

Multi-Speaker Computer Recognition
of Ten Connectedly Spoken Letters

Jianli Sun

A Thesis

in

The Department

of

Computer Science

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Computer Science at
Concordia University
Montréal, Québec, Canada

May 1985

© Jianli Sun, 1985

ABSTRACT

Multi-Speaker Computer Recognition of Ten Connectedly Spoken Letters

Jianli Sun

An automatic speech recognition system for recognizing connected letters belonging to the set

[E1 = (P, T, K, B, D, V, E, G, C, 3.)]

is described. Half of the letters in the E1 set are characterized by plosive. An expert system approach to segmentation of continuous speech which is based on a Semantic-Syntax-Directed-Transition algorithm for translating primary acoustic cues into primary phonetic features has been implemented. Acoustic property extraction and feature hypothesization are performed by a Planning System. Classification is characterized by a fuzzy algorithm which is an application of fuzzy set theory in pattern recognition. The system is tested on a protocol of 1000 connected pronunciations of symbols of the E1 set in strings of five symbols each. The strings were pronounced by five male and five female English speakers. The average recognition rate is 90%. Various experimental results are reported.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisor, Professor Renato De Mori, Department of Computer Science of Concordia University, for his cordial encouragement, excellent guidance and valuable advice in my research and in preparation of this thesis.

I also would like to thank all of the members of the speech research group and all of the speakers invited for data collection for their assistance.

I wish to express my gratitude to my parents, my elder brother and sister, for their kindness and encouragement.

I am very grateful to Chuei Ning, my dear wife, for her love, patience, encouragement and understanding making this effort possible.

CONTENTS

TITLE PAGE	...i
SIGNATURE PAGE	...ii
ABSTRACT	...iii
ACKNOWLEDGEMENTS	...iv
TABLE OF CONTENTS	...v
LIST OF FIGURES	...vii
LIST OF TABLES	...ix
CHAPTER I. INTRODUCTION	...1
1.1 Generalities on Speech Production	...2
1.2 Speech Analysis	...4
1.3 Computer Models for Speech Recognition	...13
1.4 A Brief History of Automatic Speech Recognition	...17
1.5 Motivation And Task of This Research	...25
CHAPTER II PLOSIVE SOUNDS	...26
2.1 The Acoustic Properties of Plosive Sounds	...26
2.2 Perception of Manner of Articulation	...30
2.3 Perception of Place of Articulation	...33
2.4 Stop + Vowel Stimuli	...36
2.5 Summary	...40
CHAPTER III DATA ACQUISITION AND PREPROCESSING	...42

CHAPTER IV AN EXPERT SYSTEM APPROACH TO SEGMENTATION	...46
4.1 Structure of The Expert	...47
4.2 Representation of Expert's Knowledge	...49
4.3 Auditory Experts for Interpreting Speech Patterns	...56
4.4 Syllabic Expert for Syllabic Hypotheses	...67
CHAPTER V THE PLANNING SYSTEM FOR EI-SET RECOGNITION	...71
5.1 Why Planning	...71
5.2 The System of Plans	...72
5.3 Fuzzy Algorithm for Feature Hypothesization	...81
5.4 Hypothesis Generation Rules	...97
5.5 Hierarchical Recognition Strategy	...99
CHAPTER VI SYSTEM PERFORMANCE AND EVALUATION	...102
6.1 Experimental Results And Discussions	...102
6.2 Conclusions	...108
REFERENCES	...112
APPENDIX	...121

LIST OF FIGURES

1.2	Burst Spectra of the Plosive Sound /d/ Obtained by the Spectral Smoothing Techniques	...8
1.3.1	A Passive Model for Speech Recognition	...14
1.3.2	An Active Model for Speech Understanding	...16
2.0	Phonemes in American English	...27
2.1	Place of Articulation of Plosives	...29
3.1	Signal Preprocessing System	...43
4.1	Expert Structure	...48
4.3.1	The Expert Society for Segmentation	...57
4.3.3	The Time Evolution of the Signal Energy and the Zero-crossing Counts	...64
5.2.1	An Overview of the Plan for the Recognition of the El-Set	...73
5.2.2	Envelope Curve in the Energy Dip Preceding the Onset of /b/	...75
5.2.3	The Spectrum of the Signal Shown in Figure 5.2.2	...76
5.2.4	The Burst Peak of /k/ in the 11-20 Centisecond Interval and in the 2-4 KHz Band	...79
5.2.5	Compact Burst Spectrum of /k/	...80
5.3.2	The Fuzzy Restrictions Defined over An Acoustic Cue U	...88

5.3.3	The Membership Functions of the Fuzzy Restriction Over the Parameter in PE13	...92
6.1	System Performance Improvement by Introducing More Subplans	...103

LIST OF TABLES

1.4.1	Representative Recognition Scores for Some Isolated Word Recognizers	...19
1.4.2	Performance of Recognizers of Connected Word Sequences	...22
1.4.3	Some Highlights in the History of Speech Recognition	...24
2.2	Acoustic Cues for Manner of Articulation	...32
2.3	Acoustic Cues for Place of Articulation	...34
2.4	The Characteristics of the Gross Spectral Shapes of Plosives	...39
4.2	Rule of the Frame-Structure Grammar	...54
4.3.1	Primary Acoustic Cues	...59
4.3.2	Attribute Description of PAC	...60
4.3.3	The PAC description corresponding to the signal in Figure 4.3.3	...65
4.4	Primary Phonetic Features	...66
5.5	The Similarity Evaluation for the Connectedly Spoken Letters KCBTD	...101
6.1	The Error Distributions of the Recognition of the El-Set Letters	...106
6.2	The Confusion Table	...108

CHAPTER I

INTRODUCTION

Speech is our everyday, informal, communication medium. It would be very advantageous and practical if speech could be used as machine input and output. The problems of communicating with computers through natural speech begin with the nature of speech itself. Speech communication is natural for people but it is not the simplest communication method for machines. This depends both on the physical properties and organization of the way we talk.

Now people can extract a great deal of meaning from a very small set of sounds. Although this makes for an extremely efficient, noise-immune means of communication, science hasn't yet been able to pin down completely the characteristics of speech that give our utterances meaning. And without the knowledge, making machines that can equal our own performance as listeners becomes a large-scale task.

Acoustics, the science of heard sound, and Linguistics, the science of languages, both deal with the investigation of spoken languages. Electrical Engineering and Computer Science develop equipments, methodologies and systems for automatic speech-understanding. All these fields bring contributions to the new field of speech science.

Speech recognition can be generally defined as the process of transforming the continuous acoustic speech signal into discrete representations which may be assigned proper meanings and which, when comprehended, may be used to affect responsive behavior. The ultimate goal is to understand the input sufficiently to select and produce an appropriate response..

1.1 The Generalities in Speech Production

Human vocal tract consists of the pharynx, the mouth or oral cavity and the nasal cavity. It is convenient to describe the acoustics of speech production in terms of three stages. First, through interaction between airflow from the lungs and laryngeal and supraglottal structures, a source of acoustic energy is created. This acoustic source may be one of several types, and may have several possible positions. The source acts as the excitation for the cavities above and below it. Each cavity has its own resonating characteristics.. The filtering that is imposed on the source by the vocal tract cavities is the second stage in the generation of speech sounds. Finally, speech sound is radiated from the lips and/or the nostrils [2],[8].

If the production of speech sound can be modeled as a linear system, then the pressure variations recorded outside at some distance from the lips will have a spectrum that is the product of the source spectrum, the vocal tract transfer

function, and the radiation characteristics.

Speech sounds can be classified into 3 distinct classes according to their mode of excitation [1]. VOICED SOUNDS, TENSE FRICATIVES or unvoiced sounds and PLOSIVE SOUNDS. Voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract. Unvoiced sounds are generated by forming a constriction at some point in the vocal tract (usually toward the mouth end), and forcing air through the constriction at a high enough velocity to produce turbulence, this creates a broad-spectrum noise source to excite the vocal tract. Plosive sounds result from making a complete closure (again, usually toward the front of the vocal tract), building up pressure behind the closure, and abruptly releasing it. The resonance frequencies of the vocal tract are called FORMANT FREQUENCIES or simply formants.

The formant frequencies depend upon the shape and dimensions of the vocal tract; each shape is characterized by a set of formant frequencies. 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.

1.2 Speech Analysis

Speech analysis mentioned here only means acoustic processing and phonetic analysis (APPA) in a speech recognition system. The objective of an APPA component is to accept the continuous speech signal as input and produce as output a string of discrete units that are often phonetic in size and nature. These strings are then accessed by other components of the system to perform lexical, syntactic, and semantic analysis in order to decode the utterance.

The first questions to be asked when designing an APPA component are: how the speech should be represented; what parameters should be used for phonetic processing; and how these parameters could be extracted reliably. In this section, we will answer these questions in a brief way [4], [25].

1.2.1 Time-and Frequency-Domain Representations of the Speech Signal

Due to the increasing availability of digital computer and the advantages of computer environments, most researchers in speech recognition nowadays choose digital processing techniques over analog. The speech signal is usually sampled and digitized using an analog-to-digital converter, and stored in computer. The sampling rate used varies from 6 to 20 kHz, and usually 9 to 16 bits are used

to represent the speech samples.

Given a digital representation of the speech signal, various parametric representations can then be derived. Certain parameters, such as zero-crossing density and fundamental frequency of voicing (F_0), can be derived in the time domain directly from the speech signal. On the other hand, experience has shown that frequency-domain representation of the speech signal often provides greater insights into the relationship between the articulatory and the acoustic realizations of speech. For example, spectral peaks in non-nasalized vowels can quite reliably be correlated with the resonant frequencies of the vocal tract, and the frequency location of the major energy concentration in a plosive release gives good indications about the location of the constriction in the vocal tract. It is, therefore, often desirable to obtain the short time spectrum of the speech signal.

Short time spectrum analysis has been one of the most important speech processing techniques for many years. The fundamental assumption underlying this and any other short time analysis method is that over a long time interval speech is nonstationary, but that over a sufficiently short time interval, it can be considered stationary. Thus, the Fourier transform of a short segment of speech should give a good spectral representation of the speech during that time interval. Key perceptual aspects of the speech signal are

more evident in its Fourier transform. Two methods are commonly used for implementing short time Fourier analysis. The first uses a bank of bandpass filters. The second uses a fast Fourier transform (FFT) algorithm. When implemented on a computer the FFT method is generally computationally superior to the filter bank model.

The prevailing technique used in APPA system to obtain the short-time spectrum of the signal is the linear prediction, or LPC, technique. LPC is based on a specific speech production model, namely that speech is produced by all-poles digital filter that is excited by a periodic impulse train for voiced speech and random noise for unvoiced speech. To the extent that this production model is valid, estimation of the short-time spectrum of speech can be reduced to a problem of determining the coefficients of the all-pole filter, since the filter coefficients uniquely specify the transfer function.

The reason for using the name linear prediction analysis lies in the fact that, by choosing a minimum mean-squared error criterion, estimation of the filter coefficients reduces to a solution of a set of P linear equations, where P is the order of the all-pole filter. The set of equations has certain mathematical properties that greatly reduce the computational complexity of the algorithm.

There are several advantages in choosing LPC over other spectrum analysis procedures. First, linear prediction separates the periodic excitation in voiced speech from the combined effect of the glottal characteristics. The harmonic structures in the original short-time spectrum are therefore removed. Secondly, by choosing the order of the predictor to adequately reflect the number of the formants within the frequency range, the peaks in the filter transfer function often correspond well with the actual formants. This property greatly reduces the difficulties associated with the estimation of formant trajectories in continuous speech.

Figure 1.2 compare burst spectra of a pronunciation of the plosive sound /d/ obtained by various spectral smoothing techniques : (a) and (b) by windowing (with different window length) and fourier transforming the waveform. (c) by linear prediction. In Figure 1.2(a), the effect of glottal periodicities can be seen as the ripples superimposed on the spectral envelope. These ripple are greatly reduced in Figure 1.2(b) because of the spectral smearing of the the wide frequency window. In Figure 1.2(c) the effect of the glottal excitation is removed by a homomorphic technique. Since the linear prediction analysis is based on a specific speech production model and thus limits the number of spectral peaks, there are no extraneous peaks in Figure 1.2(c). If we compare the locations of the spectral peaks

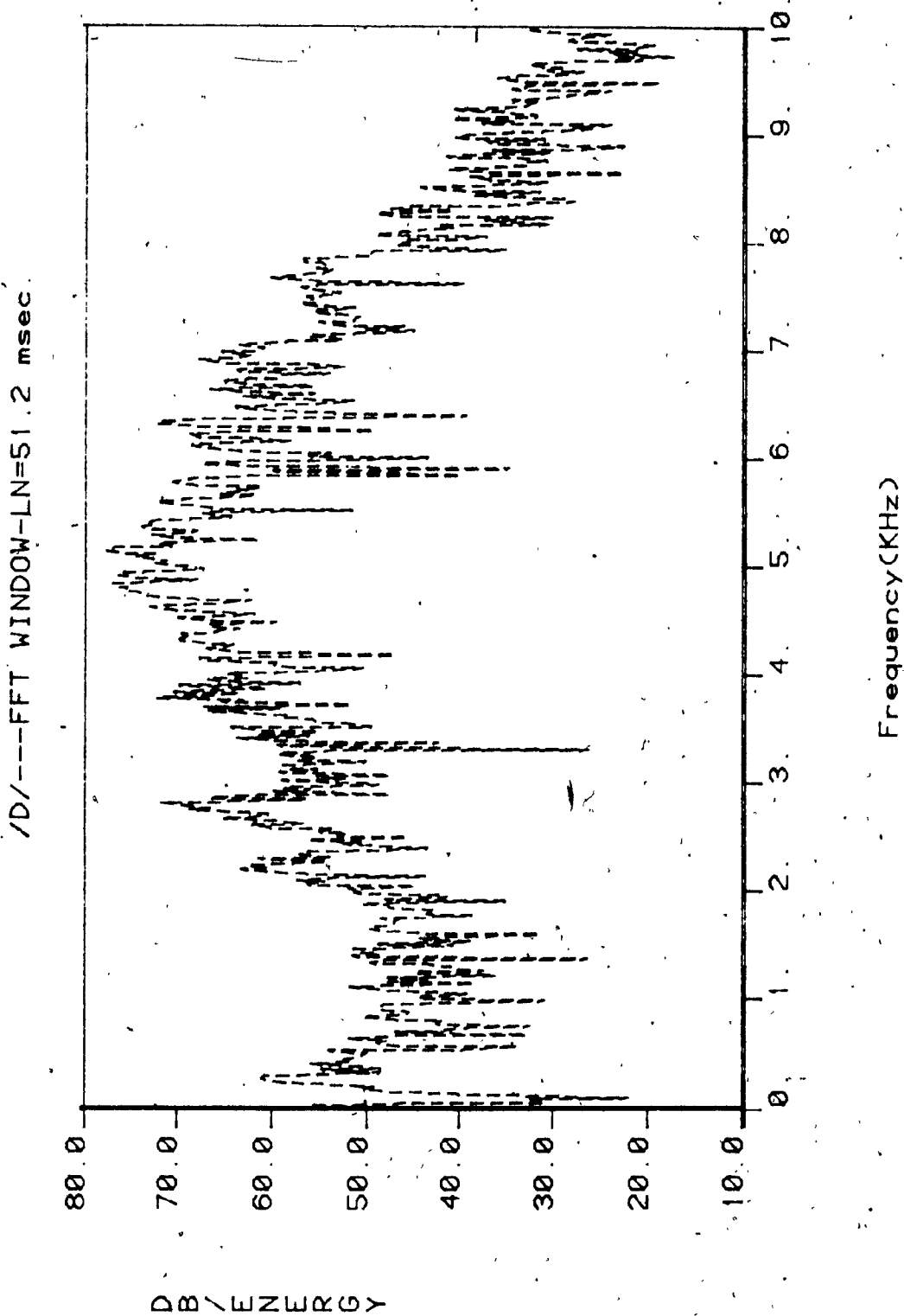
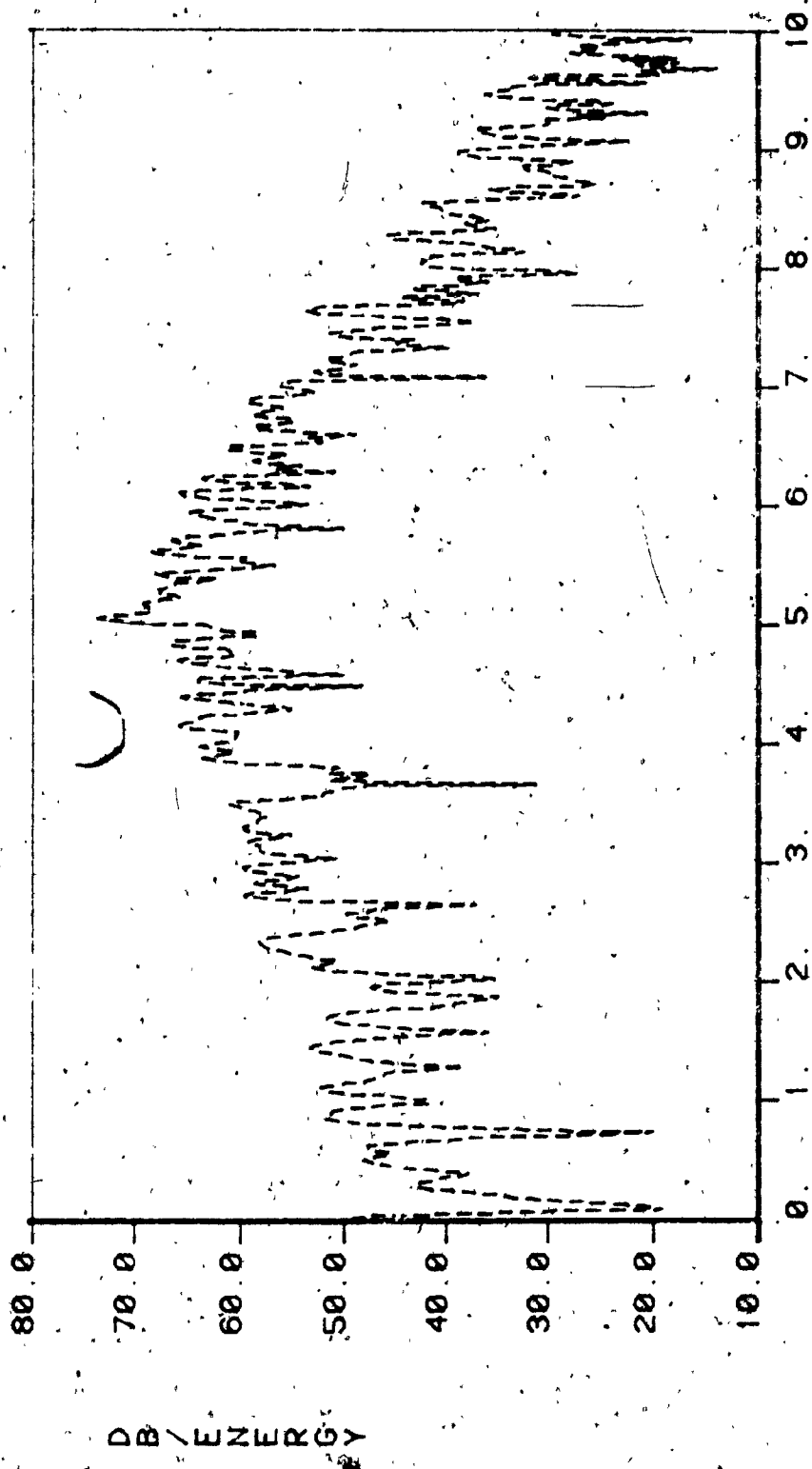


Figure 1.2(a) Burst spectra of the plosive sound /D/ obtained by the spectral smoothing techniques

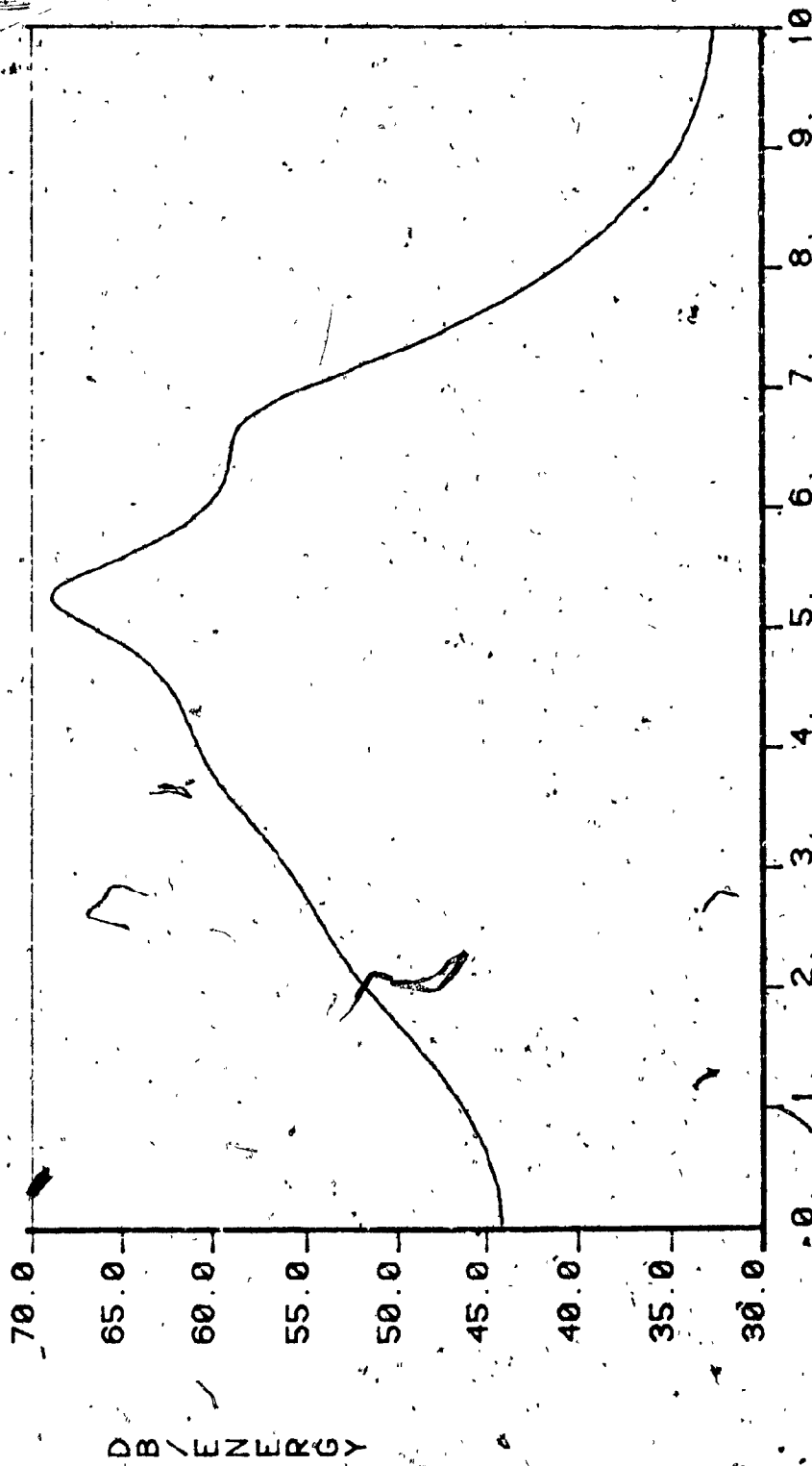
/D/---FFT WINDOW-LN=25.6 msec.



Frequency (KHz)

Figure 1.2(b) Burst spectra of the plosive sound /D/ obtained by the spectral smoothing techniques

/D/---LPC WINDOW-LN=25.6 msec



Frequency (KHz)

Figure 1.2(c) Burst spectra of the plosive sound /D/ obtained by the spectral smoothing techniques

with the actual values of the formants, spectrum derived from linear prediction provides good formant information.

1.2.2 Parameters for Phonetic Analysis

In this section, we shall present a representative, but by no means exhaustive, list of such parameters, and comment on their relative merit for phonetic analysis.

Zero-Crossing Density

Early work on recognition relied heavily on the use of zero-crossing density, i.e., the count of zero crossing of the speech signal in a given interval, to perform segmentation. This is partly due to the fact that the algorithm can easily be realized in analog hardware. Although it is difficult to associate zero-crossing density directly with the underlying acoustic and articulatory correlates of speech sound, several efforts in phonetic recognition have reported promising results using zero-crossing density, and modifications as the primary segmentation parameters.

Fundamental Frequency of Voicing

During the production of voiced sounds, the vocal cords are set into vibration. The fundamental frequency of voicing (F_0), therefore, can be used as a voicing indicator. Although the above justification is theoretically sound, fundamental frequency has not been widely used in segmental

analysis for several reasons. In the segmentation of vowels and sonorants, there exist acoustic parameters such as low-frequency energy that are just as robust and much simpler to derive. In the case of English consonants, the voicing distinction lies more in the durational difference than in the presence of FO.

Energy Related Parameters

One of the most important characteristics of the speech signal is the fact that the intensity varies as a function of time. Sharp intensity changes in different frequency regions often signify the boundaries between speech sounds. For example, low overall intensity usually signifies either a pause, a stop closure, or a weak fricative, whereas a drop in mid-frequency intensity in a vocalic segment usually indicates the presence of an intervocalic consonant.

Formant Frequencies and Trajectories

It is well known that the first three formants for vowels and sonorants carry important information about the articulatory configuration in the production of speech sound. Steady-state values of formant frequencies can be used to classify vowels and sonorant consonants. In addition, formant trajectories can be used to classify diphthongs and for characterizing the place of articulation of plosive sounds. Formant transitions in adjacent vowels can be used to determine the place of articulation of

consonants.

Gross Shape Parameters

Some of the acoustic characteristics of the speech events, such as production of fricatives and onset of plosive release, are best characterized in terms of the gross spectral shape, as opposed to the frequency locations of the spectral peaks.

1.3 Computer Model for Speech Recognition

Speech recognition is a part of a broader speech processing technology also involving computer identification or verification of speakers, computer synthesis of speech and production of stored spoken responses, computer analysis of the physical and psychological state of the speaker, efficient transmission of spoken conversations, detection of speech pathologies, and aids to the handicapped. Only the task of machine comprehension of the intended linguistic message is considered here.

According to the complexity of our task, we usually choose the model which we are going to use for speech decoding between two different types: a passive model or an active model.

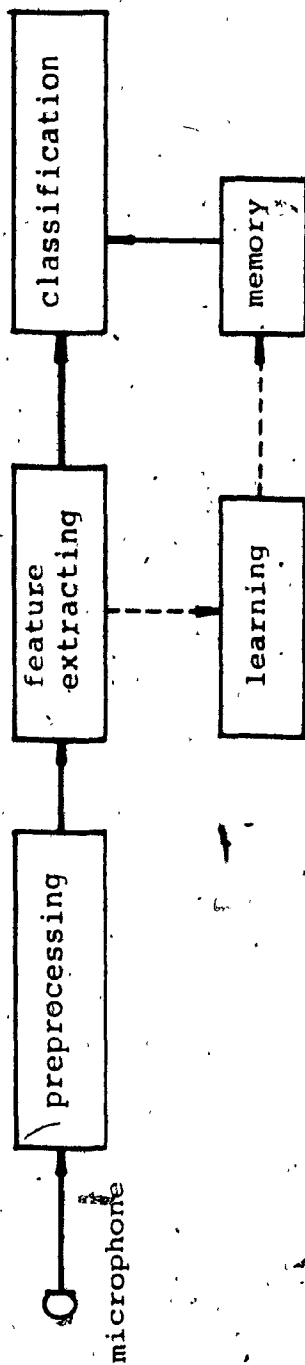


Figure 1.3.1 A passive model for speech recognition [7]

In the passive model, as shown in Figure 1.3.1, human reception of speech is viewed as consisting of sensation followed by perception, followed by cognition. Sensation deals with the raw signal, perception classifies the sensation into words or objects and cognition establishes relationships among the words or objects. This model has been used for designing systems for the recognition of isolated words and consists of the acoustic preprocessing, feature extraction, and pattern matching.

It is clear that human listeners use expectation for understanding what is being said. In a model of such behavior, all facets of the listener's knowledge, such as syntax, semantics, pragmatics, phonology, are used to aid the decoding. Human perception is thus likely to be an active process in which cognition may even guide the lower levels of decoding.

An active model, as shown in Figure 1.3.2, for speech understanding involves the representation of knowledge at various levels (Knowledge Sources : KS): a procedural knowledge containing rules on how to use the KS effectively in order to solve the problem of interpreting a signal and a set of data structures where the interpretation hypotheses are written. An essential feature of the active model is that cognition and expectation may drive decoding [7].

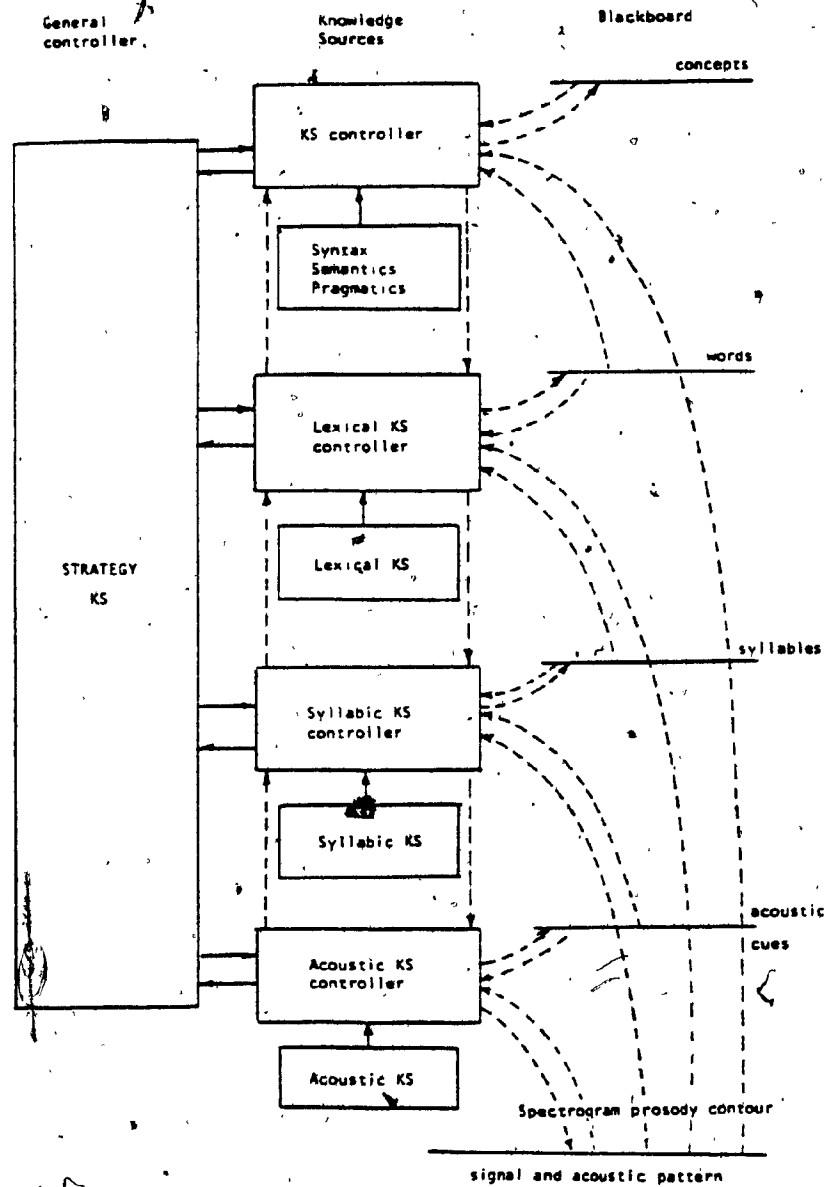


Figure 1.3.2 An active model for speech understanding [7]

1.4 A Brief History of Automatic Speech Recognition

The problems which make the development of speech recognizers so hard is that different people speak in different ways. Also, a noisy acoustic environment may interfere with reliable interpretation of the acoustic speech signal. In addition, even the same single talker will vary from time to time in his pronunciations. The problem is complicated considerably by the complexities of naturally flowing connected speech.

If the problem could be carefully limited, by restricting the population of speakers, working with good acoustic conditions, and avoiding the complexities of fluent speech, then perhaps some initial capabilities could be demonstrated. Thus, the early history of speech recognizers focused on isolated word recognizers, which could identify which word from a small vocabulary was spoken, when sufficient silence preceded and followed the word, to assure easy detection of word boundaries and avoid coarticulatory effect between neighboring words.

1.4.1 Beginning With Isolated Words Recognition

In 1950, the first attempt at automatic speech recognition was described by Dreyfus-Gräf [17]. In his 'stenosonograph', the speech signal was passed through six bandpass filters. The low-frequency sound like vowels were deflected to quite different spots from high frequency

sounds like fricatives. Different sequences of sounds gave different tracks around the screen. In 1952, Davis, Biddulph, and Balashek of Bell Telephone Laboratories, developed the first complete, speaker-dependent digit recognizer [18], which divided the frequency spectrum into two bands, above and below 900 Hertz, and counted how often the two signal levels passed through zero volts ("axis crossing"). Over 97% of the time, the machine correctly identified which of the ten words was spoken, when the pattern had been stored for that particular speaker.

A very important event in the history of isolated speech recognition happened in 1958 when Dudley and Balashek built a recognizer. The major aspect of their system was the segmentation of words into phonetic units. Almost perfect recognition accuracy for individual speaker was reported. Performance dropped drastically for other speakers [19].

The first work using a digital computer came at around 1960 [20]. Denes introduced the important concept of time normalization. From then on, speech recognition researches advance very quickly. In 1972, the first commercial products from Threshold Technology appeared. Table 1.4.1 shows some rather impressive recognition scores reported in recent years for several commercially available recognizers [22].

Table 1.4.1 Representative recognition scores [22]
for some isolated word recognizers

REFERENCE	ENVIRONMENT	VOCABULARY	NUMBER OF SPEAKERS	NUMBER OF TEST UTTERANCES	PERCENTAGE WORDS CORRECTLY RECOGNIZED
MARTIN, GRAMAS, 1975 (also MARTIN, 1975)	High Quality Speech	10 Digits 12 Words	10 10	2400 1320	99.79% 99.12%
	Pilot's mike and captive mike	12 Words	12	1440	97.15%
MARTIN, 1975	85-90dB background noise Actual baggage-handling application	34 Words	12	9180	98.5%
ISHIURA, 1975	Telephone Speech	200 Words (Japanese)	1 Male	2000	97.3%
	Telephone Speech	36 Words (Alphabet & Digits)	1 Male	720	86.6%
SCHEB, 1975	No training required, "Speaker-independent"	10 Digits	30 Males	9300	96.0%
SCOTT, 1976	Speaker-independent, Tapes and live tests	10 Digits and 4 control words	85 Males	14,200	96.0%
SCOTT, 1977	Speaker-independent, with error-correction via 1.5s. Telephone Bandwidth	10 Digits and 4 control words	139 Males 54 Females	56,000	96.0%
Coler, et al., 1977	Score PQ service	10 Digits	20 Males	20,000	87.6%
	Best NASA/Ames Extension of Score Algorithm	10 Digits	20 Males	20,000	98.9% corr- 5% rate
	Best NASA/Ames Algorithm, High Quality Speech	100 Words	10 Males	100,000	93.2% or 95.7% with 5% rate
	Speech tree dictates supvocabulary at each point in discourse	100 Words selected by speech into supvocabulary	10 Males	100,000	98.6% or 99.6% with 5% rate
Hitachi Electric Co. Announcement 1978	Speaker-dependent	10 Digits	4 Males	2400	99.8%
	Speaker-dependent	50 Japanese City Names	1 Male 1 Female	-	100%

These commercial available speech recognizers are effective within some limits. Expanding the vocabulary and the needing for speaker independency become the major tasks in this area.

Recently, prototypes are developed for the recognition of isolated words. A low cost boards compatible with personal computers have been produced by many companies. Among them it is worth remembering VOTAN and TEXAS INSTRUMENTS. Prototypes are available for dictating machines (IBM, KURTZWEILER) capable of accepting thousands of words in a speaker-dependent way and in syntactically constrained protocols.

1.4.2 Steps Toward Continuous Speech Recognition

In the late 1960's and early 1970's, several major projects were undertaken to develop appropriate techniques for recognition of connected speech. Otten [21] proposed the application of syllabic units, prosodics, and a finite state language (Markov Model) to represent the structure of speech dialogue with a machine. Several projects involved the phonetic segmentation of continuous speech. Reddy [22], for example, reportedly achieved over 80% correct identification of phonemes in nonsense strings and meaningful phrases.

In 1971, the largest project ever undertaken in speech recognition was begun, when the Advanced Research Projects

Agency (ARPA) of the United States Department of Defense started a 5-year project to develop machines that capable of "understanding" continuously-spoken sentences involving 1000-word vocabulary. ARPA SUR project called for machines that would accurately (i.e., for over 90% correctness) accept continuous speech from many cooperative speakers, with near-ideal conditions of quiet rooms and high-fidelity equipment. Around this time, major advances had been made in computer technology and "Artificial Intelligence", such as procedures to have computers make logical deductions and inferences, recognize patterns and rapidly search among thousands of alternatives to find the best solution to a problem.

Later on many systems were built with excellent phonetic segmentation results. Another very interesting highlight work is the use of Syntactic Pattern Recognition schemes, which model speech waves by structural features or units whose composition and combinations are determined by syntactic rules (R. De Mori, et al., 1975) [23]. Around this period, Itakura [24] introduced the now-popular technique of dynamic programming for time normalization, and he defined a new metric for comparing frequency spectra.

Now, there are many substantial progress toward limited versions of the challenging goal of recognizing continuous speech. Table 1.4.2 lists several studies which dealt with restricted (i.e., high formatted) word sequences [22].

Table 1.4.2 Performance of recognizers of connected word sequences [22]

REFERENCE	CONDITIONS	WORD SEQUENCES	NUMBER OF TALKERS	NUMBER OF TEST UTTERANCES	PERCENTAGE OF STRINGS CORRECTLY RECOGNIZED
Vinson, 1969	High Quality Speech	From 16-word vocabulary	2 Male	192 strings	83%
IBM, Dixon and Tappert, 1971	High Quality Speech	New Raleigh Language, 250-word vocabulary	3 Male	72 sentences	36%
IBM, Jelinek, 1976	High Quality Speech	7-Digit Strings New Raleigh Language, 250-word vocabulary	1 Male 1 Male	100 strings 143 strings	89% 81%
Bell Labs, Egger and Sauer, 1976	High Quality Speech Noisy Computer Room	7-Digit strings 3-Digit strings	5 Male, 5 Female 5 Male, 5 Female	200 strings 100 strings	71% 87%
Bell Labs, Sauer and Rabiner, 1976	Computer Room - Speaker-dependent Speaker independent	3-Digit strings	4 Male, 2 Female 10 Male	900 strings 200 strings	98.8% 95.3%
KARPT, Reddy, et al., 1976	Computer room Computer room Telephone lines Speaker independent	3-Digit strings 7-Digit strings 3-Digit strings 3-Digit strings	7 Male, 3 Female 1 Male 3 Male, 1 Female 14 Male, 6 Female	1000 strings 100 strings 400 strings 1200 strings	96% 96% 82% 83%
Ti Duddington, 1976	Good Quality Speech	6-Digit strings	NA*	NA*	99%
Kobata, et al., 1977	Good Quality Speech	Sent Reservations in Phrase Form (Japanese)	8 Male	120 reservations	86%
Seki and Masuyama, 1977	Good Quality Speech	Commands for Computer (Japanese)	10 Male	100 sentences	64%
IBM, Sakia, 1977	High Quality Speech	7-Digit strings	1 Male	NA*	95%
Nippon Electric Co. Announcement, 1977	High Quality Speech	1 to 3 Digits (Japanese)	5 Male	1500 strings	99.8%
Madros, 1978	Quiet room Vocabulary - 36 words Quiet room Vocabulary - 64 words	2 to 4 alpha-numeric 2 to 7 words	3 Male 3 Male	24 strings 23 strings	79% 87%
IBM, Bahl, et al., 1978a	High Quality Speech - Phone Model - Contiscond Model	New Raleigh language 250 word vocabulary	1 Male 1 Male	100 strings 100 strings	73% 95%
		KARPT TASK, 1010 word vocabulary	1 Male	100 sentences	99%
IBM, Bahl, et al., 1978b	High Quality Speech	Laser patent texts 1000 word vocabulary	1 Male	20 sentences	67% word recognition

*NA = information not available

1.4.3 Summary

The progress in over thirty years of work on speech recognition can not be measured by increases in recognition accuracy, though there has been some progress on that performance factor. The primary gains have been made in the complexities of the tasks that have been accurately handled. Some highlights in the history of speech recognition can be seen in Figure 1.4.3.

In isolated word recognition, vocabulary sizes have increased from ten to several hundred words; highly confusable words have been distinguished; syntax trees have been incorporated to restrict acceptable next words to be within small sub-vocabularies; improved adaption to individual speaker and speaker-independent systems have both been developed; efforts of some environmental conditions have also been produced and applied in practical interactions with machines.

Expansions to limited forms of continuous speech have yielded a few modestly successful systems for key word spotting, a number of fairly effective laboratory system and commercial products for recognizing digit strings and strictly formatted word sequence, and some limited but encouraging systems for sentence understanding.

Table 1.4.3 Some highlights in the history
of speech recognition

EARLY HISTORY

1947	SOUND SPECTROGRAPH
1952	DIGITS, USING WORD TEMPLATE, 1 SPEAKER
1956	DIGITS, USING PHONETIC SEQUENCES
1960	DIGITS, DIGITAL COMPUTER TIME NORMALIZATION
1962	IBM SHOEBOX RECOGNIZER
1964	WORD RECOGNIZER FOR JAPANESE
1965	VOWELS AND CONSONANTS DETECTED IN CONTINUOUS SPEECH, CALL FOR HIGH-LEVEL LINGUISTICS
1967	VOICE-ACTUATED ASTRONAUT MANEUVERING UNIT
1968	54-WORD RECOGNIZER, DIGIT STRING (ZIP CODE) RECOGNIZER, VICENS 50-500 WORDS

RECENT HISTORY

1969	VICENS-REDDY RECOGNIZER OF CONTINUOUS SPEECH PIERCE'S CAUSTIC LETTER OBJECTING TO SPEECH RECOGNITION WORK MAD SCIENTISTS AND UNTRUSTWORTHY ENGINEERS
1970	
1971	ARPA SPEECH UNDERSTANDING 5-YEAR PROJECT, 3 SYSTEM BUILDERS 4 RESEARCH EFFORTS UNDERSTANDING OF CONTINUOUSLY-SPOKEN SENTENCES, 1000 WORDS
1972	1st COMMERCIAL WORD RECOGNIZER 100 WORDS W/PHONOLOGICAL CONSTRAINTS
1973	
1974	DYNAMIC PROGRAMMING (200 WORDS, TELEPHONE, OXYGEN MASK, OXYGEN MASK)
1975	ALPHABET AND DIGITS; 91 WORDS W/DYN. PP MULTIPLE TALKER, NO TRAINING
1976	ARPA SYSTEMS, HARRY, HEARSAY, MWII, VICI W/182 TALKERS (97%), TELEPHONE
1977	CRT-COMPATIBLE VOICE TERMINAL; TRI-SERVICES, REVIEW
1978	IBM CONTINUOUS SPEECH RECOGNIZER
1984	IBM, KORTWEILER PHONETIC TYPEWRITERS

1.5 Motivation and Task of This Research

Our purpose is to build up a multi-speaker computer recognition system for connected letters (unformatted) belonging to the following set :

[E1 = (P,T,K,B,D,V,E,G,C,3)]

which is based on specific knowledge about the acoustic properties of the features to be extracted.

Our main considerations are in the recognition of connectedly spoken letters. We started by collecting a natural speech data base which contains stop-vowel syllables from several talkers (5 male and 5 female). An active model of speech recognition has been used in order to apply a focus of attention paradigm for recognition. The reason is that information about plosive sounds is mostly encoded in a short transient before the following vowel. As in the case of letters of the E1-Set, the vowel is the same for 9 of the elements of the set and /k/ contains a diphthong of front vowel; global methods, the other approach, may give a too high importance to vowel different as due to different speaker characteristics. This partially justifies low recognition rates (below 50%) developed so far using global methods in a multi-speaker environment.

CHAPTER II

PLOSIVE SOUNDS

Speech contains many redundant cues, which aid perception in adverse circumstances such as a noisy environment or a speaker with a foreign accent. Speakers usually take advantage of this redundancy, articulating more clearly when the need arises, but speaking more rapidly and casually in more informal conversations. Perception experiments have attempted to determine the nature of these cues in the acoustic waveforms of speech, in particular, acoustic-phonetic cues which may be invariant to context or speaker. Figure 2.0 shows the phonetic feature of each phoneme of American English. Sufficient cues have been discovered which can describe a great deal about the perception of phonemes within a language, but the cues are often not invariant with respect to phonetic context, stress, and speaking rate. The listener may be born with or develop certain feature detectors, which could be invoked to classify speech sounds into linguistic categories which differ by one or more features.

2.1 The Acoustic Properties of Plosive Sound

The plosive sounds, or stop consonants /P, T, K, B, D, G/, consist of voiced and unvoiced stop consonants. The voiced stop consonants /B/, /D/, /G/ are transient, noncontinuant

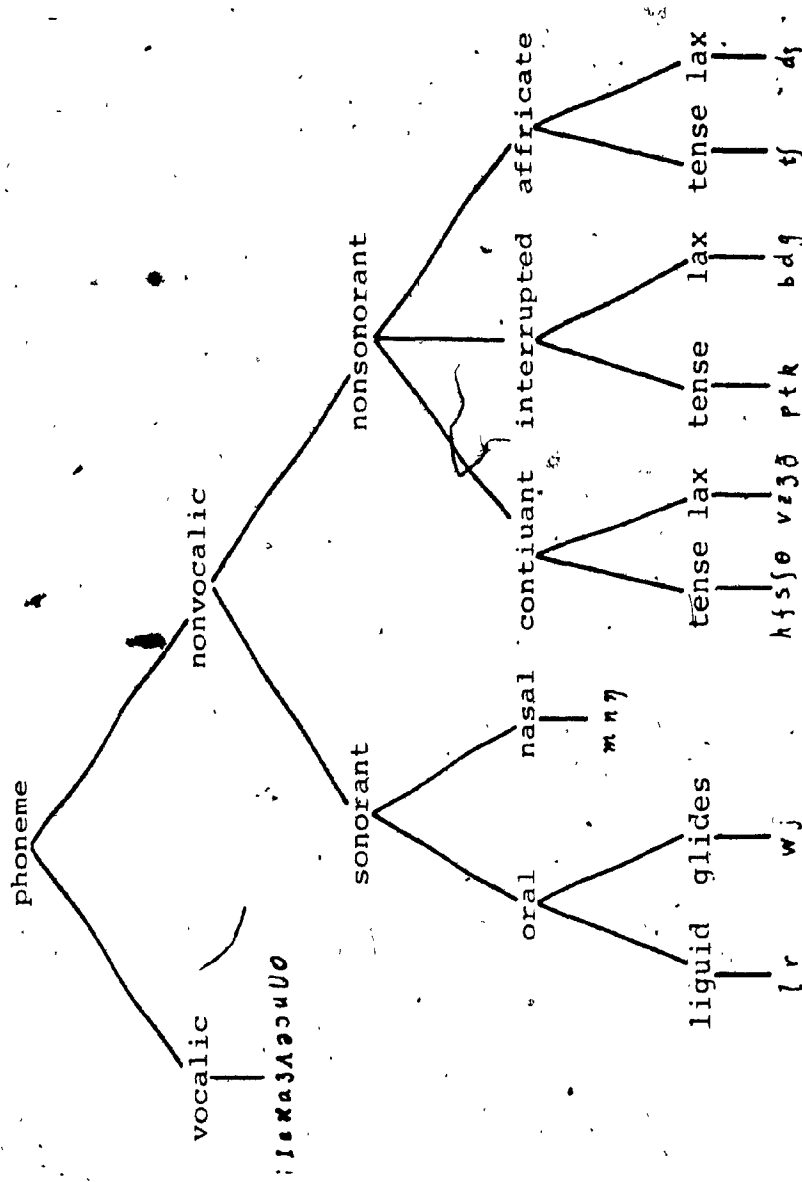


Figure 2.0 Phonemes in American English

sounds which are produced by building up pressure behind a total constriction somewhere in the oral tract, and suddenly releasing the pressure. For /B/ the constriction is at the lips; for /D/ the constriction is back of the teeth ; and for /G/ it is near the velum. During the period when there is a total constriction in the tract, there is no sound radiated from the lips. However, there is often a small amount of low frequency energy radiated through the wall of the throat (sometimes called a voice bar), this occurs when the vocal cords are able to vibrate even though the vocal tract is closed at some point.

The unvoiced stop consonants /P/, /T/, /K/ are similar to their voiced cognates /B/, /D/, and /G/ with one major exception. During the period of total closure of the tract, as the pressure builds up, the voiced cords do not vibrate; thus, following the period of closure, as the air pressure is released, there is a brief interval of friction (due to sudden turbulence of the escaping air) followed by a period of aspiration (steady air flow from the glottis exciting the resonances of the vocal tract) before voiced excitation begins. Figure 2.1 shows the place of articulation of the stop consonants.

Generation of these consonants depends upon vocal tract dynamics. Since the stop sounds are dynamical in nature, their properties are highly influenced by the vowel which follows the stop consonant [5].

Place of articulation







	labial	alveolar	palatal
voiced	<p>b (be)</p> 	<p>d (de)</p> 	<p>g (ge)</p> 
unvoiced	<p>p (pe)</p> 	<p>t (te)</p> 	<p>k (ke)</p> 

Figure 2.1 Place of articulation of plosives

2.2 Perception of Manner of Articulation

Among the distinctive features, those dealing with manner of articulation are perhaps the least controversial and simplest to explain. Manner perception concerns acoustic features which permit the listener to classify speech into one of the following categories: vowels (including liquids), glides, nasals, stops, and fricatives.

Vowels are perceived when the speech sound is voiced, with sufficient amplitude and duration and a strong formant structure (the lower formant excited with relative narrow bandwidths). Glides can usually be distinguished from vowels by weaker amplitude, briefer durations, and a greater tendency toward dynamic spectral pattern. Nasals can be distinguished from vowels by their weaker amplitude, wider bandwidths, and higher concentration of energy at low frequencies. Sound with high frequency noise of sufficient duration are perceived as fricatives. Stop, of course, are heard when a period of silence interrupts the speech signal, or a brief pause is followed by a sudden, short burst of noise.

The crucial features which separate these manner classes involve amplitude, duration, general formant structure, and the balance between low frequency voiced energy and high frequency frication. The most basic factor involves whether

the sound has an aperiodic component (stops and fricatives). We call those acoustic cues which are mainly related to the perception of manner of articulation as manner cues.

In Table 2.2 [2], we give out an example of manner cues which describe the shape and speed of all formant transitions (1 and 2), the locus or starting point of the first formant (3 to 6), presence and duration of turbulent noise (7 and 8), presence and continuancy of formant-like periodic links (9 to 11), and six cues for the voicing distinction. These six cues are the duration of a hold (12), length of the preceding vowel (13), cutback describing the elimination of the initial portion of the first formant (14), intensity of a turbulence (15), voice bar within a hold or a turbulence (16), and aspiration (17).

The most common perceptual confusions due to errors in manner of articulation involve the non-strident voiced fricatives and voiced stops (especially /b/), and to a lesser extent their unvoiced cognates. The non-strident fricatives are sufficiently weak that they are often hard to distinguish from a stop, especially a labial stop, which is usually weakly released and has formant transitions similar to those of the labial and dental fricatives.

Table 2.2 Acoustic cues for manner of articulation

1. Direct formant transitions / reverse.
2. Fast formant transition / slow.
3. F1 locus at 0 Hz / not at 0 Hz.
4. F1 locus at 250 Hz / not at 250 Hz.
5. F1 locus at 400 Hz / not at 400 Hz.
6. F1 locus at 700 Hz / not at 700 Hz.
7. Turbulence / no turbulence.
8. Short turbulence / long turbulence.
9. Low periodic link / no low per. links.
10. Discontinuous links / continuous links.
11. Short hold / long hold.
12. High periodic links / no high per. links.
13. Long preceding vowel / short prec. vowel.
14. Cutback / no cutback.
15. Weak turbulence / strong turbulence.
16. Voice bar / no voice bar.
17. Aspiration / no aspiration.

2.3 Perception of Place of Articulation

The description of the acoustic cues to place of articulation of stop in the past has been based on two sources of information. In one approach, the acoustic properties have been calculated from theoretical model of the stop consonant articulatory gestures. In the other, acoustic correlates have been measured from sound spectrograms or estimated using other analysis techniques [10]. Although a great deal of detailed knowledge has accumulated over the years, it has been difficult to verify particular sets of acoustic properties as both necessary and sufficient perceptual cues to place of articulation for the human listener [11].

The place cues in Table 2.3 describe the locus of the second formant (1 to 4), the locus of the third formant (5 to 8), and spectral energy distribution of a turbulence or burst (9 to 11).

The acoustic cues for place of articulation in stop consonants have been basically thought to lie in two readily observable acoustic segments, the release burst and the formant transitions. Many investigators have made the assumption that these two segments are independent and separable [12] [13], but the others have assumed that these acoustic segments are not separable at all but constitute a unitary or integrated acoustic stimulus for specifying.

Table 2.3 Acoustic cues for place of articulation

1. Locus of the 2. formant at 700 Hz.
2. Locus of the 2. formant at 1700 Hz.
3. Locus of the 2. formant at 2000 Hz.
4. Locus of the 2. formant at 3000 Hz.
5. Locus of the 3. formant at 2000 Hz.
6. Locus of the 3. formant at 2200 Hz.
7. Locus of the 3. formant at 2700 Hz.
8. Locus of the 3. formant at 3500 Hz.
9. Center frequency of a turbulence low.
10. Center frequency of a turbulence mid.
11. Center frequency of a turbulence high.

place in stops [8] [9].

The premise that the release burst and formant transitions are separable derives primarily from the apparent distinctiveness of their visual representations in oscillograms and same spectrograms. Typically, the spectral properties of the burst have been examined and measured primarily from sound spectrograms and perceptual cues have been described in terms of change in frequency over time. Many speech perception studies have been carried out over years to verify the role of these acoustic cues to place. The results of studies investigating formant transitions in isolation or with bursts have shown that the acoustic information for place of articulation varies with the following vowel context. In contrast, other studies have focused on the burst in a limited number of vowel contexts and concluded that most of the place information is located in the release burst. The more general conclusion has been that both the burst and the formant transitions contribute to specifying place information in a complementary way depending on the following vowel context [15].

Both Fant [26] and Stevens and Blumstein [9] [28] have argued that the acoustic information for specifying place of articulation is independent of vowel context and is located in the first 10 to 30 ms of the stop consonant waveform and invariant acoustic properties for place can be found in the gross shape of the spectrum at the onset of the release

burst. Some experimental findings by Stevens and Blumstein have been interpreted as support for this view. But their results showed that the fairly accuracy can be get only in stops in syllable-initial position, not of stops in syllable-final position. Nevertheless, these particular claims motivated several aspects of the present investigation. The other researchers further claimed that descriptions of these acoustic cues must include both spectral and temporal properties in order to capture the relevant temporal differences associated with the underlying articulatory gestures [11] [8] [26]. This conclusion is also in general agreement with the views of Liberman and other investigators at Haskins Laboratories who have emphasized the dynamic nature of the speech cues [27] [15] [29] [62].

In our research work, we have employed this new point of view that argues for the existence of invariant cues for the perception of place in stop consonants.

2.4 Stop + Vowel Stimuli

In the case of unreleased plosives in VC syllables, spectral transitions provide the sole place cues. For released plosives in CV syllables, the situation is more complex, since acoustic cues for place of articulation in the burst release, the spectral behavior during the ensuing aspiration period, and, if followed by a voiced phoneme, the

aspiration duration, as well as in spectral transitions during adjacent phonemes.

Early research found evidence of a "starting locus" of F2 for each of /b,d,g/ in the perception of two-formant CV stimuli. With F1 having a 50-ms rising transition (typical of all voiced stops), if F2 started at 1800 Hz, /d/ was heard, whereas an initial 720 Hz caused /b/ perception. A 3000 Hz start yielded /g/ for most ensuing vowels, but for high and mid back vowels, a lower locus was necessary. Generalizing, one can say that rising F2 indicates a labial stop, relatively flat F2 tends to be heard as alveolar, and a falling F2 yields velar perception. In all cases, it was necessary however to eliminate the first 50 ms of the F2 transition (so that the transition "pointed to" the locus, rather than actually started there), otherwise different stops would be heard.

The difficulty here lies in extending the results of this study to natural speech, which has more than two formants. Many two-formant CV stimuli lack natural quality, and provide listeners with ambiguous phonetic cues. There is evidence that, for CV stimuli from natural speech, stop burst and ensuing formant transitions have equivalent perceptual weight and act in complementary fashion depending upon context. When formant transition are brief, due to short articulator movements or due to anticipatory coarticulation, the release burst lies near the major

spectral peak of the following vowel and contributes significantly to place perception. Conversely, where formant transitions are extensive, the burst is distinct from the vowel spectral peaks and the formant transitions are more important for stop place perception.

When the unvoiced portion of plosive+V and V+plosive stimuli are removed, listeners identify the consonant less accurately for the CV case, since spectral transitions during the aspiration are absent. This is especially the case for unvoiced plosives, with their long VOTs (voice onset times). Recent studies have noted that, although the primary cues to place are found in spectral behavior, VOT and amplitude also have effects on place perception. When F2 and F3 transitions give ambiguous cues, VOT duration can help to distinguish labial from alveolar stops. Changes in spectrum amplitude at high frequencies (F4 and higher formants) have also been found to reliably separate labial and alveolar stops: when high frequency amplitude is lower at stop release than in the ensuing vowel, labials are perceived.

It has recently been argued that certain aspects of spectral pattern of releases in stops distinguish place of articulation. The concentration or spread of energy (diffuse vs. Compact) and whether the main spectral trend is rising, flat, or falling with frequency have been suggested as crucial spectral aspects. These

Table 2.4 The characteristics of the gross spectral shapes of plosives

PLOSIVE	PHONETIC FEATURE	BURST SPECTRAL
/P/	labial tense	diffuse falling
/T/	alveolar tense	diffuse rising
/K/	palatal tense	compact
/B/	labial lax	diffuse falling
/D/	alveolar lax	diffuse rising
/G/	palatal lax	compact

characteristics for the distinction between plosive sounds have been shown in Table 2.4. Our experiment has similar results, they can be seen from Appendix I. Manipulating buzz-bar detection, burst spectra and starting formant frequencies led to unambiguous place identification among stops when the onset spectra were either diffuse-falling, diffuse-rising, or compact.

Thus the gross properties of the spectrum over the initial 10-20 ms of a stop consonant may provide invariant cues to place perception. When the initial spectrum is ambiguous (e.g., diffuse, but flat), we can utilize formant transitions to distinguish place. Such transitions temporally link the primary place cues in the stop release to the slowly-varying vowel spectrum, with no further abrupt spectral discontinuities after the stop release. In this view, the formant pattern act as secondary cues, which are involved when the primary cues of the release spectrum are ambiguous.

2.5 Summary

The specification of the acoustic cues to place of articulation in stop consonants has continued to be a prominent issue in research on speech perception. But, it is obvious that the hope of finding a one-to-one correspondence in natural speech between invariant acoustic cues and phonemes has not been fulfilled. Therefore the

question was raised as to whether there is at least a simple relationship between acoustic cues and distinctive features. Distinctive features are characteristics of phonological segments which compose the phonemes of a language. Several systems of distinctive features which can be used for phonemically relevant distinctions have been described in the literature [30] [31]. But, the complex relations between distinctive features and acoustic cues have to be determined which is not much easier than to determine relations between cues and phonemes. It would seem to be essential to construct suitable models which take into account the various results and complex relations known from perception experiments. The application of syntactic pattern recognition algorithm could be a good tool for handling these complex relationships [2].

CHAPTER III

DATA ACQUISITION AND PREPROCESSING

The testing samples are collected from 10 native English speakers (5 male and 5 female). Each speaker was invited to our laboratory and asked to speak two different connected strings of letters to the microphone. Each string of letters was repeated 10 times. For the testing convenience, each string includes five different letters. Actually, the system does not require any limitation on the length of the string and the order of the letters. So, these collections give us 100 speech patterns for each letter in the El-Set in a noisy (from the air conditioner) environment. A multiple speaker continuous speech recognition system is then developed on a DEC Vax-780 machine .

A data base of digitized speech sentences was prepared from the input speech analog signals. The incoming speech signal is filtered in 0.10 KHz to 6.8 KHz frequency band by a programmable bandpass filter. A simple block diagram of the preprocessing part of the recognition system is shown in Figure 3.1. Where LPF stands for low pass filter, and HPF for high pass filter.

When an analogue signal is converted to digital form, it is made discrete both in time and amplitude. Discretization in time is the operation of sampling, while in amplitude it

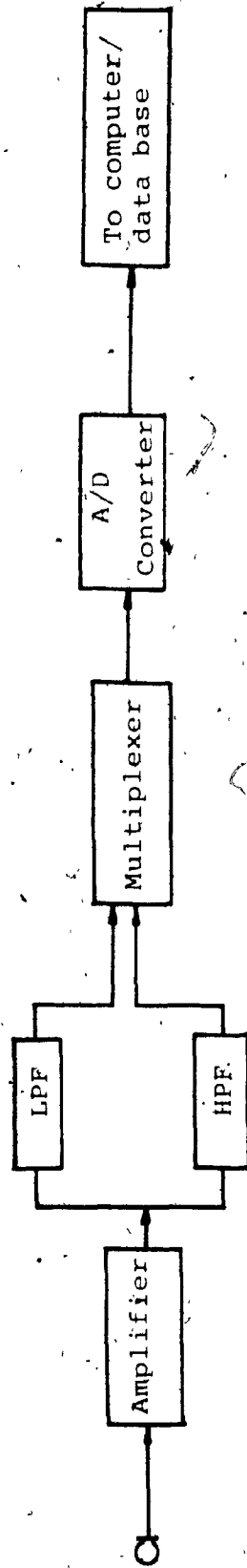


Figure 3.1 Signal preprocessing system

is quantizing. The transmission of analogue information by digital means is called " PCM " standing for " pulse code modulation ".

In applying the sampling theorem to a digital representation of speech there are two main concerns [22]. If the signal bandwidth is W Hertz, then the sampling period must be $T \leq 1/(2W)$ seconds. Since the samples of the signal generally take on a continuous range of values, they must be quantized for transmission or digital storage. If we represent the samples as B -bit binary word, then the bit rate is $2BW$ bit/s. The value of W required for speech signals depends on the ultimate use of the samples. We know from measurements and theoretical studies that speech sounds such as fricatives have rather wide bandwidths (on the order of 10 KHz). On the other hand much of information required for speech intelligibility, is contained in the variation of the first three formant frequencies of voiced speech and these are typically below 3 KHz. Thus, a sampling rate between 6 KHz and 20 KHz is generally used. No matter what the sampling rate is, the speech signal must be suitably low-pass filtered prior to the sampling process to eliminate undesired high frequencies of the speech and high frequency noise.

The choice of the number of bits per sample B is also dependent upon the intended use of the samples. If our purpose is transmission or computer storage followed by

conversion back to an analog signal, we are only concerned that the resulting analog signal be perceptually acceptable. Also, the sampling process just described is generally the first step in any digital speech analysis techniques, since errors incurred in the sampling process will propagate to more refined digital representations, we are often justified in a very generous allotment of bits and sampling rate if the sampled speech wave is to undergo further processing. However it should be noted that the amount of processing required to implement most systems is proportional to sampling rate. Thus we should try to keep the sampling rate as low as possible, consistent with other objectives.

According to the preceding discussions, we chose the sampling rate and the number of bits per sample as 20 KHz and 12 bit/sample, respectively. The bit rate is then 240000 bits/s.

Once the parametric representation of the speech signal has been obtained, the next step is segmentation. Since phone-sized segments mostly can not be localized and separated in the signal domain. It was suggested that larger units be used which are at least as long as a syllable. This corresponds with the theory that syllable can be viewed as articulatory unit [33]. In our experiments (CV syllable), it is very convenient that segmentation on consonants and the vowels within a sentence can be combined for vowel identifications.

CHAPTER IV

AN EXPERT SYSTEM APPROACH TO SEGMENTATION

Since the high complexity of the task, the system for extracting acoustic properties has been conceived in the framework of distributed problem solving in which acoustic properties are facts that drive computational processes to the achievement of goals consisting in hypothesis generation. The major considerations in favour of distributed system are the following :

- (1). A distributed processing model can be implemented with parallel computer architecture capable of reaching real-time performances ;
- (2). A distributed knowledge module can provide the convenience for separately updating each source of knowledge when new knowledge becomes available. Furthermore, different data structures and learning algorithms can be used for each source of knowledge;
- (3). A control strategy capable of scheduling the parallel execution of sensory procedures which extract new properties from the data when this is required.

Based on above considerations, an expert system proposed by De Mori [34] has been developed for extracting acoustic cues, generating syllable hypothesis from continuous speech. Part of the knowledge of such a system is a semantic syntax - directed translation (SSDT) algorithm, that segments

continuous speech into Pseudo-syllabic segments and generates hypotheses about phonetic features in each segment.

4.1 Structure of An Expert

The behaviour of an Expert is shown in Figure 4.1. To each expert EXP_j is associated with a Long Term Memory (LTM) containing the specific Expert's Knowledge and a Short Term Memory (STM) where data interpretations are written. To each expert EXP_j is also associated a message queue MQ_j containing the requests made to it from other experts. EXP_j reads sequentially these requests. If a request concerns some information which has not been requested before, then EXP_j creates an instantiation. Let $INST_{j1}, \dots, INST_{jk}$ be the instantiations created at a given time t [7].

An instantiation is a computing agent that may create other instantiations or send requests to other experts or send answers to the experts which have made requests to EXP_j . In other words, an instantiation $INST_{jk}$ can send a message $MESS_{jk}$ to other experts. Messages for experts can be stimuli coming from lower level experts or verification requests from higher level experts or General Controller (strategy KS). The experts do not communicate through a common data-base. They are provided with an elaborate control strategy.

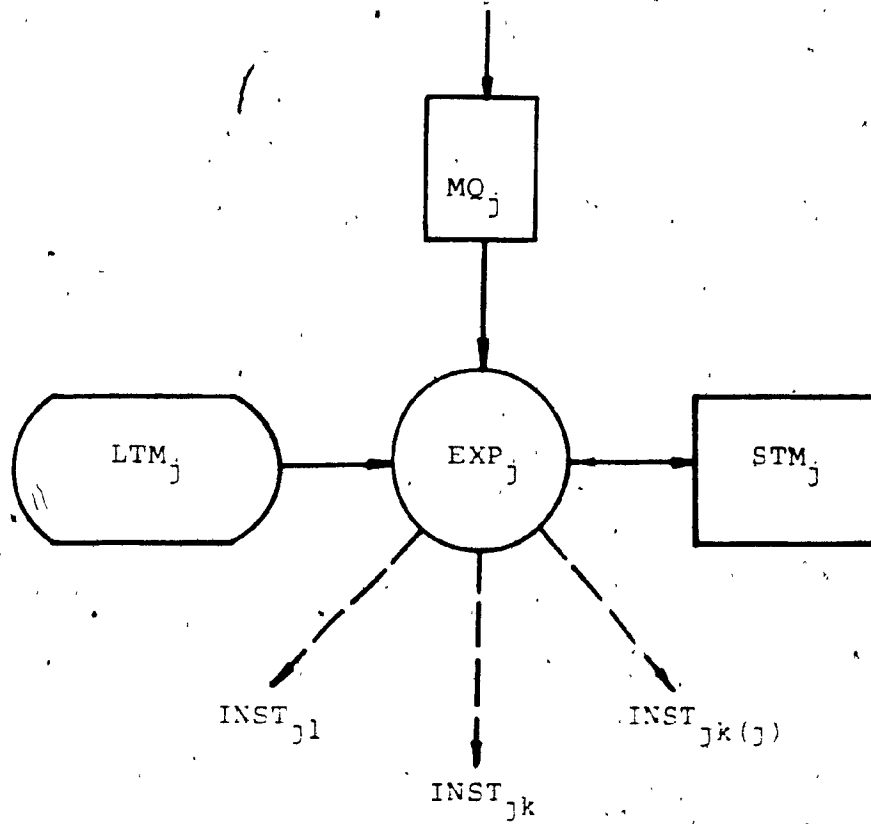


Figure 4.1 Expert structure [7]

When an instantiation has performed its task, it terminates and leaves the system. An expert receives requests for the generation of hypotheses from other experts. Experts are grouped into Societies according to their level of expertise.

The task-independent knowledge is structured on two levels, corresponding to the Auditory Expert Society (AES) and the Syllabic Expert Society (SES).

4.2 Representation of Expert's Knowledge

In the A.I. Program, knowledge is usually represented in the logical formalisms, such as predicate logic. Its major advantage is that they can be combined with simple, powerful inference mechanisms that make reasoning with the facts easy. To deal with fuzzy knowledge (for example, "you are very young." How can relative degrees of young be represented?), people have developed corresponding logic models [35], such as Fuzzy logic which provides a way of representing fuzzy or continuous properties of objects [36] [37]. Fuzzy knowledge representation will be explained in Chapter V in more detail.

But the objects in these representations are so simple that much of the complex structure of the world can not be described easily. It is often to collect these properties together to form a single description of a complex object. One advantage of such a scheme is that it enables a system

to focus its attention on entire objects without also having to consider all the other facts it knows. This is important since straightforward, uniform approaches tend to lead to combinatorial explosion if the amount of knowledge they have to deal with is very large.

A good system for the representation of complex structured knowledge in a particular domain should possess the following four properties :

- (1) Representational Adequacy ----- the ability to represent all of knowledge that are needed in that domain.
- (2) Inferential Adequacy ----- the ability to manipulate the representational structures in such a way as to derive new structures corresponding to new knowledge inferred from old.
- (3) Inferential Efficiency ----- the ability to incorporate into the knowledge structure additional information that can be used to focus the attention of the inference mechanisms in the most promising directions.
- (4) Acquisitional Efficiency ----- the ability to acquire new information easily. The simplest case involves direct insertion, by a person, of new knowledge into the database. Ideally, the program itself would be able to control knowledge acquisition.

Several techniques for accomplishing these objectives have been developed. These techniques can roughly be divided into two types :

declarative methods,

procedural methods.

For declarative methods (such as predicate logic), most of the knowledge is represented as a static collection of facts accompanied by a small set of general procedures for manipulating them; for procedural methods, the bulk of the knowledge is represented as procedures for using it. The major advantages of a declarative representation are :

- (a) Each fact need only be stored once, regardless of the number of different ways in which it can be used.
- (b) It is easy to add new facts to the system, without changing either the other facts or the small procedures.

The major advantages of a procedural representation are :

- (i) It is easy to represent knowledge of how to do things.
- (ii) It is easy to represent knowledge that does not fit well into many simple declarative schemes. Examples of this are default and probabilistic reasoning.
- (c) It is easy to represent heuristic knowledge of how to do

things efficiently.

In many domains, there is need for both kind of information. And so, in practice, most representations employ a combination of both, and the new idea of Knowledge Structure is introduced.

The Knowledge Structure is a data structure in which knowledge about particular problem domains can be stored [35]. Many of knowledge structures are composed of smaller structures. Thus the term knowledge structure will sometimes means a complete database of information about a particular domain and will sometimes refer to substructures within the larger structure. These substructures will usually correspond to such things as objects or events within the domain. There are many types of descriptions about knowledge structures, like :

frames ----- often used to describe a collection of attributes that a given object normally possesses.

rule models ----- used to describe common features shared among a set of rules in a production system.

Knowledge structures in our expert system are generated by a frame-structure grammar which defines a language for representing LTM knowledge. The knowledge stored in the LTM of an expert is a collection of algorithms. The algorithms for generating descriptions of acoustic data and for

generating hypotheses of corresponding phonetic features are expressed in a frame language. Frame language is particularly suitable for integrating structural and procedural knowledge, and making inference.

A frame is an information structure made of a frame-name and a number of slots. A slot is the holder of information concerning a particular item called "slot-filler" [38]. Slot-fillers may be descriptions of events, relations or results of procedure. Attempts to fill the slots are made during a frame instantiation. A frame instantiation can be started by a simple reasoning program of an expert after having received a message. After a frame is instantiated, a copy of its LTM structure is created into STM. At the beginning all the slots in the STM are empty and the expert which created the instantiation attempts to fill the slots sequentially [39] [40] [41].

Frame structures are precisely defined by the rules of a grammar defining all the acceptable composition of the attribute relations [42]. Table 4.2 shows the rules of this frame-structure grammar. The exponent $k > 1$ of an expression means that the expression can be rewritten any number of times greater than 1. Brackets in Table 4.2 contain optional items which can be repeated any number of times. The terminal symbols are written in lower case letters and the non-terminals in upper case. The starting symbol is $\langle \text{frame} \rangle$.

Table 4.2 Rules of the frame-structure grammar

```

<FRAME>      := (<NAME> <SLOT-LIST>)
<SLOT-LIST>  := (<NAME> [( <DESCRIPTION> )])K>0
<DESCRIPTION> := (DESCRIBED-AS <CHDES>)
              := (<CONNECTIVE> <DESCRIPTION>K>1)
              := (not <DESCRIPTION>)
              := (filled-by <FRAME>)
              := <CONDITIONAL>
              := (result-of <proc>)
<CONDITIONAL> := (WHEN <PREDICATE EXPRESSION>
                  <DESCRIPTION>
                  [(else <DESCRIPTION>)])
              := (unless <DESCRIPTION><DESCRIPTION>)
              := (case <NAME> of
                  (<DESCRIPTION> filled-by
                   <FRAME>)K>1)
<CONNECTIVE> := OR
              := and
              := xor
              := sequence

```

Table 4.2 (continued)

<PREDICATE EXPRESSION>	:=<PREDICATE> :=(not <PREDICATE>) :=(<CONNECTIVE>(<PREDICATE> ^K >1))
<PROC>	:=F-<FUNCTION> :=P-<procedure>
<NAME>	:=ANY string of characters
<CHDES>	:=ANY cue or hypothesis description

The slot described as < CHDES > gets filled by generating descriptions of acoustic cues or interpreting hypotheses. The execution of a procedure can be initiated by trying to fill that particular slot of the frame. A procedure in a given instantiation has access to all the slots which have already been filled for that particular instantiation.

The slot filled-by < CHDES > corresponding to the instantiation of a frame represented by its NAME. The slots with connective descriptions may cause the invocation of other frames and execution of procedures for extracting new cues if necessary for evidence. The connective sequence implies that time consistency must be maintained while describing the temporal sequence of events such that the $(i+1)^{st}$ event must begin at the end of the i^{th} one.

4.3 Auditory Experts for Interpreting Speech Patterns

Interpretation and segmentation of the speech waveform are generated by an Expert Society. Its structure is shown in Figure 4.3.1. This Expert Society contains two parts : Auditory Experts (AE) and Syllabic Expert (SE). SE will be introduced in next section.

The extraction of acoustic cues from spectrograms is performed by a group of experts referred to as the Auditory Experts. Actions of writing into and reading from Short Term Memories are represented by dashed arrows. Requests

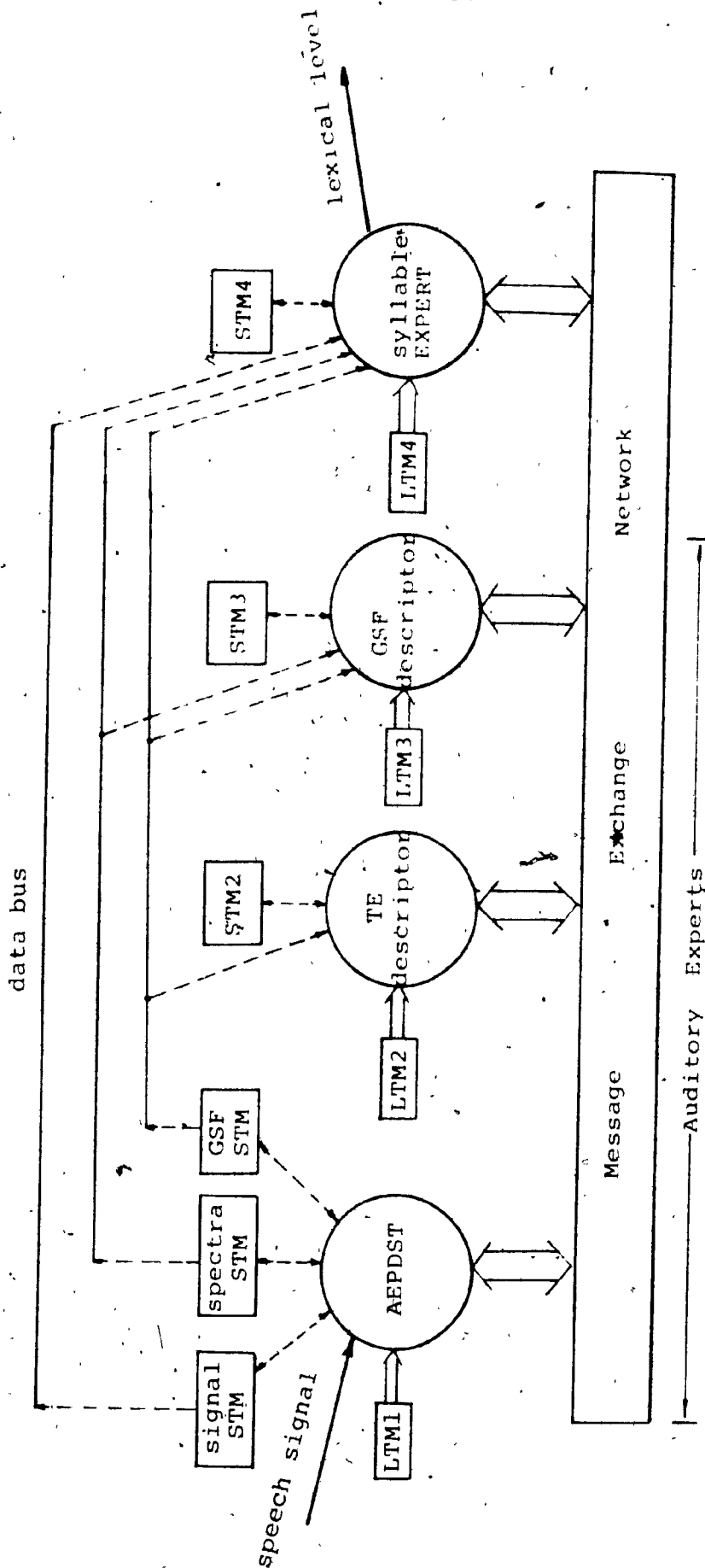


Figure 4.3.1 The Expert Society for segmentation

and control messages are exchanged among Experts through the "message exchange network". Data, cues, descriptions and hypotheses are written by an Expert into its own Short Term Memory (STM), only the Expert which owns the STM can write into it, but any Expert can read any STM.

The speech signal is sampled, quantized, stored into a "SIGNAL-STM" and transformed by an expert called "Auditory Expert for End-Point Detection and Signal Transformation" (AEPDST). AEPDST looks for the starting point of a sentence by using a set of rules for end-points detection. When this point has been detected, AEPDST starts transforming the signal in order to obtain a frequency-domain representation of it which is stored into the "SPECTRA-STM". Some gross spectral features (GSF) are computed from the spectra and stored into the "GSF-STM". The LTM of AEPDST, denoted LTM1, contains rules for end-point detection and spectral transformation. After a long enough part of the signal has been transformed, a synchronization signal is sent to the Expert for the description of the time evolution of the total energy (TE-DESCRIPTOR).

TE-DESCRIPTOR has the task of describing the time evolution of the total energy of the signal (TE) in terms of peaks and valleys based on zero-crossing densities. These descriptions are so-called Primary Acoustic Cues (PAC), as shown in Table 4.3.1 and attributes in that table are described in Table 4.3.2. At the same time, AEPDST

Table 4.3.1 Primary Acoustic Cues

<u>Symbol</u>	<u>Attributes</u>	<u>Description</u>
LPK	tb,te,ml,zx	long peak of total energy (TE)
SPK	"	short peak of TE
MPK	"	peak of TE of medium duration
LOWP	"	low energy peak of TE
LNS	tb,te,zx	long nonsonorant tract
MNS	"	medium nonsonorant tract
SNS	"	short nonsonorant tract
LVI	tb,te,ml,zx	long vocalic tract adjacent to a LNS or a MNS in a TE peak
MVI	"	medium vocalic tract adjacent to a LNS or a MNS in a TE peak
LDD	emin,tb,te,zx	long deep dip of total energy
SDD	"	short, deep dip of total energy
LMD	"	long dip of TE with medium depth
SMD	"	short dip of TE with medium depth
LHD	"	long non-deep dip of TE
SHD	"	short non-deep dip of TE

Table 4.3.2 Attribution Description

<u>Attribute</u>	<u>Description</u>
tb	time of beginning
te	time of end
pi	maximum signal energy in the peak
emin	minimum total energy in a dip
zx	maximum zero-crossing density of the signal derivative in the tract

continues to transform another portion of the signal and send a message to TE-DESCRIPTOR. This operation is repeated until a sentence end point is detected. The LTM of TE-DESCRIPTOR, denoted LTM2, contains a grammar GTE that controls a coding of TE in terms of peaks and valleys.

The attributed grammar GTE has the following form [7, [34] :

$$GTE = \{T_1, N_1, \langle \text{sentence} \rangle, RRI\}$$

The terminal alphabet T_1 of GTE is made of symbols representing peaks and valleys :

$$T_1 = (\text{peak}, \text{dip})$$

The nonterminal alphabet N_1 is made of the starting symbol $\langle \text{SENTENCE} \rangle$ and the term $\langle \text{ZETA} \rangle$. The rewriting rules RRI are :

$$\begin{aligned} \langle \text{SENTENCE} \rangle &::= \langle \text{ZETA} \rangle \\ &::= \langle \text{ZETA} \rangle \langle \text{SENTENCE} \rangle \\ \langle \text{ZETA} \rangle &::= \text{dip peak} \quad \text{alg (Z1)} \\ &::= \text{dip} \quad \text{alg (Z2)} \end{aligned}$$

The time of beginning and the time of end are attributes associated to the detected terminal symbols. Other attributes like the coordinates of the maximum value of a peak and of the minimum value of a dip are also associated to the descriptions generated under the control of GTE. Attribute extraction for TE description is performed by algorithms (Z1) and (Z2) associated to rules of GTE.

Algorithms (Z1) and (Z2) also provide a translation of phrases composed of symbols of the alphabet T1 into more detailed descriptions of peaks and valleys using the symbols introduced in Table 4.3.1.

The algorithms represent an augmentation of grammar GTE. A brief description of the cues they produce is given in the following.

Algorithm (Z1) analyzes each DIP of TE and describes it as LONG or SHORT depending on its duration. Furthermore, for each dip, the difference between the minimum value of the energy in the dip and the energy level of the background noise is computed. If this difference is small, then the dip is described as DEEP; if the difference is large, then the dip is described as HIGH, otherwise it is described as MEDIUM. Deep dip characterizes pause, plosive and continuant consonants and, sometimes, the nonsonorant affricate V. Deep dip tends to be short for voiced (Lax) plosives and long for unvoiced (tense) plosives. High dips characterize sonorant consonants. Medium dip characterizes a large variety of consonants.

Alg (Z1) describes dip and invokes alg (Z2) for peak descriptions. Peaks are described as short (SPK), medium (MPK) or long (LPK) depending on their durations. Nonsonorant tracts inside a peak are described by 'LNS' if they are long, otherwise, they are described by MNS and even

SNS. If a peak contains a nonsonorant and a sonorant track, the latter is described as LVI if long, MVI otherwise. A nonsonorant tract is characterized by high zero-crossing density and low low-to-high frequency energy ratio. The latter parameter is computed by a sensory procedure invoked by alg (Z2) when the value of the zero-crossing density does not show to make a reliable decision about the nature of a tract.

Examples of various types of PACs are shown in Figure 4.3.3. The two curves in Figure 4.3.3 represent the time evolution of the signal energy (—) and the zero-crossing counts (---) in successive intervals of 10 msec of the first derivative of the signal. The phrase is the sequence of letters KCBTD. Table 4.3.3 shows the corresponding PAC description. Time unit is 0.01 sec.

Description of the signal energy (TE) are sent to another expert, called "GSF-DESCRIPTOR", which provides the acoustic cues for segmentation. The organization of the knowledge stored into the Long-Term Memory of the GSF-DESCRIPTOR is described by means of a frame language introduced in [42].

Acoustic Experts can perform various types of signal transformations, extracting and describing acoustic cues. The acoustic cues will be used for indicating spectral or signal properties describing aspects that are relevant for

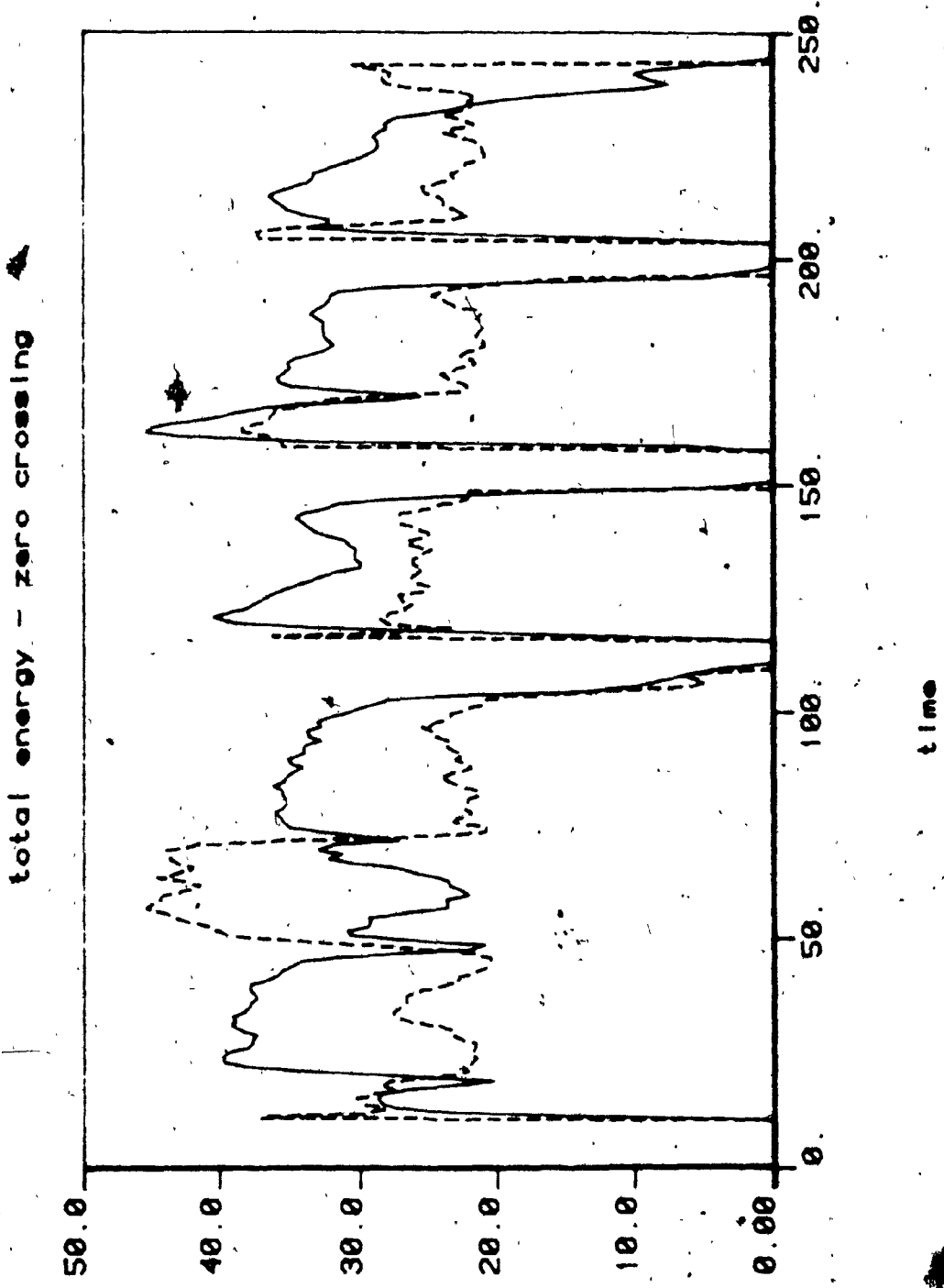


Figure 4.3.3 The time evolution of the signal energy and the zero-crossing counts

Table 4.3.3 The PAC description corresponding to the signal in Figure 4.3.3

<u>PAC</u>	<u>t_b</u>	<u>t_e</u>
LDD	1	10
SSI	10	14
SNS	14	17
LPF	19	48
SHD	48	49
MCS	50	57
LHD	57	63
LNS	63	72
LPF	72	111
LDD	111	116
LPF	116	132
LHD	132	138
LPK	138	151
LDD	151	158
LNS	161	170
LPK	170	198
LDD	198	204
SNS	207	209
LVI	209	238
SMD	238	240

Table 4.4 Primary phonetic features

<u>SYMBOL</u>	<u>FEATURE DESCRIPTIONS</u>
VF	Front vowel
VC	Central vowel
VB	Back vowel
VFC	Front or central vowel
VBC	Back or central vowel
VW	Uncertain vowel
NI	Non-sonorant interrupted consonant
NA	Non-sonorant affricate consonant
NC	Non-sonorant continuant consonant
SON	Sonorant consonant
NIV	The /v/ or a NI consonant
SONV	A sonorant or the /v/ consonant

hypothesizing phonetic features. Except the peaks and valleys of signal energy, these acoustic cues could be formant loci, characteristics of burst spectra, etc.

4.4 The Syllabic Expert for Syllabic Hypotheses

The primary acoustic cues generated by Acoustic Experts are sent to the SE which determine pseudo-syllabic bounds, generate some primary phonetic feature (PPF) using context-independent rules and extract detailed acoustic cues to be used for pseudo syllabic segments (PSS) hypothesization. Table 4.4 shows the alphabet of PPF.

The organization of knowledge stored into the LTM of the Syllabic Expert are described by a frame language. It contains a Semantic Syntax - Directed Transition (SSDT) algorithm. The SSDT algorithm receives at the input PAC descriptions, applies the rules of an attributed grammar and gives PSS hypotheses containing PPF hypotheses as a result of translation. Generation of hypotheses about segments between two successive vocalic intervals and placement of syllabic bounds is also performed under the control of SSDT whose definition is given in the following [46].

A SSDT is a 5-tuple [44]:

$$\text{SSDT} = \{ N, \Sigma, A, S, P \}$$

where :

N is a set of nonterminal symbols;

Σ is a set of input symbols;

Δ is a set of output symbols;

$S \in N$ is the start symbol;

P is the set of rules.

The set of input symbols is given in Table 4.3.1 plus SN and NS . Every symbol is associated a vector of attributes. The meaning of the attributes is given in Table 4.3.2. The set of output symbols is given in Table 4.4. The start symbol S is denoted PSS . The rewriting rules have the following general form :

$$PK \quad X := YB; YG; \quad \text{Alg. (K)}$$

where $X \in N$ is a nonterminal symbol; $Y \in N^*$ is a (possible empty) string of nonterminal symbols; $B \in \Sigma^*$ is a (possible empty) string of input symbols; $G \in \Delta^*$ is a set of strings of output symbols.

The sequences YB, YG can appear in the reverse order, i.e. BY, GY . In any case, Y is in the same position in both expressions.

Alg. (K) is an algorithm that may contain a condition made of a logical expression of predicates defined by semantic attachments; the rule can be applied only if the condition is verified.

Each symbol of N or Σ is associated a vector of attributes. The attributes associated with X belong to the vector $A(X)$. In a similar way, the attributes associated with the symbols of B are grouped into $A(B)$.

The algorithm Alg(K) may contain a semantic rule $f_k(A(Y), A(B))$ which allows to compute the attributes of A(X) of X given the attributes of A(Y) and A(B). Another semantic rule $f'_k(A(B))$ allows to compute the attributes of the output hypotheses G, given the attributes of symbols in B which have been translated into G. The details of the segmentation grammar can be seen in [45]; [34].

Referring back to Figure 4.3.1 on the Expert system, there is an inter-expert communication link between any two Experts through the message exchange network. The Syllabic Expert repeatedly interacts with the AEPDST. The invocation of AEPDST occurs as a result of a frame instantiation in SE. The AEPDST carries out various signal transformations depending on the message it receives from SE. The SE could also retrieve information stored in GSF-STM. This is necessary when the cues generated by TE and GSP are insufficient to make a hypothesis of PSS or PPF.

Basically, Auditory Experts may perform "spontaneous" data-driven activities and expectation-driven activities based on requests issued by other Experts. The Syllabic Expert has a spontaneous activity in which it receives primary acoustic cue descriptions and generates PSS hypotheses with Primary Phonetic Feature hypotheses. PPF hypotheses are sent to the lexical level and are used to access a lexical subset, based on which requests for a detailed detection of the place and manner of articulation

of some phonemes will be issued. SE also generates scores of syllabic hypotheses. This can be explained in the next chapter.

CHAPTER V.

THE PLANNING SYSTEM FOR E1-SET RECOGNITION

5.1 Why Planning

After segmentation, hierarchical feature extracting is performed. Hypothesis generation should be performed at the highest possible level as long as matches are good, dropping down toward the acoustic level only when the higher-level matching process gets in trouble [51]. For example, in our experiment, in order to distinguish between /p/ and /t/ the place of articulation is the only distinctive feature and its detection may require the execution of special procedures on a limited portion of the signal with a time resolution finer than 10 msec. This suggested to introduce plans for hypotheses generation and disambiguation [49] [50].

The planning concept comes from A.I. Techniques. For complicated problem solving, it can be worked out on small pieces of a problem separately and then combine the partial solutions at the end into a complete problem solution. Planning just focuses on ways of decomposing the original problem into appropriate subparts and on ways of recording and handling interactions among the subparts as they are detected during problem-solving process.

The recognition of unconstrained sequences of connected letters is a problem unsolved so far. Using a redundant set of acoustic properties for characterizing place and manner of articulation of some sounds makes it possible to have an accurate phoneme hypothesization even in difficult protocols. The hierarchical application of recognition algorithms for plosive sounds recognition has been introduced in our system by using planning. In this system, computer perception of speech is modelled with perceptual plans containing operators. These operators may translate a description of acoustic properties into more abstract descriptions or they may extract useful properties. Operators may also contain the execution of sensory procedures. Operator application is conditioned by the verification of some preconditions depending on already generated descriptions [48].

5.2 The System of Plans

The speech signal is first analyzed on the basis of loudness, zero-crossing rates and broad-band energy using the expert system described in the last chapter. The result of this analysis is a string of symbols and attributes. Symbols belong to an alphabet of Primary Acoustic Cues (PAC), after that a SSDT algorithm operates on PAC descriptions and through the use of sensory procedures identifies the vocalic and the consonantal segments of syllable nuclei and for the vocalic segments hypothesizes

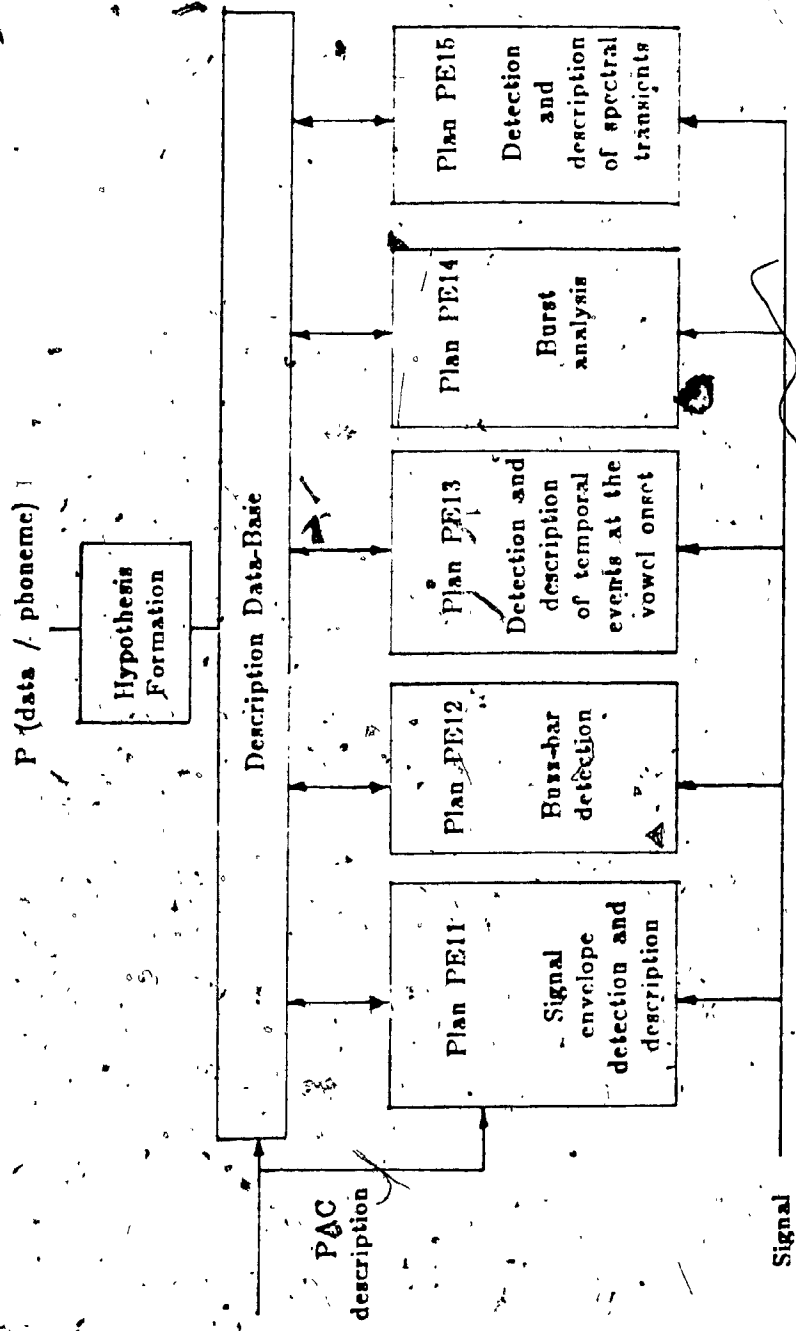


Figure 5.2.1 An overview of the plan for the recognition of the L1-set

the place of articulation of the vowel. Plans are then applied for interpreting the consonantal segment of every syllable. An overview of the plan for the recognition of the EI-SET is shown in Figure 5.2.1.

The plan is subdivided into sub-plans (PE11, PE12, PE13, PE14, PE15).

PE11 produces an envelope description by analyzing the signal amplitude before and after preemphasis. Envelope samples are obtained every msec by taking the absolute value of the difference between the absolute maximum and the absolute minimum of the signal in a 3 msec interval. The envelope description is based on the following alphabet (\sim represents negation):

AL = {SHORT-STEP (ST),
NO-STEP (NST),
STEP WITH HIGH-LOW-FREQUENCY ENERGY (NZ),
BURST-PEAK (BUR),
POSSIBLE-BURST (PBU),
NBU = \sim BUR,
NBZ = \sim BZ,
NPB = \sim PBU.}

Figure 5.2.2 shows the envelope curve in the energy dip preceding the onset of /b/. The time reference 0 in Figure 5.2.2 corresponds to time 115 in the Figure 4.3.3(a). Because of the shape of the envelope, this segment is

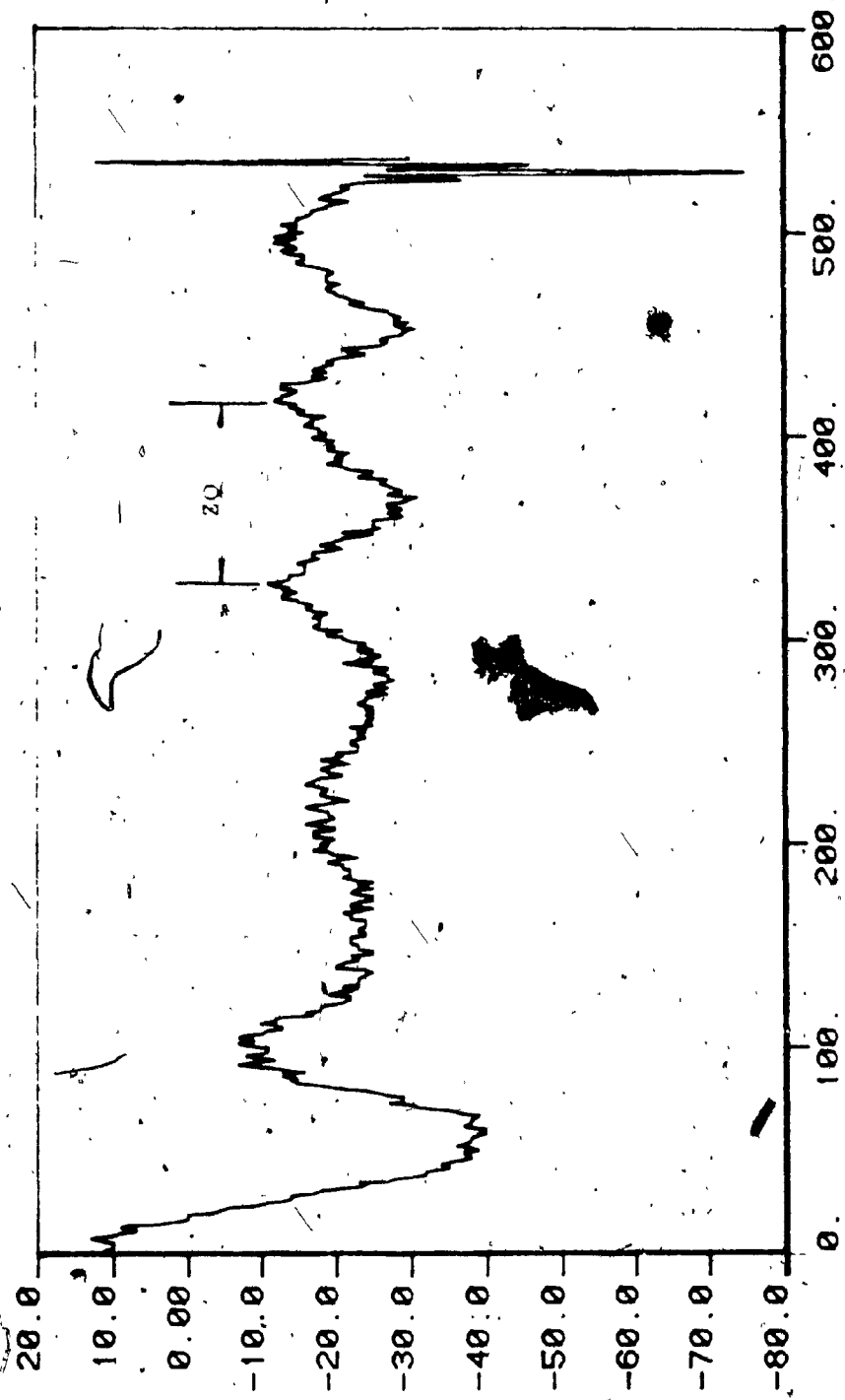


Figure 5.2.2 Envelope curve in the energy dip preceding the onset of 'B'

buzz-bar spectrum

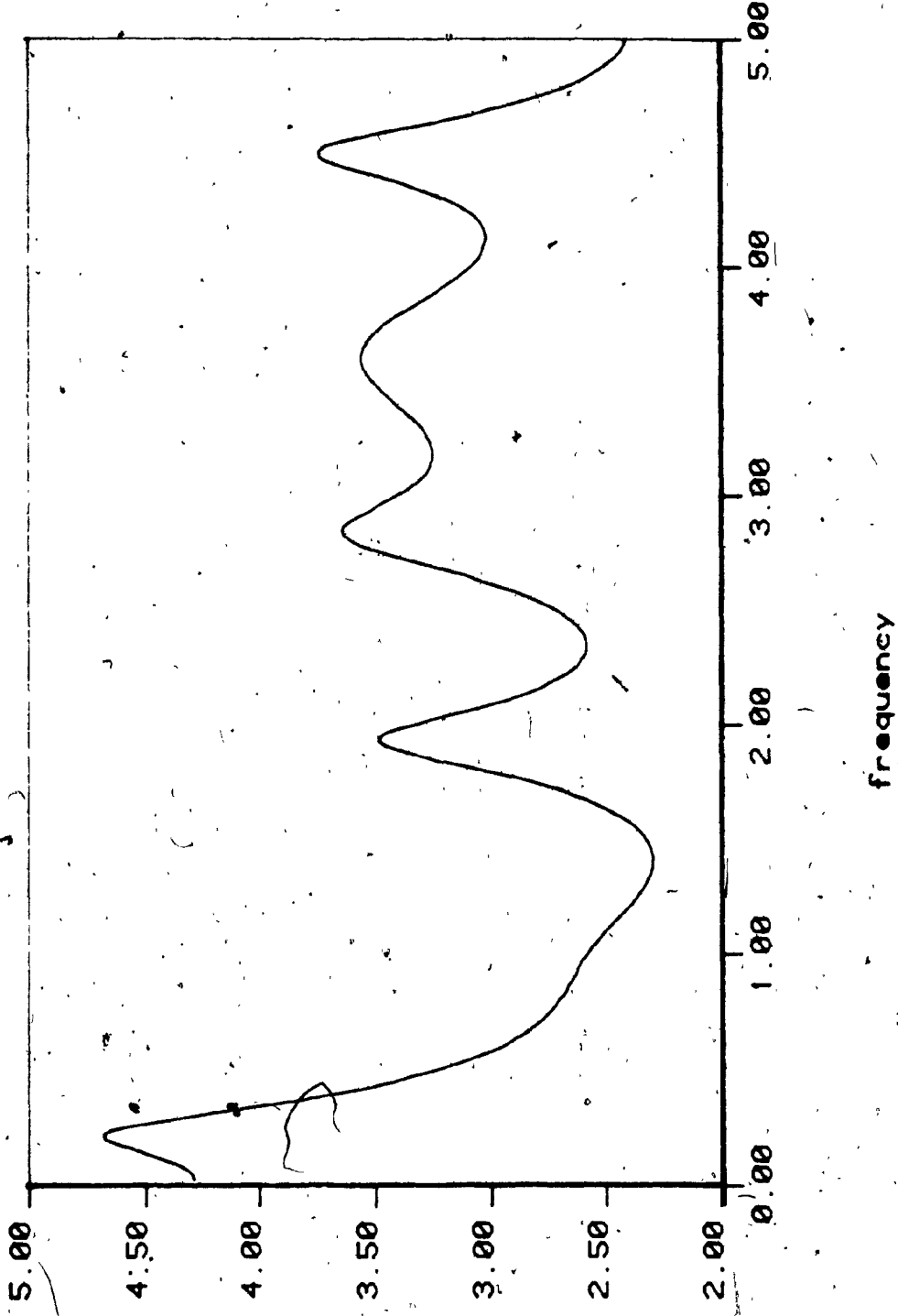


Figure 5.2.3 The spectrum of the signal shown in Figure 5.2.2

described as BZ.

PE12 describes the buzz-bar by analyzing the shape of the time waveform and of the spectra before the voice onset.

The alphabets of the descriptions it produces are:

BZA1 = { NOB, BU1, BU2, BU } /

for time waveform and

BZA2 = { NBP, BP1, BP2, BP3, BP }

for the spectra.

NOB and NBP mean no buzz and the other three symbols describe degree of buzz-bar evidence (BU1, BP1 : little evidence, BU, BP : strong evidence). Based on the waveform of Figure 5.2.2 the segment is described as BU. Figure 5.2.3 shows the spectrum of the signal shown in Figure 5.2.2. It is described as BP.

PE13 analyzes temporal events at the voice onset. These events are related to voice onset time. They are:

D ----- the delay between the onset of low and high frequency energies.

ZQ ----- the duration of the largest zero-crossing interval of the signal at the onset.

ZR ----- the number of zero-crossing counts in the largest (sequence of successive zero-crossing intervals with duration less than 0.5 msec).

Figure 5.2.2 shows an example of ZQ.

PE14 performs burst detection and analysis. Burst is characterized by the detection of a short energy peak (SPK in Table A.3(2)) or of short envelope peak or of a short peak in 2-4 KHz energy at the voice onset. PE14 also detects frication intervals at the voice onset. They are characterized by peaks of the zero-crossing density of the signal first derivative and dips of the ratio of low to high frequency energy. Burst and frication spectra are computed from the coefficients of an 8-pole model of the vocal tract only when burst and frication are detected. The spectra are described with the following alphabet:

BSA = { COMPACT,
 DIFFUSE FALLING,
 DIFFUSE RISING,
 SONORANT-LIKE.}

SONORANT-LIKE spectra are those presenting resonances with frequencies and band-width comparable to those of the vocalic sounds. For the compact spectra some parameters are further computed, such as the frequency corresponding to gravity center of the highest peak and the ratio between the maximum energy and the average in the 0.2-1.0 KHz band. Figure 5.2.4 shows the burst peak of /k/ in the 11-20 centisecond interval and in the 2-4 KHz band. Figure 5.2.5 shows the compact burst spectrum of /k/.

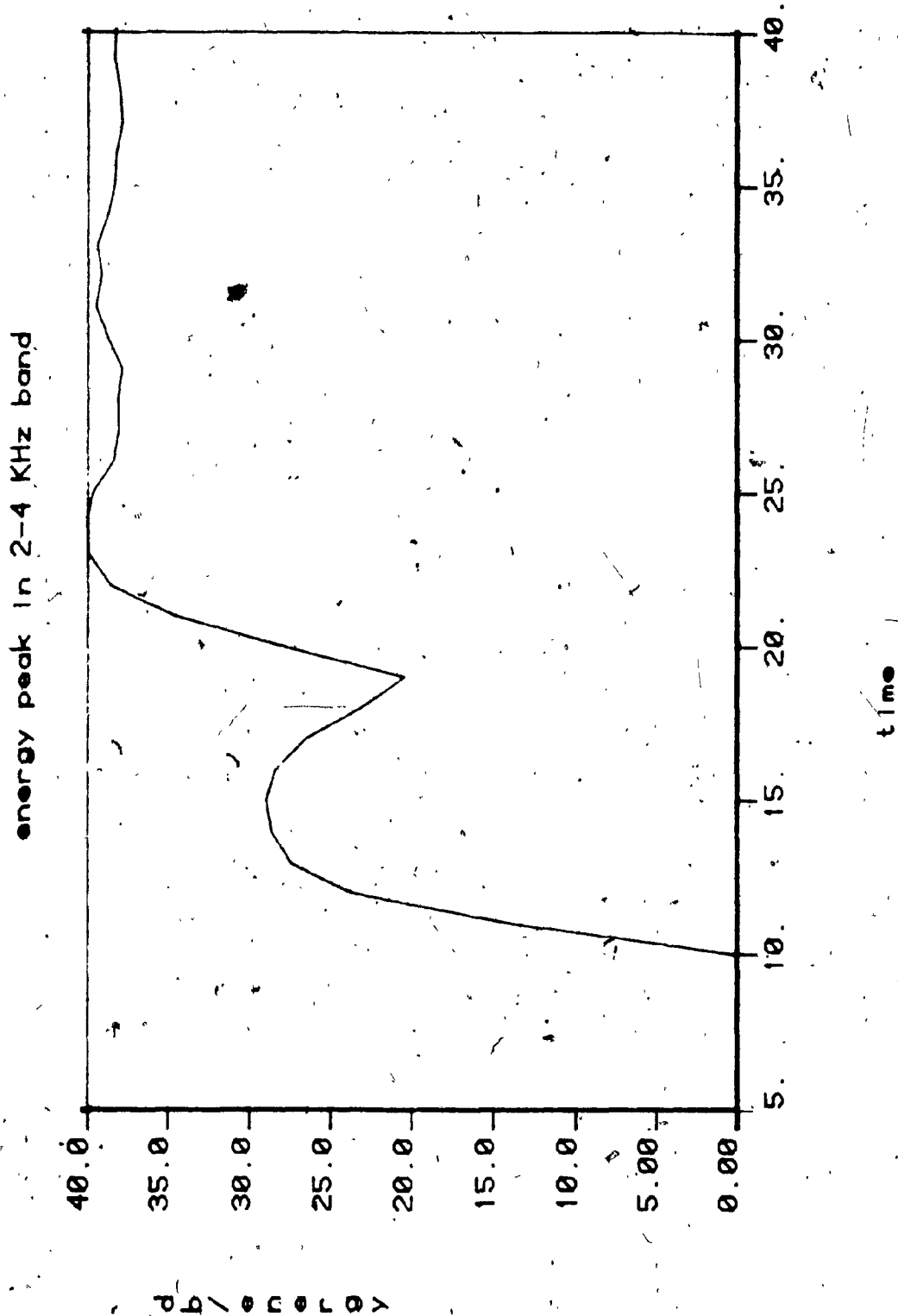


Figure 5.2.4 The burst peak of /K/ in the 11-20 centisecond interval and in the 2-4 kHz band

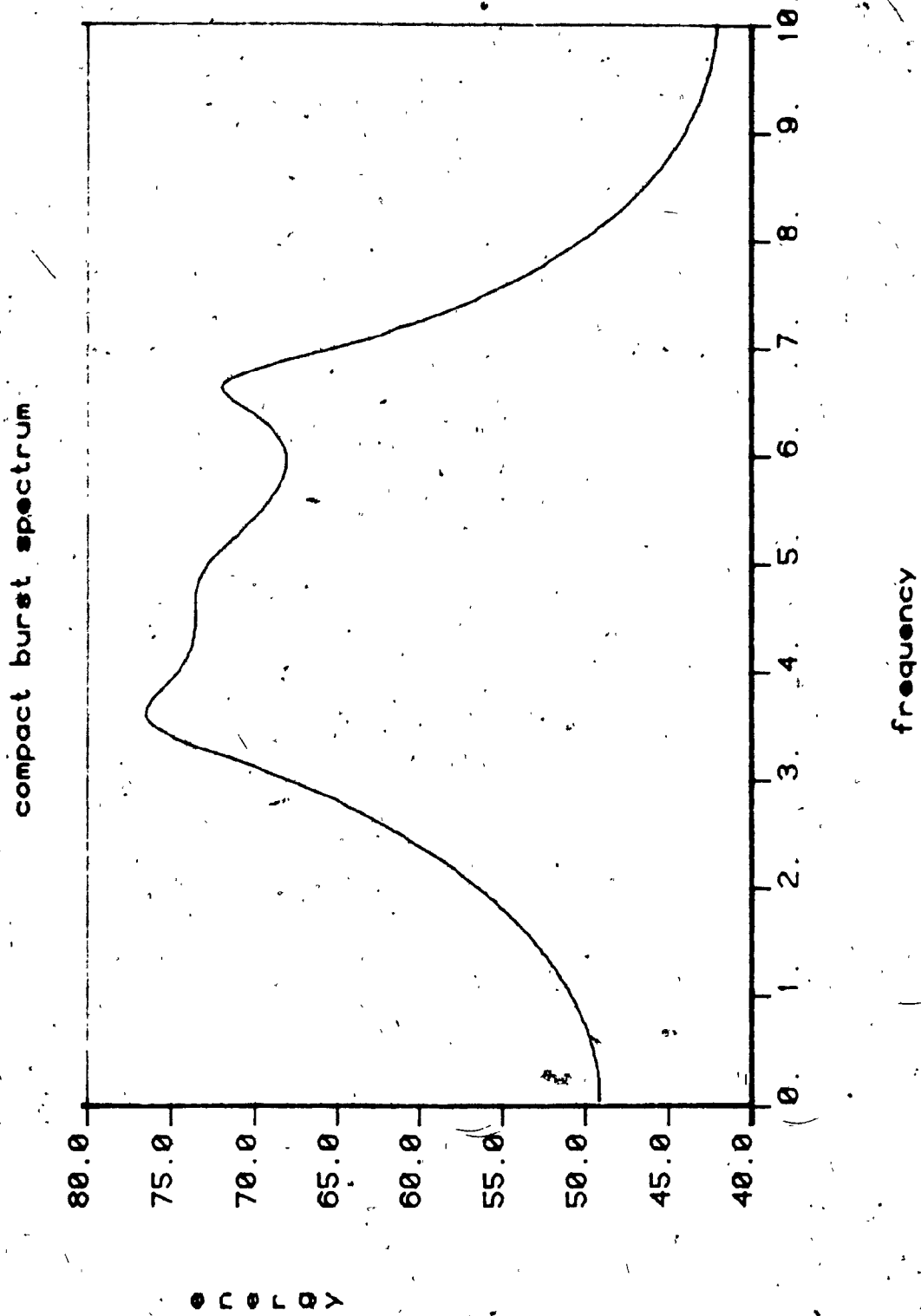


Figure 5.2.5 Compact burst spectrum of /K/

PE15 describes transitions of the second and third formant at the voice onset. Formants are tracked backwards from the steady-state portion of the vowel. Formant transitions are described by the following alphabet :

FTA = { ASCENDANT,
 QUASI-ASCENDANT,
 HORIZONTAL,
 DESCENDANT. }

Preconditions for plan execution are learned with a general-purpose algorithm whose details are given in [52]. The algorithms of parameter extracting and analyzing used by subplans will be introduced in the next section.

5.3 Fuzzy Algorithm for Feature Hypothesisization

The parameters extracted by PE13, PE14, and PE15 are used in fuzzy relations for feature hypothesisization. This use of fuzzy algorithms is in order to model, to some extent the fact that most of the acoustic-phonetic properties of speech sounds are only known with a degree of vagueness. For example, the concentration or spread of energy (diffuse vs. Compact) and whether the main spectral trend is rising, flat, or falling with frequency are crucial spectral aspects; etc. In this section, some basic Fuzzy Set Theory [53] [7] will be introduced before the explanations of the methods for feature hypothesisization using fuzzy relations.

5.3.1 Introduction to Fuzzy Set Theory

Definition Of Fuzzy Set

The theory of fuzzy sets deal with a subset S of the universe of discourse U , where the transition between full membership and no membership is gradual rather than abrupt. Fuzzy set theory, introduced by Zadeh in 1965 [36], is a generalization of abstract set theory. A fuzzy set is a class that admits the possibility of partial membership in it.

Let $U = \{ u \}$ denote a space of objects. Then a fuzzy set F in U is a set of ordered pairs, expressed as :

$$F = \{ (\mu, u) \}, \quad u \in U$$

where μ is termed " the grade of membership of u in F ".

Usually, the grades of membership are assumed to be real numbers lying in the interval $[0,1]$, with 0 and 1 denoting no membership and full membership respectively.

A fuzzy set can be also defined introducing a function $\mu_F(u)$ mapping a set U into the unit interval $[0,1]$ as follows :

$$F = \int_U \mu_F(u)/u$$

where \int_U represents the union of the FUZZY SINGLETON $\mu_F(u)/u$ for all $u \in U$.

The grades of membership reflect an "ordering" of objects in the universe. It is interesting to note that the

grade of membership value $\mu_F(u)$ of an object u in F can be possible to interpret $\mu_F(u)$ as the degree of possibility that u is the value of a parameter fuzzily restricted by F .

Operations On Fuzzy Sets

If F_1 and F_2 are fuzzy subsets of U , then their union (\cup), intersection (\cap), and complement, ($\text{COMPL}\{ \}$) operations are defined as follows :

$$(1) \text{ union } F_1 \cup F_2 = \int_U (\mu_{F_1}(u) \vee \mu_{F_2}(u)) / u$$

where \vee means "by definition" and \vee is an max operator which takes the maximum of $\mu_{F_1}(u)$ and $\mu_{F_2}(u)$.

$$(2) \text{ intersection } F_1 \cap F_2 = \int_U (\mu_{F_1}(u) \wedge \mu_{F_2}(u)) / u$$

where \wedge is the min operator which takes the minimum of the operands.

(3) complement $\text{COMPL}\{F\}$ of a fuzzy set F of U is defined by :

$$\text{COMPL}\{F\} = \int_U (1 - \mu_F(u)) / u$$

It is clear that a fuzzy algebra can be introduced based on above operations which satisfy with the strict algebra axiom system. It is interesting that Boolean algebra is a specific case of a fuzzy algebra [53] [7].

Fuzzy Restrictions And Possibility Distributions

A fuzzy set F can be used to define the extent to which an element $u \in U$ possesses a certain property X . This

property can be represented by a variable taking values in U .

The property X may define a binary valued restriction (hard bounded interval) on U ; in this case a variable takes value 1 for every element of U having the property X , 0 otherwise.

A property X may also induce a fuzzy restriction defined over U ; in this case the restriction is represented by a fuzzy subset F of U acting as an elastic constraint (soft bounded interval) on the elements of U which may possess the property X . Based on fuzzy restriction, the concept of possibility distributions can be introduced.

Let x be a variable taking value in U and F represents a fuzzy restriction, $R(x)$ associated with x . Then the proposition "x is F" which translates into :

$$R(x) = F$$

associates a possibility distribution π_x with x which is postulated to be equal to $R(x)$.

A possibility distribution π_x associated with x is defined as :

$$\pi_x(u) = \mu_F(u) \quad \forall u \in U.$$

$\pi_x(u)$ is the possibility that $x=u$.

Possibility relates to the perception of the degree of feasibility whereas probability is associated with the

concept of frequency of occurrence. The only connection between them is that impossibility (zero possibility) implies improbability but not vice versa.

Although possibilities are non-statistical in nature, this does not prevent one from using statistics in the estimation of membership functions. But this estimation does not require necessarily as large a number of experiments as the estimation of a probability density. The induction of a possibility density can be largely influenced by the a priori knowledge of human experience and is less constrained.

In investigating models for speech recognition, we are interested in evaluating the possibility that a feature t is in a pattern P , i.e.;

POSS (t is in P).

The algorithm used for evaluating the possibility of a hypothesis will be referred as fuzzy algorithm.

Fuzzy relations and fuzzy languages are also important concepts which will be used in the fuzzy algorithm in the later. They are omitted for the sake of brevity. It is also easier to understand after these.

5.3.2 The Basic Mathematical Model of Fuzzy Algorithms

In this section, the fuzzy algorithm model will be explained through an example in the following.

Let $H_p(p)$ be a hypothesis about a plosive sound in the acoustic pattern p . $H_p(p)$ takes values on the following set P of phoneme labels for plosive sound :

$$P = [p, t, k, b, d, g] .$$

Hypotheses in P are assigned by a fuzzy algorithm, which is executed whenever a consonant is hypothesized in an interval $[t_i, t_j]$ of the acoustic pattern and the possibility that the consonant may be nonsonorant and interrupted is high enough.

The fuzzy algorithm for generating plosive hypothesis is based on a set of fuzzy composite questions of the type :

$$Q(H_p) = \text{"is } H_p \text{ in } P(t_i, t_j)\text{"}$$

where H_p takes values in P .

The universe of discourse U of $Q(H_p)$ is the set of all possible acoustic patterns, the body B is a structured linguistic variable having a label belonging to P and the answer set A of $Q(H_p)$ is a set of linguistic a posteriori possibilities, expressing the evidence of H_p in p , based on the evaluation of the possibilities :

$$\text{POSS} \{ H_p \text{ is in } p(t_i, t_j) \} \quad \forall H_p \in P .$$

Each structural linguistic variable representing a phonemic hypothesis is a triple $H' = (H_p, U, R(H))$, where U is the previously defined universe of acoustic patterns, H_p is a label in P , and $R(H)$ is a fuzzy restriction of U .

associated with H_p ; $R(H)$ defines the meaning of H_p .

So, each plosive phoneme can be regarded as a fuzzy linguistic variable defining a set of P on which H takes values.

The fuzzy restriction $R(H)$ defined over an acoustic cue U could be seen as a fuzzy relation which can be characterized by a vector of break-points:

$$V(u) = [u_1, u_2, \dots, u_n], \quad u_i \in U \quad (i = 1, 2, \dots, n).$$

let the corresponding labels of the fuzzy restrictions are:

$$(K_{1u}, K_{2u}, \dots, K_{mu}), \quad m \leq n .$$

Figure 5.3.2 shows an example of fuzzy restriction $R(H)$. Where K_{1u} covers the lowest part of the interval over the U -axis. The membership of K_{1u} takes value 1 for $u \leq u_2$ and decreases linearly, taking the value 0 on u_3 . The membership of fuzzy restriction K_{2u} assumes value 1 for $u_2 \leq u \leq u_3$ and decrease linearly to 0 from u_2 to u_1 and from u_3 to u_4 . It is possible that there are some intervals on which memberships are all equal to 0.

These fuzzy restrictions over acoustic cue U can be get by some observations of histogram and designer's knowledge.

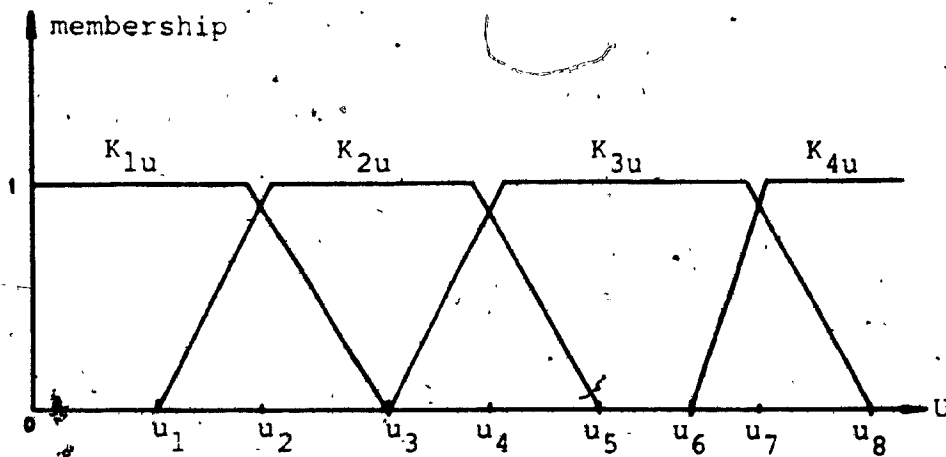


Figure 5.3.2 Fuzzy restrictions defined over an acoustic cue U

5.3.3 Parameter Extracting and Feature Hypothesization

The phonemes involved in El-Set contain plosive (in K, B, T, D, P), fricative (in S, V, C), affricate (in G) and vacalic (in E) sounds, etc.

Plosives are characterized acoustically by a period of prolonged silence, followed by an abrupt increase in amplitude at the consonant release; the release is accompanied by a burst of fricative noise. Fricatives are detectable by the presence of turbulence noise. The affricate are often considered to be a plosive followed by a fricative. Vowels can be detected by the presence of substantial energy in the low- and mid-frequency regions. They are characterized mainly by the steady state value of the first three formant frequencies.

So, the important time intervals for extracting acoustic cues could be Pre-Onset Time, Voice Onset Time (VOT) defined as the duration between the release and the onset of normal voicing for the following vowel, and transition interval from consonant to the following vowel.

In the pre-onset time, the voiced stops are OFTEN prevoiced, they create the voice-bar or buzz-bar in the low-frequency region during the closure interval. It is important to detect the occurrence of quiescence or buzz-bar in this interval for recognition of plosives. Subplan PE12 is of this task.

In the voice onset time, subplan PE13 analyzes temporal properties and PE14 achieves the burst or frication spectra analysis. The duration of silence following the burst is critical to plosive detection. Voiced stops tend to lengthen the duration of the preceding vowel [54] [27]. A silence duration exceeding 70 ms means that the sound is tagged as a fricative [7] [57]. The duration of the frication in affricate is typically half as long as in other occurrences of those fricatives. Recognition according to place of articulation for plosive is done by finding the frequency location and the relative strength of the major concentrations of energy in the burst spectrum. Fricatives can often be determined by examining the gross spectral shape during the fricative. [57] [58].

The acoustic cues analysed during the temporal of transitions are concerned with formant transition, i.e. : formant pseudo-loci and formant slopes. Formant pseudo-loci are defined as the first formant samples which are detected at the beginning of the transition, before the vowel following the plosive. Formant slope is defined as the difference between the frequency of the pseudo-locus and the frequency of the second formant when the formant amplitude reaches the absolute maximum on a vowel. Subplan PE15 will give the descriptions about formant slopes as mentioned in the last chapter.

The parameters extracted by PE13, PE14, PE15 are used in fuzzy relations. The features are described by a set of fuzzy sets which will be introduced in the following.

Feature Hypothesization By PE13

The parameters extracted by PE13 are D, ZQ, ZR (see section 5.2). The labels for the fuzzy restrictions over D are

$$\{ SD, MD, LD \}$$

with the corresponding vector of break-points :

$$V(D) = [3, 10, 12, 16, 20] .$$

where SD means "short delay", and LD means "long delay", etc. The membership function of the fuzzy restriction is shown in Figure 5.3.3 (a).

The labels of fuzzy restrictions over ZQ are

$$\{ SZQ, ZQ1, ZQ2, LZQ \}$$

with the corresponding vector of break-points :

$$V(ZQ) = [10, 25, 40, 50, 60, 70, 80] .$$

where SZQ means "short ZQ", and LZQ means "long ZQ", etc. The membership function of the fuzzy restriction is shown in Figure 5.3.3 (b).

The labels of fuzzy restrictions over ZR are

$$\{ LZR, ZR1, ZR1, HZR \}$$

with the corresponding vector of break-points :

$$V(ZR) = [10, 15, 20, 25, 35, 45, 60, 70] .$$

where LZR means "little ZR", and HZR means "high ZR", etc.

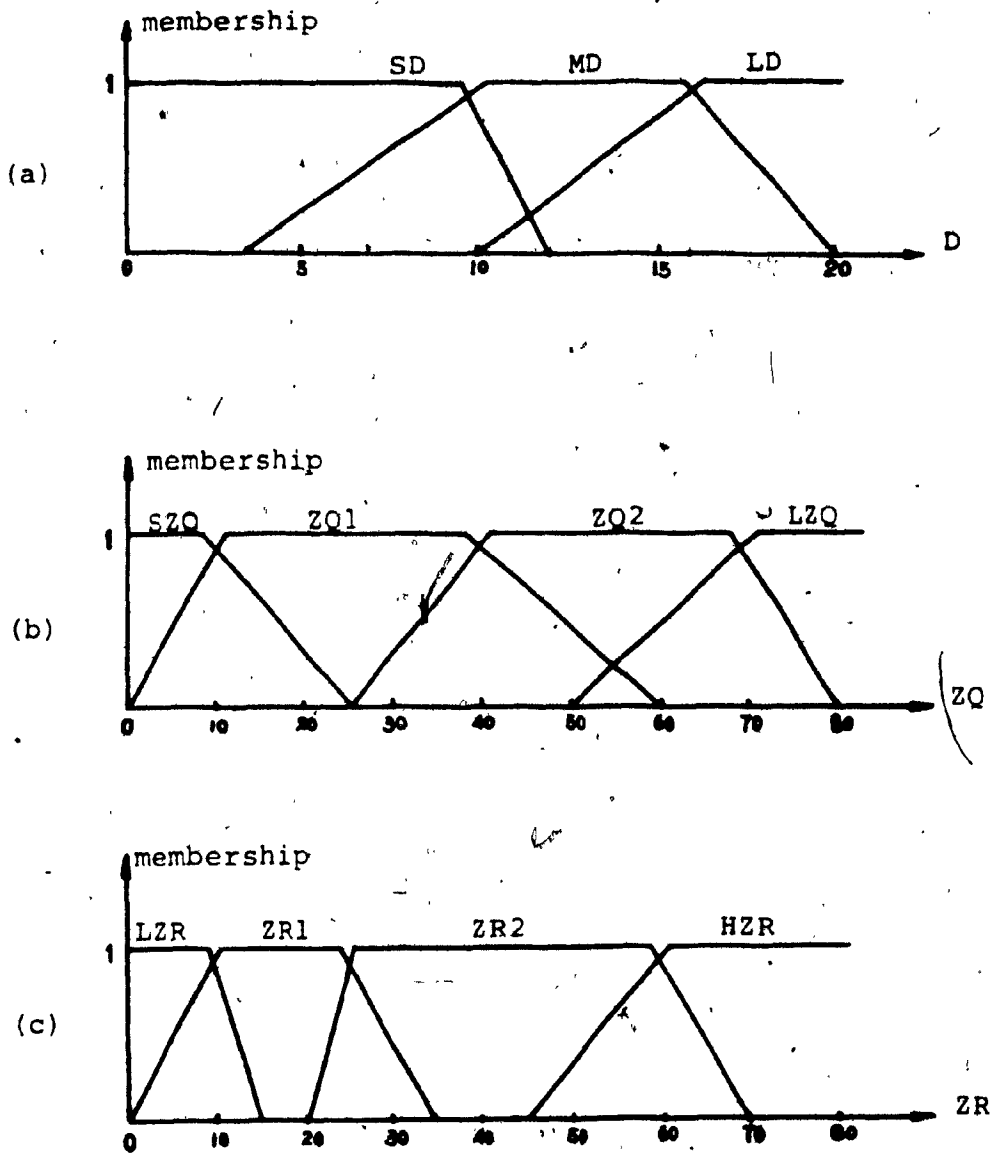


Figure 5.3.3 The membership function of the fuzzy restrictions over the parameters in PE13

the membership function of the fuzzy restriction is shown in Figure 5.3.3 (c).

Feature Hypothesisization By PE14

PE14 provides a fuzzy language FL for burst analysis. This fuzzy language FL is a fuzzy relation from the set of terms BSA (see section 5.2) to a Cartesian product space

$$U = U_1 \times U_2 \times \dots \times U_n,$$

each U_i , $i=1,2,\dots,n$, is one kind of parameter set and is characterized by a membership function :

$$\mu_{FL} : BSA \times U \rightarrow [0,1]$$

for any element $E = (u_1, u_2, \dots, u_n) \in U$, where $u_i \in U_i$ and $i=1,2,\dots,n$, there is a

$$\mu_{FL}(x/E), \forall x \in BSA,$$

which defines a fuzzy set $D(E)$ in BSA called "descriptor" and serves to characterize the extent to which each term in BSA describes a given element E of U . According to the relations between BSA and plosive sounds or others, we can get the similarity evaluation for each sound.

In the above, FL has been described as a binary relation, actually it could be a n -ary relation. Anyhow the point is how to determine the μ_{FL} . In the following, this fuzzy relation will be analyzed in more detail.

On the other hand, the fuzzy relation FL can also be seen as a fuzzy restriction induced by BSA defined over U . In this case the restriction is represented by a fuzzy

subset D of U acting as an elastic constraint on the element of U which may possess the property

$$x \in BSA.$$

Now let D_1, D_2, \dots, D_n be fuzzy subsets of U_1, U_2, \dots, U_n respectively, each D_i can be seen as the fuzzy restriction defined over U_i , $i = 1, 2, \dots, n$. According to the theory of fuzzy sets, their Cartesian product is :

$$\begin{aligned} D &= D_1 \times D_2 \times \dots \times D_n \\ &= \int_{U_1 \times U_2 \times \dots \times U_n} \mu_{D_1}(u_1) \wedge \mu_{D_2}(u_2) \wedge \dots \wedge \mu_{D_n}(u_n) / (u_1, u_2, \dots, u_n), \end{aligned}$$

and D is a fuzzy subset of $U = U_1 \times U_2 \times \dots \times U_n$ which can be defined as the fuzzy restriction over U .

Where $u_i \in U_i$, $i = 1, 2, \dots, n$.

This result is very useful, it means that the fuzzy restriction defined over $U = U_1 \times U_2 \times \dots \times U_n$ can be get from each fuzzy restriction defined over U_i , $i = 1, 2, \dots, n$.

So far, the mathematical model of feature hypothesization for PE14 has been built up. In the following, we will introduce various parameter sets or elements in U and its corresponding fuzzy restriction, which can be further combined from a set of fuzzy restrictions in it [7].

In our system $U = U_1 \times U_2 \times U_3$, this means that three kind of parameters have been considered for burst spectra

analysis, etc. They are :

U_1 : frequency corresponding to the gravity center of the highest peak in the 2-7 KHz band.

U_2 : the ratio between the maximum energy and the average energy in the 0.2-1.0 KHz band.

U_3 : frequency corresponding to the highest peak over the spectra.

Their fuzzy restrictions are based on the feature distributions in this 3-ary space of each plosive sound or others so that the property $x \in BSA$ for each plosive sound could be represented by a string of labels or properties inducing these fuzzy restrictions. In this way, the fuzzy relation for each sound can be generated flexibly.

The labels (not need to be strictly meaningful in this level) for the fuzzy restrictions over U_1 are :

{ A1, B1, C1, D1, E1 }

with the corresponding vector of break-points :

A1	0.8	1.2	1.8	2.2
B1	1.6	2.1	3.9	4.3
C1	3.6	4.1	5.1	5.4
D1	4.8	5.2	5.9	6.4
E1	5.6	6.0	7.5	8.0

The labels for fuzzy restrictions over U_2 are.

{ A2, B2, C2, D2, E2, F2 }

with the corresponding vector of break-points :

A2	8.0	10.0	15.0	17.0
B2	12.0	15.0	20.0	27.0
C2	15.0	20.0	27.0	32.0
D2	25.0	30.0	38.0	41.0
E2	35.0	40.0	50.0	55.0
F2	46.0	52.0	60.0	70.0

The labels for the fuzzy restrictions over U_3 are

{ A3, B3, C3, D3, E3, F3, G3 }

with the corresponding vector of break-points :

A3	2.4	2.7	3.0	3.2
B3	2.8	3.1	3.5	3.7
C3	3.3	3.5	3.8	4.0
D3	3.8	4.0	4.4	4.6
E3	4.2	4.5	4.8	5.0
F3	4.85	5.0	5.4	5.6
G3	5.3	5.5	6.2	6.5

The figures of the membership functions of the fuzzy restrictions are omitted here since they are similar to Figure 5.3.3.

Feature Hypothesization By PE15

The algorithm model of PE15 is same as that of PE14. The fuzzy language presented by PE15 is exactly a binary FUZZY relation. Based on the parameter distributions of some statistics of the sounds, the corresponding fuzzy restrictions could be get and the fuzzy relations for the

sounds can be generated. The details of the fuzzy restrictions are omitted here for the sake of brevity.

5.4 Hypothesizing Generation Rules

Learning rules from examples can be seen as the process of generalizing descriptions of positive and negative examples and previous learned rules to form new candidate rules. When applied incrementally, this methodology can produce results which depend on the order in which examples are supplied and on the occurrence of examples which are exceptions to the relevant rules. Incremental learning of rules has to come out with a set of rules that is the most consistent with the examples encountered so far [52].

Expressions made of symbols extracted by subplans PE11 and PE12 and representing positive and negative examples have been inferred for each PAC description and for each phoneme using the learning algorithm presented in [52].

An example of such rules is given in the following :

E := NOB NBP NBZ NST NBU NPB

B := BU BP BZ NST NBU PBU

There are 96 of such rules in the system [48].

A PAC description is used for indexing a set of rules that is matched against the input description produced by the plan. As rules and descriptions contain the same number of symbols, a similarity index S_1 between a rule and a

descriptions is computed by the following algorithm.

Algorithm Similarity (rule,description)

Begin

 C := 0

 D := 0

 for each I do

 begin

 if rule-symbol(i) matches description-symbol(i)

 then D := D + 1

 else D := D - 1 ;

 C := C + 1

 end;

 similarity := (D + C) / 2C

End.

Another similarity S2 is computed from PE13 by using MAX operator for disjunctions and by summing the contributions of each clauses and dividing the sum by the number of clauses.

An example of clauses involved in fuzzy relations is the following :

E := (Short D) (Short ZQ) (LOW ZR)

K := (Long D) (Short ZQ) (HIGH ZR)

where "short, long, high, low" are defined by corresponding fuzzy restrictions over the cues. There are 43 of such relations.

Similarly, the fuzzy relation from PE15 compute the similarity S3. An example of these relations is the following :

T := (Horizontal 1) (Horizontal 2)

B := (Ascendant 2) (Ascendant 3)

there are 8 of such relations in the system.

The fuzzy relation from PE14 compute the similarity S4, the relations are more complicated. They reflect the rules between each phoneme and a set of parameters under certain preconditions. They have a mathematical background in the fuzzy algebra [55] [56]. An example of these fuzzy relations is the following :

POSS{K} := POSS{C2} * POSS{E1} + POSS{B3} * POSS{B2}

where the operators * and + are MAX and MIN operators, respectively. There are 36 of such relations.

5.5 Hierarchical Recognition Strategy

Recognition of the sounds in E1-SET is Bottom-Up. At the first level, if the PAC description is unambiguous then a decision is made immediately. If not, the PAC description will invoke a set of rules and recognition process goes to the next level.

At the second level, similarity measurements are computed for the hypotheses generated by PE11, PE12 and PE13 and averaged. The parameter

$$S12 = (S1 + S2)/2$$

is used for selecting the three candidates having the highest similarity with the data. If the algorithm finds only one candidate, then decision is made. If not, formant transitions are analyzed by PE15 at the third level and a new similarity value S3 is computed for the three candidates. S3 is used for changing or confirming the ordering established by S12. If the algorithm for analyzing formant transitions does not find acceptable formants, then S3 is not used. In this case, if preconditions (see 5.2) for executing PE14 are found, then a last similarity value S4 is computed. Usually S4 is very reliable. It seems that burst and transitions complement each other in the sense that when one cue is weak, the other is strong.

During the whole recognition process, expectations are built up using a-priori knowledge and parameter histograms. Candidates are then ranked according to how well do they match expectations based on a voting criterion. Table 5.5 shows the similarity S1, S2, S3, S4 in a testing experiment.

Table 5.5 The similarity evaluations for the connectedly spoken letters KCBDT

SYMBOL		PRECONDITION	SIMILARITIES				AVERAGE
K	LDD SNS LPK	S1(K) = -	S2(K) = -	S3(K) = -	S4(K) = 0.732	S ₁₂₃₄ = 0.732	
		S1(T) = -	S2(T) = -	S3(T) = -	S4(T) = 0.000	S ₁₂₃₄ = 0.000	
C	SHD LNS LVI	Identified as C based on preconditions					
B	LDD LPK	S1(B) = 0.929	S2(B) = 0.800	S3(B) = 0.500	S4(B) = -	S ₁₂₃₄ = 0.866	
		S1(V) = 0.333	S2(V) = 0.870	S3(V) = -	S4(V) = -	S ₁₂₃₄ = 0.602	
		S1(G) = 0.333	S2(G) = 0.750	S3(G) = -	S4(G) = -	S ₁₂₃₄ = 0.542	
		S1(P) = 0.429	S2(P) = -	S3(P) = -	S4(P) = -	S ₁₂₃₄ = 0.429	
		S1(D) = 0.500	S2(D) = 0.333	S3(D) = -	S4(D) = -	S ₁₂₃₄ = 0.417	
T	SDD SNS SHD LPK	S1(T) = 0.929	S2(T) = 0.973	S3(T) = -	S4(T) = 0.519	S ₁₂₃₄ = 0.879	
		S1(G) = 0.929	S2(G) = 1.000	S3(G) = -	S4(G) = 0.182	S ₁₂₃₄ = 0.834	
		S1(D) = 0.917	S2(D) = 0.343	S3(D) = -	S4(D) = 0.000	S ₁₂₃₄ = 0.525	
		S1(E) = 0.563	S2(E) = 0.321	S3(E) = -	S4(E) = -	S ₁₂₃₄ = 0.442	
D	LDD SNS SHD LPK	S1(D) = 1.000	S2(D) = 1.000	S3(D) = -	S4(D) = -	S ₁₂₃₄ = 1.000	
		S1(T) = -	S2(T) = 0.370	S3(T) = -	S4(T) = -	S ₁₂₃₄ = 0.370	

note: "--" means that similarity is not computed in that condition.

CHAPTER VI

SYSTEM PERFORMANCE AND EVALUATION

6.1 Experimental Results and Discussions

The system has been tested on a protocol of 1000 connected pronunciations of symbols of the E1 set in strings of five symbols each. The strings were pronounced by five male and five female English speakers. The voice from two male and two female speakers has been used for deriving the rules. The average recognition rate is around 90% which is preeminent [59]. The system response time for each complete string is around 1.5 minutes. The results are satisfactory since implementation is on a timesharing machine. In addition, the system has been conceived in a distributed processing model, but now it is implemented in a classical sequential machine. Real-time operation would be reached by implementation on a parallel computer architecture.

In the following, we will analyze the elements which affect the recognition accuracy or performances.

Figure 6.1 shows the system performance improvement when more subplans are introduced. A set of experiments have been carried out to check the quality of subplans. In the beginning, only subplans PE11, PE12 and PE13 were used, the

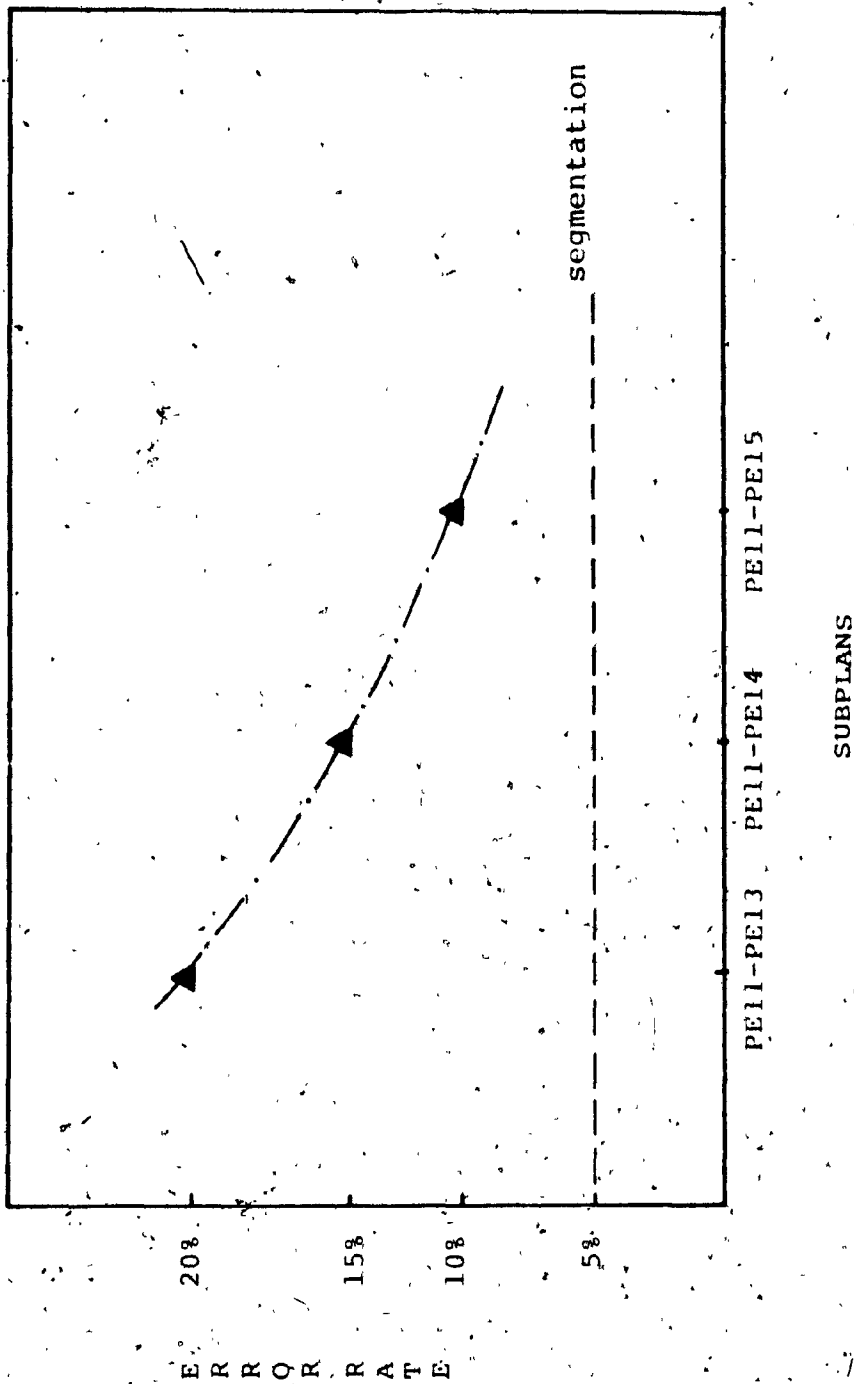


Figure 6.1 The system performance improvement by introducing more subplans

system achieved 80% correct recognition of the letters of E1 set. When PE14 was put into the system, the correct recognition rate reached 86%. Finally the system arrives at around 90% correct recognition when all of designed subplans worked cooperatively. Further great improvement was limited by the system segmentation accuracy which achieved 95% correct detections of syllables.

The experiments have shown that the planning system is a powerful tool for extracting features in continuous speech and an a priori syllable segmentation followed by recognition, the recognition accuracy is limited by the algorithm's ability to segment continuous speech into syllable-sized units [60]. Even so, the expert system approach to segmentation has shown an excellent result instead of the following conditions :

- (1) Some letters in the string had been pronounced by error because of the carelessness of long time repeat.
- (2) The letters in the middle of the strings had been pronounced too weak because of the customs with some speakers. This case happened very often since the speakers usually pronounced very heavy in the beginning and the end, the data acquisition instruments still accepted that kind of speech data as "good" one.
- (3) Some speech signals had been missed since speaker spoke earlier than the onset of the acquisition.
- (4) The heavy noise and other persons walking in and out

disturbed the data acquisition quality.

- (5) No any speaker had been trained for the way of speaking. The speaking speed is quite different from one to another among the ten speakers.

Since some of the above cases are avoidable, the system performance could be improved in the further experiments by no means of any refinements of the algorithms.

Another set of experiments on testing the system are shown in Table 6:1. Each row in the Table 6.1 shows the absolute error recognition rate of corresponding letter or letters(segmentation error is not included), and especially its error distribution among subplans or what aspects and how extent that errors result from. For example, the recognition of letters B or D has the absolute error rate 1.5% and 70% of these errors come from the mistakes of PE11, PE12 and PE13; 15% of these errors come from PE14 and so on.

From Table 6.1 we can see that the burst analysis and gross spectra description about the properties in transient time are very reliable for the recognitions of plosive and fricative segments. The formant transition analysis is also very important for the recognition of plosive sounds. Another aspect we can see is that the acoustic-phonetic decoding of a sentence is a major bottleneck in continuous speech recognition.

Table 6.1 The recognition error and distribution

	PE11, PE12, PE13	PE14	PE15,	error rate
** ***				
P, T, K	66 %	13 %	20 %	1.9 %
B, D	70 %	15 %	15 %	1.5 %
E	60 %		40 %	0.4 %
G	82 %	17 %		0.3 %
V	99 %			0.7 %
C, 3	90 %	10 %		0.3 %

*: subplan

**: error distribution

***: El-set letter

Other important facts observed are that the burst of plosive sounds is often missed in continuous speech and for voiced stops, the buzz-bar is not easily detected. This is probably the reason why voiced stops have relatively higher error recognition rate in the first part of Table 6.1.

In addition, Table 6.1 has shown that /E/, /G/ and /V/ have considerable high error recognition rate. From experiments we have found that /E/ is usually confused with /T/, /P/, /B/, and /G/ is usually confused with /P/, /D/, /T/, and vice-versa. The former observation can be explained as follows. In many cases the prevocalic transient of /T/ and /P/ or /B/ is missed and the plan analyzing the beginning of the vowel classifies it as a /E/. In some other cases, the onset of /E/ from the silence is preceded by a transient that makes it appear similar to the transient of /T/ or /P/. Probably, a better burst characterization will reduce this error [34]. /G/ is an affricate with characteristics between plosive and fricative sounds. Introducing ~~formant~~ transition analysis rules for /G/ and /V/ will probably reduce the errors in the recognition of this sound.

Other recognition errors in /C/ and /3/ are very small. The confusion table is shown in Table 6.2 which does not include 5% segmentation error.

Table 6.2 The confusion table

		PRONOUNCED									
		P	T	K	B	D	G	C	V	E	3
R E C O G N I Z E D	P	93		1	2		2			2	
	T		92	4		2	2			1	1
	K		2	94						1	
	B	5			93	4			3	1	
	D		2		2	90			2		
	G		3		1	2	95		1		1
	C						1	100			
	V				2	2			92		1
	E	2		1					1	95	
	3		1						1		97

Segmentation Error Rate: 5%.

Learning for knowledge acquisition now mainly consists of automatic collecting and classifying the facts, such as break-point data for fuzzy sets, etc. It is possible to introduce complete automatic learning. This can be achieved soon.

6.2 Conclusions

The multi-speaker computer recognition system for the recognition of connectedly spoken letters in the E1 set has been successfully implemented. A set of algorithms for extracting acoustic cues and phonetic features has been introduced. These algorithms have been embedded into a planning system. The experimental results have shown that these features can be useful for segmenting continuous speech into syllabic segments and for a bottom-up generation of hypotheses. The system recognition accuracy is satisfactory. The artificial intelligence methodology of incorporating in the system some kind of reasoning on the basis of speech knowledge at each level appears very promising. Some of the characteristics of the system can be simply summarized as follows.

- (1) As the recognition algorithm is syllable based, the recognition is not constrained by the number of syllables and the order they appeared in the string.
- (2) A distributed knowledge base system allows one easy to find a detailed explanation of the errors indicating along which directions the system should grow.

- (3) Burst analysis and formant transition analysis complement each other for providing important cues of characterizing plosive sounds.
- (4) Phonetic features are characterized by acoustic properties. Redundancies in this representation improve the recognition accuracy.
- (5) Statistical algorithm has been used for clustering parameters used by plan. Clusters are characterized by fuzzy sets which provide useful methods for describing speaker independent knowledge.
- (6) The excellent segmentation performances partition the difficulty of continuous speech recognition in speaker-independent circumstances.
- (7) Various methods for knowledge representation have been used in a distributed knowledge base. Frames and rule models structured knowledge uses different types of acoustic analysis and of phonetic features in phonetic decoding of continuous speech and controlling the recognition process in planning system which is more sophisticated control than that in classical system.
- (8) Distributed processing model makes it possible to reach real-time performance in parallel computer architectures.

Further work shown in the following is probably useful for improving system performance.

- (a) The acoustic-phonetic decoding of a sentence is critical

both to the segmentation and the hypothesization in continuous speech recognition. The results generated from Acoustic Expert and Syllabic Expert have to be as good as possible.

- (b) More parameters can be extracted in burst analysis in order to improve the accuracy of description about burst property. Before doing this, the trade off between the dimensions of the n-ary fuzzy relation and the processing complexity should be considered.
- (c) The accurate formant tracking of natural continuous speech is still an open question. Better algorithms for formant analysis will improve the quality of the analyses. The experimental results have shown that stop burst and ensuing formant transitions have equivalent perceptual weight.
- (d) The strategy for controlling the recognition process in the planning system could be improved. Results have shown that the acoustic cues are rich enough for a reliable hypothesization.
- (e) The total software system structure can be refined in order to decrease the response time. Even though, the most important speed-up can be achieved by introducing parallel processing for achieving real-time performance.
- (f) Automatic learning by machine is very important and has to be performed. Its realization will allow acquisition of large population of speakers :

REFERENCES

- [1] L.R. RABINER, R.W. SCHAFER, Digital Processing Of Speech Signals. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.
- [2] C.Y. SUEN and R. DE MORI (ED.), Computer Analysis And Perception Of Visual And Auditory Signals. Boca Raton, Fla: CRC Press, 1982.
- [3] A. HOLBROOK and G. FAIRBANK, Diphthong Formants And Their Movements. J. Of Speech and Hearing Research, vol.5, No.1, pp 38-58, 1962.
- [4] J.L. FLANAGAN, Speech Analysis, Synthesis And Perception. 2nd Ed., Springer-Verlag, New York, 1972.
- [5] P.C. DELATTRE, A.M. LIBERMAN, and F.S. COOPER, Acoustic Loci And Transitional Cues For Consonants, J. Acoust. Soc. Am., Vol., 27, No. 4, pp 769-773, July 1955.
- [6] O. FUJIMURA, Analysis Of Nasal Consonants, J. Acoust. Soc. Am., Vol. 34, No. 12, pp 1865-1875, December 1962.
- [7] R. DE MORI, Computer Models Of Speech Using Fuzzy Algorithms, Plenum Press, New York, 1983.
- [8] FANT, G., Acoustic Theory Of Speech Production, The

Hague : Mouton, 1960.

- [9] STEVENS, K.N. And BLUMSTEIN, S.E., Invariant Cues For Place Of Articulation In Stop Consonants. Journal of the Acoustical Society of America, 1978, 64, pp 1358-1368.
- [10] COOPER, F.S., DELATRE, P.C., LIBERMAN, A.M., BORST, J.M. And GERSTMAN, L.J., Some Experiments On The Perception Of Synthetic Speech Sounds. Journal of the Acoustical Society of America, 1952, 24, pp. 597-606.
- [11] Research On Speech Perception, Department of Psychology, Indiana University, Technical Report No.3, Dec. 1980.
- [12] LIBERMAN, A.M., etc. The Role Of Consonant-Vowel Transitions In The Perception Of The Stop And Nasal Consonants, Psychological Monographs, 1954, 68, pp. 1-13.
- [13] COLE, R.A. And SCOTT, B., The Phantom In The Phoneme : Invariant Cues For Stop Consonants, Perception & Psychophysics, 1974, 15, pp. 101-107.
- [14] ZUE, V.W., Acoustic Characteristics Of Stop Consonants : A Controlled Study, Technical Report No. 523, Lincoln Laboratory, M.I.T., May 1976.
- [15] DORMAN, M.F., STUDDERT-KENNEDY, M. And RAPHAEL, L.J., Stop-Consonant Recognition : Release Burst And Formant

Transitions As Functionally Equivalent,
Context-Dependent Cues. Perception & Psychophysics,
1977, 22, pp: 109-122.

[16] C. LARIVIERE, H. WINTZ and E. HERRIMAN, Vocalic
Transitions In The Perception Of Voiceless Initial Stops
J. Acoust. Soc. Amer., Vol.57, No.2, 470-475, 1975.

[17] DREFUS-GRAF, J. (1949), Sonograph And Sound Mechanics,
Journal of the Acoustical Society of America, Vol.22,
pp. 731-739.

[18] DAVIS, K.H., R. BIDDULPH, and J. BALASHEK, Automatic
Recognition Of Spoken Digits, Journal of the Acoustical
Society of America, Vol. 24, pp. 637-645, 1952.

[19] DUDLEY, H., and S. BALASHEK, Automatic Recognition Of
Phonetic Patterns In Speech, Journal of Acoustical
Society of America, Vol. 30, pp. 721-739, 1958.

[20] DENES, P. And M.V. MATHEWS, Spoken Digit Recognition
Using Time-Frequency Patterns Matching, Journal of the
Acoustical Society of America, Vol. 32, pp. 1450-1455.

[21] OTTEN, K.W., Automatic Recognition Of Continuous
Speech, Technical Report AFAL-TE-66-408, AF Avionics
Laboratory, Wright-Patterson AFB, Ohio, 1966.

[22] WAYNE A. LEA, Speech Recognition : Past, Present, And
Future, TRENDS IN SPEECH RECOGNITION, W.A. Lea (Ed.),

pp. 39-98, 1980.

- [23] R. DE MORI, S. RIVOIRA, and A. SERRA, A Speech Understanding System With Learning Capability, Proc. Of the 4th International Joint Conference Artificial Intelligence, Tbilisi, USSR, 1975.
- [24] ITAKURA, F., Minimum Prediction Residual Principle Applied To Speech Recognition, IEEE Trans. On Acoustic Speech, and Signal Processing, Vol. ASSP-23, 67-72, 1975.
- [25] SCHAFER, R.N. And L.R. RABINER, Parametric Representations Of Speech, SPEECH RECOGNITION, REDDY (Ed.), 1975, pp. 99-150.
- [26] FANT, G., Stops In CV-SYLLABLES. In G. Fant SPEECH SOUNDS AND FEATURES. Cambridge, MA: MIT, 1973, 110-139.
- [27] ZUE, V.W., Acoustic Processing And Phonetic Analysis, TRENDS IN SPEECH RECOGNITION, W.A. Lea (Ed.), 1980.
- [28] STEVENS, K.N. And BLUMSTEIN, S.E. , The Search For Invariant Acoustic Correlates Of Phonetic Features. In P.D. Eimas and J. Miller (Eds.), PERSPECTIVES ON THE STUDY OF SPEECH, 1980.
- [29] STUDDERT-KENNEDY, M. , Perceiving Phonetic Segments . In T.F. Myers, J. Lover, and J. Anderson (Eds.), THE COGNITIVE REPRESENTATION OF SPEECH, Amsterdam :

North-Holland, 1980.

- [30] CHOMSKY, N. And HALLE, M., The Pattern Of English, Harper & Row, New York, 1968.
- [31] FANT, G., The Native Of Distinctive Features, STL-QPSR4/1, 1966.
- [32] RONALD W. SCHAFFER, LAWRENCE R. RABINER, Digital Representations Of Speech Signals, Proc. Of the IEEE, vol. 63, no. 4, April 1975.
- [33] STUDDERT-KENNEDY, M., Speech Perception, Status Report On Speech Research SR-59/60, 1979.
- [34] R. DE MORI, P. LAFACE, Y. MONG, Parallel Algorithms For Syllable Recognition In Continuous Speech, IEEE Trans. On Pattern Analysis And Machine Intelligence, 1984.
- [35] ELAINE RICH, Artificial Intelligence, Mc Graw-Hill, 1983.
- [36] ZADEH, L.A., Fuzzy Sets, Information and Control, vol. 8, 1965.
- [37] ZADEH, L.A., Fuzzy Logic And Approximate Reasoning, Synthese, vol. 30, 1975.
- [38] MINSKY, M., A Framework For Representing Knowledge, Psychology For Computer Vision, Winston, P. (Ed.),

— McGraw Hill, 1975.

- [39] R. DE MORI, Y. MONG, M. PALAKAL, C.Y. SUEN, Network System For Generating Syllabic Hypothesis In Continuous Speech Recognition In A Speaker-Independent Environment, International Conference On Systems, Man, and Cybernetics, Bombay, 1983.
- [40] J.B. KUIPERS, A Frame For Frames : Representing Knowledge For Recognition, in REPRESENTATION AND UNDERSTANDING, Bobrow and Collins (Eds.), Academic Press.
- [41] R. DE MORI, Automatic Speech Recognition, in APPLICATION OF PATTERN RECOGNITION, K.S. FU (Eds), CRC Press, 1982.
- [42] R. DE MORI, A. GIORDANA, P. LAFACE and L. SAITTA, An Expert System For Interpreting Speech Patterns, Proc. AAAI Conference, Pittsburgh, PA, pp.107-110, 1982.
- [43] TAI, J.W. • And FU, K.S., Semantic-Syntax-directed Translation For Pictorial Pattern Recognition, Purdue University Report, TR-EE 81-38, 1981.
- [44] K.C. YOU and K.S. FU, A Syntactic Approach To Shape Recognition Using Attributed Grammars. IEEE Trans. On System, Man and Cybernetics SMC-9, pp. 334-345, 1979.
- [45] R. DE MORI, A. GIORDANA, P. LAFACE : Speech

Segmentation And Interpretation Using A Semantic Syntax-Directed Translation. Pattern Recognition Letters, vol. 1, no. 2, pp. 121-124, 1982.

[46] R. DE MORI, A. GIORDANA, P. LAFACE, Phonetic Feature Hypothesization In Continuous Speech, IEEE ICASSP vol. 1, pp. 316-319, 1983.

[47] LOWERRE, B.T., The Harpy Speech Recognition System, Ph.D. Thesis Carnegie-Mellon University, Pittsburgh, PA.

[48] R. DE MORI, G. ROSSI and JIANLI SUN, Multi-Speaker Computer Recognition Of Ten Connectedly Spoken Letters, IEEE ICASSP, 1985.

[49] DE MICHELIS P., R. DE MORI, LAFACE P. And O'KANE M., Computer Recognition Of Plosive Sounds Using Contextual Information, IEEE Trans. ASSP, vol. ASSP-31, no. 2, pp.359-377, April 1983.

[50] KOPEC G.E., Voiceless Stop Consonant Identification Using LPC Spectra; Proc. IEEE Conferece on ASSP, San Diego, Col. 4211-4214.

[51] E.P. NEUBURG, Philosophies Of Speech Recognition, in SPEECH RECOGNITION, D. R. REDDY (Eds.), 1975.

[52] R. DE MORI and GILLOUX M. : Inductive Learning Of Phonetic Rules For Automatic Speech Recognition, Proc.

CSCSI-84.

[53] ABRAHAM KANDEL, Fuzzy Techniques In Pattern Recognition, Wiley-Interscience, Press, 1982.

[54] House, A.S. and FAIRBANKS, G., The Influence Of Consonant Environment Upon The Secondary Acoustical Characteristics Of Vowels, J. Acoust. Soc. Am., vol. 25, pp. 105-113, 1953.

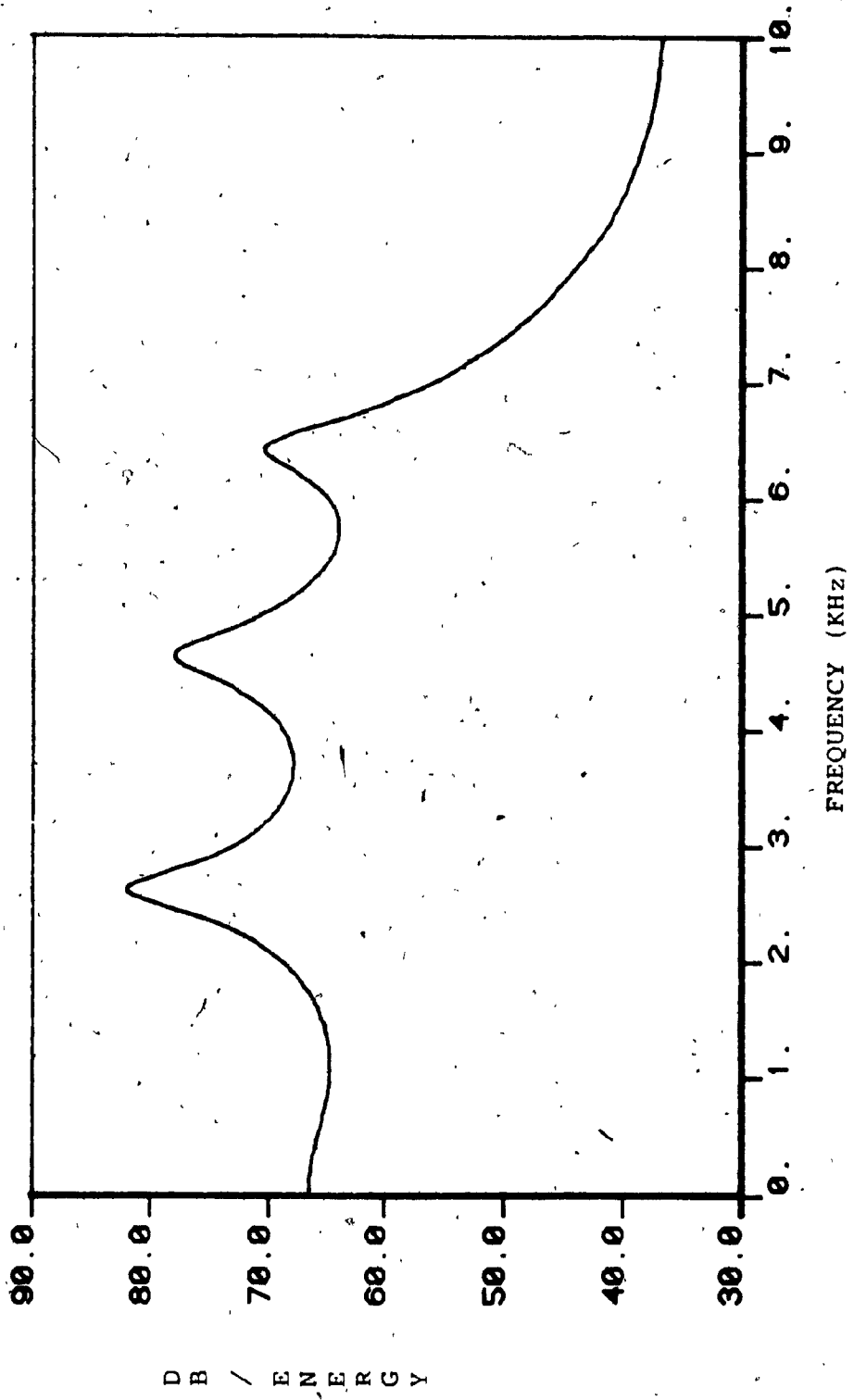
[55] R. DE MORI, P. LAFACE : Use Of Fuzzy Algorithm For Phonetic And Phonemic Labeling Of Continuous Speech, IEEE Trans. On Pattern Analysis And Machine Intelligence, vol. PAMI-2, no. 2, March, 1980.

[56] P. DEMICHELIS, R. DE MORI, P. LAFACE, and M. O'KANE, Computer Recognition Of Plosive Sounds Using Contextual Information, IEEE Trans. On Acoustic, Speech, and Signal Processing, vol. ASSP-31, no. 2, 1983.

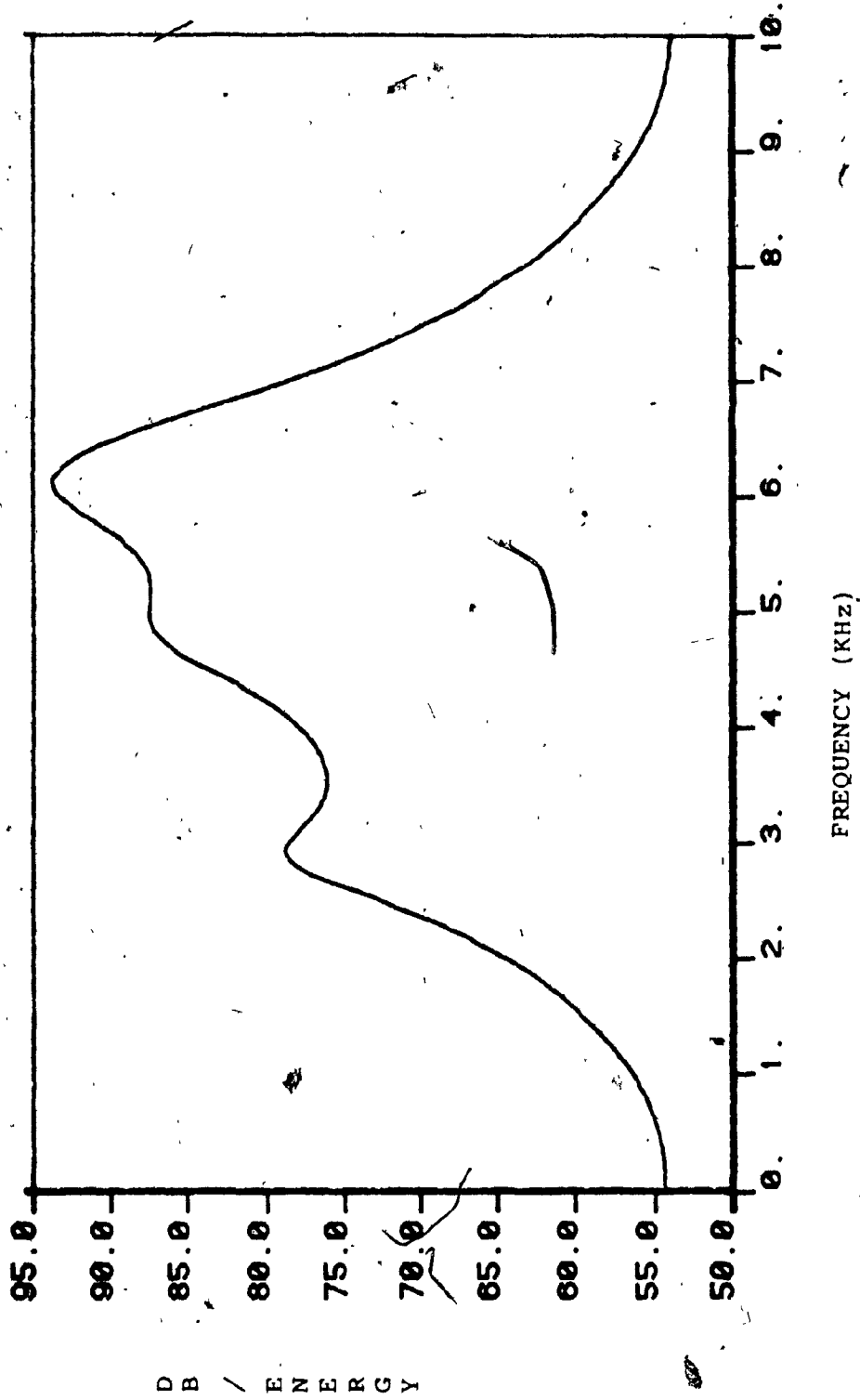
[57] WEINSTEIN, C.J., McCANDLESS, S.S., MONDSHEIN, L.F. And ZUE, V.W., A System For Acoustic-Phonetic Analysis Of Continuous Speech. IEEE Symposium On Speech Recognition, Carnegie-Mellon University, Pittsburgh, PA, pp.89-100.

[58] SCHWARTZ, R.M., and V.W. ZUE, Acoustic-Phonetic Recognition In BBN SPEECHLIS, IEEE ICASSP, pp. 21-24, 1976.

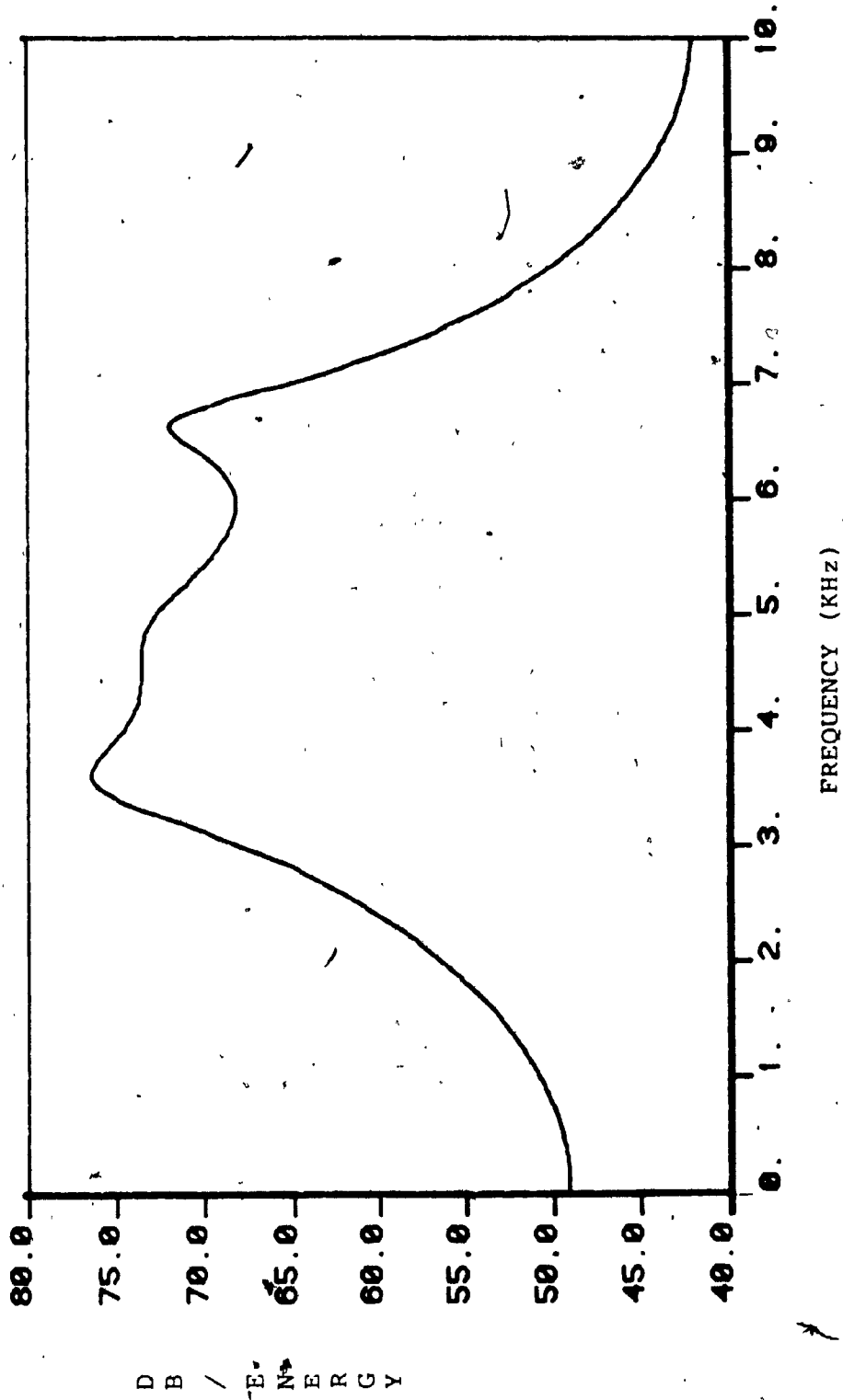
- [59] JEAN-PAUL Haton and JEAN-PAUL DAMESTOY, A Frame Language For The Control Of Phonetic Decoding In Continuous Speech Recognition, IEEE ICASSP, vol. 4, 1985, pp. 1565-1568.
- [60] JEAN-PAUL BRASSARD, Integration Of Segmenting And Nonsegmenting Approaches In Continuous Speech Recognition, IEEE ICASSP, vol. 4, 1985, pp. 1217-1220.
- [61] ANNA MARIA COLLA, etc, A Connected Speech Recognition System Using A Diphone-Based Language Model, IEEE ICASSP, vol. 4, 1985, pp. 1229-1232.
- [62] LIBERMAN, A.M., COOPER, F.S., etc, Perception Of The Speech Code, Psychological Review, 1967, 74, pp.431-461.



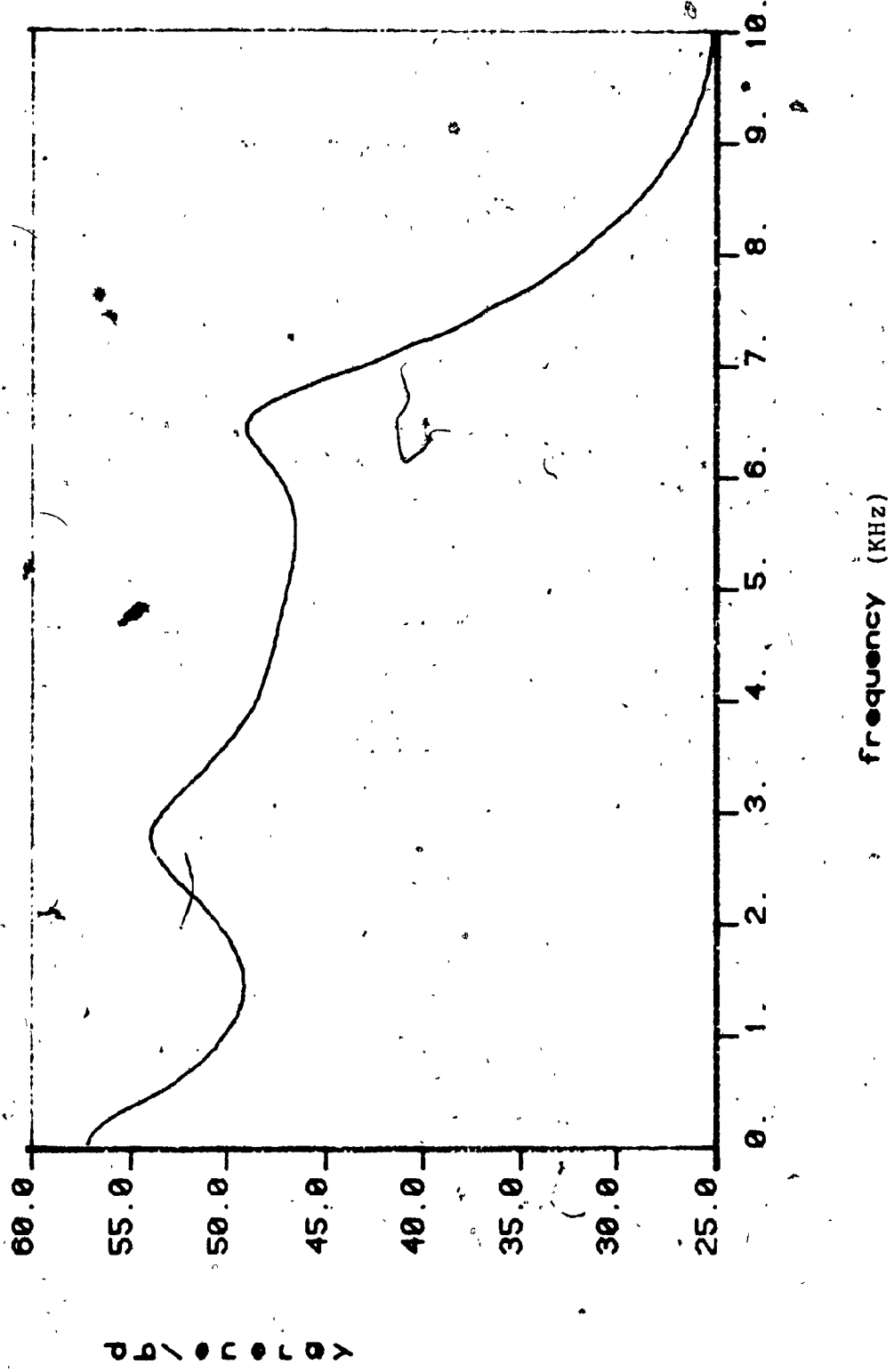
Appendix I Burst spectra of the plosive sound /p/



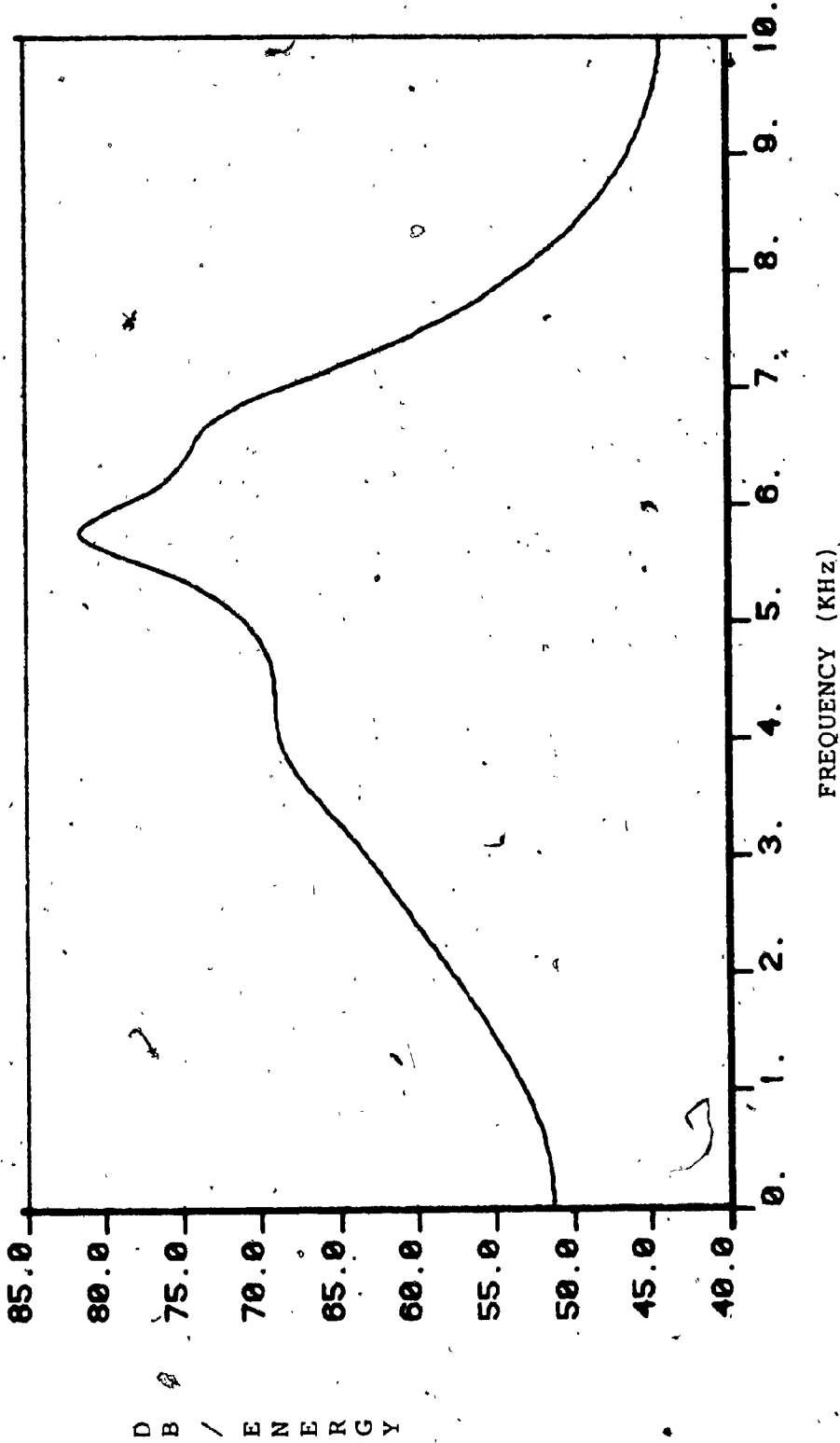
Appendix I Burst spectra of the plosive sound /t/



Appendix I Burst spectra of the plosive sound /k/



Appendix I Burst spectra of the plosive sound /b/



Appendix I Burst spectra of the plosive sound /d/

Handwritten scribbles and marks on the right side of the page.