connecting the World., 1996 IEEE International Symposium on Pages: 45-48 vol.2.

[65]. Logan, J.; Gowdy, J., "Adaptive Pitch Period Decimation and its Application in Speech Compression", Southeastcon '96. Bringing Together Education, Science and Technology., Proceedings of the IEEE,pp : 220 – 222, 1996 ISBN: 0-7803-3088-9,

[66]. Janer, L., "New Pitch Detection Algorithm Based on Wavelet Transform, Time-Frequency and Time-Scale Analysis", 1998. Proceedings of the IEEE-SP International Symposium, pp : 165 – 168, 1998 ISBN: 0-7803-5073-1

[67]. Obaidat, M.S.; Lee, T.; Zhang, E.; Khalid, G.; Nelson, D., "Wavelet Algorithm for the Estimation of Pitch Period of Speech Signal", Electronics, Circuits, and Systems, 1996. ICECS '96., Proceedings of the Third IEEE International Conference, pp: 471 - 474 vol.1, 1996 Vol. 1 ISBN: 0-7803-3650-X

[68]. Janer, L.; Bonet, J.J.; Lleida-Solano, E., "Pitch Detection and Voiced/Unvoiced Decision Algorithm based on Wavelet Transforms", Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference , pp: 1209 - 1212 vol.2 .

[69]. Nemat Sayed Abdel Kader, Amr M. Refat , "End Points Detection using Wavelet Based Algorithm", Eurospeech’99

[70]. Long, C.J.; Datta, S., "Wavelet Based Feature Extraction for Phoneme Recognition, Spoken Language", 1996. ICSLP 96.

Proceedings., Fourth International Conference, pp : 264 - 267 vol.1., 1996 ISBN: 0-7803-3555-4

[71]. Stegmann, J.; Schroder, G.; Fischer, K.A., "Robust Classification of Speech Based on the Dyadic Wavelet Transform with Application to CELP Coding", Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference , pp: 546 - 549 vol. 1,1996 Vol. 1 ISBN: 0-7803-3192-3

[72]. Nemat Sayed Abdel Kader ," Arabic Text-to-Speech Synthesis by Rule", Ph.D thesies , Cairo university, faculty of Eng., communication dept., 1992. Page 165.

ملخص الرسالة

التوسع في استخدام التطبيقات التي تعتمد على التحكم عن طريق الصوت مثل آلة الإملاء الآلية و تطبيقات التعرف على الأشخاص عن طريق الصوت و غيرها جعل كثير من الباحثين يتجهون إلى البحث في هذا المجال لرفع أداء هذه التطبيقات و تطويرها.

يهدف هذا البحث إلى محاولة حل مشكلة التعرف على وحدات الصوت الأساسية بطريقة آلية عن طريق محاولة تطوير أداء الدعامات الرئيسية التي تستخدمها الآلة لفهم الموجة الصوتية. هذه الدعامات التي تناولها هذا البحث هي:

1. تحديد حدود الكلمة و ذلك في وسط عادي و وسط عالي الشوشرة.
2. تحديد الأجزاء الصوتية و الأجزاء غير الصوتية بداخل الكلمة و ذلك في وسط عادي و وسط عالي الشوشرة.
3. تحديد التردد الرئيسي داخل الأجزاء الصوتية في الكلمة و ذلك في وسط عادي و وسط عالي الشوشرة.
4. تحديد الوحدات الصوتية الرئيسية (أحرف ساكنة و حركات) داخل الكلمة.
5. التعرف على الوحدات الصوتية.

بداء البحث بشرح طبيعة الموجات الصوتية. ثم تطرق البحث لشرح الموجات المحدودة و كيفية الاستفادة منها لفهم موجة الصوت حيث تستطيع الموجات المحدودة التعبير عن التغير في المحتوى الترددي لموجة الصوت على مدار زمن النطق.

استخدمت طريقة الموجات المحدودة لرفع أداء الأربع دعامات التي تم سردهم سابقا. تم عمل مقارنات مع الطرق التقليدية لحل هذه المشاكل.

يتكون الباب الأول من مقدمة و شرح مبسط للمويجات و كيفية الاستفادة منها في تطبيقات متعلقة بالموجات الصوتية. كذلك يتعرض الباب لشرح بعض الطرق المستخدمة في التصنيف و هم التصنيف باستخدام الشبكات العصبية و التصنيف بعمل نموذج رياضي خطي يعتمد على بيانات إحصائية.

يتعرض الباب الثاني لمشكلة استنباط المحتوى الكلامي من الكلام المنطوق في وسط عادي أو وسط عالي الشوشرة. يبتدئ الباب بشرح مبسط لأحد الطرق شائعة الاستخدام في هذا المجال و هي الطريقة التي تعتمد على الطاقة و معدل تغير الإشارة. ثم يقدم هذا الباب ثلاثة خوارزميات تعتمد على المويجات لحل هذه المشكلة. الطريقة الأولى مبنية على استنباط دالة اعتما دية من معاملات المويجات المختارة في حيز ترددي معين. هذه الدالة قادرة على تتبع المحتوى الكلامي من خلال الكلام المنطوق. و تم تجربة الطريقة في وسط عادي و وسط عالي الشوشرة. أما الطريقتين الأخريين فيعمدان على التصنيف بواسطة الشبكات العصبية أو النموذج الخطي لمعاملات المويجات في كل الحيز الترددي المتاح.

الباب الثالث يستعرض مشكلة التصنيف لأجزاء الكلام إلى أجزاء صوتية و أجزاء غير صوتية و تحديد التردد الأساسي للأجزاء الصوتية. يبتدئ الباب بشرح خصائص الأجزاء الصوتية و الغير صوتية. يتطرق الباب بعد ذلك للطرق المستخدمة في عملية التصنيف. يقدم الباب بعد ذلك ثلاث خوارزميات جديدة تعتمد على المويجات لعمل التصنيف. تعتمد الطريقة الأولى على استنباط دالة من معاملات المويجات هذه الدالة تتتبع التغيرات الصوتية. أما الطريقة الثانية فتعتمد على اختلاف الخصائص الإعتمادية بين معاملات المويجات للأجزاء الصوتية عنها للأجزاء غير الصوتية. أما الطريقة الثالثة فتعتمد على التصنيف الإحصائي الخطي عن طريق معاملات المويجات.

يتطرق الباب الثالث بعد ذلك لمشكلة تحديد التردد الرئيسي للأجزاء الصوتية. يقوم الباب بعرض طريقتين لتحديد التردد باستخدام المويجات.

الباب الرابع يتناول معالجة مشكلة التصنيف للوحدات الصوتية للغة العربية إلى ساكنة و متحركة. و بعد ذلك يتعرض لمشكلة التعرف على الوحدة المتحركة. و قد تمت معالجة مشكلة التصنيف باستخدام طريقتين يعتمدا على المويجات. كما تمت معالجة التعرف على الوحدات المتحركة باستخدام النموذج الإحصائي الخطي لمعاملات المويجات.

الباب الخامس يعرض النظام المتكامل الذي تم تنفيذه و يحتوي على كل الخوارزميات السابق شرحها في الأبواب السابقة.

الباب السادس عبارة عن ملخص للبحث و استنتاجات. كما يحتوي على بعض الإمتدادات المستقبلية للبحث.

التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على المويجات

إعداد

عمرو مـــحمد رفعت مـــحمد جودي

رسالة مقدمه إلى كلية الهندسة جامعة القاهرة

كجزء من متطلبات الحصول على درجة الدكتوراه

في

هندسة الإلكترونيات و الاتصالات الكهربية

كلية الهندسة ـ جامعة القاهرة

الجيزة ـ جمهورية مصر العربية

1999

التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على المويجات

إعداد

عمرو مـــحمد رفعت مـــحمد جودي

رسالة مقدمه إلى كلية الهندسة جامعة القاهرة

كجزء من متطلبات الحصول على درجة الدكتوراه

في

هندسة الإلكترونيات و الاتصالات الكهربية

تحت إشراف

أ.د. أمين محمد نصار

قسم هندسة الإلكترونيات و الاتصالات الكهربية

كلية الهندسة–جامعة القاهرة

د. نعمت سيد عبد القادر

قسم هندسة الإلكترونيات و الاتصالات الكهربية

كلية الهندسة–جامعة القاهرة

كلية الهندسة – جامعة القاهرة

الجيزة – جمهورية مصر العربية

التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على المويجات

إعداد

عمرو مـــحمد رفعت مـــحمد جودي

رسالة مقدمه إلى كلية الهندسة جامعة القاهرة

كجزء من متطلبات الحصول على درجة الدكتوراه

في

هندسة الإلكترونيات و الاتصالات الكهربية

يعتمد من لجنة الممتحنين:

أستاذ دكتور/ أمين نصار	المشرف الرئيسي
أستاذ دكتور/ سلوى حسين الرملي	عضو
أستاذ م. دكتور/ محسن رشوان	عضو
دكتور/ نعمت سيد عبد القادر	المشرف الرئيسي

كلية الهندسة – جامعة القاهرة

الجيزة – جمهورية مصر العربية

التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على المويجات

إعداد

عمرو مـــحمد رفعت مـــحمد جودي

رسالة مقدمه إلى كلية الهندسة جامعة القاهرة

كجزء من متطلبات الحصول على درجة الدكتوراه

في

هندسة الإلكترونيات و الاتصالات الكهربية

يعتمد من لجنة الممتحنين:

أستاذ دكتور/ أمين نصار	المشرف الرئيسي
أستاذ دكتور/ مجدي فكري	عضو
أستاذ دكتور/ سلوى حسين الرملي	عضو

كلية الهندسة – جامعة القاهرة

الجيزة – جمهورية مصر العربية

الأستاذ الدكتور/ رئيس قسم هندسة الإلكترونيات و الاتصالات الكهربية كلية الهندسة-جامعة القاهرة،

تحية طيبة و بعد،

أحيط سيادتكم علما بأن المهندس/ عمرو محمد رفعت محمد جودي قد أتم رسالة الدكتوراه و قد ألقى محاضرة عامة Seminar يوم السبت 1999/3/6 و الرسالة تحت عنوان ،

"التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على الموجات المحدودة"

"Speech processing using wavelet-based algorithms"

و اللجنة المقترحة هي:

1-أ.د. سلوى حسين الرملي	قسم الإلكترونيات و الاتصالات الكهربية كلية الهندسة-جامعة عين شمس
2- أ.م.د. محسن رشوان	قسم الإلكترونيات و الاتصالات الكهربية كلية الهندسة-جامعة القاهرة
3-أ.د. أمين محمد نصار	مشرفين
د. نعمت سيد عبد القادر	قسم الإلكترونيات و الاتصالات الكهربية كلية الهندسة-جامعة القاهرة

و لكم جزيل الشكر،

أ.د. أمين محمد نصار

1999/3/

تقرير صلاحية

عن رسالة الدكتوراه المقدمة من المهندس/عمرو محمد رفعت محمد جودي إلى قسم الإلكترونيات و الاتصالات الكهربية كلية الهندسة–جامعة القاهرة تحت عنوان،

"التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على الموجات المحدودة"

"Speech processing using wavelet-based algorithms"

الرسالة المقدمة من المهندس/عمرو محمد رفعت محمد جودي تمت و هي الآن صالحة للمناقشة و شكرا،

د. نعمت سيد عبد القادر

1999/3/

أ.د. أمين محمد نصار

1999/3/

التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على المويجات

إعداد

عمرو مـــحمد رفعت مـــحمد جودي

رسالة مقدمه إلى كلية الهندسة جامعة القاهرة

كجزء من متطلبات الحصول على درجة الدكتوراه

في

هندسة الإلكترونيات و الاتصالات الكهربية

تحت إشراف

أ.د. أمين محمد نصار

قسم هندسة الإلكترونيات و الاتصالات الكهربية

كلية الهندسة–جامعة القاهرة

د. نعمت سيد عبد القادر

قسم هندسة الإلكترونيات و الاتصالات الكهربية

كلية الهندسة–جامعة القاهرة

كلية الهندسة ـ جامعة القاهرة

الجيزة ـ جمهورية مصر العربية

1999

التعامل مع الموجات الصوتية باستخدام خوارزميات تعتمد على المويجات

إعداد

عمرو مـــحمد رفعت مـــحمد جودي

رسالة مقدمه إلى كلية الهندسة جامعة القاهرة

كجزء من متطلبات الحصول على درجة الدكتوراه

في

هندسة الإلكترونيات و الاتصالات الكهربية

يعتمد من لجنة الممتحنين:

أستاذ دكتور/ أمين نصار	**المشرف الرئيسي**
أستاذ دكتور/ مجدي فكري محمد رجائي	**عضو**
أستاذ دكتور/ سلوى حسين الرملي	**عضو**

كلية الهندسة ـ جامعة القاهرة

الجيزة ـ جمهورية مصر العربية

1999