# USE OF PROSODY IN SPEECH RECOGNITION

by

Lale Akarun

B.S. in E.E. Boğaziçi University, 1984

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science
in
Electrical Engineering

Boğaziçi University

1986

# USE OF PROSODY

# IN SPEECH RECOGNITION

APPROVED BY:

Doç. Dr. Bülent Sankur ..B Sankur...

(Thesis Supervisor)

Doç. Dr. Yusuf P. Tan ....................

Y. Doç. Dr. Eser Taylan ....E. Taylan....

Y. Doç. Dr. Emin Anarım ....Emin..........

DATE OF APPROVAL: 12.12.1986

## ACKNOWLEDGEMENTS

I am very grateful to my thesis supervisor Dr. Bülent Sankur for his help, guidance and for his encouraging supervision at the preparation of this thesis.

I am also very grateful to Dr. Yusuf P. Tan for his help and guidance while working with the PDP 11/23 microcomputer.

I wish to express my thanks to Dr. Eser Taylan and Dr. Sumru özsoy for their interest and assistance.

Finally, I would like to express my very sincere thanks to my friends, Nedim Kender, Nükhet Azman, Murat Taşman, and to Cem Ersoy, without whose help I could do nothing.

# ABSTRACT

Automatic speech recognizers which were once considered as "a dream of mad scientists" have shown considerable success in the last decade. What has made this success possible has been the use of sophisticated mathematical tools along with speech knowledge at various levels. Future success seems to depend on the exhaustive use of the latter.

This thesis is an attempt at using prosodic information, which conveys speech knowledge at various levels, in recognition systems. Programs have been developed to extract physical correlates of prosodic features from the speech signal. Results of analyses with Turkish words and sentences point out some methods to detect linguistic cues from the speech signal. Based on these results, strategies are outlined for a Turkish speech recognizer. Some of these are integrated in an isolated word recognizer and improvements are obtained.

# ÖZETÇE

Son on yılda söz tanıma alanında büyük bir atılım yapılmış, somut başarılar elde edilmiştir. Bu başarının altında gelişmiş matematik modellerin kullanımı ile birlikte insan kavramasının çeşitli evrelerindeki ses bilgisinin söz tanıyıcılara uyarlanması yatmaktadır. Gelecekteki başarılara ses bilgisinin daha çok kullanılması ile ulaşılacağı anlaşılmaktadır.

Bu tezde çeşitli düzeylerde ses bilgisi taşıyan bürün bilgisinin söz tanımadaki kullanım alanları araştırılmaktadır. Bürün özelliklerinin fiziksel karşılıklarını sesten elde etmek için izlenceler geliştirilmiştir. Değişik sözcük ve tümcelerle yapılan çözümlemeler sonucunda dilbilimsel yapıların bulunması için bazı yordamlar ve bunların Türkçe söz tanıyıcılarda kullanım yolları önerilmiş, bazı yordamlar ayrık bir söz tanıyıcıda kullanılmış ve ilerlemeler elde edilmiştir.

## TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS .................................................... III

ABSTRACT ............................................................ IV

OZETCE .............................................................. V

LIST OF FIGURES ..................................................... IX

LIST OF TABLES ...................................................... XI

LIST OF SYMBOLS ..................................................... XII

I. INTRODUCTION ..................................................... 1

II. THE SPEECH SIGNAL ............................................... 4

2.1 SPEECH PRODUCTION ............................................... 4

2.1.1 HUMAN NATURAL SPEECH PRODUCTION ............................... 4

2.1.2 DIGITAL MODEL OF THE SPEECH SIGNAL ............................ 6

2.2 SPEECH PERCEPTION ............................................... 7

III. THE SOUND SYSTEM ................................................ 10

3.1 PHONEMES ........................................................ 10

3.1.1 VOWELS ........................................................ 11

3.1.2 DIPHTHONGS .................................................... 12

3.1.3 SEMIVOWELS .................................................... 12

3.1.4 NASALS ........................................................ 12

3.1.5 FRICATIVES .................................................... 13

3.1.6 STOPS ......................................................... 13

3.1.7 AFFRICATIVES .................................................. 14

3.1.8 FLAPS ......................................................... 14

Page

3.1.9 LATERALS .............................. 14

3.2 SYLLABLES ................................ 15

3.3 PROSODY .................................. 17

3.3.1 DURATION .............................. 17

3.3.2 STRESS ................................ 18

3.3.3 TONE AND INTONATION .................. 19

3.3.4 VOWEL HARMONY ........................ 20

IV. EXTRACTING ACOUSTIC PROSODIC FEATURES ..... 23

4.1 PITCH PERIOD DETECTION METHODS .......... 23

4.1.1 AUTOCORRELATION METHOD ............... 25

4.1.2 AVERAGE MAGNITUDE DIFFERENCE FUNCTION ....... 28

4.1.3 PARALLEL PROCESSING METHOD ........... 33

4.1.4 SIMPLIFIED INVERSE FILTERING TECHNIQUE ...... 37

4.1.5 CEPSTRUM METHOD ...................... 41

4.1.6 HARMONIC PATTERN MATCHING METHOD ........... 46

4.2 MEASUREMENT OF ENERGY AND DURATION ...... 48

V. PROSODY IN SPEECH RECOGNITION ............. 49

5.1 SPEECH RECOGNITION SYSTEMS .............. 49

5.1.1 ISOLATED WORD RECOGNITION SYSTEMS ........... 52

5.1.2 CONNECTED WORD RECOGNITION SYSTEMS .......... 53

5.2 PROSODIC AIDS TO SPEECH RECOGNITION ..... 58

5.2.1 STRESS ................................ 58

5.2.2 TIMING ................................ 60

5.2.3 INTONATION ........................... 61

Page

5.2.4 VOWEL HARMONY .......................... 63

5.3 PROSODICALLY BASED SPEECH RECOGNITION ............ 64

VI. RESULTS ..................................... 68

6.1 PITCH DETECTION ALGORITHMS ...................... 68

6.1.1 THE ARTIFICIAL SPEECH SIGNAL ................ 68

6.1.2 COMPARISON OF THE ALGORITHMS ................ 69

6.2 VOICE INPUT AND ANALYSIS SYSTEM ................. 72

6.3 SYLLABLE ........................................ 73

6.3.1 SYLLABLE AS A UNIT OF RECOGNITION ........... 73

6.3.2 SYLLABLE SEGMENTATION ....................... 76

6.4 PROSODIC FEATURES ............................... 81

6.4.1 INTONATION ................................. 81

6.4.2 STRESS ...................................... 81

6.4.3 DURATION .................................... 87

6.5 PROSODICALLY AIDED ISOLATED WORD RECOGNITION SYSTEM 89

6.6 SOME STRATEGIES FOR CONTINUOUS SPEECH RECOGNITION . 92

VII. CONCLUSION ................................... 95

7.1 SUGGESTIONS FOR FURTHER WORK .................... 96

REFERENCES ........................................ 97

## LIST OF FIGURES

| | | Page |
|---|---|---|
| FIGURE 2.1 | Anatomical structures involved in speech production | 5 |
| FIGURE 2.2 | General discrete-time model for speech production | 6 |
| FIGURE 2.3 | Model of stages in speech perception | 8 |
| FIGURE 4.1 | Block diagram of the AUTOC pitch detector | 25 |
| FIGURE 4.2 | Example of voiced speech and its autocorrelation function | 27 |
| FIGURE 4.3 | A frame of voiced speech and its AMDF | 30 |
| FIGURE 4.4 | Block diagram of the AMDF pitch detector | 31 |
| FIGURE 4.5 | AMDF pitch extraction logic flowchart | 32 |
| FIGURE 4.6 | Block diagram of parallel processing pitch detector | 33 |
| FIGURE 4.7 | Input signal and the impulse trains generated from the valleys and peaks | 34 |
| FIGURE 4.8 | Operation of the pitch period estimator | 35 |
| FIGURE 4.9 | Flowchart of the voiced-unvoiced decision | 36 |
| FIGURE 4.10 | Block diagram of the SIFT pitch detector | 39 |
| FIGURE 4.11 | Decision algorithm for voiced-unvoiced decision and error correction | 40 |
| FIGURE 4.12 | Block diagram of the CEP pitch detector | 42 |
| FIGURE 4.13 | Cepstrum of the frame in Fig. 4.3 | 43 |
| FIGURE 4.14 | Flowchart of the peak picking algorithm | 45 |
| FIGURE 4.15 | Block diagram of the harmonic pattern matching method pitch detector | 46 |
| FIGURE 5.1 | Dimensions of performance in speech recognition | 50 |

Page

FIGURE 5.2 Block diagram of a word recognition system 52

FIGURE 5.3 Processes involved in recognition and understanding 57

FIGURE 5.4 Tune I and tune II intonation contours 62

FIGURE 5.5 Clause and syntactic phase boundaries on $F_o$ contours 63

FIGURE 5.6 Prosodically guided analysis-by-synthesis system 65

FIGURE 6.1 The artificial speech model 69

FIGURE 6.2 Voice input and analysis system 72

FIGURE 6.3 Frequency of syllables in text 75

FIGURE 6.4 Fundamental frequency and energy curves for the word "BIRLESTIR" 78

FIGURE 6.5 Flowchart of the syllabic nuclei detection algorithm 79

FIGURE 6.6 Fundamental frequency and energy curves for "CEVIR" 80

FIGURE 6.7 Example of Tune I contour 82

FIGURE 6.8 Example of Tune II contour 83

FIGURE 6.9 Second syllable stressed : koNUSma 85

FIGURE 6.10 Last syllable stressed : konusMA 86

FIGURE 6.11 Overall block diagram of the word recognition system 89

# LIST OF TABLES

| | | Page |
|---|---|---|
| TABLE 3.1 | Phonemes of Turkish | 11 |
| TABLE 3.2 | Vowels of Turkish | 12 |
| TABLE 3.3 | Turkish initial syllables | 15 |
| TABLE 3.4 | Turkish middle syllables | 15 |
| TABLE 3.5 | Turkish final syllables | 15 |
| TABLE 3.6 | Frequency of occurance of syllables in Turkish | 16 |
| TABLE 5.1 | The regulations introduced by vowel harmony | 64 |
| TABLE 6.1 | Pitch period errors for the algorithms | 70 |
| TABLE 6.2 | Advantages and disadvantages of recognition units | 74 |
| TABLE 6.3 | Frequency of syllables in text | 75 |
| TABLE 6.4 | Frequency of syllables in certain positions | 75 |
| TABLE 6.5 | Variation of duration, amplitude and pitch with stress | 87 |
| TABLE 6.6 | Duration of syllables | 88 |
| TABLE 6.7 | Classification of the syllables | 90 |
| TABLE 6.8 | Improvements due to prosodic aids to recognition | 91 |

## LIST OF SYMBOLS AND ABBREVIATIONS

| | |
|---|---|
| ASR | Automatic speech recognition |
| AUTOC | Autocorrelation method using clipping |
| $R_x(m)$ | Short-time autocorrelation function |
| $x(n)$ | Sampled speech signal |
| AMDF | Average magnitude difference function |
| $D_n$ | Average magnitude difference function |
| $s(n)$ | Speech samples |
| $u(n)$ | Voice excitation |
| $h(n)$ | Vocal tract impulse response |
| $e(n)$ | Prediction error |
| $\tilde{s}(n)$ | Predicted speech signal |
| $A(z)$ | LPC inverse filter |
| CEP | Cepstral pitch detection method |
| SIFT | Simplified inverse filtering technique pitch detector |
| PPROC | Parallel processing method pitch detector |
| $A_c(p)$ | Harmonic pitch estimator function |
| $G(z)$ | Glottal pulse model |
| $R(z)$ | Radiation model |
| $V(z)$ | Vocal tract model |
| A/D | Analog-to-digital converter |
| LPC | Linear predictive coding |
| DTW | Dynamic time warping |
| LPF | Low-pass filter |

| | |
|---|---|
| E | Energy |
| $F_o$ | Fundamental frequency, inverse of pitch period |
| NN | Nearest neighbor rule |
| KNN | K-nearest neighbor rule |

# I. INTRODUCTION

Speech is the basic device humans use for communication. The information to be transmitted is first encoded at a discrete level according to the rules of the language used by the speaker, and then, through the complex process of physiological speech production, this information is converted to an acoustic signal. This signal, when received by another speaker of the same language, is converted to another discrete sequence and decoded to extract the information transmitted.

Speech production mechanism in humans is better understood than the perception mechanism. There are many mathematical models of speech production which enable the construction of synthesis systems. It is generally accepted that the message is conveyed both locally by spectral features and globally, by a hierarchy of structural features. The relationship of these spectral and structural features to linguistic units and concepts has been thoroughly investigated and fairly well understood.

The perception mechanism, on the other hand, is not fully understood yet. Models that use abstract formalisms instead and physical and physiological correlates are used.

Automatic speech recognition is the process of transforming the acoustic speech waveform into a sequence of discrete representations and assignment of meanings to these sequences by a machine. The first attempts for machine recognition of speech date back some 30 years. The first attempts ordinarily reflected speech production viewpoints, for this was the best understood. These systems could recognize with a fair rate of performance a small vocabulary of words spoken in isolation by a trained speaker. With the advent of mathematical tools in the 70's, systems capable of recognizing larger vocabularies of words independent of the speaker with good rates of performance were built. Recent trends have added successes in recognition of continuous speech such as strings of digits and spoken sentences related to a restricted task domain, and the technology is currently expanding rapidly . What has made this success possible may have been the integration of more speech knowledge into the system. The main trend in some recent systems [21] is making use of all linguistic information at various levels. Among these, the prosodic level is one level which gives cues to the other levels as well, and is easy to extract from the acoustic waveform.

Prosodic features are parts of the way humans encode information in the speech signal. As contrasted to the sounds which occupy short time segments each, prosodic features are of longer duration, and they are imposed on the sounds that follow each other. For this reason, they are sometimes called "suprasegmental features". The main prosodic features are stress, tone, intonation, duration and harmony. These features have functions specific to the language.

Many recognition systems have attempted to use prosodic information in their systems, mainly for the purpose of error detection. When used in this way, prosodic information can help improve the performance of a system, but to fully take the benefit of this information, it should be used. much earlier in the recognition process, at the step of hypothesization. Some strategies for this task were outlined [22],[34] and a part of these were integrated in the Sperry Univac recognition system [35]. There is still much to do in this field. It is agreed by the specialists of this field that prosodic cues to linguistic structures needs further investigation and effort to use this knowledge in real systems [36].

The purpose of this thesis is to investigate the prosodic features of Turkish, develop algorithms for their detection and outline strategies to incorporate this knowledge to automatic recognizers of Turkish speech. For this purpose, algorithms have been developed to detect the main physical correlate of prosodic structures, namely, fundamental frequency of speech. With the help of these algorithms, some linguistic structures of Turkish are investigated and some strategies developed for use in an isolated word recognition system. Other strategies are suggested for use in connected recognition systems of Turkish speech.

The speech production and perception mechanisms in humans will be summarized and two models will be given in Chapter II. In Chapter III, some units and rules of the Turkish linguistic system will be summarized. The concept of prosody will be introduced, and prosodic features will be discussed. Chapter IV presents the description of algorithms to detect fundamental frequency and energy. In Chapter V, after a review of some concepts in ASR (Automatic Speech Recognition), use of prosodic features in recognition systems is discussed. Chapter VI is a presentation of the results and Chapter VII, conclusions and possible areas of future research in this field.

# II. THE SPEECH SIGNAL

Speech signals are composed of a sequence of sounds. These sounds and the transitions between them servé as a representation of information. In processing speech signals to extract information, it would be useful to have knowledge about the production and perception of speech in humans.

## 2.1. SPEECH PRODUCTION

### 2.1.1. Human Natural Speech Production

The accoustical speech waveform is an accoustic pressure wave which originates from the voluntary movements of the human vocal system (Figure 2.1). Speech is the acoustic wave that is radiated from this system when air is expelled from the lungs and the resulting flow of air is perturbed by a constriction somewhere in the vocal tract. During the generation of voiced sounds, the air pushed toward the lips causes the vocal cords to open and close at a rate dependent upon the air pressure in the trachea and the length, thickness and tension of the vocal cords. The greater the tension, the higher the perceived pitch of the voice. The opening between the vocal cords is defined as the glottis. The subglottic air pressure and the time variations in the glottal area determine the glottal volume velocity waveform which defines the driving function to the vocal tract.

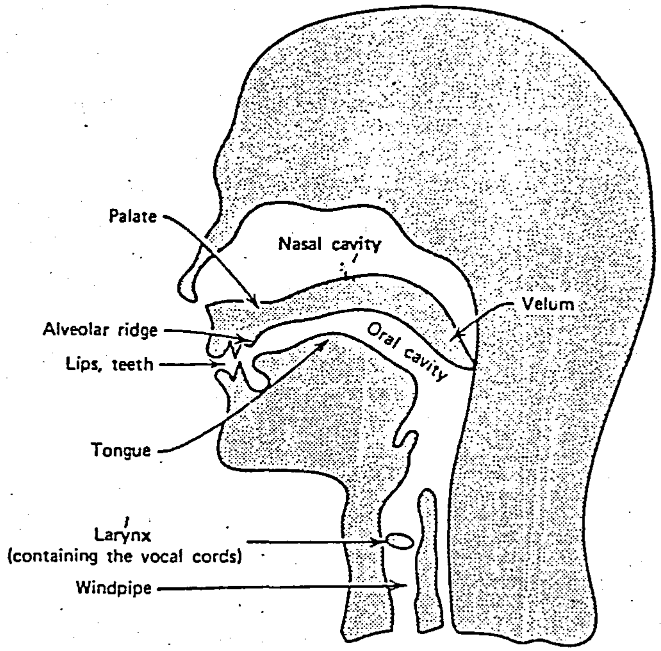


Figure 2.1. Anatomical structures involved in speech production

The vocal tract is a nonuniform accoustic tube which extends from the glottis to the lips and varies in shape as a function of time. The components causing this change are the lips, jaw, tongue, and velum. For example, the cross sectional area of the lip opening can be varied from 0 cm$^2$ to about 20 cm$^2$. The nasal cavity which begins at the velum and ends at the nostrils constitutes an additional accoustic tube for sound transmission used in the generation of the nasal sounds. As sound propagates in the vocal and nasal tracts, its frequency spectrum is shaped by the resonances of these tracts. The resonance frequencies of the vocal tract are called formant frequencies. The formant frequencies depend upon the shape and dimensions of the vocal tract; each shape is characterized by a set of formants. Different sounds are formed by varying the shape of the vocal tract. Thus, the spectral properties of the speech signal vary with time as the vocal tract shape varies.

### 2.1.2. Digital model of the speech signal

Detailed mathematical representations of the accoustics of speech production have been derived [1],[2]. These models mimic the physics of speech production. Here, we will consider a slowly time varying linear system excited by a signal whose basic nature changes from quasiperiodic pulses for voiced speech to random noise for unvoiced speech.

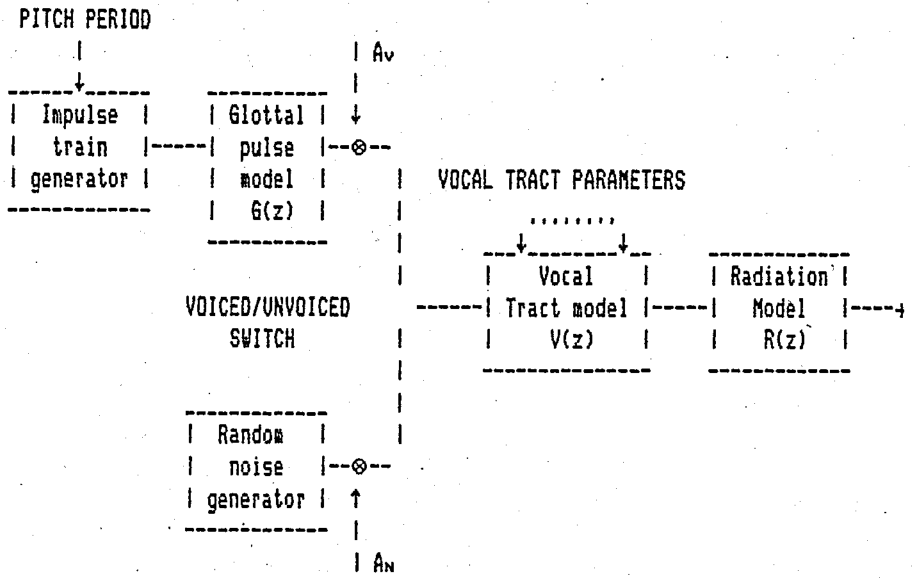


Figure 2.2. General discrete-time model for speech production

In the model shown in Figure 2.2, the changing mode of excitation is modelled by switching between the voiced and unvoiced excitation generators. In the case of voiced speech, the impulse train generator produces a sequence of unit impulses which are spaced by the desired pitch period. This signal in turn excites a linear system whose impulse response g(n) has the desired glottal wave shape. For unvoiced sounds a source of random noise is all that is required.

The effect of the vocal tract is modelled by an all-pole digital filter V(z) which has the formants as its poles. V(z) relates volume velocity at the source to volume velocity at the lips and finally, the radiation model takes care of the radiation at the lips.

The parameters of the model are assumed to be constant over time intervals typically 10-20 ms. long. This model is quite appropriate for sounds whose parameters change slowly with time, namely, vowels. It fails to represent voiced fricatives, for which both sources are involved at the same time. A second limitation is in the representation of nasals, because of the lack of zeros in V(z). Against all its limitations, this is a model that works sufficiently well and is widely used.

## 2.2. SPEECH PERCEPTION

Perception of speech in humans is a complicated process; it involves the perception of accoustic quantities at the ear and transformation of those into learned quantities like phonemes, syllables, words, phrases, sentences, and the association of those with certain meanings. We will not study these individually but will see a model [2] which will give a view of the process of speech perception.

The development of a model for human speech perception is the same problem as the development of an automatic speech recognizer. The proposal for such a model involves the hierarchial structure shown in Figure 2.3. The model is envisioned as a chain of transformations in which each stage acts as an information filter to reduce the dimensionality of the signal. For example, the first three blocks transform an accoustic signal into a succession of words where each word is described by a set of lexical and grammatical features and by prosodic characteristics. Syntax and finally semantic analysis complete the transformations necessary for message understanding. The natures of the transformations are not known, but perceptual experiments suggest certain characteristics of the first two stages.

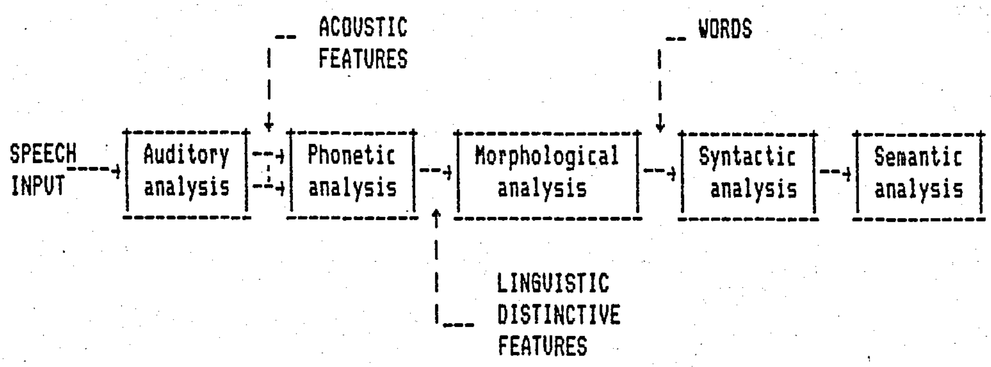


Figure 2.3. Model of stages in speech perception

The peripheral auditory analysis is such that features of the short-time spectrum, i.e. changes in spectral distribution, periodicity (or non-periodicity) and intensity of the input signal are preserved. This is shown by experiments on perception of changes in pitch, formants or intensity of speech and speech-like sounds. That this information is reduced in dimensionality for later processing is supported by experiments which show that consonant perception is influenced only by the rate and direction of the change in formant transitions, and not by their absolute values. Similar perceptions of the direction and rate of change of fundamental frequency have also been observed in other experiments.

The reduction of dimensionality performed in the phonetic analysis is likely to be a feature analysis than one of comparison to a stored reference pattern. This view is supported by data on syllable recognition where features such as manner of production may be correctly perceived while place of production is perceived incorrectly. Similarly, prosodic features may be perceived without discrimination of phonetic factors. Experiments show that some phonematic features can be recognized and produced even before a listener hears a whole syllable.

Experiments also point out that the phonemic analysis window is shorter than average word length. In an experiment with nonsense syllables, it was observed that a man cannot remember sequences longer than 7 to 10 syllables. This fact gives an idea of the size of the time window through which the message is seen by the morphological analysis stage.

On the other side it is clear that a listener does not make seperate decisions about every phoneme in continuous speech. The units with which he operates is likely to correspond to words or phrases. Information handed from the morphological analysis to the syntactic and semantic analysis can, consequently be reduced in dimensionality to this extent; auditory segments need not coincide with the phonemes.

Experiments on recall show that a listener remembers phonemes as a set of features. Therefore, the phonemic information at the output of the phonetic analysis block should be represented by abstract, distinctive features. Several different accoustic features may contain information about one and the same distinctive feature.

# III. THE SOUND SYSTEM

The arrangement of the sounds of speech is governed by the rules of language. The study of these rules and their implications in human communication is the domain of linguistics. In processing speech signals to extract information, it is helpful to have as much knowledge as possible about the way in which information is encoded in the signal, so it will be useful to give a brief review of the classification and arrangement of the sounds of Turkish.

## 3.1. PHONEMES

Most languages can be described in terms of a set of distinctive sounds, or phonemes. The phoneme is an abstract symbol that is used to represent the total collection of sounds that function similarly and do not make meaningful distinctions among themselves within a given language. With this definition, a phoneme encompasses a group of sounds, each called an allophone, that do not cause a change in the meaning of a word when substituted for each other in that word. The study and classification of these sounds is called phonetics. For our purposes it is most appropriate to discuss the acoustic characterization of the various sounds with the place and manner of articulation.

The sounds of Turkish can be broken into phoneme classes as shown in Table 3.1 (For convenience, letters of the alphabet have been used instead of the actual phonemic symbols). The four broad classes of sounds are vowels, diphthongs, semivowels and consonants. Each of these classes may be broken down into sub-classes, which are related to the manner and place of articulation of the sound within the vocal tract.

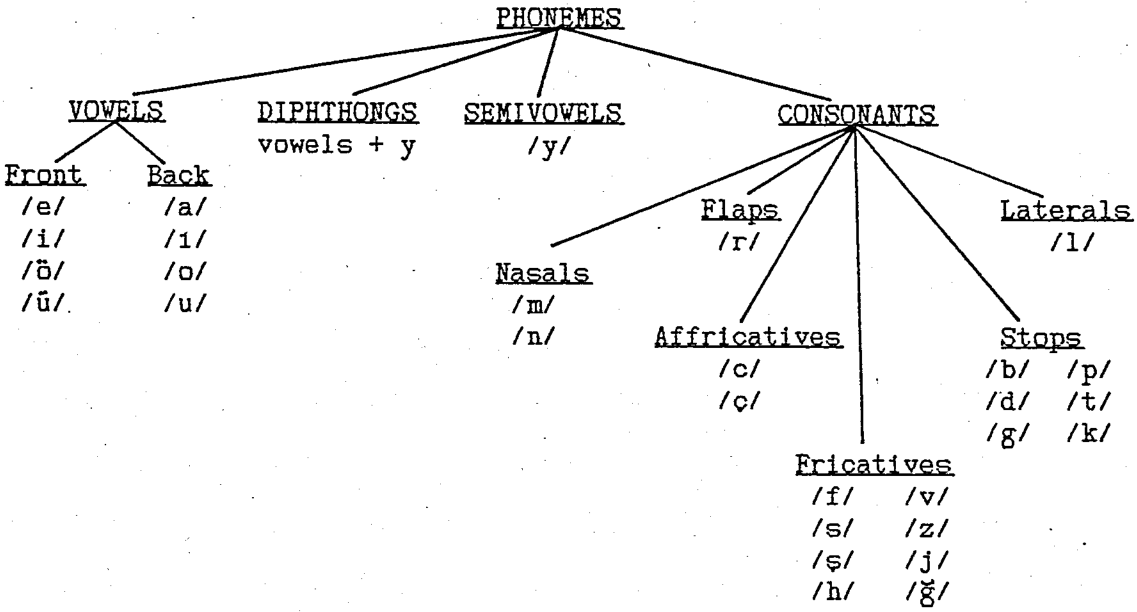


Table 3.1 Phonemes of Turkish

### 3.1.1. Vowels

Vowels are produced by exciting the vocal tract with quasi-periodic pulses of air caused by the vibration of the vocal cords. The position of the tongue, jaw, and lips changes the cross sectional area of the vocal tract, which, in turn determines the resonant frequencies of the tract (formants), and thus the sound that is produced.

In Turkish, vowels are classified according to (see Table 3.2):

- Angle of the jaws : wide (a,e,o,ö) or close (ı,i,u,ü)
- Shape of the lips : rounded (o,ö,u,ü) or unrounded(a,e,ı,i)
- Position of the tongue : back (a,ı,o,u) or front (e,i,ö,ü)

| | UNROUNDED | | ROUNDED | |
|---|---|---|---|---|
| | WIDE | CLOSE | WIDE | CLOSE |
| BACK | A | I | O | U |
| FRONT | E | İ | Ö | Ü |

Table 3.2 Vowels of Turkish

### 3.1.2. Diphthongs

A dipthong is a gliding monosyllabic speech item that starts at or near the articulary position for one vowel and moves to the position for another. The dipthongs are produced by varying the vocal tract smoothly between vowel configurations. In Turkish, vowels gliding to /y/ can be classified as dipthongs [4]. (e.g., *oy, bay*)

### 3.1.3. Semivowels

Semivowels are transitional, vowel-like sounds which are characterized by a gliding transition in vocal tract area function between adjacent phonemes. /y/ is a semivowel of Turkish and its nature is influenced by where it occurs; it is voiced in the beginning and in the middle of words, and semi-voiced in the end. (e.g., *yine,iyi,say*)

### 3.1.4. Nasals

The nasal consonants /m/ and /n/ are produced with glottal excitation and the vocal tract constricted at a point along the oral paasageway with the velum lowered so that the nasal tract is coupled.The effect of this coupling is to produce zeros of the transfer function of the vocal tract. Furthermore, nasal consonants and vowels preceding or following nasal consonants are characterized by resonances which are spectrally broader, due to the fact that the nasal tract has a different area function. The two nasal consonants are distinguished by the place along the vocal tract at which a constriction is made. For /m/, the constriction is at the lips, and for /n/, the constriction is at the back of the teeth.

### 3.1.5. Fricatives

Fricatives are produced by exciting the vocal tract by a steady air flow which becomes turbulent in the region of a constriction. For the voiced fricatives, the vocal cords are also vibrating; two excitation sources are involved in their production, thus, the spectra of voiced fricatives displays two distinct components. /v/,/z/,/j/,/ğ/ are the voiced fricatives and /f/,/s/,/ş/,/h/ are the voiceless fricatives of Turkish. The location of the constriction serves to determine which fricative sound is produced. The places of constriction are, lips and teeth for /v/ and /f/, near the middle of the vocal tract for /z/ and /s/, alveloar ridge and the palate for /j/ and /ş/, back of the tongue for /ğ/, and the back of the vocal tract for /h/.

### 3.1.6. Stops

Stops are transient sounds produced by building up pressure behind a constriction in the oral tract, and suddenly releasing the pressure. During the period when there is a total constriction in the tract there is no sound radiated from the lips. In voiced fricatives, the vocal cords are able to vibrate during this closure and a small amount of low frequency enrgy is radiated through the walls of the throat. For voiceless fricatives, the vocal cords do not vibrate, and following the period of closure, there is a brief interval of friction followed by a period of aspiration. Since the stop sounds are dynamical in nature, their properties are highly influenced by the vowel which follows the stop consonant. /b/,/d/ and /g/ the are voiced stops and /p/,/t/ and/k/ are the voiceless stops of Turkish. The places of constriction are, lips for /b/ and /p/, alveloar ridge for /d/ and /t/, and the velum for /g/ and /k/.

### 3.1.7. Affricatives

Affricatives can be modelled as the concatenation of a stop and a fricative. There are two affricatives in Turkish, /c/, voiced, and /ç/, voiceless. The place of constriction for both is at the middle of the oral tract.

### 3.1.8. Flaps

/r/ is produced by flapping the tongue to the front of the oral cavity. It is voiced, and the place of constriction is at the back of the teeth.

### 3.1.9. Laterals

The voiced lateral /l/ is produced by flow of air on both sides of the tongue. The place of constriction is at the back of the teeth.

## 3.2. SYLLABLES

In Turkish, we can roughly define the syllable as a unit containing one vowel only, which may be preceded and followed by a number of consonant units. In Turkish, all combinations of 0 to 1 consonant units preceding the vowel unit with 0 to 2 following are found. However, the types of consonant combinations are restricted to the order of their appearance in a word; initial,middle, or final. (V:vowel, C:consonant):

| | | |
|---|---|---|
| V | : | *O, o-(yun), u-(yum)* |
| CV | : | *bu, su, ge-(lin)* |
| VC | : | *al, at, in-(di)* |
| VCC | : | *alt, ilk* |
| CVC | : | *gir, tut* |
| CVCC | : | *kurt, sarp* |

Table 3.3. Turkish initial syllables

| | | |
|---|---|---|
| VC | : | *(bil-di)-(ğ)in-(den)* |
| CV | : | *(se)-vi-ne-mi-yo-(rum)* |
| CVC | : | *(se)-vin-(ce)* |
| CVCC | : | *(u)-çurt-(ma)* |

Table 3.4. Turkish middle syllables

| | | |
|---|---|---|
| V | : | *(gel-di)-ği* |
| CV | : | *(ol)-du, (git)-ti* |
| CVC | : | *(ge)-len, (ya)-tak* |
| CVCC | : | *(u)-tanç, (se)-vinç* |

Table 3.5. Turkish final syllables

There are other types of syllables, usually in loan words, of the form CCV, CCVC, or CCCV, but these are rather rare, as indicated by the frequency of occurance data in Table 3.6. The data in this table are the results of a study on a Turkish text consisting of 59000 syllables [15].

| Syllable type | Frequency of occurance |
|---|---|
| V | 4.950 % |
| VC | 4.201 % |
| CV | 52.912 % |
| VCC | 0.124 % |
| CVC | 37.144 % |
| CVCC | 0.502 % |
| CCV | 0.070 % |
| CCVC | 0.097 % |
| CCCV | 0.002 % |

Table 3.6. Frequency of occurance of syllables in Turkish

In the first syllable of a word, any of the 8 vowels may appear, while in the second syllable, 4 of them (o,ö,(u,ü)) are ruled out because of vowel harmony. In the suffixes, the number of vowels which may appear further reduces to two groups within which the vowel may be predicted (with the exception of the suffix *-yor*). If consonants are considered, in the initial position of a syllable, the number is 20, while in the final position 5 are ruled out (b,c,d,g and j). In the suffixes, these numbers further reduce to give 15 consonants in the initial position and 11 in the final position. It has been observed that in the phonemic represantation of polysyllabic words, there is a redundancy of about 50 % [4].

In many languages, syllable division is not uniquely defined. In Turkish, rules for syllable division are clearly set. In general, there will be as many syllables as vowels in a word. For syllable division in polysyllabic words, the C unit is taken with the following V, e.g. CV-CVC. Where two Cs come together, the syllable division comes between them, e.g. CVC-CVC and when three Cs come together, the division comes between the second and third C of the group, e.g. CVCC-CVC.

## 3.3. PROSODY

Although much of the message in speech is conveyed by the segmental phonemes, additional information is carried by the suprasegmental phonemes. Prosodic features, or, suprasegmental phonemes are properties of articulation that encompass more than one phoneme. Duration, stress, tone, intonation and harmony are the prosodic features used in Turkish. The physical parameters of the speech wave which signal the prosody of an utterance are the durations and intensities of the syllables, and the fundamental frequency contours.

### 3.3.1. Duration

Each sound has a certain average duration, but this may change according to the environment of that sound. The consonant following a vowel influences the vowel duration, with voiced consonants in particular causing a lengthening of their preceding vowels. In Turkish, (g) is a special case: It is considered not as a sound but as a lengthening of the vowel it is next to, e.g. the duration of /a/ is short in (*ak,yak*) and much longer in (*dağ,yağ*).

Another factor which must be taken into account when predicting phonetic durations is the stress of the syllable. A stressed /i/ is longer than an unstressed /i/. Similarly, at fast speaking rates, stressed syllables will be somewhat shorter, and unstressed syllables will be substantially shortened.

### 3.3.2. Stress

Stress is associated with the relative prominence of a syllable in speech. In Turkish, stress on base forms is usually on the final syllable with quite clearly defined exceptions [14],[16]-[20].

One common class of exceptions is the names of places. When a common word which has stress on the last syllable is used as a place name, the location of the stress is changed, the new location being usually the initial syllable:

| | |
|---|---|
| *ulus* | *Ulus* |
| *kumla* | *Kumla* |

Another class of exceptions is the loan-words. In this class, stress is usually placed on the middle syllables:

| | |
|---|---|
| *abone* | *sinema* |
| *sempati* | *çikolata* |

With inflected forms, stress usually stays in the final position,

| | |
|---|---|
| *somun* | *somunlarınız* |
| *somunlar* | *somunlarınızda* |
| *somunları* | |

with the exception of some suffixes (*-ca*, *-yor*, *-mi*, etc), which cause the placement of stress on the syllable preceding them:

| | |
|---|---|
| *yapacak* | *yapmayacak* |
| *yapıyor* | *yapmıyor* |

*Note how stress becomes a distinguishing feature for the two homonyms "konuşma*" and "*konuşma*".

The physical correlates of stress are duration, intensity and fundamental frequency. In a sentence, the syllable with the longest duration and highest fundamental frequency is perceived as stressed.

### 3.3.3. Tone and intonation

Tone refers to the pitch variation over a single syllable. In some languages called tone languages, there are several pitch levels or patterns, and the lexical meaning of a word changes according to which tone is used. For example, certain words which are represented by identical phonemic strings will have different meanings depending upon whether they are spoken with a rising or falling pitch.

There are 12 patterns of pitch variation which usually appear in one-word sentences [4]. These are:

-decrease from low pitch

-decrease from high pitch

-decrease from middle pitch

-decrease with pauses

-increase with pauses

-increase

-increase with large slope

-decrease/increase

-increase/decrease

-decrease/increase/decrease

-neutral tone

-long neutral tone

These patterns are usually associated with the mood of the speaker, such as interest, indifference, anger, confusion, impatience, etc. One can give a positive meaning to a negative sentence or express a question with a declarative sentence using the appropriate tones.

Intonation is the pitch variation over the whole sentence or a part of it. In all languages, intonation has a grammatical or syntactic function. The distinction between a statement and a question, between a question and a command and so on, can be signalled by a difference of pitch contour. The grammatical function of intonation is an important part of language structure, part of the common knowledge shared by speakers of the language.

In Turkish, for declarative sentences and YES/NO questions

(questions with the suffix *-mı*), a falling pitch at the end of the sentence is observed:

*Bu kitabı okudum.* ↓

*Denize girdin mi?*↓

While for questions with interrogative words (*kim, ne, hangi*, etc.) and and the phrases (*değil mi?, öyle mi?*), pitch increases at the end.

*Onu sevdiniz, değil mi?*↑

*Bunu kim getirdi?*↑

Another function of intonation is connected much more with individual psychology. For example, a decreasing pitch contour at the end of the sentence indicates the end of conversation, while an increasing pitch contour at the end of a sentence would mean that more is going to be said.

*Sabahleyin (↓ okula ) gittim.*

*Sabahleyin okula ( gittim ↑). Derslere ( girdim ↑)...*

Intonation patterns, like stresses, serve also to designate certain syllables as prominent relative to others. This prominence may or may not coincide with that of the stresses. The interference of intonational phenomena and stress may be confusing, but it seems to bring one very interesting peculiarity. Turkish sentence rythm seems to assign approximately equal time to each syllable as contrasted to the Germanic system which seems to assign equal time to stress groups clustered about each successive strong stress [5].

### 3.3.4 Vowel harmony

Any phonemically based analysis of Turkish recognizes eight vowels: /a,e,ı,i,o,ö,u,ü/. Any one of these vowels may occur in monosyllabic words. In words of more than one syllable, however, there are systematic restrictions on the co-occurance of the several vowel phonemes. Thus, in words of native Turkish origin, front vowels, /i,ü,e,ö/, and back vowels, /ı,u,a,o/, do not occur together. And then, there are the rounded vowels /o,ö,u,ü/, and unrounded vowels /a,e,ı,i/. If a word contains an unrounded vowel in its first syllable, it cannot contain rounded vowels in its other syllables (However, there is a clearly defined class of exceptions to this case, e.g. *havlu*). Moreover, the phonemes /o/ and /ö/ occur generally only in the first syllable of a word (with the exception of the suffix *-yor*). This is generally called "vowel harmony" in Turkish. A phonemic represantation of polysyllabic words is therefore highly redundant, since it represents each vowel in the structure as a selection from eight contrasting units, whereas all but two of the eight vowel phonemes are excluded from occurance by the occurance of any other given vowel phoneme of the word.

A different approach to this problem is by prosodic analysis which introduces two binary prosodic contrasts of front/back and rounding/non-rounding, and admits only two contrasting segmental phonematic units, high/low. This way, a much more economic and satisfying description of the language is obtained-one based on the patterns actually operative in the language [7].

Let us clearly define the prosodies:

F:B prosody characterizing words having front/back vowels and consonants with palatalization/without palatalization.

R:N prosody where there is lip labiality (lip ronding) throughout the articulation of the whole syllable, e.g. *kol* /there is absence of labiality thoroughout the whole syllable or where labiality is initial or final only, e.g. *bal*.

A two-term vowel system is set up by defining: Δ (denoting openness) and Γ (denoting closeness). Now rules can be defined as:

F,B and U prosodies operate over the whole word with Δ or Γ, e.g.

| | | | |
|---|---|---|---|
| *bekletmek* | UF: CΔC-CΔC-CΔC | *istek* | UF: ΓC-CΔC |
| *azalmak* | UB: Δ-CΔC-CΔC | *ılınmak* | UB: Γ-CΓC-CΔC |

R prosody operates with Γ in all syllables of the word but with Δ only in the first syllable, e.g.

| | | | |
|---|---|---|---|
| *yolumuz* | RB: CΔ-CΓ-CΓC | *yollarımız* | RB: CΔC-CΔ-CΓ-CΓC |
| *önü* | RF: Δ-CΓ | *önünden* | RF: Δ-CΓC-CΔC |

Another advantage of prosodic representation is its use in predicting the form of suffixes. There are six type-2 prosodies, and these and the type-1 prosodies stated above enable one to handle all suffixed forms of the word [7].

The prosodic approach, by rejecting the phonemicist view that the phonology of any language is a uniform system to be analyzed independently of its grammar, considers the phonology of a language as a set of subsystems, each relevant for different phonological structures. The segmentation in prosodic analysis reveals a horizontal type of segmentation which has the advantage of preserving the syllabic pattern of the language analyzed. Turkish seems to exhibit such a structure that by applying prosodic analysis, both clarity and economy of phonological elements is achieved in its representation. For this reason, it has repeatedly been an example for phonologists adopting prosodic analysis [7],[13].

# IV. EXTRACTING ACOUSTIC PROSODIC FEATURES

In this chapter tools developed to measure accoustic prosodic parameters like pitch, energy, and duration will be discussed. Among those, pitch is the most difficult parameter to measure, and many methods have been developed for its extraction from the accoustic speech waveform. Measurement of energy and duration is a trivial task, but rather sophisticated approaches must be used to incorporate these findings to explore the underlying prosodic structures.

## 4.1. PITCH PERIOD DETECTION METHODS

As discussed in the previous chapter, the speech waveform can be modelled as the response of the vocal tract filter to a source which is a periodic sequence of pulses during voiced segments or a random noise during unvoiced segments. The periodic pulses occur as a consequence of the opening and closing of the glottis, and the frequency of the periodicity is often referred to as the pitch.

Accurate and reliable measurement of the pitch period of a speech signal from the accoustic pressure waveform is often difficult for several reasons:

-The glottal excitation waveform is not a perfect train of periodic pulses. Measuring the period of a speech waveform, which varies both in period and the detailed structure of the waveform within a period can be quite difficult.

-In some instances the formants of the vocal tract can alter the structure of the glottal waveform so that the actual pitch period is difficult to detect.

-Distinguishing between unvoiced speech and low-level voiced speech, and the detection of the pitch period during transitions between voiced and unvoiced sections is often hard.

A pitch detector is a device which makes a voiced-unvoiced decision, and during periods of voiced speech, provides a measurement of the pitch period. As a result of the numerous difficulties in pitch measurements, many pitch detection methods have been developed. Some methods which have been used in this study will be discussed here.

The usual realization of a pitch detector may be considered to be consisting of three main blocks which are passed through successively:

-*the preprocessor*

-*the basic extractor*

-*the postprocessor*

The function of the preprocessor is data reduction in order to increase the ease of pitch extraction. Some examples of preprocessing are computation of the AMDF, computation of the cepstrum, etc. The basic extractor operates on this altered signal to convert it into a sequence of pitch estimates. The postprocessor is a block which performs the tasks of error detection and correction, smoothing of an obtained contour, time-to-frequency conversion and display of the parameters.

### 4.1.1. Autocorrelation Method

One of the difficulties in pitch period estimation is the effect of the formant structure on measurements related to the periodicity of the waveform. Thus, it is desired to remove the spectral shaping in the waveform due to the formants. A way to achieve this spectral flattening is using centre clipping by which signal values below the clipping level are set to zero and those above the clipping level are offset by the clipping level. If the clipping level is appropriately chosen, most of the waveform structure due to the formants can be eliminated. AUTOC [44] uses this approach combined with autocorrelation analysis. (Figure 4.1)

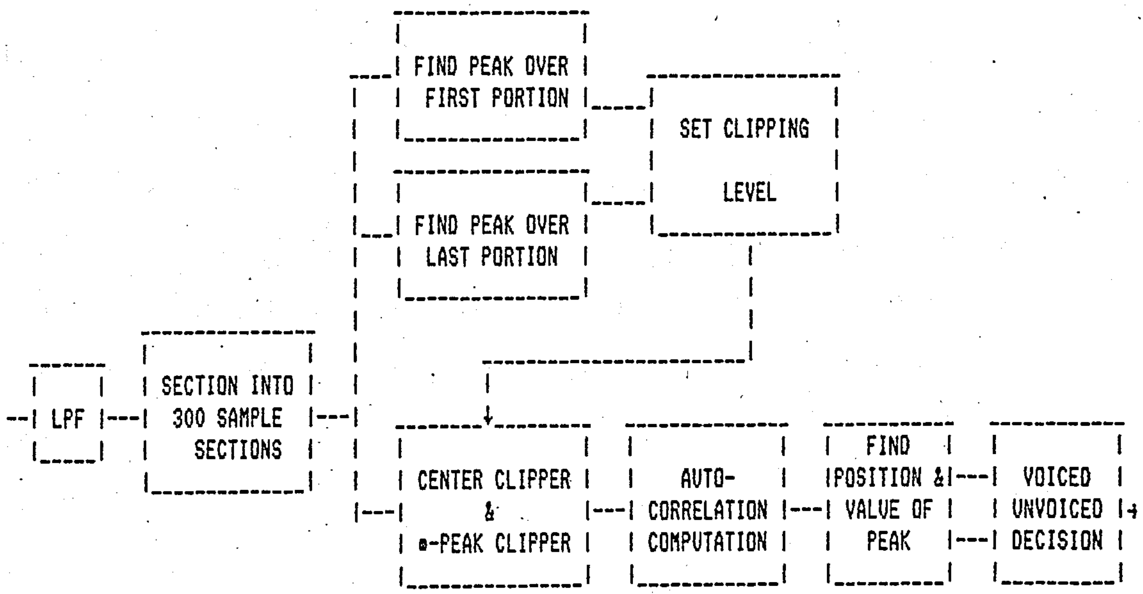

FIND PEAK OVER FIRST PORTION
SET CLIPPING LEVEL
FIND PEAK OVER LAST PORTION
LPF
SECTION INTO 300 SAMPLE SECTIONS
CENTER CLIPPER & ●-PEAK CLIPPER
AUTO- CORRELATION COMPUTATION
FIND POSITION & VALUE OF PEAK
VOICED UNVOICED DECISION

Figure 4.1. Block diagram of the AUTOC pitch detector

The analog speech signal is sampled at a 8 kHz. sampling rate using a 12-bit A/D converter. The digital signal is low-pass filtered to a bandwidth of 900 Hz using a 99-point FIR filter. The output of the filter is then sectioned into 300 samples overlapping by 100 samples for processing. Each section of 300 samples is called a frame.

The first stage of processing is the computation of the clipping level. Because of the wide dynamic range of speech, the clipping level must be carefully chosen so as to prevent loss of information when the waveform is either rising or falling in amplitude within a frame. Such cases occur when voicing is just beginning or ending, as well as during voicing transitions, e.g., from a vowel to a voiced fricative, or a nasal. For the selection of $C_L$, the clipping level, the first and third 100 samples of the frame is searched for maximum absolute peak levels. The clipping level is then set as 80 percent of the smaller of these two levels.

Following the determination of the clipping level, the speech section is then both center clipped, and infinite peak clipped, resulting in a signal which assumes one of three possible values; +1 if the sample exceeds the positive clipping level, -1 if the sample falls below the negative clipping level, and 0 otherwise. The use of infinite peak clipping greatly reduces the computational complexity of the autocorrelation measurement, because no multiplications are required in the computation.

The next stage in processing is the autocorrelation computation. The short-time autocorrelation function of the 300-samples frame is defined as:

$$R_x(m) = \sum_{n=0}^{299-m} x(n)x(n+m) \qquad m=M_i, M_{i+1}, \ldots, M_f \qquad (4.1)$$

where $M_i$ is the initial lag and $M_f$ is the final lag for which the autocorrelation function is computed. For the frequency range of 100 to 500 Hz, these values are 16 and 80 respectively. Additionally, $R_x(0)$ is computed for the normalization of the autocorrelation function.

Figure 4.2 shows an example of an anlysis frame, the infinite peak and center clipped version, and the short-time autocorrelation function. For this example the pitch period of the section is at 64 samples, which corresponds to 125 Hz at 8 kHz sampling rate.

In the computation of the autocorrelation function (Eq'n 4.1), it is assumed that samples outside the current frame are assumed to be zero. This effectively weights the autocorrelation function by a linear taper which starts at 1 at m=0 and goes to 0 at m=300. That property is desired, because it enhances the peak at the pitch period with respect

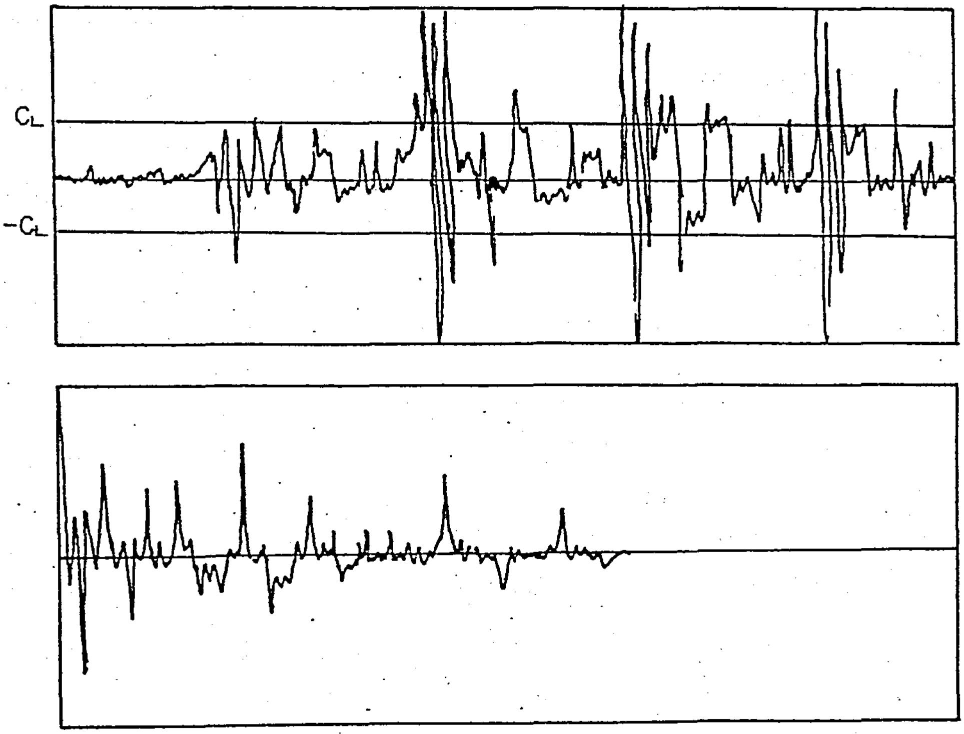


Figure 4.2. Example of voiced speech and its autocorrelation function

to peaks at multiples of the pitch period, thereby reducing the possibility of doubling or tripling the pitch period estimate.

For voiced-unvoiced decision, the autocorrelation peak is compared to the energy, $R_x(0)$. If this ratio exceeds a voiced-unvoiced threshold of around 30 %, The frame is classified as voiced and the pitch period is the position of the autocorrelation peak. If the peak value falls below the threshold, the interval is classified as unvoiced.

The decision for the current interval is modified by the decisions for the preceding and succeeding intervals. If these are both voiced (unvoiced), then the current interval is forced to be declared voiced (unvoiced).

### 4.1.2 Average Magnitude Difference Function

The AMDF (average magnitude difference function) is a variation on autocorrelation analysis where, instead of correlating the input speech at various delays, a difference signal is formed between the delayed speech and the original and, at each delay,the absolute magnitude of the difference is taken. The difference function is always zero at delay = 0 and exhibits deep nulls at delays corresponding to the pitch period of voiced sounds.

An approximate expression that provides a relationship between the AMDF and the autocorrelation function will be developed. The AMDF for a sequence of samples {x(m)} is defined by the relation

$$D_n \equiv \sum_{k=n}^{N-1} |x(k)-x(k-n)| \qquad n = -(N-1),\ldots\ldots,+(N-1) \qquad (4.2)$$

We can approximate $D_n$ in the form

$$D_n \equiv \sum_k |x(k)-x(k-n)| \simeq \beta_n \left( \sum_k (x(k)-x(k-n))^2 \right)^{1/2} \tag{4.3}$$

where the coefficient $\beta_n$ is a scale factor. By expanding the squared term in braces under the square root sign in (4.3) we can express $D_n$ in the form,

$$D_n \simeq \beta_n \left( \sum_k x(k)^2 + \sum_k x(k-n)^2 - 2\sum_k x(k)x(k-n) \right)^{1/2} . \tag{4.4}$$

The third sum in the braces can be identified as $-2R_n$. Assuming that the sequence {x} corresponds to a stationary process, it is evident that the first two sums are simply the autocorrelation function evaluated at n=0. Then, we can rewrite $D_n$ as

$$D_n \simeq \beta_n [ 2(R_0 - R_n) ]^{1/2} . \tag{4.5}$$

The properties of the AMDF are accurately characterized by (Eq'n 4.5). Figure 4.3 shows a frame of speech samples and their AMDF, which is seen to be zero at zero delay and varies as the square root of the autocorrelation function that has been negated and dc shifted by $R_0$. Nulls will appear in $D_n$ at those points where $R_n$ is large compared with $R_0$. This occurs when the sequence {x} is taken from a voiced speech sound. The separation of the nulls is a direct measure of the pitch period.

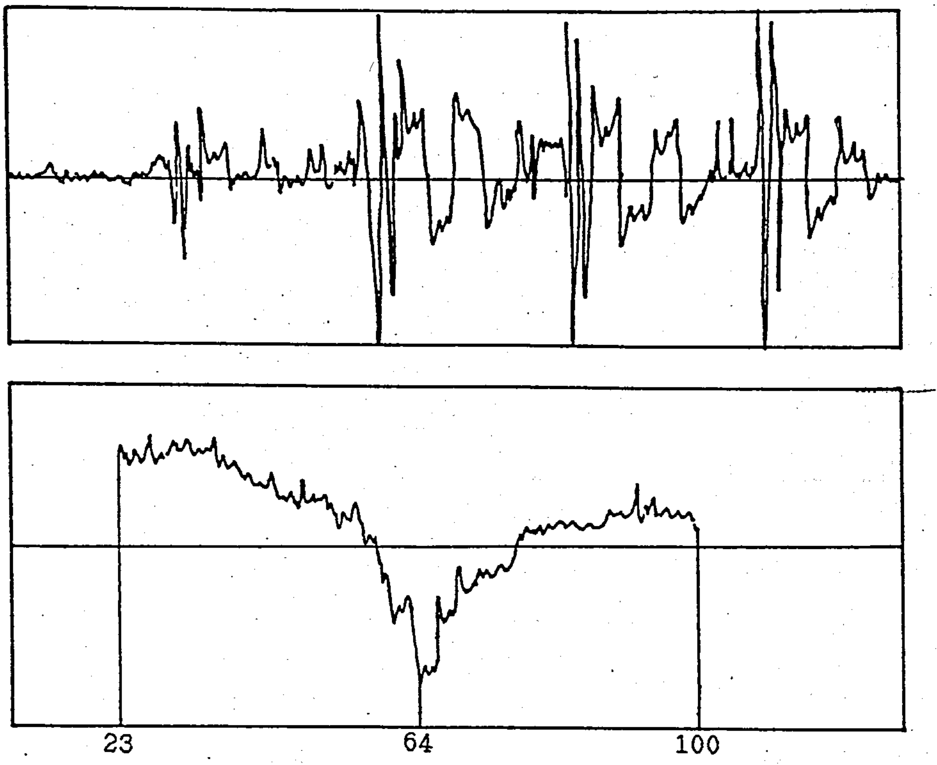


Figure 4.3. A frame of voiced speech and its AMDF

The block diagram of a method of pitch detection using the AMDF is given in Figure 4.4 [45]. The first stage in processing is A/D conversion, followed by lowpass filtering. At the output of the filter, the input speech samples are divided into frames consisting of 300 samples that overlap by 200 samples. The samples are then summed up to find the energy of the section. If this value exceeds a fixed threshold, the frame is classified as voiced. The next step is the computation of the AMDF in the range of the pitch period.

At this step, the effect of the formants is still inherent in the spectral envelope of the signal because no preprocessing has been done prior to computation of the AMDF. For this reason,decision logic and prior knowledge of voicing are used along with the function itself to help make the pitch decision more reliable. Figure 4.5 shows the set of logical rules developed for extraction of pitch information from the

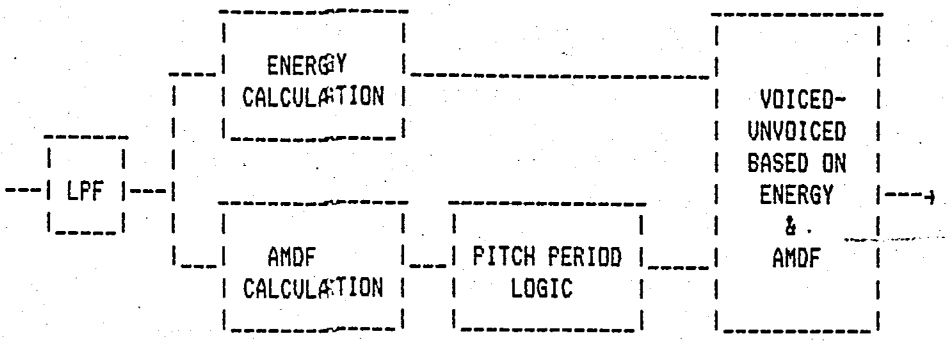


Figure 4.4. Block diagram of the AMDF pitch detector

AMDF. There are five seperate logic paths, each of which are selected, based on the three most recent voiced/unvoiced (VUV) decisions.

In path A, the present VUV decision is unvoiced and the logic asks whether this decision should be changed to voiced. A change is justified by the presence of a strong periodic waveform within the interval.

In path B, the present VUV decision is voiced. However, this decision can be changed to unvoiced, if either the maximum AMDF value is not sufficiently strong or the ratio of the maximum to minimum AMDF value is below a certain threshold.

In path C, the nth and (n-1)th VUV decisions are voiced but the (n-2)th interval was unvoiced. This is an indication of the onset of voicing; the pitch extractor changes to voiced and chooses the minimum value of the AMDF as the pitch.

In path D, voicing is extended an additional frame when the VUV decision indicates unvoicing after an extended period of voicing.

Path E is the normal path for sustained voicing and uses a feature for pitch tracking in a window of 12 samples about the last measured pitch period. The logic will change to the nontracking position if the amplitude of the minimum outside the tracking range is less than 1/2 of the tracking amplitude minimum. For higher frequencies, more nulls are present in the AMDF, so a null outside the tracking window is required to be less than 1/8 the minimum in the tracking window to be chosen. There is also a path for changing the VUV decision from voiced to unvoiced, and for extending the previous pitch value.

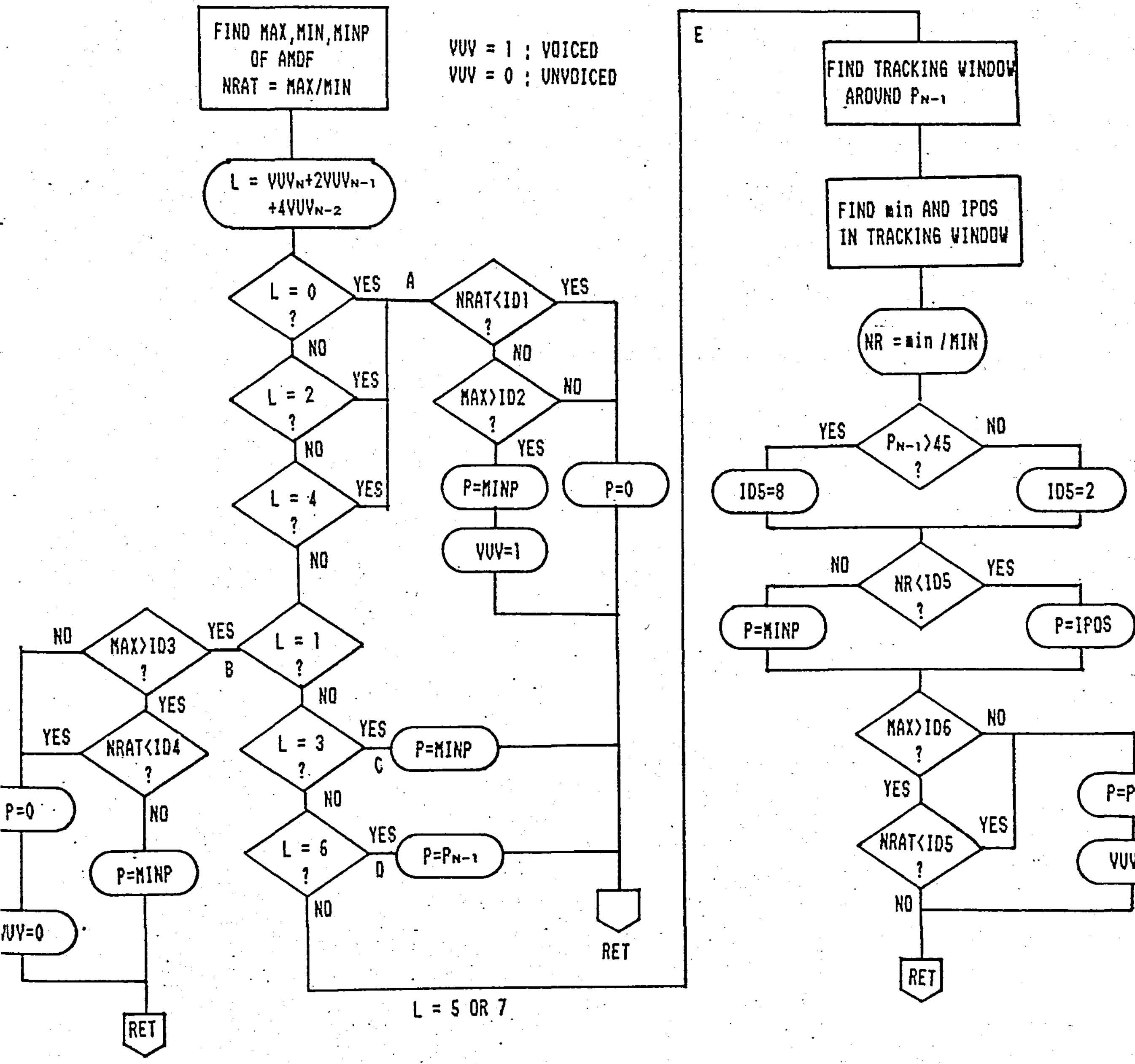


Figure 4.5. AMDF pitch extraction logic flow chart

### 4.1.3 Parallel Processing Method

The basic idea in parallel processing is that an improvement in accuracy can be obtained by combining the outputs of more than one elementary pitch period estimators. The speech signal is processed so as to create a number of impulse trains which retain the periodicity of the original signal and discard features which are irrelevant to the pitch detection process. A pitch estimate is obtained using a simple pitch detector from each of the impulse trains. These estimates are then logically combined to infer the period of the speech waveform.

The block diagram of the pitch detector in [46] is given in Figure 4.6. After A/D conversion, the speech is lowpass filtered with a cutoff of about 900 Hz. Following the filtering, local minima and maxima are located, and from their locations and amplitudes, several impulse trains are derived from the filtered signal (Figure 4.7).

The impulse trains $p_i$ are generated at the location of peaks (local maxima) and the impulse trains $v_i$ are generated at the locations of valleys (local minima). These are defined as [1]:

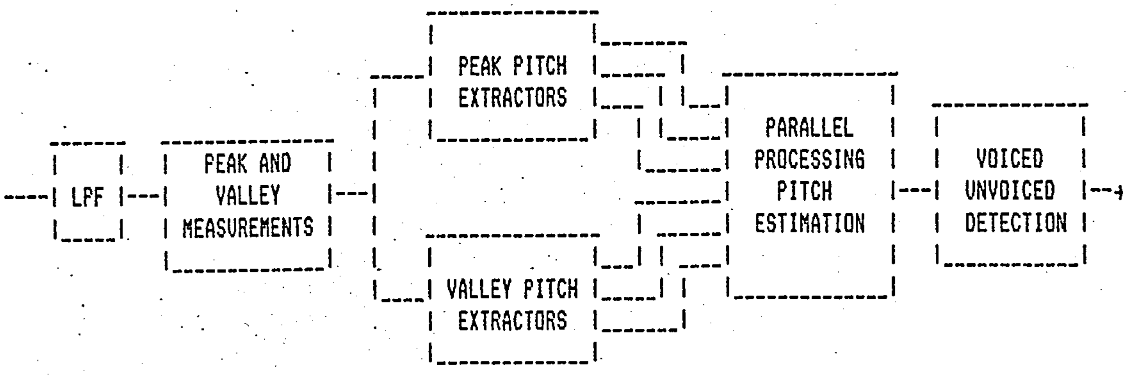

LPF
PEAK AND VALLEY MEASUREMENTS
PEAK PITCH EXTRACTORS
VALLEY PITCH EXTRACTORS
PARALLEL PROCESSING PITCH ESTIMATION
VOICED UNVOICED DETECTION

Figure 4.6. Block diagram of the parallel processing pitch detector

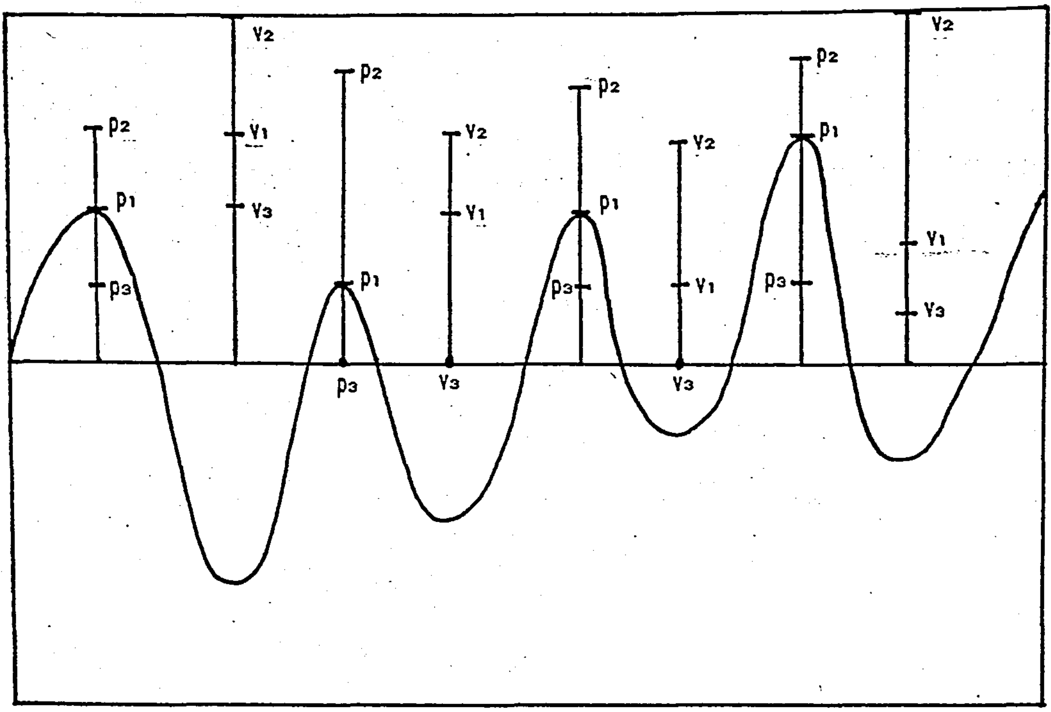


Figure 4.7. Input signal and the impulse trains generated from the peaks and valleys

-$p_1(n)$ : An impulse equal to the peak amplitude.

-$p_2(n)$ : An impulse equal to the difference between the peak amplitude and the preceding valley amplitude.

-$p_3(n)$ : An impulse equal to the difference between the peak amplitude and the preceding peak amplitude.

-$v_1(n)$ : An impulse equal to the negative of the amplitude of a valley.

-$v_2(n)$ : An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding peak.

-$v_3(n)$ : An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding valley.

The six sets of pulse trains are applied to the six individual pitch detectors. The operation of the detector is illustrated in Figure 4.8. Following each detected pulse, there is a blanking interval (during which no pulses can be detected), followed by an exponential decay. Whenever a pulse 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 the pitch period. The result is a kind of smoothing of the impulse train, producing a quasi-periodic sequence of pulses as shown in Figure 4.8. The length of each inter-pulse interval is an estimate of the pitch period.

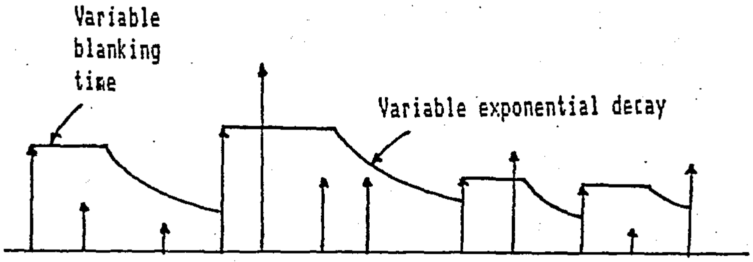


Figure 4.8. Operation of the pitch period estimator

This technique is applied to each of the six impulse trains producing six estimates of the pitch period. These six estimates are combined with two of the most recent estimates of each of the six pitch detectors. These estimates are then combined and the one with the most coincidences within a specified tolerance is declared the pitch period at that time.

For unvoiced speech, there is a distinct lack of consistency among the estimates. For each analysis interval, the coincidence count of the pitch period estimate with the greatest number of occurances is tested against a certain threshold. The value of this threshold is lowered if the preceding frame is voiced. A final voiced-unvoiced decision is given on the basis of the VUV decisions of the preceding and following frames - Isolated voiced and unvoiced frames are not allowed. These principles are illustrated in the flowchart in Figure 4.9.

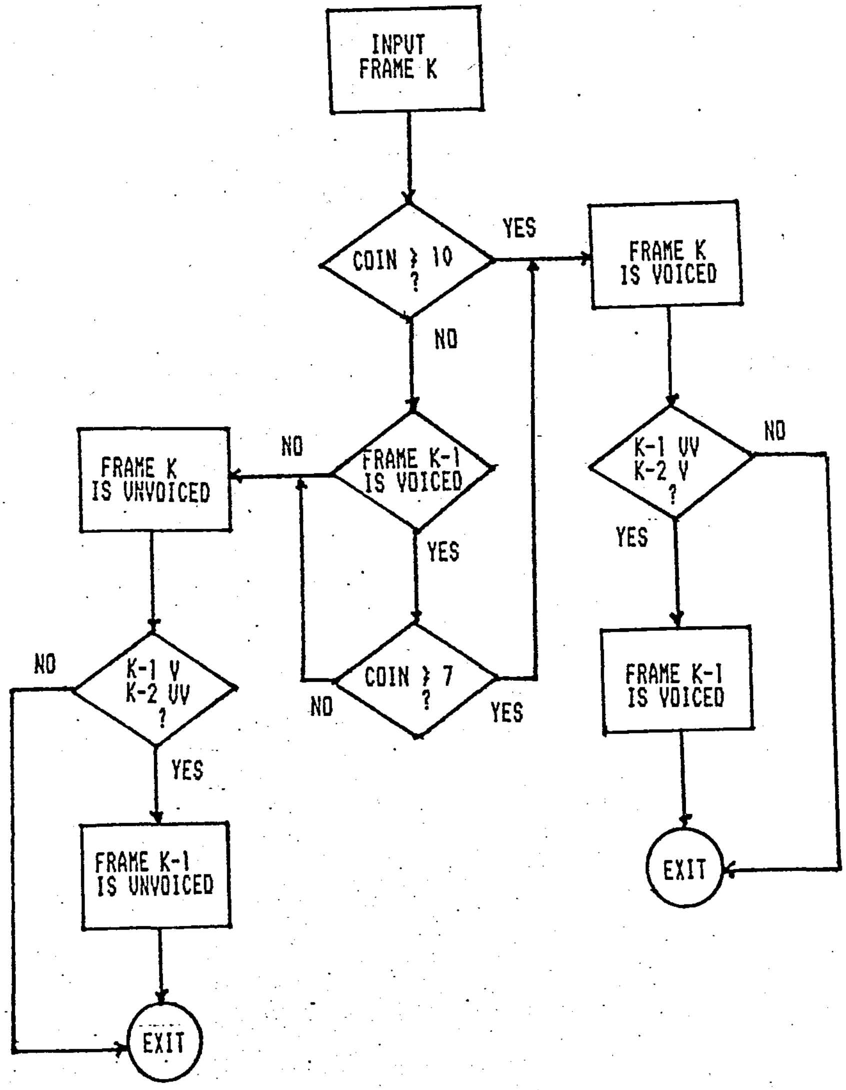


Figure 4.9. Flowchart of the voiced-unvoiced decision

### 4.1.4 Simplified Inverse Filtering Technique

Before describing the SIFT (simplified inverse filtering technique) pitch detection, let us review some basic principles of linear predictive analysis.

The digital speech production model of section 2.1.2 can be further simplified to represent the accoustic speech waveform as the convolution of the voice source with the impulse response of an all-pole digital filter whose steady-state transfer function is of the form

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (4.6)$$

The speech samples s(n) are related to the excitation u(n) by the difference equation

$$s(n) = \sum_{k=1}^{p} a_k s(n-k)+Gu(n) \qquad (4.7)$$

A linear predictor with prediction coefficients, $\alpha_k$ is defined as a system whose output is

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \qquad (4.8)$$

The prediction error e(n) is defined as

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \qquad (4.9)$$

It is seen that the prediction error sequence is the output of a system whose transfer function is

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} . \qquad (4.10)$$

The basic problem of LPC analysis is to determine a set of predictor coefficients $\{\alpha_k\}$ directly from the speech signal in such a manner as to obtain a good estimate of the spectral properties of the speech signal. The basic approach is to find a set of predictor coefficients that will minimize the mean-squared prediction error over a short segment of speech waveform. It can be seen that if $\alpha_k = a_k$, then $e(n) = Gu(n)$.For voiced speech this means that $e(n)$ would consist of a train of impulses; i.e., $e(n)$ would be small most of the time.

Thus the purpose of the linear predictive analysis is to spectrally flatten the input signal, similar to the clipping method discussed before. Based on the reasoning that $e(n)$ is a good approximation to the excitation source, it is expected that the prediction error will be large at the beginning of each pitch period, so that the pitch period can be estimated by performing an autocorrelation analysis on $e(n)$ and detecting the largest peak in the appropriate range. Figure 4.10 shows the block diagram of the SIFT pitch detector which is based on these ideas [47],[3].

The input signal is first prefiltered by a lowpass filter with a cutoff at 800 Hz. Then, the sampling rate is reduced to 2 kHz by a decimation process. The samples are differenced to accentuate the region of the second formant , and multiplied by a Hamming window. A fourth

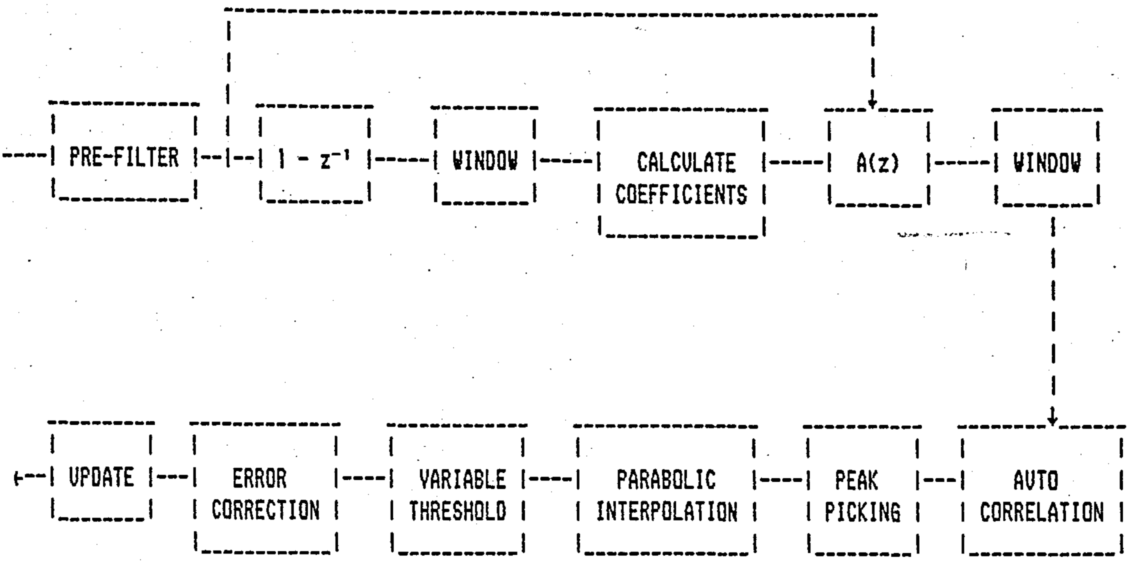


Figure 4.10. Block diagram of the SIFT pitch detector

order inverse filter A(z) is then designed using the autocorrelation method. A fourth order filter suffices to remove the formant structure because only two formants can be present in the frequency range (0,1 kHz). After inverse filtering, the signal is then multiplied by a second hamming window.

The autocorrelation function of the frame is computed and the peak of the sequence is searched in the range in which pitch period is expected. Parabolic interpolation is applied to provide greater resolution. A variable threshold is defined. 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. An attempt at error detection and correction is made by storing several pitch period candidates. The algorithm for the voiced-unvoiced decision and the error correction step can be seen clearly in Figure 4.11. After this operation, the pitch period estimate with maximum delay is output.

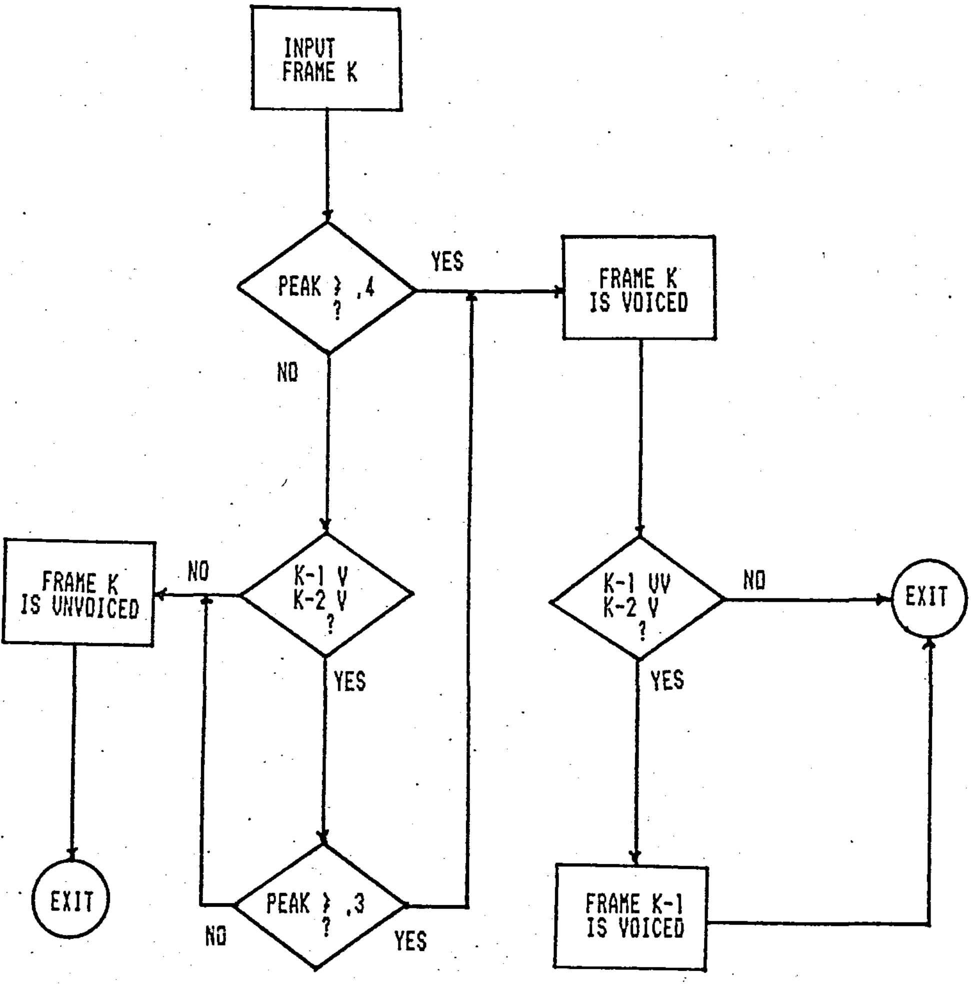


Figure 4.11. Decision algorithm for voiced-unvoiced decision and error correction

### 4.1.5 Cepstrum Method

In its most basic form, the system for producing voiced speech sounds consists only of the vocal source and the vocal tract. The output speech signal s(t) is given by

$$s(t) = h(t) * u(t) \qquad (4.11.a)$$

$$S(w) = H(w) \,.\, U(w) \qquad (4.11.b)$$

where the sorce signal is denoted by u(t), the impulse response of the vocal tract is h(t), and '*' represents convolution.

The source signal, and therefore, the speech signal are quasi-periodic for voiced speech sounds. If the period is $\tau$ seconds, then the spectrum of the speech signal consists of harmonics spaced $\tau^{-1}$ Hz. Thus, the spectrum of a voiced speech signal is periodic along the frequency axis with period equal to the reciprocal of the period of the time signal being analyzed. The obvious way to measure this 'period' in the spectrum is to take the Fourier transform of the spectrum. This will result in a waveform having a peak corresponding to the 'period'. But the effect of the vocal tract is still superimposed on the signal. The solution is to take the logarithm of the amplitude of the spectrum and take another Fourier transform. This function,defined as the 'cepstrum' seperates the effects of the vocal source and tract. The reason for this is the property that the logarithm of a product equals the sum of the logarithms of the multiplicands:

$$\log |S(w)| = \log |H(w).U(w)| = \log |H(w)| + \log |U(w)| \qquad (4.12)$$

The Fourier transform of the logarithm of the spectrum preserves the additive property and the source and tract effects become additive. The effect of the vocal tract is to produce a low frequency ripple in the logarithm spectrum, while the source produces a high frequency ripple. Therefore, the cepstrum has a sharp peak corresponding to the high frequency source ripple and a broader peak corresponding to the low frequency formant structure in the logarithm spectrum.

Another difficulty arises from the fact that a time limited speech signal is used. The effect of time limiting the speech signal with a multiplicative time window w(t) is a convolution of the corresponding spectral window W(w). Hence, the complex spectrum is not strictly frequency limited, but can be described as being approximately frequency-limited if W(w) has very small side lobes. The Hamming window which has a maximum side lobe 44 dB below its peak response is a good choice.

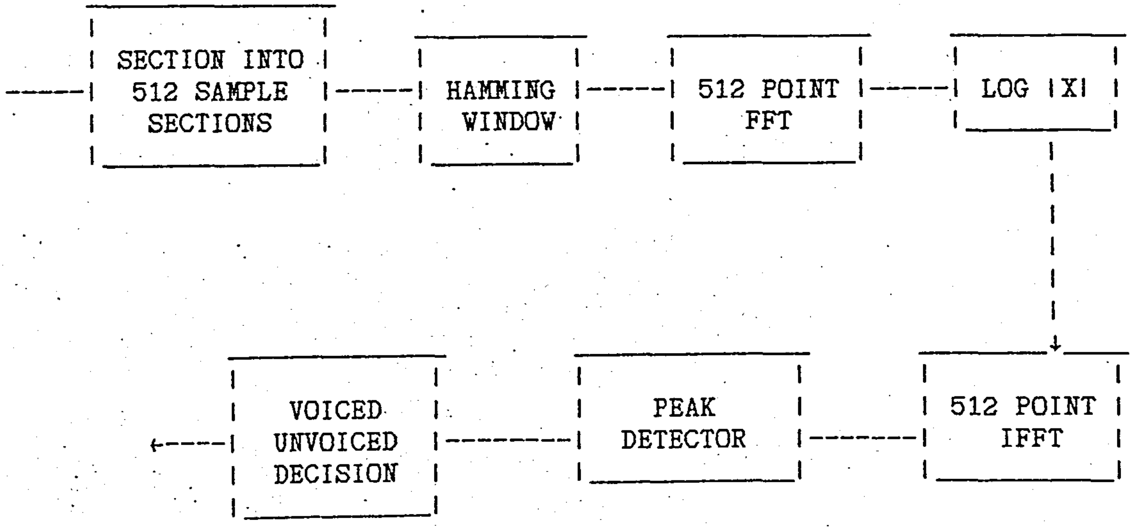


Figure 4.12. Block diagram of the CEP pitch detector

These properties of the cepstrum have been used as basis for pitch detection algoritms [48],[39]. The block diagram of a cepstral pitch detector is given in Figure 4.12. The speech signals are first sectioned into 512 sample frames, and then multiplied by a Hamming window. The cepstrum of the windowed signal is computed using FFT techniques. Due to the effect of the time window, the cepstral peaks decrease in amplitude with increasing quefrency. To overcome this effect, a linear multiplicative weighting is applied. The range that is searched for the peak of the cepstrum is 1-15 ms., since pitch periods outside this range are not usually encountered. Figure 4.13 shows the cepstrum of the frame in Figure 4.3. Since no weighting is applied, a decrease in amplitude can be seen as the quefrency increases.

The peak value of the cepstrum is searched and compared to a threshold. The cepstral peaks at the end of a voiced speech segment usually decrease in amplitude and would fall below the peak threshold. The solution is to decrease the threshold by some factor over the

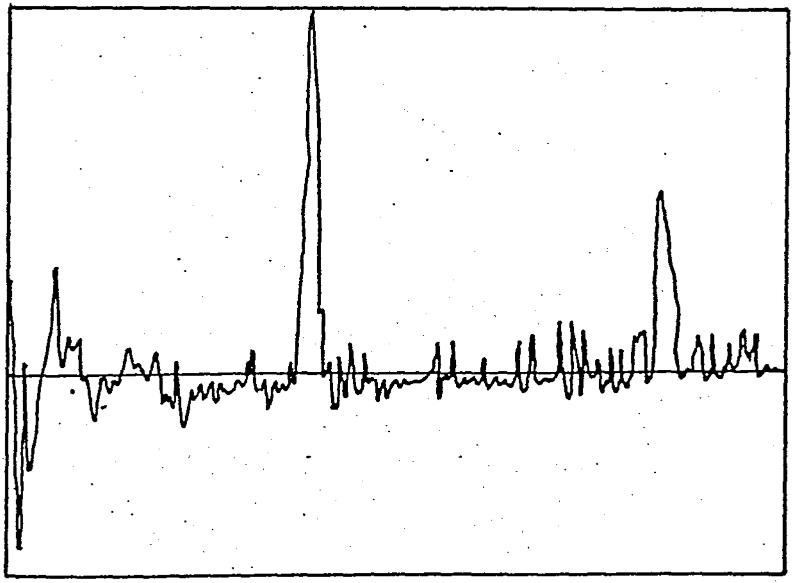


Figure 4.13. Cepstrum of the frame in Fig. 4.3.

quefrency range of ±1 msec of the immediately preceding pitch period when tracking the pitch in a series of voiced speech segments. The threshold reverts to its normal value over the whole cepstrum range after the end of the series of voiced segments.

There is also the possibility that an isolated cepstral peak might exceed the threshold, resulting in a false indication of a voiced speech segment. Such peaks are disregarded.

Another problem often encountered is pitch doubling. The second rahmonic of a cepstral peak sometimes exceeds the fundamental, and the second rahmonic should not be chosen as representing the pitch period. Thus, the peak picking algorithm should eliminate false pitch doubling caused by a second rahmonic but should also allow legitimate pitch doubling. For legitimate doubling, there is no cepstral peak at one-half quefrency, but for erroneous doubling, there is such a peak at one-half quefrency since this is the fundamental.

A flow-chart of the peak picking algorithm is given in Figure 4.14. The algorithm determines whether the cepstral peak of the $N^{th}$ cepstrum represents a voiced speech segment. Information about the $N-1^{th}$ cepstrum is stored, and the $N+1^{th}$ cepstrum is peak picked before deciding about the $N^{th}$ cepstrum. If pitch tracking is in effect, the threshold is reduced if the quefrency of the peak is within ±1 msec of the quefrency of the previous pitch peak. The peak is compared with the threshold, and pitch doubling is investigated whether the peak exceeds or does not exceed the threshold. The information about the $N+1^{th}$ cepstrum and $N-1^{th}$ cepstrum is then used to decide if the $N^{th}$ cepsral peak represents an isolated voiced segment or an isolated absence of voicing in a series of voiced speech segments.

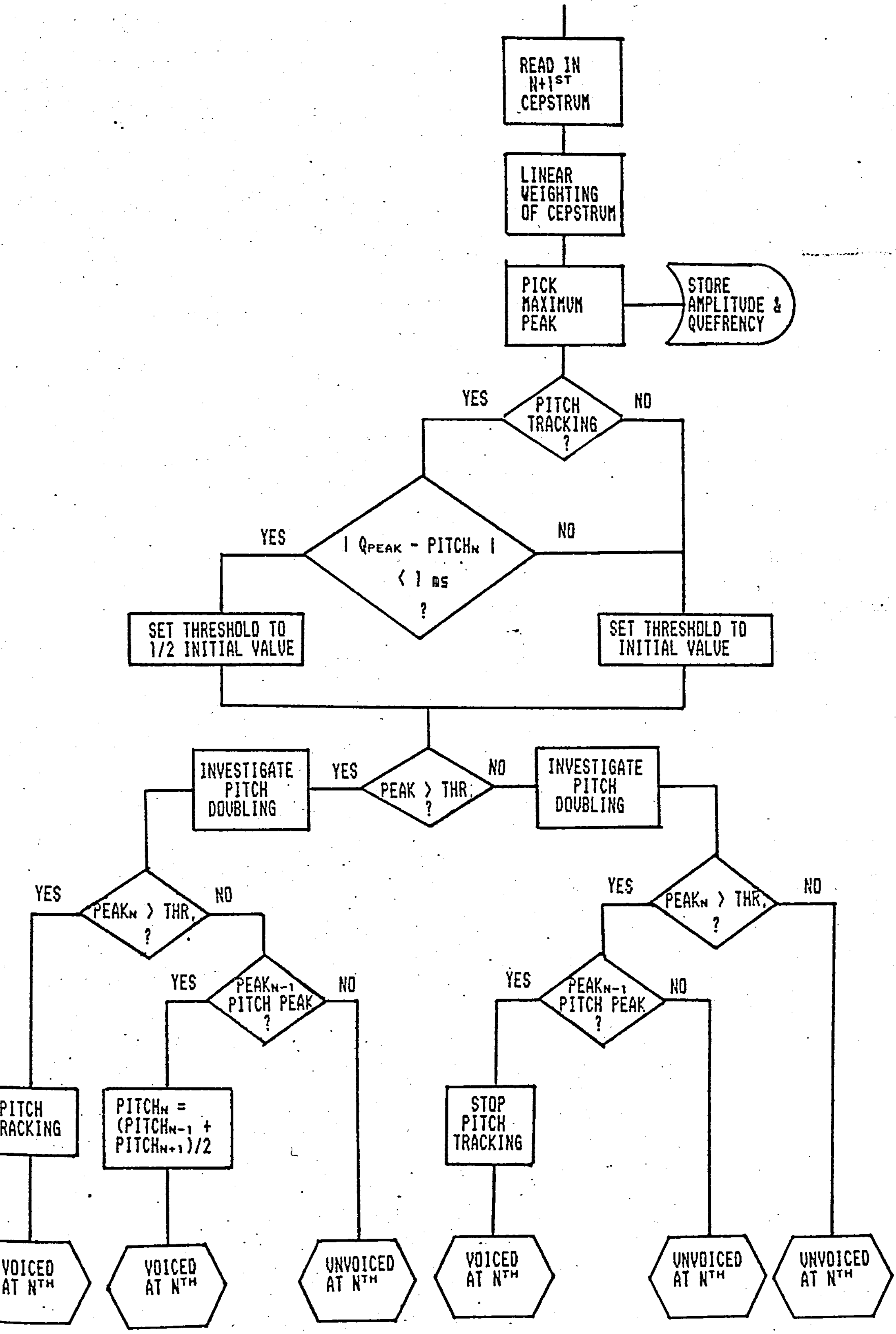


Figure 4.14. Flowchart of the peak picking algorithm.

### 4.1.6 Harmonic Pattern Matching Method

In the Fourier representation, the excitation for voiced speech is manifested in sharp peaks that occur in integer multiples of the fundamental frequency. This fact has served as the basis of a number of pitch detection schemes. The harmonic pattern matching—approach of Martin discussed in [39] will be reviewed here. Figure 4.15 presents the block diagram of the pitch detector.

Since only frequency components below 2 kHz are taken into account, the signal is first downsampled in the time domain to a sampling rate of 4 kHz. A frame of 32 ms length (128 samples) is then windowed and transformed into the frequency domain. In the amplitude spectrum all the values are set to zero except the peaks that exceed a threshold of -35 dB relative to the global maximum of the spectrum, and their immediate neighbors. The original spectral resolution of 33 Hz guarantees that the

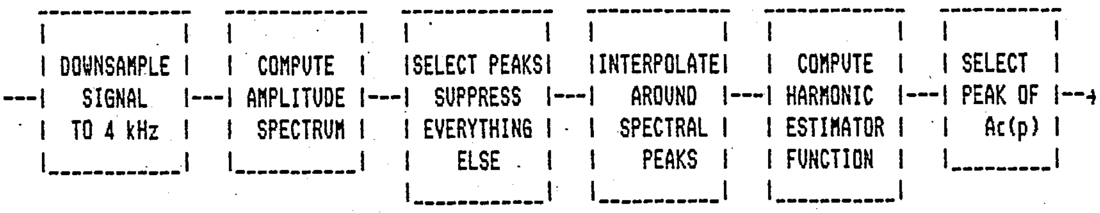

DOWNSAMPLE SIGNAL TO 4 kHz
COMPUTE AMPLITUDE SPECTRUM
SELECT PEAKS SUPPRESS EVERYTHING ELSE
INTERPOLATE AROUND SPECTRAL PEAKS
COMPUTE HARMONIC ESTIMATOR FUNCTION
SELECT PEAK OF Ac(p)

Figure 4.15. Block diagram of the harmonic pattern matching method pitch detector

information on fundamental frequency is present in the spectrum for values of $F_o$ down to 70 Hz. This resolution is then increased to 1 Hz interpolating the missing points. From this short-time amplitude spectrum, a harmonic estimator function is derived by applying a comb filter. The principle behind the comb method consists in the search for values of the spectrum situated at harmonic frequencies, and whose sum is a maximum for a given frequency interval. The fundamental corresponding to the harmonic structure giving the largest sum is then taken to be the fundamental frequency of the signal, as long as this sum differs sufficiently from the values obtained for other structures in the same spectrum (which would correspond to the case of voiceless signals).

The spectral comb is given as an impulse sequence provided with weights that decrease with increasing frequency; the distance of the individual pulses equals the trial fundamental frequency p,

$$C(m,p) = \begin{cases} k^{-1/s} & m=kp; \quad k=1,2,\ldots \\ 0 & \text{otherwise} \end{cases} \qquad (4.13)$$

For each value of p of the amplitude spectrum A(m) is weighted by the spectral comb C(m,p), and the spectral components that pass the comb are added up to form the harmonic estimator function $A_c(p)$,

$$A_c(p) = \sum_{k=1}^{N/2p} A(kp)C(kp,p) \qquad (4.14)$$

The value of p where $A_c$ reaches its maximum is then taken as the estimate of the fundamental frequency $F_o$.

## 4.2. MEASUREMENT OF ENERGY AND DURATION

Measurement of the total energy of a speech signal, either by digital or analog techniques, is straightforward. This contour alone serves as an important cue to the determination of word or syllable endpoints, or it can be used in combination with other data to give more reliable results.

Other energy components that prove useful are the energy contours in certain frequency bands [21]. The energy in the band 60 to 3000 Hz is called sonorant energy. Sonorant energy has been shown to be more useful for prosodic analysis than the broadband total energy since total energy remains high during obstruents while the sonorant energy dips at obstruents (which occur at syllable boundaries). Other bandwidth-limited energy functions have also been shown effective. The energy in the band 650 to 3000 Hz is useful for seperating vowel nuclei from surrounding nasals, liquids and glides. A very low frequency energy function in the bandwidth 60 to 400 Hz can provide an independent decision about the voicing state of the speech.

Duration is usually measured as the seperation between two marked points on the speech waveform, and these points are usually marked by an energy or pitch measurement of the waveform. Some durations that are of importance are the duration of sounds, duration of syllables, duration of words, and duration of phrases. The averages of these durations may give information about the speech rate of the speaker.

The algoriths that combine these measurements to identify the underlying prosodic phenomena will be discussed in the next chapters.

# V. PROSODY IN SPEECH RECOGNITION

The human speech perception system is such that a native speaker uses, subconciously, his knowledge of the language, the environment, and the context in understanding a sentence. These sources of knowledge include the characteristics of speech sounds (phonetics), variability in pronounciations (phonology), the stress and intonation patterns of speech (prosodics), the sound patterns of words (lexicon), the grammatical structure of language (syntax), the meaning of words and sentences (semantics), and the context of conversation (pragmatics). To approach human performance, a machine must also use all the available knowledge sources effectively. The prosodic features in speech carry valuable information that can be used in this process. Up to date, very little of this information has been used in automatic speech recognition. In this chapter, after a brief review of recognition systems, strategies to incorporate this knowledge source into various recognition systems will be discussed.

## 5.1 SPEECH RECOGNITION SYSTEMS

Speech recognition can be described as the process of transforming the continuous acoustic speech signal into discrete representations which may be assigned proper meanings. The ultimate goal is to understand the input sufficiently to select an appropriate response.

Speech signals convey information about *who* spoke *what* message in *what* manner and *what* environment. There is an extensive amount of information in the speech signal, only some of which is related to selecting correct machine responses. The critical task is to extract all and only those parts that convey the message, and ignore the rest.

Some dimensions of performance in speech recognition are the size of the command vocabulary, whether or not the system can accommadate any talker, or only those who have trained the system, and whether or not input speech can be continuous connected utterances, or must be isolated individual commands. These dimensions are illustrated in Figure 5.1.

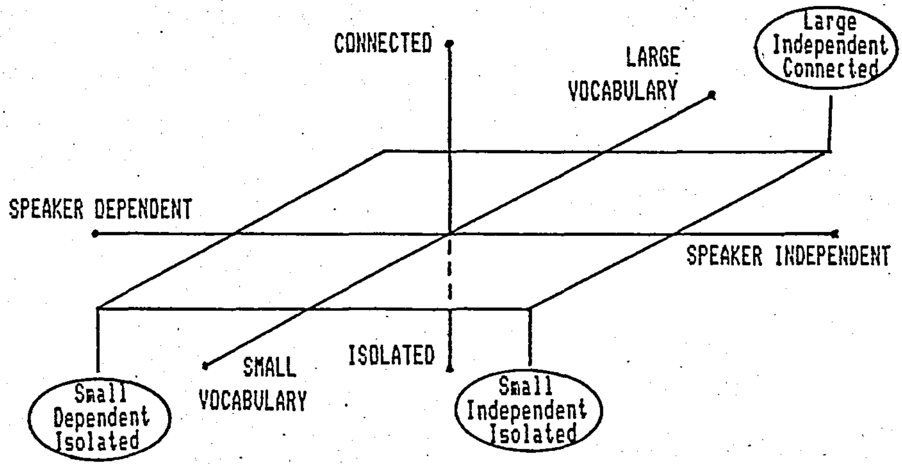


Figure 5.1. Dimensions of performance in speech recognition

At present, highly reliable automatic recognition can be achieved for relatively small vocabularies of single words spoken in isolation by a talker to whom the system is trained. By contrast, automatic recognition of unconstrained fluent speech by any talker on any subject is nowhere near reality.

Approaches to speech recognition are differentiated by whether or not recognition is effected by a template matching to vocabulary items that originally were measured and derived from human speech, or whether or not recognition is effected purely by computation using programmed rules that analyze the unknown input and which utilize no vestige of stored human speech. Most practical success to date are with the former because it is easier. But the greatest promises may be with the more sophisticated latter.

In template matching, human spoken utterances (typically phrases, words, syllables, or phonemes) are typically represented in the form of spectral sequences as a function of time. Recognition is achieved by using a pre-defined similarity measure to compare the unknown token against stored templates. In many cases, time-alignment algorithms are used to account for some variability in speech rate. While template matching systems can achieve high performance with a small set of accoustically distinct words, they are limited in their ability.

In the feature-based approach to speech recognition, a set of acoustic features that capture the phonetically relevant information in the speech signal are identified. With this knowledge, algorithms can be developed to extract the features from the speech signal. A classisfier is then used to combine the features and arrive at a recognition decision.

Drawing a sharp division between these two approaches is somewhat arbitrary and perhaps unnecessary. Actual systems may make use of both techniques with a varying mixture. What is most important is whether and to what extent speech-specific knowledge is being utilized for recognition.

### 5.1.1. Isolated Word Recognition Systems

In these recognition systems, the human must command the machine in single utterances. The vocabulary is usually small (in the order of 20, 100, or 1000 words depending on the application). They can be made talker independent at the cost of computational complexity. Figure 5.2 shows the block diagram of a typical word recognition system.

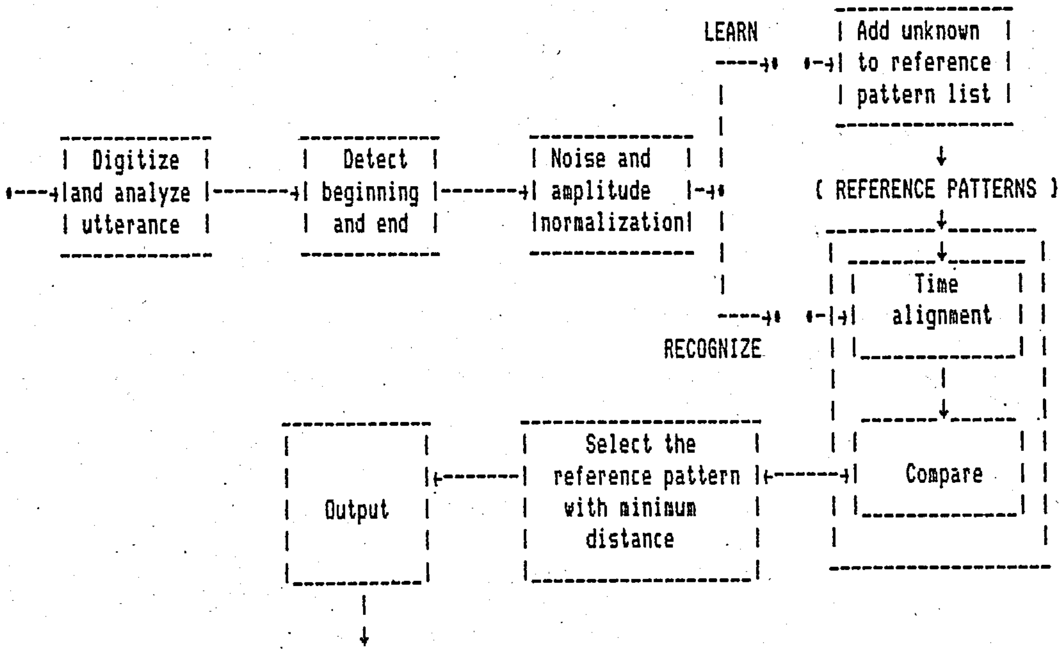

LEARN
Add unknown to reference pattern list
Digitize and analyze utterance
Detect beginning and end
Noise and amplitude normalization
REFERENCE PATTERNS
Time alignment
RECOGNIZE
Output
Select the reference pattern with minimum distance
Compare

Figure 5.2. Block diagram of a word recognition system

A common approach is to measure a time pattern of features of the frequency spectrum for the input human utterances, and compare these to a vocabulary of human derived stored patterns, one for each single acceptable utterance. If the system is talker-dependent, these vocabulary patterns have been provided previously for the given talker. The linear prediction coefficient (LPC) parameters are the most common features and prove to be a useful set. The vocabulary pattern corresponding most closely to the unknown input is judged by the machine to have been the spoken command. Various distance measures may be used for the closeness of fit judgement, but one of particular appropriateness for speech is the so called maximum likelihood LPC ratio [49]. In making the distance measure, to take account of different rates of speech, a procedure called dynamic time warping (DTW) is applied [50]. Often the output of the model is a set of estimates of the words in the output, ordered by similarity, allowing the final decision of what was actually spoken to be deferred to a higher level in the recognition system.

By storing multiple patterns that characterize a large population of talkers for each utterance in the vocabulary, the system can be made speaker independent. Statistical clustering analyses then determine the set of multiple patterns.

### 5.1.2. Connected Word Recognition Systems

In the production of continuous speech, pronounciation is less careful, speaker differences are underlined, speaking rate is less constant, co-articulation effects exist between words as well as within them. There is even little evidence of word boundaries. Stress and intonation change due to the importance of a word in the message. Thus, the task of recognition becomes impossible to achieve with the weapons of isolated word recognition. Note that in an attempt to recognize the utterance as a whole, one would be faced with storing $10^{10}$ reference templates even with a 10-word vocabulary and given a 90 % word recognition accuracy, the overall performance would drop to 35 %.

Instead, an active system which makes use of all the known constraints of language and varies its analysis in the light of this knowledge is required.

Systems which recognize words with the pattern recognition techniques of isolated word recognition and group these together to form larger units have been proposed, but as the number of words in the vocabulary and the number of different contextual variations per word get large, the storage and computation time become enormous. However, taking into account the advances in the VLSI technology, this is a solution to be considered for very small-sized vocabularies or speaker-dependent systems.

For more sophisticated systems, what is needed is a more compact representation of the sound patterns of the words such as those used by linguists, i.e. repesentation of words as a sequence of phonemes, allophones, or syllables. This change from signal space representation of the words to a symbol space representation requires segmenting the continuous speech signal into discrete acoustically invariant parts and labeling each segment with phonemic or feature labels. A phonemic dictionary of the words could then be used to match at a symbolic level and determine which word was spoken.

Feature detection usually represents the detection of silence, voicing, stress, LPC or spectrum parameters, and so on. The purpose of segmentation is to divide the continuous speech signal into discrete units based on some measure of acoustic similarity. Energy in certain bands is the most important measure in segmentation. There is no simple algorithm that gives phonemic boundaries. Usually boundaries associated with significant changes in acoustic characteristics are used for segmentation. Labeling schemes associate a phonemic (or some other) symbol with each segmental unit. Before this symbol sequence can be used in matching, it is necessary to apply phonological rules to combine segments, change labels based on context, delete segments, and so on.

One of the most important problems in continuous speech recognition is that of detecting the boundaries of words. In systems that operate without this knowledge, the analysis proceeds from left to right, matching at each step. In this case, one must find techniques for terminating the match when an optimal match is found.

Matching and verification of hypothesized words is basic to all recognition systems. Three different word verification techniques are

-Heuristic matching,

-Stochastic matching,

-Analysis-by-synthesis.

*Heuristic matching* involves aligning the phonemic spelling of the word to be matched with the segmental labels while allowing for the possibility that errors may have occured. Alignment is usually based on the notion of anchor points in which stressed vowels which are much less likely to be missed are aligned first, followed by other vowels and consonants. Degree of similarity is defined as a weighted sum of the individual phoneme versus segment label similarity values which are available as a confusion matrix generated by experiments. In *stochastic matching*, given a finite-state representation of alternative pronounciations of a word with associated transition probabilities, a dynamic programming technique is used to perform matching left-to-right.The best phonemic match and the corresponding likelihood are determined by matching all the possible phonemic variations of the word with the unknown segmental phoneme string. The basis for the *analysis-by-synthesis* method is the observation that phonological phenomena such as vowel reduction, flapping, palatalization, etc. are basically generative in nature and cannot be easily captured in terms of analytic rules. In this method, the abstract representation of a word is transformed into an acoustic representation suitable for matching with the acoustic parametrization of the unknown utterance.

Human listeners make use of linguistic cues and constraints in recognizing continuous speech. A listener's application of linguistic knowledge often enables him to guess the remainder of a sentence after hearing the first few words. If machines are to approach human performance, this linguistic expertise must be built into them. Such systems are sometimes referred to as *speech understanding systems*. In addition to the problems of having to recognize connected speech, these systems tend to have the additional requirement that they must do so even when the utterance is not grammatically well formed, and in the presence of speech-like noise. The requirement is somewhat relaxed by the fact that what matters in the end is not the recognition of each and every word in the utterance but rather the intent of the message. Figure 5.3 shows the processes involved in *recognition* and *understanding*. Syntactic analysis refers to testing whether a hypothesized word is syntactically consistent with words already recognized,and using syntactic constraints to predict likely upcoming words. The meaningfulness of hypothesized word sequences are then tested (semantic analysis) and likely future words are predicted based on prior discourse and the specific task (pragmatic analysis).

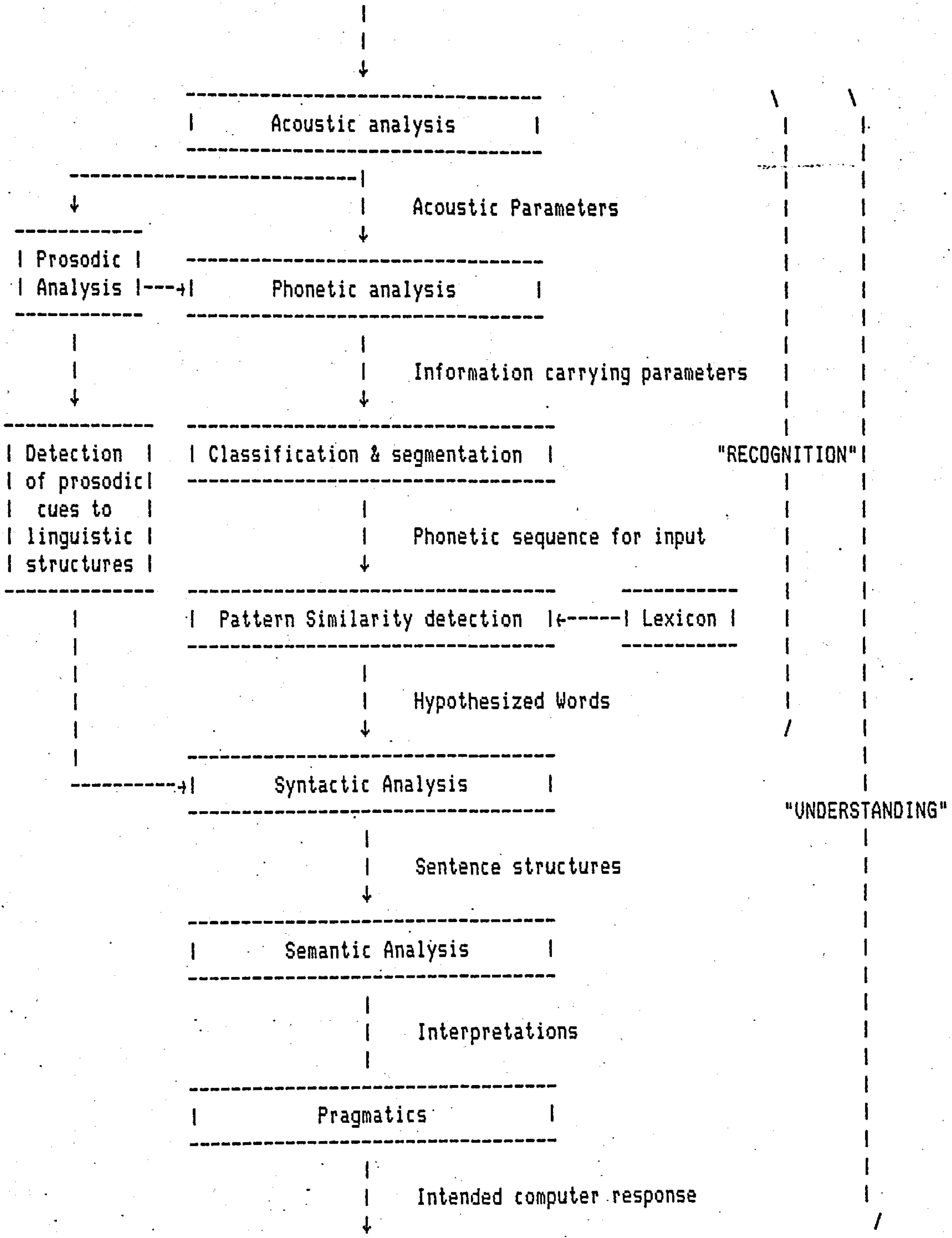


Figure 5.3 Processes involved in "recognition" and "understanding"

## 5.2. PROSODIC AIDS TO SPEECH RECOGNITION

While discussing recognition systems, it has already been outlined that prosodic features give many cues to the speech knowledge at various layers. In the present recognition systems, some of this knowledge is being utilized, but still, prosodic analysis is one of the gaps in speech recognition technology.

There are certain prosodic features, like stress, timing, and intonation, which give very useful cues to the inherent structures of the speech waveform, and are relatively easy to extract from the speech waveform. However, since the linguistic functions of these features may change in each language, a strategy specific to the language must be used. These features, when utilized by any isolated or connected recognition system, lead to an improvement in performance at various steps of the recognition process. This section will be an account on present and potential uses of these features in recognition systems.

### 5.2.1. Stress

Stress is usually considered to be the most basic abstract prosodic feature. Since the linguistic functions of stress may change in each language, there is no absolute way of defining how stress may be used in recognition. However, some notions that are common may be adopted directly, while others must be devised according to the stress characteristics of the specific language.

One of the most common uses of stress is its providing "an island of phonetic reliability". Stress usually has the effect of lengthening a vowel and enhancing its pronounciation, so that stressed vowels are expected to be clearer. Furthermore, it is known that as the rate or style of speech changes, it is unstressed syllables that experience the largest variations, stressed syllables remain more or less constant in their pronounciation. For this reason, they have been accepted as anchor points around which the alignment process is usually done in heuristic matching methods.

Phonological distortions and errors in automatic phonetic analysis get more frequent as time intervals between stresses are shortened. This observation has been a justification of the hypothesis that the interstress intervals serve as a direct correlate of the speech rate.

Stress patterns are closely associated with specific syntactic structures. This is true in general for English, where certain words like the articles and propositions are pronounced reduced and other groups like the command verbs and quantifiers are stressed all the time. This is not the case in Turkish, where each word can be said to have one strong stress [5],[14],[16-20]. What can make a distinction here may be the place of the stress. For example, adverb particles are one class of words that take stress on their first syllables (e.g. *artik*, *henuz*, *hemen*), while words usually take it on their final syllables.

The placement of stress may be a distinguishing feature for homonyms (e.g. *yalnız* : only , *yalnız* : alone), or words with high probability of confusion (e.g. *yarım*, *yarın*).

Stress usually marks contrasts, emphasis or important words in a sentence. However, this is usually accompanied by a reordering of the words in a sentence. The word to be emphasized is brought in front of the verb in addition to being more highly stressed:

*Bugün annem geliyor;* *Annem bugün geliyor.*

### 5.2.2. Timing

One of the most important problems in speech recognition are that of finding boundaries, or segmenation. Timing information may be used to help overcome these problems.

Phonemes usually show a characteristic duration, for example, /a/ usually tends to be longer then /ı/, while all the vowels are usually longer than most of the consonants. This property is used in labeling schemes. The durations of detected phonetic segments are compared with expected durations for various phones to aid labeling.

It has been observed that phrase-final and pre-pausal syllables have vowels whose durations are lengthened by 20 to 50 % over their values in other positions. Although this may be a way for detecting phrase boundaries, it has not been used in any speech recognition system due to the complexity involved in its implementation. A more practical device for finding phrase boundaries has been the interstress intervals, as mentioned above.

One other major factor that influences time intervals in speech is the rate of speech. This information is crucial in a recognition system to compensate the effects of change in pronounciation or vary the expected phone durations. Rate of speech is also essential in determining what phonological rules should apply at various regions of an utterance, since some rules apply for fast speech while others are applicable only to slow speech. The following have all been referred to as a measure of speech rate:

-The total duration of a specific spoken text;

-The average measure of the number of words per unit time;

-The average number of stresses per unit time;

-The number of syllables per unit time;

-The average or local number of phones per unit time.

The Germanic and English sentence rythms assign approximately equal time to stress groups clustered around each stressed syllable, so for these languages, the average number of stresses serves as a good indication of speech rate. This is not true for Turkish, where each word is stressed. However, it is claimed that the Turkish intonation assigns longer time to unstressed syllables, thereby equalizing the time for each syllable [5]. If this is the case, the average number of syllables per unit time can serve as a good measure of speech rate.

The main method in finding clause and sentence boundaries is by the duration of pauses in speech. These pauses are the spoken equivalents of written punctuation marks. Since speech recognizers usually receive only one sentence at a time, only clause boundaries are of importance in this respect. These can be detected from 200 millisecond or longer periods of silence, or from 350 millisecond or longer periods of unvoicing.

### 5.2.3 Intonation

Intonation is a vital aspect of speech which conveys information about the type of sentence spoken, the divisions and categories in phrase structures, paragraphing and topic change, semantics, and emotion. Children learn intonational cues to phrase structure and sentence type even before they learn words, so that they understand what is a question and what is a command.

Each language has its intonation system. The overall intonation of English sentences has been characterized in terms of two alternative contours as shown in Figure 5.4. Tune I contour has a characteristic rising of $F_o$ until the first stressed syllable in the sentence is reached, and a falling of the pitch from the first stressed syllable to the last. Sentence final intonation falls rapidly. This type of intonation accompanies declarative sentences, exclamations, and questions with interrogative words. Tune II is like Tune I, but is terminated by a brief rise in pitch. Tune II marks yes/no questions, uncertainty or indifference in expression, and incompleteness.

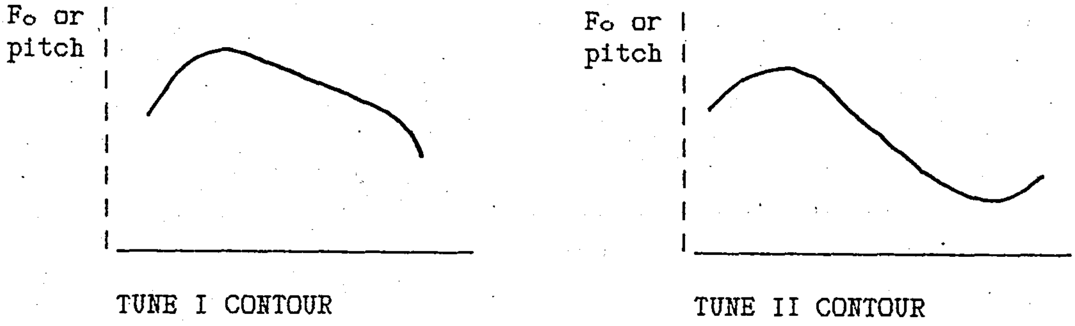


Figure 5.4. Tune I and Tune II intonation contours

As discussed in section 3.3, Turkish exhibits the same intonation contours, with the difference that Tune I marks yes/no questions and Tune II, questions with interrogative words. These contours may be used either in the sentence hypothesizing or error detection steps.

Observations of pitch contours of English sentences have shown that boundaries between clauses are detectable from very large (e.g., more than 90 %) increases in $F_o$ at the beginning of the new clause and that boundaries between major syntactic phrases are detectable from substantial (7 % or more) increases in $F_o$ [21]. Although exact locations of the boundaries are difficult to detect, this property can aid the syntactic parser in a speech understanding system (Figure 5.5).

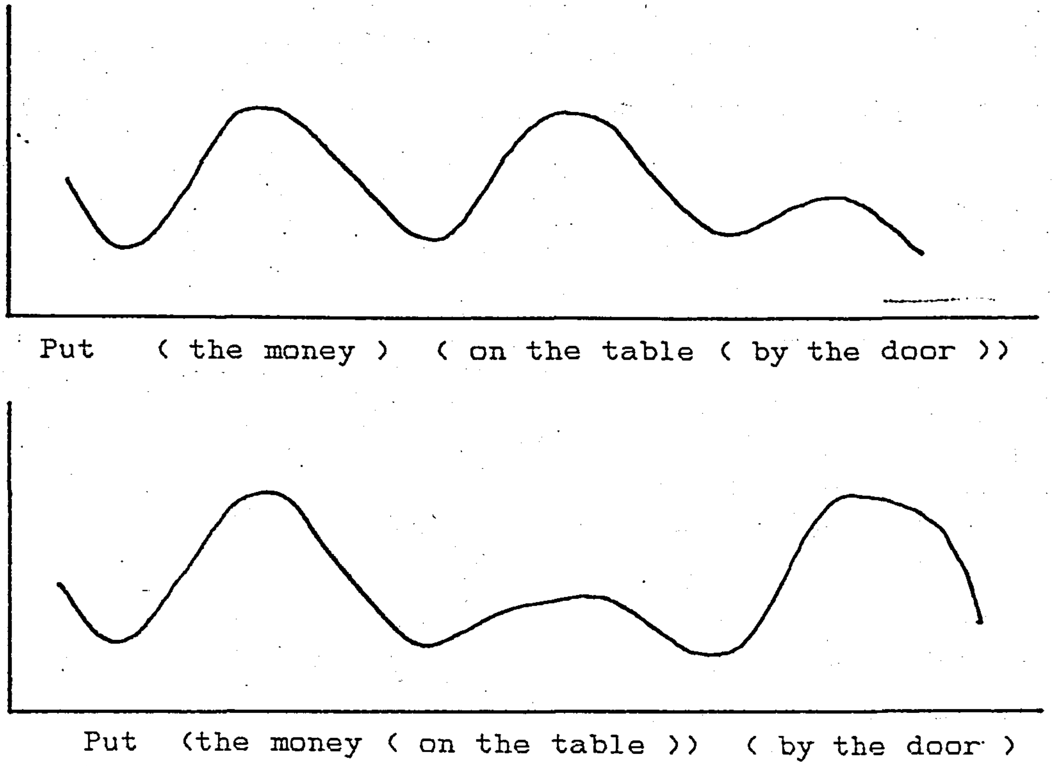


Figure 5.5. Clause and syntactic phrase boundaries on $F_o$ contours

### 5.2.4. Vowel Harmony

About 90 % of words in Turkish obey the "vowel harmony" rules that have been outlined in section 3.3 [4]. Those that do not obey these rules are usually loan-words. So, a reasonable approach would be to form a vocabulary consisting only of words of native Turkish origin to benefit from this regularity. If this is not possible, one can still make use of these rules and handle the exceptions seperately.

Out of the 8 types of vowels in Table 3.1, vowel harmony reduces the possible types in the non-initial syllable of a word to 2. Table 5.1 shows these classes and summarizes the rules. Considering the fact that Turkish is a language which consists mostly of polysyllabic words, it is clear that this will bring great amount of redundancy of the phonemic representation. The most common word type in Turkish is tri-syllabic [15]. If one takes this as an average word length, it can be concluded that the saving introduced will be about 50 %.

| GROUP OF THE FIRST SYLLABLE | POSSIBLE GROUPS FOR PRECEDING SYLLABLES |
|---|---|
| 1 | 1 , 2 |
| 2 | 1 , 2 |
| 3 | 1 , 4 |
| 4 | 1 , 4 |
| 5 | 5 , 6 |
| 6 | 5 , 6 |
| 7 | 5 , 8 |
| 8 | 5 , 8 |

|  | UNROUNDED |  | ROUNDED |  |
|---|---|---|---|---|
|  | WIDE | CLOSE | WIDE | CLOSE |
| BACK | 1 | 2 | 3 | 4 |
| FRONT | 5 | 6 | 7 | 8 |

Table 5.1 The regulations introduced by vowel harmony

## 5.3. PROSODICALLY BASED SPEECH RECOGNITION

In most of the speech recognition systems up to date, prosodic, syntactic, semantic and pragmatic analyses have served an "after the fact" role of weeding out the unlikely word sequences, based on pre-compiled information about acceptable, meaningful, and task-related sentences. The hierarchy can be seen in Figure 5.2. Based on incoming acoustic data, words are hypothesized throughout an utterance, to account for the phonetic data in all regions of the signal, and thus many of the hypothesized words overlap in position or compete as alternative hypotheses on the same portion of the utterance. To allow for possible errors that may have occurred, all combinations of these are fed into linguistic analysis which has to select the most likely sequences of non-overlapping words which form grammatical, meaningful and relevant strings. However, because of the structural redundancy present in a listener's linguistic knowledge, a speaker does not have to encode into the acoustic waveform all of the features describing an utterance, and the features that are encoded can vary from time to time. In some utterances, whole phonemes or syllables may be missing. A speech recognition system based on the acoustic manifestation of all phonemes or all distinctive features would thus frequently fail.

In contrast, prosodic analysis offers an independent way of acoustically detecting some aspects of syntactic structure, without depending upon the potentially errorful sentences of hypothesized words derived from the incoming acoustic phonetic information. In the system

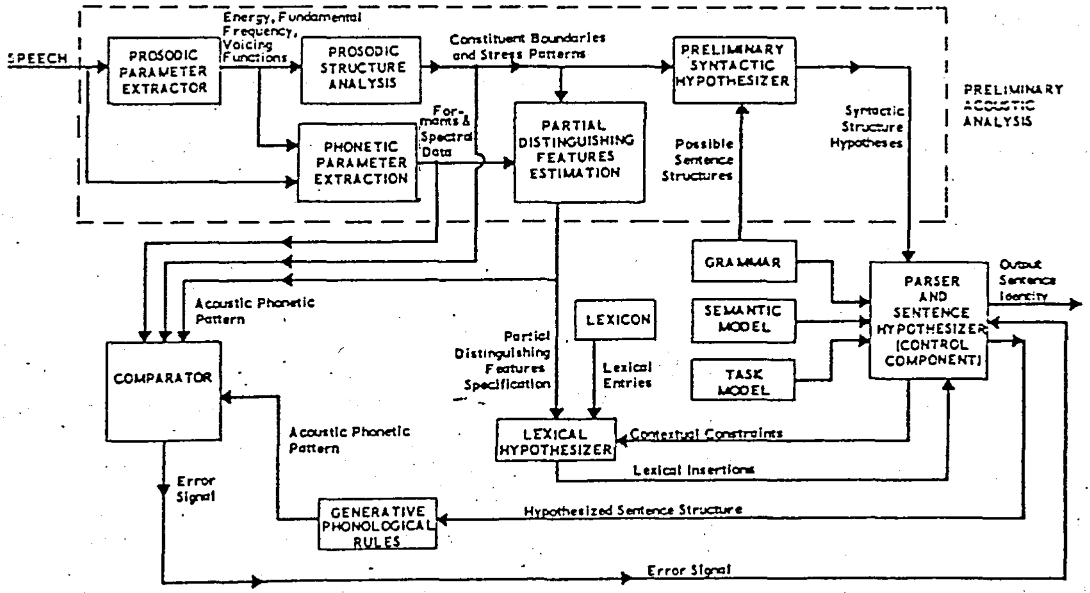

SPEECH
PROSODIC PARAMETER EXTRACTOR
Energy, Fundamental Frequency, Voicing Functions
PROSODIC STRUCTURE ANALYSIS
Constituent Boundaries and Stress Patterns
PRELIMINARY SYNTACTIC HYPOTHESIZER
PRELIMINARY ACOUSTIC ANALYSIS
For-mants & Spectral Data
PHONETIC PARAMETER EXTRACTION
PARTIAL DISTINGUISHING FEATURES ESTIMATION
Possible Sentence Structures
Syntactic Structure Hypotheses
GRAMMAR
SEMANTIC MODEL
TASK MODEL
PARSER AND SENTENCE HYPOTHESIZER (CONTROL COMPONENT)
Output Sentence Identity
Acoustic Phonetic Pattern
COMPARATOR
LEXICON
Partial Distinguishing Features Specification
Lexical Entries
LEXICAL HYPOTHESIZER
Contextual Constraints
Lexical Insertions
Acoustic Phonetic Pattern
Error Signal
GENERATIVE PHONOLOGICAL RULES
Hypothesized Sentence Structure
Error Signal

Figure 5.6. Prosodically guided analysis-by-synthesis system (After Lea et. al. [22])

proposed in [22], prosodics is used throughout the system. The block diagram of this system is given in Figure 5.6. In the preliminaryhypothesis stage, prosodic features are used to segment continuous speech into sentences and phrases and locate stressed syllables. Such prosodic information is coupled with acoustic-phonetic and structural information in an analysis-by-synthesis system. Here, what is usually called the "preliminary analisis" block is broken down into :

-a component for extracting prosodic features (energy,pitch,voicing)

-a component for extracting phonetic parameters (formants, etc.)

-a prosodic structure analysis which obtains phrase boundaries, rhythms, and stress patterns

-a component for obtaining a partial representation of the phonetic segment structure (distinguishing features) within stressed syllables

-a preliminary syntactic hypothesizer which uses phrase boundaries, rythms and stress patterns to predict likely syntactic structures.

Following such preliminary analyses, the lexical hypothesizer proposes possible lexical entries for insertion in the sentence structure, based on the closest match between the partial distinguishing features representation of the input and the lexical entries in the lexicon. Contextual constraints, such as lexical categories that could occur at certain positions in the sentence structure, and likely words in certain semantic and task contexts, are used to guide the lexical hypothesizing. Grammatical, semantic and task constraints combine together with the lexical and syntactic hypotheses to yield a total hypothesis about the identity of the sentence spoken. The sentence hypothesizer controls the order in which acoustic-phonetic patterns are generated, by phonological rules, for comparison with the input acoustic-phonetic patterns. The error signal of the comparator is fed back to the sentence hypothesizer to formulate new hypotheses or to announce an output.

Only some aspects of this strategy were implemented in a recognition system [35]. Although it has not been realized, the overall recognition strategy deserves attention in the sense that it shows how many aspects of recognition systems discussed so far can be combined to produce a prosodically guided system which operates in a sense which is more close to the perception in humans.

# VI. RESULTS

In this study, tools have been developed to investigate prosodic features of Turkish. Analyses have been made on a total of 200 s of speech, in the form of 2-s utterances by four different speakers. Some of the curves from these analyses have been presented throughout the text. The phenomena observed on these curves are discussed, and some recognition strategies are outlined. Some of these strategies have been integrated in an isolated word recognition system of Turkish speech. The description of the system is given together with performance scores obtained.

## 6.1. PITCH DETECTION ALGORITHMS

The six pitch detection algorithms described in section 4.1 have been realized. A comparison between these six algorithms had been intended, but two of the algorithms had to be discarded because their memory requirements did not comply with the available user memory of the system described in section 6.2. The remaining four algorithms were tested using an artificial signal generated by a speech model.

### 6.1.1. The Artificial Speech Signal

The model used is a simplification of the general discrete-time model of Figure 2.2. The model consists of a glottal pulse generator, a vocal tract filter, and a whitening filter (Figure 6.1). Fundamental frequency and vocal tract parameters are held constant. A whitening filter H has been used to flatten the output spectrum. The contributions of the glottal pulse, vocal tract and radiation are all included in the filter F. Finally, an envelope has been used to simulate the effect of the alteration of voiced and unvoiced segments. Details of the model can be found in [51].

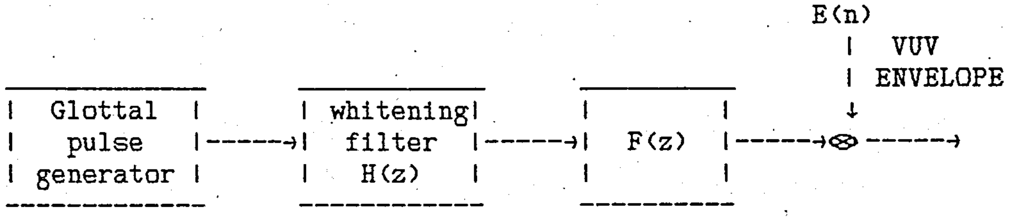


Figure 6.1 The artificial speech model

### 6.1.2. Comparison of the Algorithms

To make a performance evaluation of the algoritms, there was need for a speech signal for which the true pitch contours were known. The most adequate signal to use for this task would be the ouput of a speech synthesis system. Since this was not available, the artificial signal produced as described above was used. This signal is a rough approximation to the speech signal, and the results of this analysis may not reflect the true behaviour of the algorithms with real speech, but still, they gave an idea about their performances and speeds.

In Table 6.1, the error measurements for the pitch detectors is given. Four types of errors were classified as gross pitch errors, fine pitch errors, voiced-to-unvoiced errors and unvoiced-to-voiced errors. If $p_s(m)$ represents the standard pitch for $m^{th}$ frame, and $p_j(m)$, the pitch detected for the $m^{th}$ frame by the $j^{th}$ pitch detector, different types of errors result from the following situations:

$-p_s(m) = 0,\ p_j(m) \neq 0$ : Unvoiced-to-voiced error

$-p_s(m) \neq 0,\ p_j(m) = 0$ : Voiced-to-unvoiced error

$-p_s(m) = p_1 \neq 0,\ p_j(m) = p_2 \neq 0$ : In this case a voiced frame is correctly classified as voiced by the algorithm. For this case two types of errors can exist, depending on the values of $p_1$ and $p_2$. If we define the voiced error $e(m)$ as the the difference :

$$e(m) = p_1 - p_2$$

then, if $|e(m)| \geqslant 10$ samples the error is classified as a gross pitch error. For such cases, the pitch detector has failed in estimating the pitch period. All of the gross errors encountered were due to pitch doubling. The second type of pitch error is fine pitch error in which case $|e(m)| < 10$ samples. For this case the pitch detector has estimated the pitch period sufficiently.

The parameters in Table 6.1 are, the mean $e_{av}$ and the standard deviation $\sigma_e$ for the fine pitch errors, and error rates for the other types of errors.

| | $e_{av}$ | $\sigma_e$ | VUV errors | UVV errors | Gross errors |
|---|---|---|---|---|---|
| AUTOC | 0 | 0.46 | 0 | 0.07 | 0.07 |
| AMDF | 0.07 | 0.88 | 0.07 | 0 | 0 |
| CEP | 0.77 | 1.12 | 0 | 0.18 | 0 |
| PPROC | 0.13 | 1.17 | 0 | 0.13 | 0.13 |

Table 6.1. Pitch period errors for the algorithms

It is observed that the errors are often in the voiced-unvoiced transitions. However, it should be noted that there are no voiced-to-unvoiced errors in three of the algorithms, while in the fourth one, there are no unvoiced-to-voiced errors, which shows that all of the algorithms are biased toward voiced or unvoiced. The means and standard deviations for the fine errors are too small to be considered. There are few gross errors, which are due to pitch doubling. None of these are severe errors, and they can be corrected using logic similar to the error correction and tracking logic used in AMDF. So, the main criterion for selection is the speed. Exact values cannot be given, but PPROC performed the fastest, with AUTOC and AMDF about 10 times slower and CEP 100 times slower. So, PPROC was chosen as a first alternative to be used in the analyses. The voiced-unvoiced detection logic was developed as described in section 4.1.3. However, during its use with real speech, problems arose because of the wide pitch frequency range required (100-500 Hz, for both male and female speakers). The algorithm produced acceptable results when parameters were adjusted to a smaller interval for analysis with only male speakers, but since speech from both male and female speakers was to be analysed, this method had to be discarded. AUTOC and AMDF were taken as possible alternatives. With real speech, AMDF also presented some errors which were found to be due to the tracking logic. Similar tracking logic was used with AUTOC and it was observed that it produced the same type of errors. So, it was concluded that tracking was inadequate in regions of high-slope pitch changes. Since pitch peaks and valleys played an important role in the detection of many of the prosodic features, that was unacceptable. So, AUTOC was chosen to be used in the rest of the analysis. This selection agrees with many comparisons of pitch detectors ([40], [41]) in that the autocorrelation method after center clipping provides a simple and reliable way of detecting pitch. The pitch range searched was readjusted according to the pitch ranges of the speakers and this reduced the computation time by a factor of ½. A tracking logic was integrated to further reduce the computation time, but that was later discarded since this caused errors.The autocorrelation method with center clipping as described in section 4.1.1 was used in the rest of the analysis.

## 6.2 VOICE INPUT AND ANALYSIS SYSTEM

The analysis is performed on a PDP 11/23 microcomputer. Voice input to the system is via analog circuitry which consists of a standard carbon microphone of the type used in telephones, a lowpass filter which has 6-dB point at 3.5 kHz, followed by an amplifier. The otput of the amplifier is designed to be between ± 10 V. The ouput of this analog system is fed to the 12 bit A/D converter of the PDP 11. The sampling is

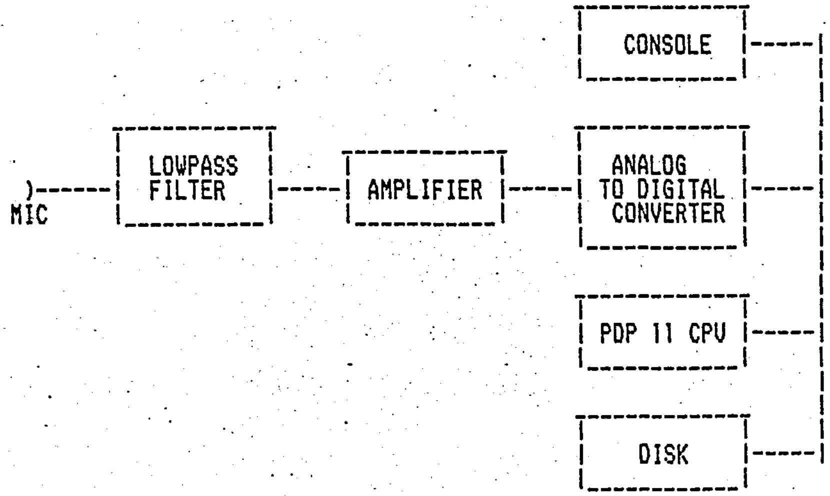


Figure 6.2 Voice input and analysis system

performed at 8 kHz, the minimum allowed by the Nyquist criterion, to make the most of the limited memory capacity. At this sampling rate, the maximum duration of speech that can be stored is 2 seconds. At a normal speech rate, single words or 2-3 word sentences can be uttered in this time. So, analysis was performed on those. The block diagram of the voice input/analysis system is given in Figure 6.2.

## 6.3 SYLLABLE

### 6.3.1 Syllable As a Unit of Recognition

There are several alternatives for a recognition unit: phoneme, allophone, diphone, syllable and word. All of these have been used as units in different recognition systems, but none of them has proved ideal. In fact, all have their advantages and disadvantages, and a recognition system may use a combination of these units. The advantages and disadvantages of these units have been summarized in Table 6.2.

In present recognition systems, the most often used units have been the phoneme and the word. The syllable, being halfway between these two units, has the advantages of both to a degree. It is indeed the only unit which is easy to detect in continuous speech, and one in which the context dependence is somewhat eliminated. One additional advantage of using the syllable is its being a prosodic unit; it is the smallest unit that prosodic features are carried on. Stressed syllables are of great importance in recognition, and using the syllable as a unit enables one to get easy access to this information. For these reasons, attempts have been made to use the syllable as a unit in some recognition systems [21]. The main drawback to using the syllable is its being a unit not

| Unit | Advantages | Disadvantages |
|---|---|---|
| Phoneme | 1. Total number is small<br>2. Suitable representation for lexical entries | 1. Hard to detect acoustically<br>2. Some sounds belong to more than one phoneme<br>3. Many rules are needed at lower and higher levels |
| Allophone | 1. Easily identifiable<br>2. No rules needed at lower level | 1. Total number is excessively large<br>2. Dependent on their environment |
| Diphone | 1. Transitional information is included<br>2. Some coarticulation rule is included | 1. Total number is large<br>2. Phonological rules are not easy to apply |
| Syllable | 1. Easy to locate<br>2. Much coarticulation rule is included<br>3. Phonological rules easier to apply<br>4. Easy access to prosodic features | 1. Total number rather large<br>2. Precise boundaries are difficult to detect |
| Word | 1. Eliminates an entire level of recognition activity | 1. Matching is difficult with large vocabularies<br>2. Junctural phonological rules are hard to characterize in lexicon entries |

Table 6.2. Advantages and disadvantages of recognition units

uniquely defined in English. In Turkish, syllable is a more basic unit and many rules of the Turkish language act upon the syllable as a whole.

A possible disadvantage is that the syllable inventory can become very large with extensive vocabularies. The size does not approach that of allophones or words ordinarily, but it far exceeds that of phonemes. To give an idea on the size of the syllable inventory, some results of a study on the count of units in a Turkish text [15] will be given. The text consists of 22,216 words (58,992 syllables). In this text, the number of different syllables was found to be 1506, 807 of them appearing in word initial position, 873 in the middle and 759 in word final position. The frequency of occurence of these syllables was such that a small number of them (60) formed about half of the text. The amount of text that can be formed with the most frequent syllables is tabulated in Table 6.3. If the probability of occurance of syllables is used in the search process, it is clear that the search time will be much less than that required for 1506 syllables. More economy can be made if the knowledge of position in the word is used. It has been shown that some syllables exist only in certain positions in the word. Data on this knowledge is given in Table 6.4.

| # syllables | amount of text | # syllables | amount of text |
|---|---|---|---|
| 10 | 16.61 % | 90 | 59.95 % |
| 20 | 26.81 % | 100 | 62.26 % |
| 30 | 34.44 % | 150 | 71.12 % |
| 40 | 41.56 % | 200 | 77.21 % |
| 50 | 46.05 % | 250 | 81.41 % |
| 60 | 50.85 % | 300 | 84.40 % |
| 70 | 54.01 % | 400 | 87.01 % |
| 80 | 57.60 % | 500 | 92.10 % |

Table 6.3. Frequency of syllables in text

| Position in the word | # syllables | amount of text |
|---|---|---|
| initial | 10 | 24.04 % |
| | 45 | 52.42 % |
| middle | 10 | 25.55 % |
| | 40 | 51.83 % |
| final | 10 | 19.29 % |
| | 40 | 51.37 % |
| monosyllabic word | 5 | 44.22 % |
| | 10 | 56.28 % |

Table 6.4 Frequency of syllables in certain positions

The last disadvantage of using the syllable in recognition has been the lack of methods to detect syllable boundaries, namely, syllable segmentation. A method has been developed for syllable segmentation.

## 6.3.2 Syllable Segmentation

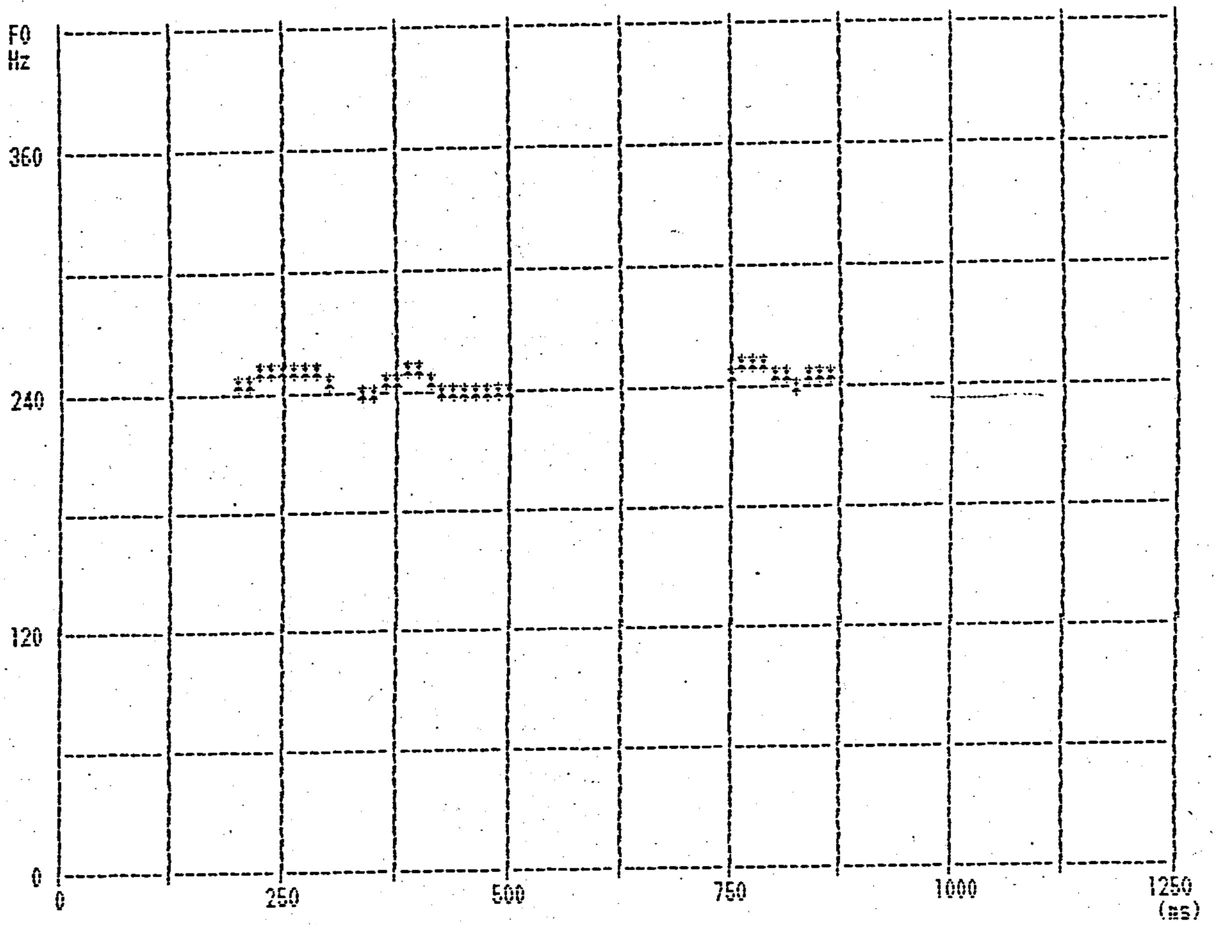
Syllables are usually defined as high energy chunks which correspond to voiced sections. Detection of the syllabic nuclei is straightforward based on this definition; they are manifested as voiced regions which last long enough (30 ms or more).

One existing system for finding syllables [32] locates syllabic nuclei by detecting high sonorant energy (energy in the 70-300 Hz band) regions bounded by substantial (4 or 5 dB) dips in energy. It then detects beginning and ending points of the syllabic nuclei as the halfway points in the dips. This algorithm is reported to detect 91 % of the syllables with only 1 % false detections of nuclei. Another similar program which uses a spectrally weighted loudness function was reported to detect 92 % of the syllabic nuclei.

A different approach which is based on the same principles but instead of filtering the speech signal, makes use of the fundamental frequency in finding the syllabic nuclei has been used in this study. The syllable structure of Turkish is such that there will be a vowel at the nucleus of each syllable, and these vowels will be manifested by long sections of voicing. The algorithm uses these sections as candidates of syllabic nuclei and the energy waveform to find the syllable endpoints; it accepts each local minimum between two sections of voicing as a syllable boundary point. This algorithm usually works, because the voiced consonants (which are causes of possible false detections of nuclei) are always next to a vowel, and during articulation of the vowel and the voiced consonant next to it, no discontinuity in voicing long enough to be detected occurs, and even if this occurs, there will be no local minimum in the energy waveform corresponding to this discontinuity.

One example word is given in Figure 6.4, where the the fundamental frequency and energy curves are plotted for the utterance "birlestir". It is observed that the sections of voicing coincide with the syllabic nuclei. The algorithm detects the endpoints of the syllabic nuclei that last long enough, and the endpoints for each nucleus are used by another algorithm that uses these data to find the exact endpoints of syllables from the energy contour. The flowchart of the syllabic nuclei detection routine is given in Figure 6.5.

This algorithm has been used to segment into syllables the 19 words (consisting of 37 syllables) of the vocabulary of the syllable-based isolated word recognition system described in Section 6.5, where it detected 81 % of the syllables (with no false detections), failing only in those words that consisted of all-voiced sequences, where all the consonants are voiced, and no discontinuity in voicing is detected. An example word where this occurs is in the word "cevir", for which the fundamental frequency and energy contours are given in Figure 6.6. In this case, the syllable boundary is marked by a sharp increase in $F_o$. When this property is utilized by the syllable detection algorithm, the performance increases to 89 %, with no false detections.



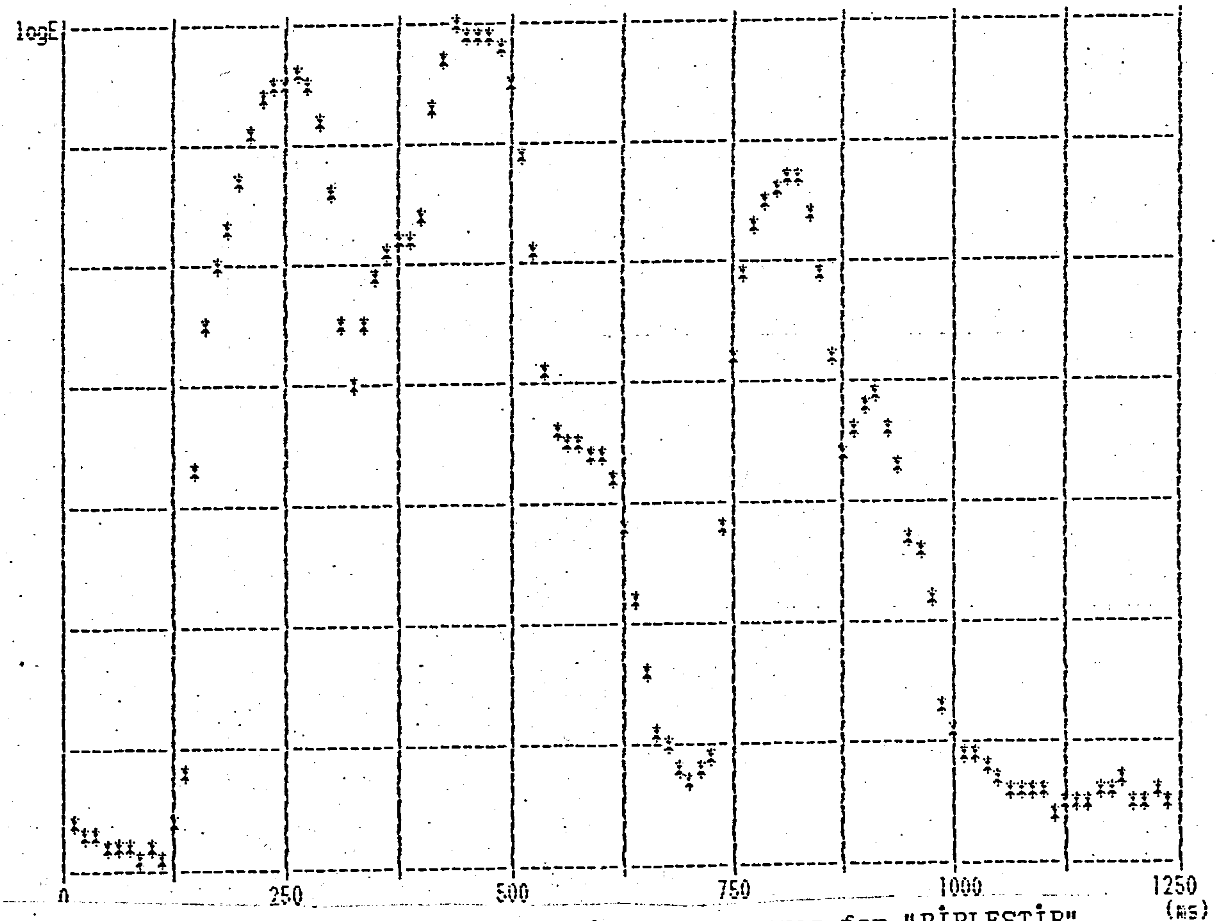


Figure 6.4. Fundamental frequency and energy curves for "BİRLEŞTİR"

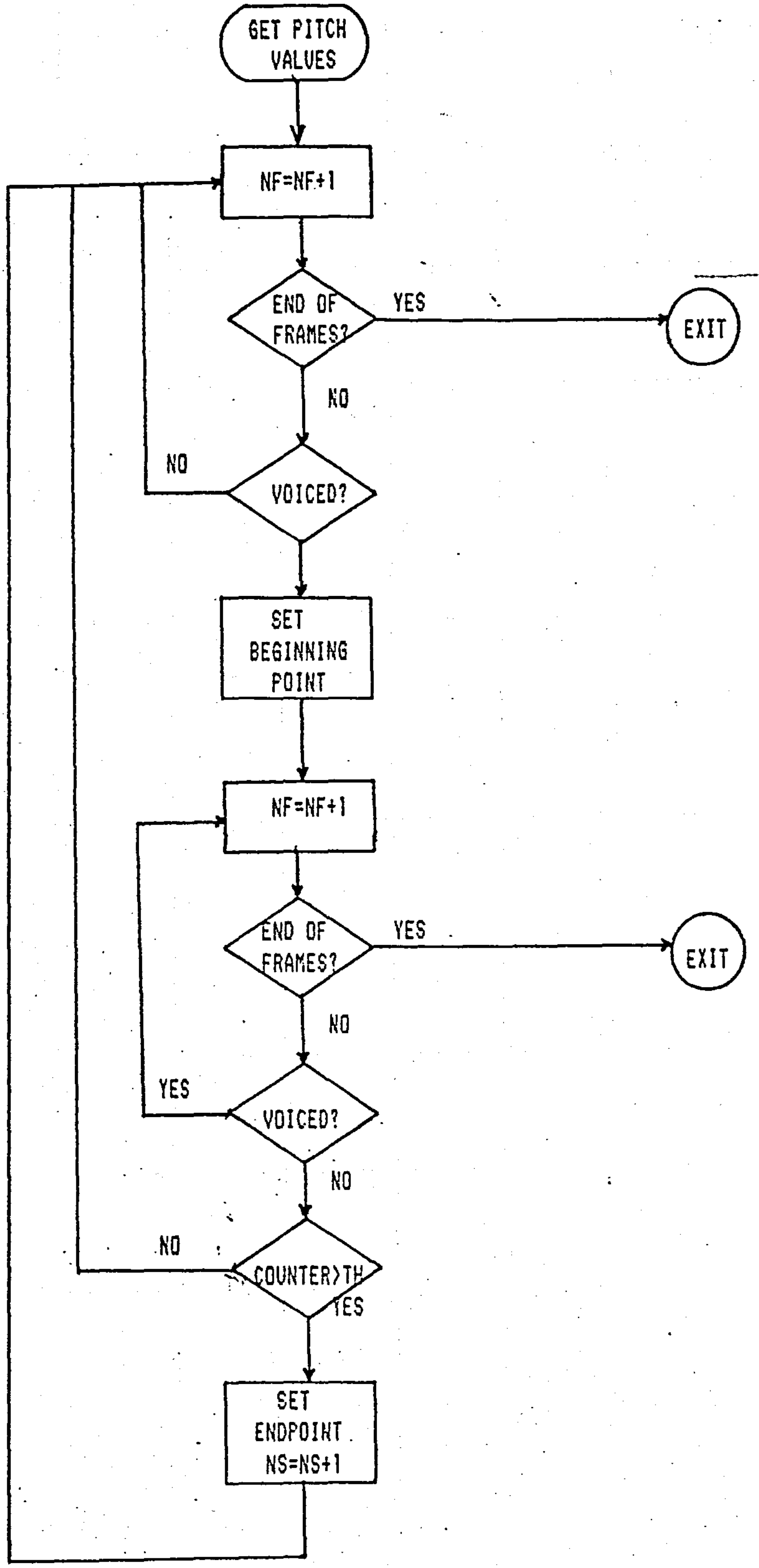


Figure 6.5. Flowchart of the syllabic nuclei detection algorithm

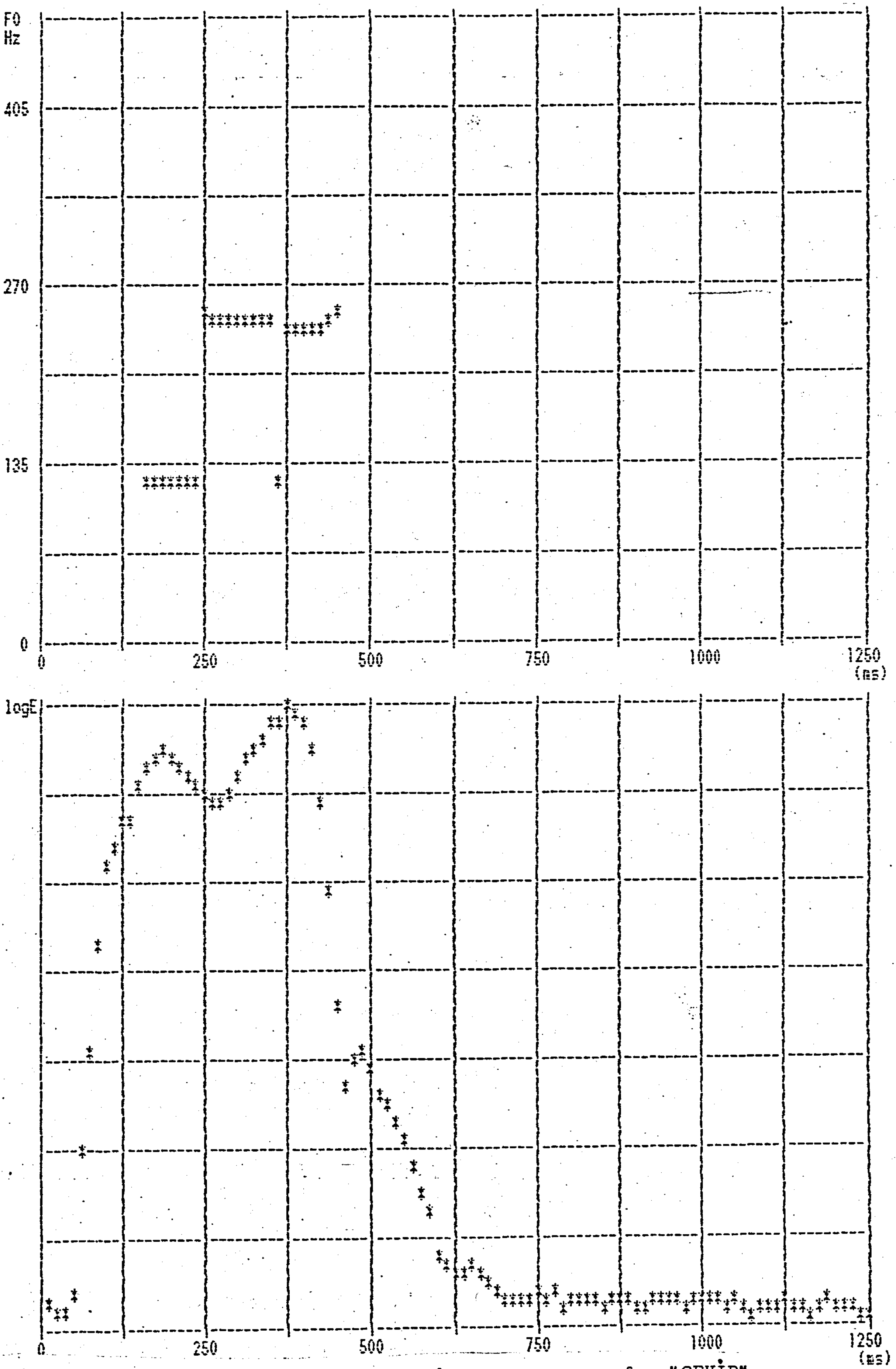


Figure 6.6. Fundamental frequency and energy curves for "ÇEVİR"

## 6.4 PROSODIC FEATURES

The possible uses of prosodic features in recognition systems was discussed in Section 5.2. Here, methods will be described to detect those features in speech, and results will be presented.

### 6.4.1 Intonation

Possible uses of intonation in recognition systems were outlined as:

-Segmentation of continuous speech into phrases;

-Extracting grammatical cues about sentences.

As continuous speech was not avaliable, segmentation could not be investigated. However, the grammatical intonation contours (Tune I and II contours of Figure 5.4) were observed in short sentences.

As noted in Section 5.2, Tune I contour marks yes/no questions (Figure 6.7), and Tune II contour marks questions with interrogative words (Figure 6.8).

### 6.4.2 Stress

Stress is the most important prosodic parameter in use in recognition systems. It is being used in many present systems for the purposes discussed in section 5.2. The main physical correlates of stress are fundamental frequency, duration and energy. It has been observed that stressed syllables are usually articulated with longer duration and higher intensity as well as an increasing fundamental frequency. Although energy and duration show a characteristic increase in stressed syllables, these parameters have proved insufficient for the

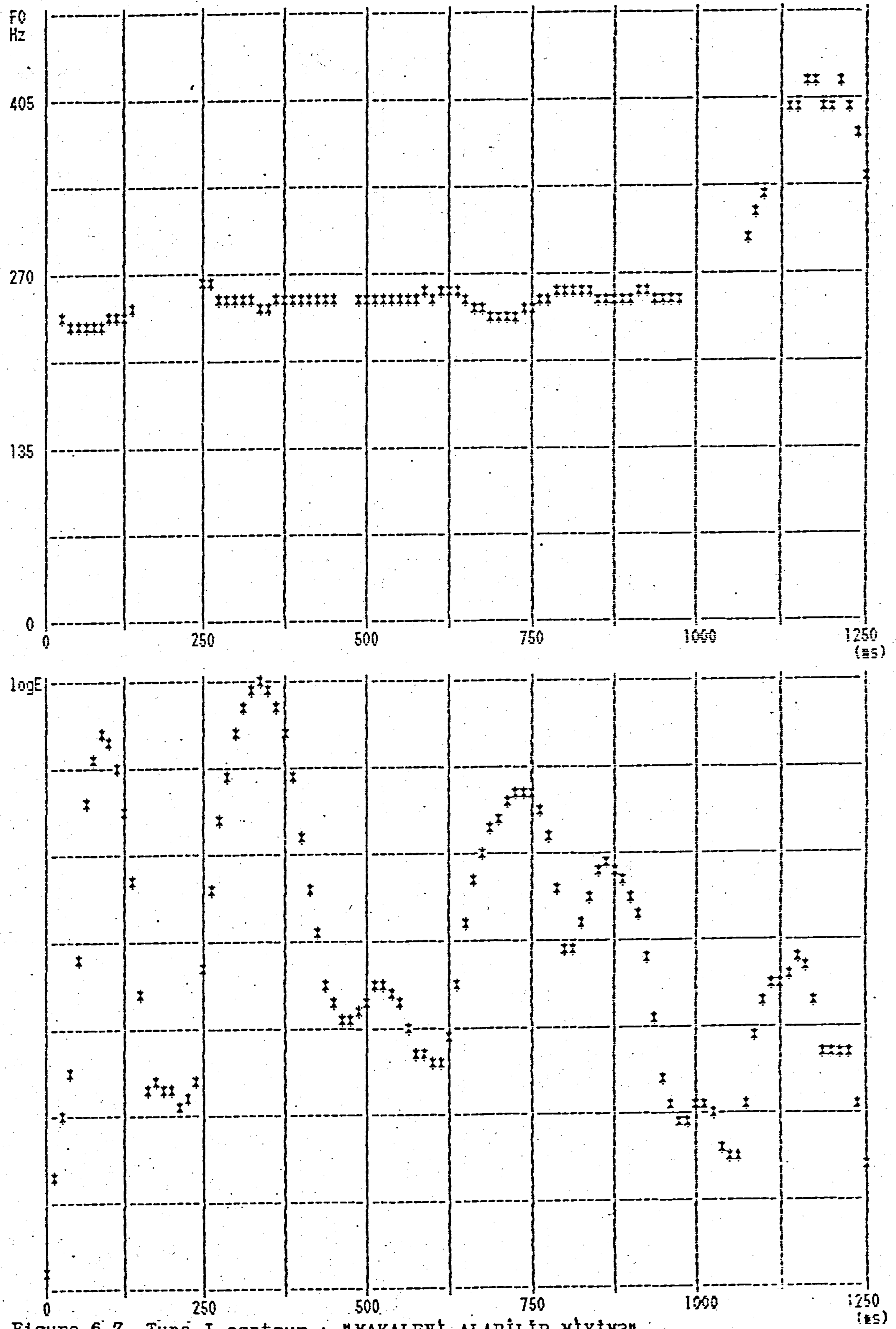


Figure 6.7. Tune I contour : "MAKALENİ ALABİLİR MİYİM?"

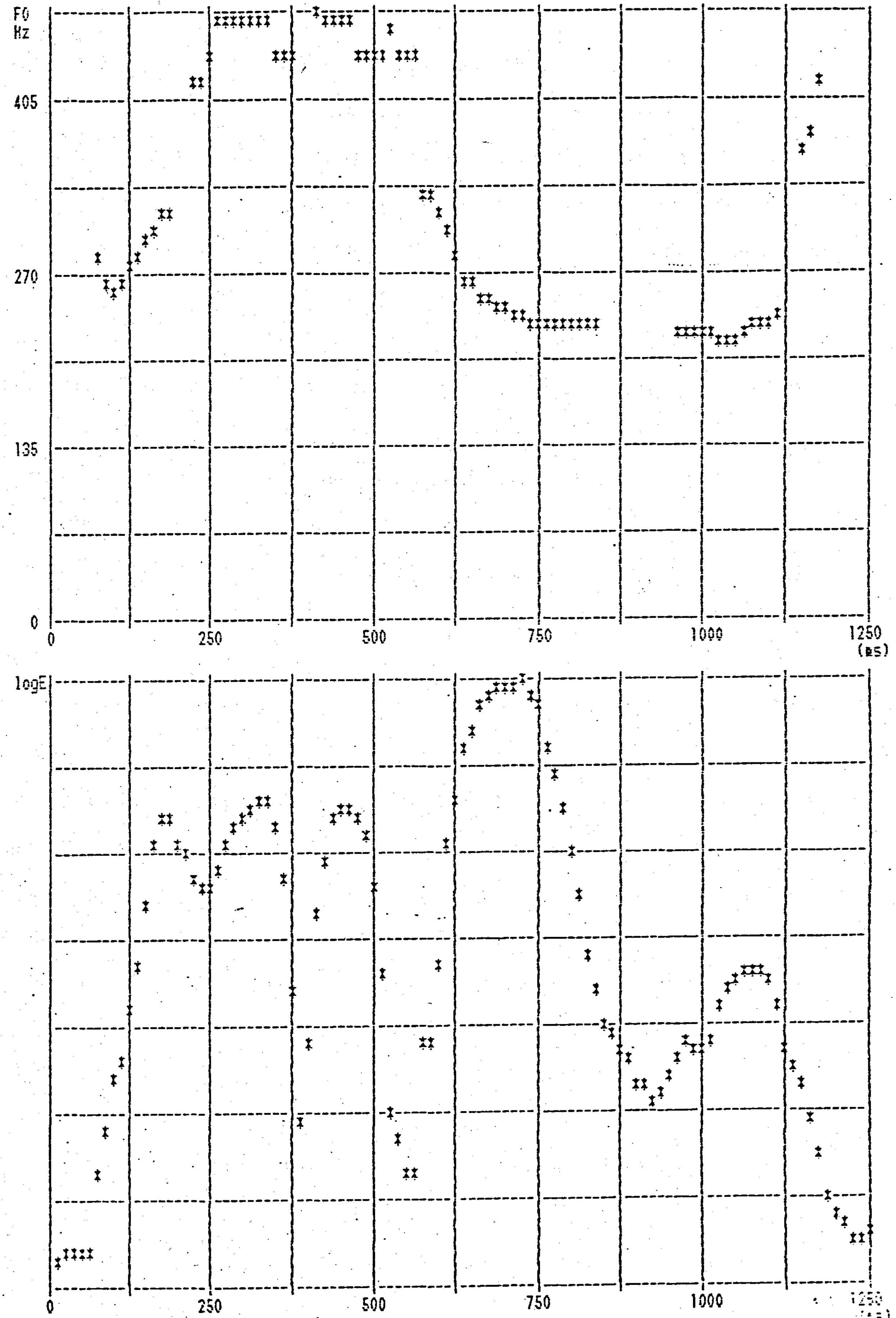


Figure 6.8. Tune II contour : "NEREDE KALDINIZ?"

detection of stressed syllables. It seems that the main characteristic that enables their detection is a local rise in the fundamental frequency contour which is naturally accompanied by increases in energy and duration. If isolated from the effects of intonation, stress will usually be revealed by a local rise in $F_o$. However, with intonation imposed on the stress patterns, the situation will slightly change; the intonation patterns will cause the pitch contour to take the general form of Tune I or Tune II contours (Figure 5.4) . The stresses will then have the effect of increasing $F_o$ locally at stressed syllables. If this happens during the time intonation contours are rising, this will show as a sharp rise in $F_o$; and if this occurs during fall of intonation contours, stresses wil show as local rises above the gradually falling $F_o$ contour; even if $F_o$ does not rise absolutely near the stressed syllable.

Figures 6.9 and 6.10 show the fundamental frequency and energy contours for the two different pronounciations of the same word, "*konuşma*". In Figure 6.9, the second syllable is stressed due to the negation suffix, while in Figure 6.10, the third syllable is stressed, because the suffix in this word is a regular one. It is observed that both amplitude, duration and fundamental frequency contours are different in the two graphs. These values have been averaged for different utterances of the same words and the results are given in Table 6.5.

It can be seen clearly that the amplitude and duration show cues to the stressed syllables, but these are hard to use in the detection of stressed syllables. In the word "*konuşma*", although the amplitude and duration of the second syllable is relatively higher than it is in

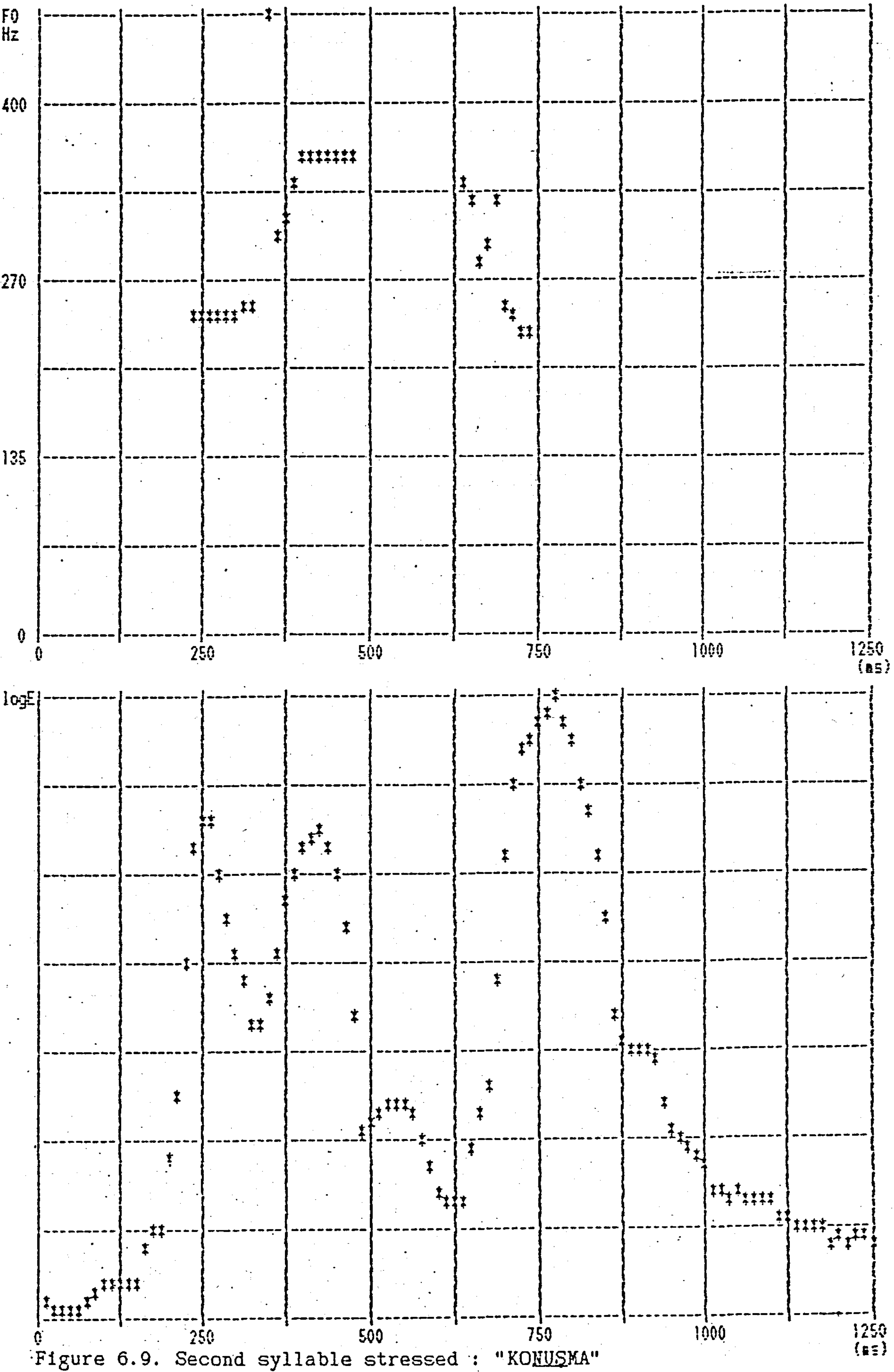


Figure 6.9. Second syllable stressed : "KO<u>NUŞ</u>MA"

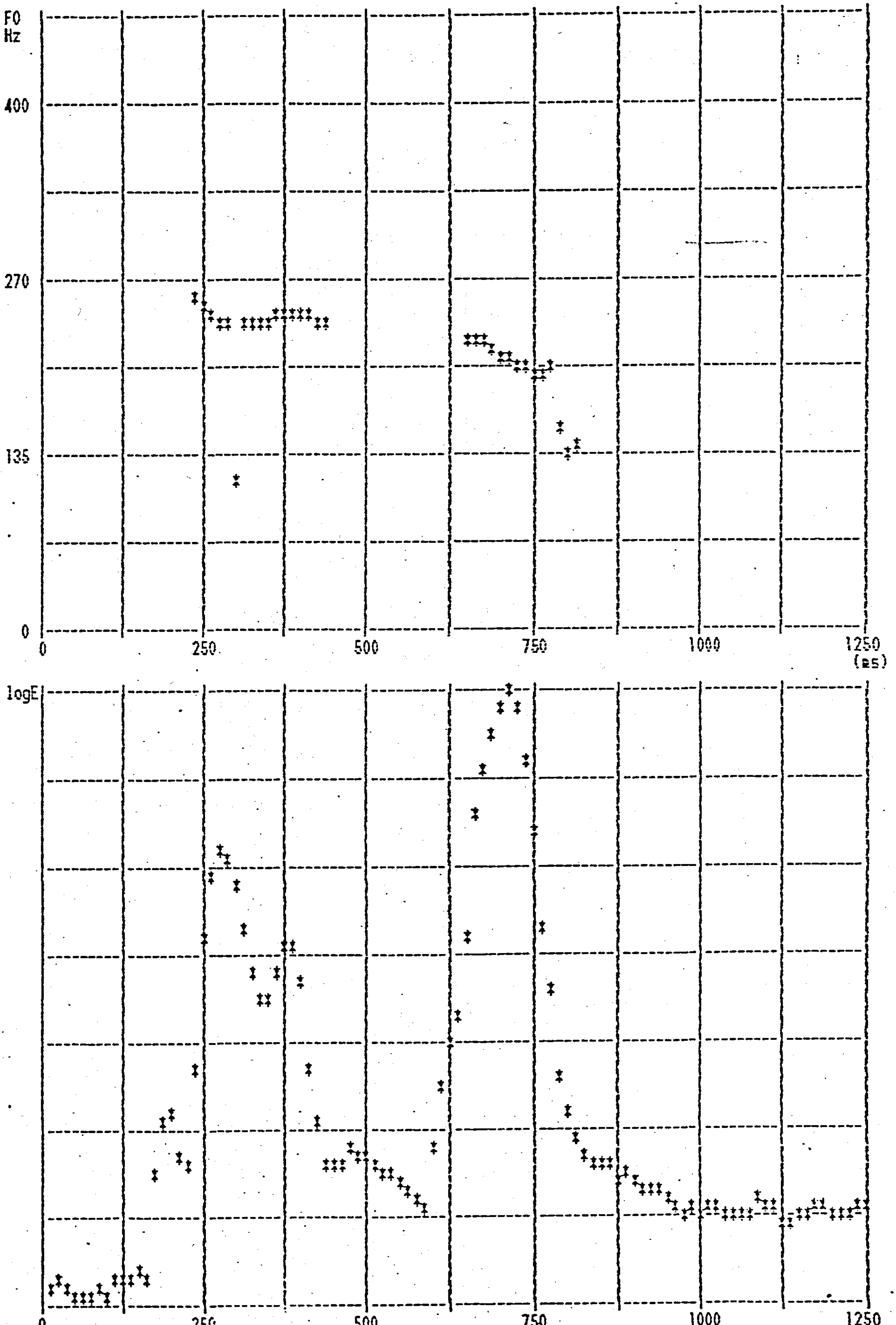


Figure 6.10. Last syllable stressed : "KONUŞMA"

"*konusma*",the syllable with the highest amplitude and longest duration is still the third syllable; it is the fundamental frequency contour that shows useful for the detection of stressed syllables.

| | Average Duration ( % ) | | | Average Amplitude | | | Pitch Contour | | |
|---|---|---|---|---|---|---|---|---|---|
| | syl, 1 | syl, 2 | syl, 3 | syl, 1 | syl, 2 | syl, 3 | syl, 1 | syl, 2 | syl, 3 |
| syl, 2 stressed | 26 | 29 | 45 | 0,8 | 0,8 | 1,0 | falling | rising | falling |
| syl, 3 stressed | 28 | 25 | 47 | 0,7 | 0,5 | 1,0 | falling | level | rising |

Table 6.5. Variation of duration, amplitude, and pitch with stress

An increase in fundamental frequency is taken as the main indication of a stressed syllable. There are cases where no significant increase in fundamental frequency is observed, and in these cases, taking the syllable that is highest in amplitude and longest in duration as the stressed syllable gives good results.

### 6.4.3. Duration

Duration information is used in segmentation and labeling schemes. The expected values of sounds are important parameters that are used in various steps of a recognition process. These values, however, are highly context-dependent and must be obtained considering all sorts of environments. In a syllable-based system, duration of syllables may be a more reliable measure because this dependency is included.

The average durations of syllables for the vocabulary of the isolated word recognition system (Section 6.5) have been calculated and this information has been used in the matching step.

| SYLLABLE | Dav | σ | SYLLABLE | Dav | σ | SYLLABLE | Dav | σ |
|---|---|---|---|---|---|---|---|---|
| İ- | 10 | 1,9 | SAK- | 19 | 2,1 | -TI | 33 | 6,9 |
| SI- | 10 | 2,1 | GİR- | 25 | 2,9 | -VİR | 33 | 4,9 |
| -Nİ- | 12 | 1,1 | BAŞ- | 26 | 1,1 | -Kİ | 35 | 2,2 |
| SE- | 13 | 1,7 | AL- | 26 | 2,4 | -KUZ | 35 | 5,5 |
| GE- | 13 | 0,7 | -KİZ | 26 | 8,1 | -LA | 36 | 6,6 |
| DO- | 14 | 1,2 | -LEŞ- | 27 | 2,6 | -TİR | 36 | 6,6 |
| -RA- | 14 | 0,9 | BİR | 30 | 10,5 | BEŞ | 38 | 3,7 |
| YE- | 17 | 2,4 | -FIR | 30 | 2,1 | DÖRT | 40 | 4,1 |
| ÇIK- | 17 | 2,3 | -Dİ | 32 | 5,6 | ÜÇ | 40 | 5,6 |
| ÇE- | 18 | 2,0 | -DEN | 32 | 4,5 | | | |

Table 6.6. Duration of syllables

The average durations and standard deviations of syllables have been given in Table 6.6. All the figures in the table are number of 12.5 ms frames. It is observed that the syllables show a characteristic duration with small standard deviations. Cases where the standard deviation is relatively high are generally the final syllables and this uncertainty is due to final breath noise. Also, the syllables are slightly longer in final position than in initial position. This has been the cause of the relatively large standard deviation observed in the syllable "BIR".

These regularities have been utilized in the matching step of the isolated word recognizer. The average duration of each syllable is stored and the unknown utterance is not tested against those syllables for which the ratio of unkown duration to syllable duration is outside specified limits. It is assumed that the duration of a syllable can only change 40 % from the average and this ratio is used in matching. The use of durations in this way provides a time saving of about 52 % .

## 6.5. PROSODICALLY AIDED ISOLATED WORD RECOGNITION SYSTEM

A speaker independent isolated word recognizer based on the discussed ideas has been realized. The minimum recognition unit has been chosen as the syllable. Syllable segmentation has been performed by the technique described in the previous section. Dynamic time warping technique has been used in syllable verification. Duration and vowel harmony information have been used in syllable matching. The block diagram of the system is given in Figure 6.11.

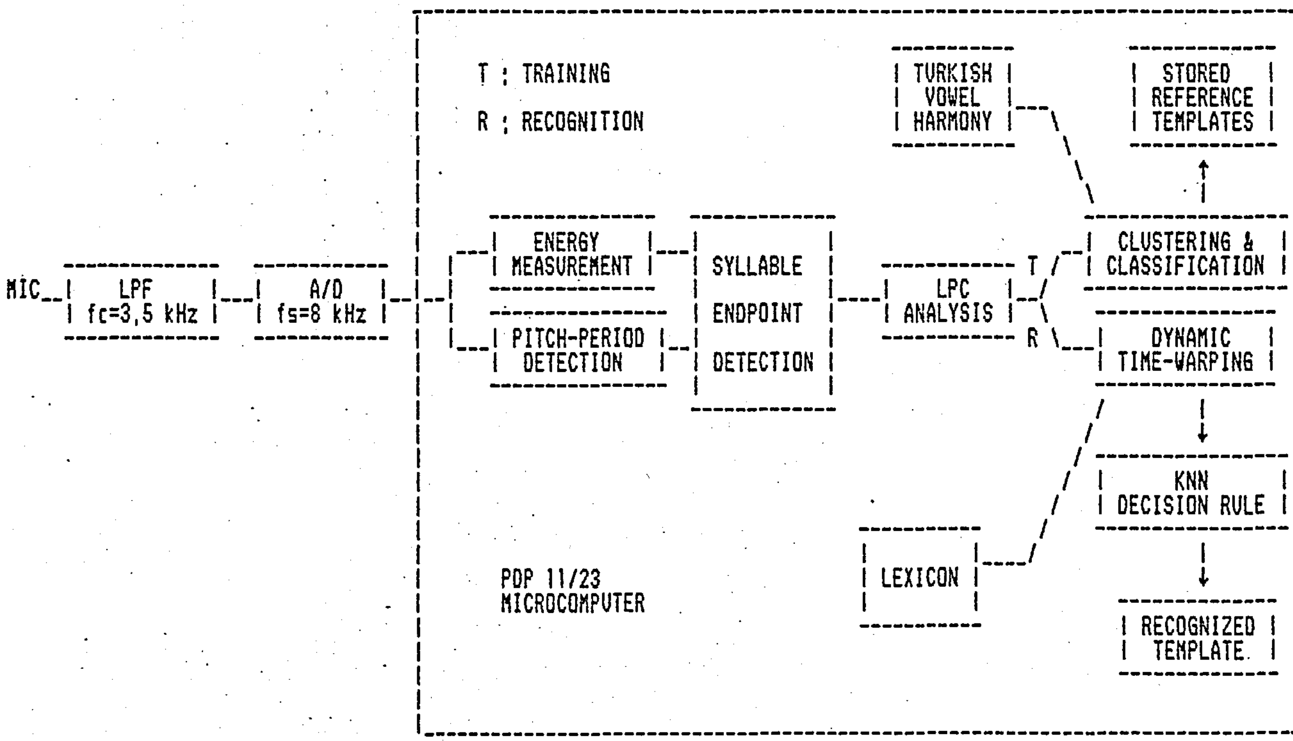


Figure 6.11. Overall block diagram of the word recognition system

The vocabulary consists of 19 words, 10 digits and 9 polysyllabic commands. The total number of syllables is 37, and the number of different syllables is 29. The redundancy is much less than it is normally, because of the small size of the vocabulary. The saving introduced in memory because of using syllables instead of words is 22%. Both a time saving and improvement in performance is also achieved due to using syllable as a unit. These are summarized in Table 6.8.

Syllable segmentation is performed with the technique described in Section 6.3. Syllables are classified in 8 groups according to the vowel they contain. Classification of syllables can be seen in Table 6.7.

|  | UNROUNDED |  | ROUNDED |  |
|---|---|---|---|---|
|  | WIDE | CLOSE | WIDE | CLOSE |
| BACK | -1-<br>BAŞ<br>SAK<br>AL<br>LA<br>RA | -2-<br>ÇIK<br>FIR<br>TI<br>SI | -3-<br>DO | -4-<br>KUZ |
| FRONT | -5-<br>BEŞ<br>LEŞ<br>DEN<br>YE<br>SE<br>ÇE<br>GE | -6-<br>KİZ<br>VİR<br>TİR<br>BİR<br>GİR<br>Kİ<br>Nİ<br>Dİ<br>İ | -7-<br>DÖRT | -8-<br>ÛÇ |

Table 6.7. Classification of the syllables

The system has two modes of operation; training mode and recognition mode. In the training mode, recordings are made of the different utterances (by different speakers) of the vocabulary words. These are then analyzed; syllable segmentation is made and a feature set of LPC coefficients (p=10) is prepared for each syllable. Among the different sets representing seperate utterances of the same syllable, a subset which optimally represents the syllable is chosen. This process is called 'clustering'. The feature sets of the chosen utterances are stored in memory and these are called 'reference templates'.

In the recognition mode, an unknown utterance is input to the system. This unkown utterance is processed in the same way to segment it into its syllables and extract features to prepare a 'test template'. This test template is then compared with the reference templates in the memory and a 'score' is associated with each according to a distance measure. Both Euclid distance measure and LPC log likelihood distance measure have been used. In the comparison, linear time warping followed by dynamic time warping is used to perform time normalization. Duration and vowel harmony information is used to eliminate unlikely matches. This substantially reduces the computation time (Table 6.8). After the test template is compared with all permissible reference templates in memory, a decision is made according to K-nearest neighbor decision rule; the K minimum distances for each compared syllable are added and the one with the total score is announced to be the recognized syllable.

| | MEMORY SAVING INTRODUCED | TIME SAVING IN DTW | IMPROVEMENT IN PERFORMANCE |
|---|---|---|---|
| SYLLABLE AS A UNIT | 22 % | 3-8 % | around 10 % |
| 40% DURATION THRESHOLD | - | 52 % | 2-3 % |
| VOWEL HARMONY | - | 35 % | around 5 % |

Table 6.8. Improvements due to prosodic aids to recognition

## 6.6. SOME STRATEGIES FOR CONTINUOUS SPEECH RECOGNITION

In this study, analyses were made on 2-s utterances which consisted mainly of single isolated words and a few short sentences. The observations on these sentences led to some strategies for a continuous speech recognition system. These will be shortly discussed here. However, extensive analyses must be made for their justification.

In continuous speech, there is very little (if any) evidence of word boundaries. Due to the increased difficulty of endpoint detection and increased size of the vocabulary, the word as a unit loses its appeal. The phoneme may be another alternative, but in addition to the disadvantages discussed before, there is the additional problem that in continuous speech, especially when the speaking rate is high, many phonemes may be missing. These considerations lead to the conclusion that syllable is the most convenient unit to be used in continuos speech recognition. Syllable boundaries can still be extracted from the energy waveform, so the speech can again be segmented into syllables using similar techniques. So, the syllable can again be used as a unit with some methods to deal with the above problems. Rules must be incorporated into the system to account for missing sounds in fast speech. Allowing for different pronounciations of the syllables might be one way of dealing with this problem.

In continuous speech, since the word boundaries are not known, complicated procedures must be used for word matching. In a phoneme or syllable based system, word matching is done at a symbolic level. Once a sequence of syllables have been recognized, these are compared at the symbolic level with the words in lexicon. To do this, one has to find a way to hypothesize non-overlapping sets of words from these sequences. What is usually done is to proceed from left to right and go on trying all possible sequences while allowing for missing or errorful segments. Then, assuming that sentence boundaries are known, competing hypotheses will be formed for each sentence. These are tested so as to prove grammatically and semantically meaningful, and the ones that do are compared. The hypothesis with the minimum total distance is chosen to be the recognized sentence. The stress structure of Turkish has one regularity which can introduce a very important convenience to this procedure. In Turkish, one strong stress is assigned to each word, and this is ordinarily placed on the final syllable. So, a stressed syllable usually marks the final syllable of a word; and words in a sentence can be segmented using this property. There are, of course, exception words which must be handled carefully. Exceptions are usually clearly defined; the most common class is certain suffixes which cause stress to be assigned on the syllable preceding them. There are some other classes of loan words or names of places which are more difficult to handle; so, a more convenient approach would be to again hypothesize words from left to right, but to score the hypotheses according to their stress structures. Those hypotheses in which some words have more than one stress can be ruled out. This will reduce the number of hypotheses substantially. In this context, vowel harmony can also be used. It is known that words of Turkish origin contain vowels from certain groups throughout. If the syllables are grouped as done in Table 6.7, this information can be used in a syllable-based recognition system. If the structure of the vocabulary is suitable, hypotheses that contain words which do not obey vowel harmony rules can be ruled out, or some convenient score may be assigned accordingly. The two properties proposed for use in word segmentation, stress and vowel harmony, have also been classified by linguists as properties that define words [6].

The rate of speech is another parameter that is used in a continuous speech recognition system to account for different rules for fast speech. In English, the most common measure used for rate of speech is the number of stresses in unit time. As discussed above, each word is marked with a stress in Turkish, and for this reason, the number of stresses per unit time is a measure of word rate, not speech rate. Some linguists [5] point out that syllables are assigned equal time in Turkish. If this is true, number of syllables per unit time can be used as an indication of speech rate.

Prosodic aids are, in fact, part of a linguistic framework used for speech recognition system. They have proved to be very useful for the isolated word recognition system discussed. A linguistic framework in which all the prosodic aids discussed with the addition of many others is more essential in a continuous speech recognition system. A linguistic framework, in addition to improving system performance, makes the system easily expandable with the enhancement of syntactic and semantic analysis capabilities.

# VII. CONCLUSION

In this study, algorithms have been developed to extract prosodic parameters from the speech signal. The prosodic structures of Turkish have been investigated for use in speech recognition systems and some of the ideas have been realized in an isolated speech recognition system. The basic conclusions drawn in each step of the analysis can be summarized as follows.

The Autocorrelation method using center clipping (AUTOC) provides a simple and reliable method of pitch period detection. Parallel processing (PPROC) method is also remarkable for its speed in implementation, but AUTOC has been favored in this study because of its reliability.

Syllable is a very suitable unit for automatic recognition of Turkish. It has many advantages both in isolated and connected speech. The algorithm developed for syllable segmentation has shown considerable success; it has detected 89 % of the syllable endpoints.

The prosodic structures of Turkish, namely, duration, stress, intonation, and vowel harmony can be used in automatic speech recognition of Turkish in the following ways:

-Duration of a syllable changes very little from an expected duration. This property can be used in word matching.

-Stress can be detected using fundamental frequency, energy, and duration. This information can be used in word hypothesization.

-Intonation contours give cues about grammatical functions of sentences.

-Vowel harmony can be used to group syllables. Matching and verification can be made within these groups. This reduces the computation time substantially. Vowel harmony information can also be used in word hypothesization.

Some of the above prosodic aids have been incorporated in a syllable-based isolated word recognition system. Both time and memory savings and an improvement in performance have been obtained due to these.

## 7.1. SUGGESTIONS FOR FURTHER WORK

All of the above prosodic aids may be incorporated in a continuous speech recognition system. The method of word hypothesization suggested in Section 6.6 can be tried in such a system. The method suggested for finding rate of speech in Turkish should be tested with carefully prepared test data to prove its validity. Intonation contours should be carefully examined for possible uses in continuous speech recognition systems. Possible use of intonation contours in segmenting Turkish sentences into their grammatical constituents should be investigated.

The performance of the syllable segmentation method may be improved if smaller segments of analysis are used. More complicated algorithms may also be used to deal with those phenomena using the information on the energy waveform only.

# REFERENCES

[1] L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Englewood Cliffs, NJ : Prentice-Hall, 1978.

[2] J.L. Flanagan, Speech Analysis, Synthesis and Perception, Heidelberg, Berlin: Springer-Verlag, 1972.

[3] J.D. Markel and A.H. Gray, Linear Prediction of Speech, New York : Springer-Verlag, 1976.

[4] Ö. Demircan, Türkiye Türkcesinin Ses Düzeni Türkiye Türkcesinde Sesler, Ankara : TDK Yayınlari, 1979.

[5] R.B. Lees, The Phonology of Modern Standard Turkish, Indiana University Publications, 1961.

[6] S. Barav, Yapısal Dilbilimi, Edebiyat Fakültesi Basımevi, 1969

[7] F. R. Palmer, Prosodic Analysis, London : Oxford University Press, 1970.

[8] W.A. Ainsworth, Mechanisms of Speech Recognition, Oxford : Pergamon Press, 1976.

[9] P. Ladefoged, Preliminaries to Linguistic Phonetics, Chicago : The University of Chicago Press, 1971.

[10] G. Fant, Speech Sounds and Features, Cambridge, Massachusetts : The MIT Press, 1973.

[11] W.E. Jones and J. Laver, Phonetics in Linguistics, London : Longman, 1973.

[12] B. Malmberg, Manual of Phonetics, Amsterdam : North Holland Publishing Company, 1974.

[13] V.B. Makkai, Phonological Theory, Chicago : Holt, Rinehart and Winston, 1972

[14] E. Sezer, "On non-final stress in Turkish", J. Turkish Studies, vol.5, pp.61-70, 1981.

[15] G. Gönenç and E. Töreci, "Türkçenin bazı özelliklerinin bilgisayarla çözümlenmesi", Bilişim Dergisi.

[16] Ö. Demircan, "Bileşik sözcükler ve bileşik sözcüklerde vurgu"

[17] Ö. Demircan, "Türkiye yer adlarinda vurgu"

[18] Ö. Demircan, "Türkçe ezgilemeye giriş"

[19] Ö Demircan, "Türkiye Türkçesinde vurgulama ve odaklama"

[20] Ö. Demircan, "Türk dilinde ek vurgusu"

[21] W.A. Lea, Trends in Speech Recognition, Englewood Cliffs, NJ : Prentice-Hall, 1980.

[22] W.A. Lea, M.F. Medress and T.E. Skinner, "A Prosodically Guided Speech Understanding Strategy", IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-23, pp.30-38, February 1976.

[23] O. Fujimura, "Syllable as a unit of speech recognition", IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-23, pp. 82-87, February 1975.

[24] D.H. Johnson and C.J. Weinstein, "A phrase recognizer using syllable based acoustic measurements", IEEE Trans. Acoust., Signal Speech Processing, vol. ASSP-26, pp. 409-418, October 1978.

[25] D.R. Reddy, "Speech recognition by machine", Proc. IEEE, vol.64, pp.501-531, April 1976.

[26] V.W. Zue, "The use of speech knowledge in automatic speech recognition", Proc. IEEE, vol.73, pp.1602-1615, November 1985.

[27] S.E. Levinson, "Structural methods in automatic speech recognition", Proc. IEEE, vol. 73, pp. 1625-1650, November 1985.

[28] L.R. Rabiner and S.E. Levinson, "Isolated and connected word recognition-theory and selected applications", IEEE Trans. Commun., vol.COM-29, pp. 621-659, May 1981.

[29] J. Allen, "A perspective on man-machine communication by speech", Proc.IEEE vol. 73, pp. 1541-1550, November 1985.

[30] J.L. Flanagan, "Computers that talk and listen : Man-machine communication by voice", Proc. IEEE, vol.64, pp.405-415, April 1976.

[31] J.L. Flanagan, "Talking with computers : Synthesis and recognition of speech by machines", IEEE Trans. Biomed. Engineering, vol.BME-29, pp. 223-232, April 1982.

[32] W.A. Lea, "Speech recognition: past,present and future", [21].

[33] J.E. Shoup, "Phonological aspects of speech recognition", [21].

[34] W.A. Lea, "Prosodic aids to speech recognition", [21].

[35] M.F. Medress, "The Sperry Univac system for continuous speech recognition", [21].

[36] W.A. Lea, "Speech recognition: What is needed now?", [21]

[37] R.W.Schafer and L.R. Rabiner, "Digital representation of speech signals", Proc. IEEE, vol.63, pp.662-677, April 1975.

[38] J. Makhoul, "Linear Prediction : A tutorial review", Proc. IEEE, vol.63, pp.561-580, April 1975.

[39] W. Hess, Pitch Determination of Speech Signals, Berlin,Heidelberg: Springer-Verlag, 1983.

[40] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal, "A comparative performance study of several pitch detection algorithms",IEEE Trans. Acoust.,Speech, Signal Processing, vol.ASSP-24, pp. 399-418, October 1976.

[41] C.A. McGonegal, L.R. Rabiner, A.E. Rosenberg, "A subjective evaluation of pitch detection methods using LPC synthesized speech", IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-25, pp. 221-231, June 1977.

[42] M. Sondhi, "New methods of pitch extraction", IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 262-266, June 1968.

[43] L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection", IEEE Trans Acoust., Speech, Signal Processing, vol. ASSP-25, pp. 24-33, February 1977.

[44] J.J. Dubnowski, R.W. Schafer and L.R. Rabiner, "Real-time digital hardware pitch detector", IEEE Trans. Acoust. Signal, Speech Processing, vol. ASSP-24, pp. 2-8, February 1976.

[45] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg and H.J. Manley, "Average Magnitude Difference Function Pitch Extractor", IEEE Trans Acoust., Speech, Signal Processing, vol. ASSP-22, pp. 353-361, October 1974.

[46] B. Gold, L.R. Rabiner, "Parallel Processing Techniques for estimating pitch periods of speech in the time domain", Speech Analysis, IEEE Press, 1978.

[47] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation", IEEE Trans. Audio Electroacoust., vol. AU-20, pp. 367-377, December 1972.

[48] A.M.Noll, "Cepstrum pitch determination", Speech Analysis, IEEE Press, 1978.

[49] F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Trans. Acoust., Speech, Signal Processing, vol ASSP-23, pp 67-72, February 1975.

[50] H. Sakoe and S. Chiba, "Dynamic Programming algorithm optimization for spoken word recognition",IEEE Trans. Acoust., Speech, Signal-Processing, vol ASSP-26, pp 43-49, February 1978.

[51] S. Bingöl, "A nonuniform sampling approach to data compression of speech and voiceband data signals", M.S. thesis, Boğaziçi University, 1985.

[52] C. Ersoy, " Prosodically guided speaker independent isolated Turkish word recognizer", M.S. thesis, Boğaziçi University, 1986.