SPEECH RECOGNITION OF HINDI STOP CONSONANTS

THESIS
submitted for the award of the degree
of
DOCTOR OF PHILOSOPHY

S. RAMAN

Research Guides

Dr. H. N. MAHABALA
Dr. B. YEGNANARAYANA

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
MADRAS 600 036
INDIA

JULY 1985

To

SAMPATH

SHARADA

ARUN

V IJAY

CERTIFICATE

This is to certify that the thesis entitled "SPEECH RECOGNITION OF HINDI STOP CONSONANTS" submitted by Mr. **S.Raman** to the Indian Institute of Technology, Madras, for the award of the degree of Doctor of Philosophy, is a record of bonafide research work carried out by him. The contents of this thesis have not been submitted to any other University or Institution for the award of any degree.

(Dr. **B. Yegnanarayana**)

(Dr. H.N. **Mahabala**)

Research Guides

ACKNOWLEDGEMENT

The author gratefully recalls his association with Dr. H.N.Mahabala for over a decade, during which there were various developmental activities culminating in this thesis. He is equally grateful to Dr. B. Yegnanarayana, who was directly behind this work from start to finish.

He sincerely thanks his numerous students - Mr. R.Sundar, Mr. T.Sreekumar and Mr. V.V.Chandrasekaran, to name just a few - who helped him in the laboratory in many ways and at different stages of this work. If it were not for the timely help of Dr. Y.Nagendra, in enhancing the author's understanding of Statistics, and that of Dr. M.Anthony Reddy, in the patient correction of the language of the manuscript, the author might have easily ended up writing one thing and meaning another. He is thankful to them both for their considerate help.

He wishes to place on record his gratitude to the system-staff of the Computer Centre and the Microprocessor Laboratory, o r their ever-cheerful help. For having given shape and form to this thesis through patient typing and editing, the author especial1y thanks his friends, Mr. D.S.Chandrasekaran and Mr. N.Alwar.

CONTENTS

Abstract vi

Prologue x

Chapter 1 DESCRIPTION OF THE SPEECH RECOGNITION PROBLEM 1

1-1 Motivation for this work 1

1-2 The Hindi alphabet 6

1-3 Approach to the speech recognition of Hindi stop consonants 10

1-4 Organization of the thesis 15

Chapter 2 REVIEW OF ISOLATED WORD RECOONITION STRATEOIES 19

2-1 State of the art of speech recognition 19

2-2 Review of isolated word recognition methods 22

2-2.1 The pattern recognition model, 22

2-2.2 Parameter extraction, 24

2-2.3 Matching of test and reference patterns, 28

2-2.4 Decision rule for recognition, 34

2-2.5 Performance of isolated word speech recognition systems, 35

2-3 Need for a different approach for Hindi consonants 39

Chapter 3 SIGNIFICANCE OF PERFORMANCE INDEX IN THE EVALUATION OF SPEECH RECOGNITION SYSTEMS 43

3-1 Need for a new measure of performance evaluation 43

3-2 Definition of Performance Index 45

3-3 Comparative study of Performance Index (PIX) and Percentage Recognition Score (PRECS) 52

3-3.1 Methodology of the study, 52

3-3.2 Relevant statistics for small sampling, 54

3-3.3 Statistical significance of PIX for Number-set, 56

3-3.4 Statistical significance of PIX for Hindi stop consonants-set, 58

3-4 Conclusions 60

Chapter 4 SIGNAL-DEPENDENT MATCHING FOR SPEECH RECOONITION 63

4-1 Introduction 63

4-2 An algorithm for signal-dependent matching 65

4-3 Speech material and representation 66

4-4 Experiments with other vocabularies 69

4-5 Experiments with Hindi stop consonants 70

Chapter 5 KNOWLEWE-BASED PROCESSING FOR CONFUSABLE VOCABULARY.. 71

5-1 Introduction 71

5-2 Speech-specific knowledge 72

5-3 Effect of knowledge on recognition results 73

5-4 Speech recognition method for confusable vocabulary ... 76

5-4.1 Segmentation based on transient analysis, 76

5-4.2 Processing of the utterance by parts, 85

5-4.3 Choice of appropriate parameters +or the respective parts, 86

5-4.3.1 Experiments with same vowel-ending, 87

5-4.3.2 Experiments with different vowel-endings, 89

5-4.4 Evaluation of segmentation based on different strategies, 89

Chapter 6 REFINEMENTS BASED ON EXPERIMENTS WITH LARGE DATA-SETS 91

6-1 Introduction 91

6-2 Choice of reference set 92

6-3 Considerations in the determination of warping path ... 96

6-3.1 Influence of the dominating vowel-part, 96

6-3.2 Addition of prosodic features to the knowledge-base, 99

6-4 Effect of conditions of recording on the system performance 101

6-5 Performance for vocabulary having other vowel-endings.. 102

Chapter 7 A SCHEME FOR RECOGNITION OF HINDI STOP CONSONANTS ... 106

7-1 Introduction 106

7-2 Determination of the vowel-ending 107

7-3 Determination of the manner of production 109

7-4 An overall scheme of recognition 111

Chapter 8 CONCLUSIONS 129

8-1 Summary 129

8-2 Conclusions 133

REFERENCES 141

LIST OF AUTHOR'S PUBLICATIONS RELATED TO THIS WORK 151

APPENDIX-I Analog 1/0 subsystem 152

APPENDIX-I1 Description of program modules 154

APPENDIX-I11 An output showing the PIX-values for overall and consonantal part 156

APPENDIX-IV An output demonstrating the easy recognition of consonants due to different manners of production.. 157

APPENDIX-V An output demonstrating the dominant features due to different manners of production 158

APPENDIX-VI Outputs showing the improvements in the system-performance for selective processing on GA-BA sat.. 159

ABSTRACT

The objective of this thesis is to study some issues connected with the development of speech recognition systems for discrete utterances of Hindi stop consonants. There are small but systematic changes in the characteristics of the speech producing mechanisms, in the production of these consonants. The problems due to these changes, and which are encountered during recognition, have motivated the author to take up this study. Waveforms of these utterances are shown in **Fig.1.** The non-nasalised consonants Are characterized by different places of articulation and manners of production. **Specifically,** the utterances in a given row in **Fig.1** are due to different manners of production. The utterances in a **given** column are due to different places of articulation. Characteristics such as voicing, aspiration and combinations of these contribute to the different manners. The utterances in a given row from left to right belong to the following categories, in that order:

(i) unvoiced and unaspirated, **(ii)** unvoiced and aspirated, **(iii)** voiced and unaspirated, and, **(iv)** voiced and aspirated.

In a given column, from top to bottom, the utterances are due to constrictions at different points in the vocal tract, starting from velum outwards towards lips. Thus the five stop consonants are velar, palatal, **retroflex**, dental and bilabial in that order. These consonants with the vowel-ending अ [ʌ] are shown in **Fig.1.**

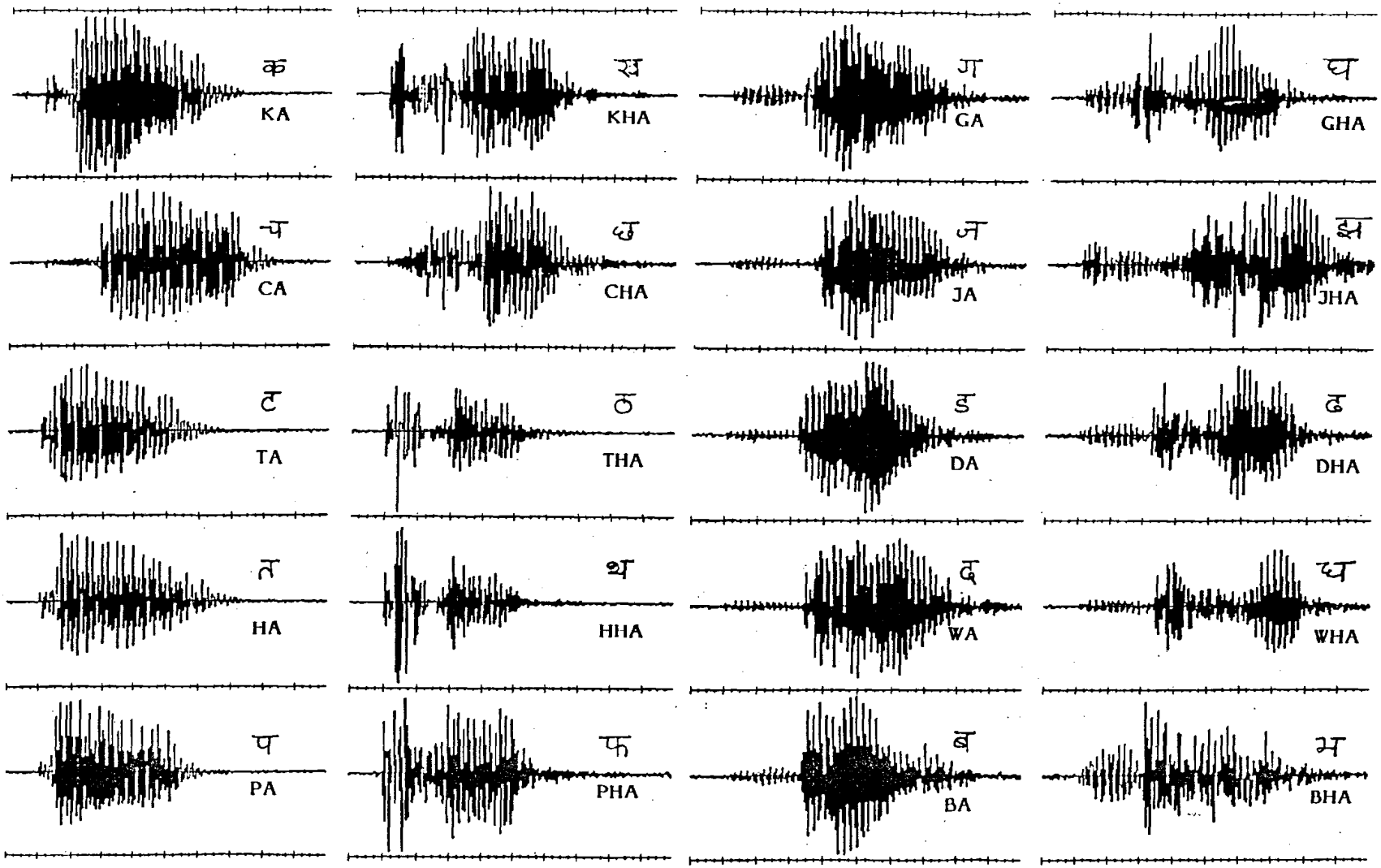


FIG. 1 THE WAVEFORMS OF DIFFERENT UTTERANCES OF HINDI STOP CONSONANTS

(for pronunciation refer to Table - 1.2)

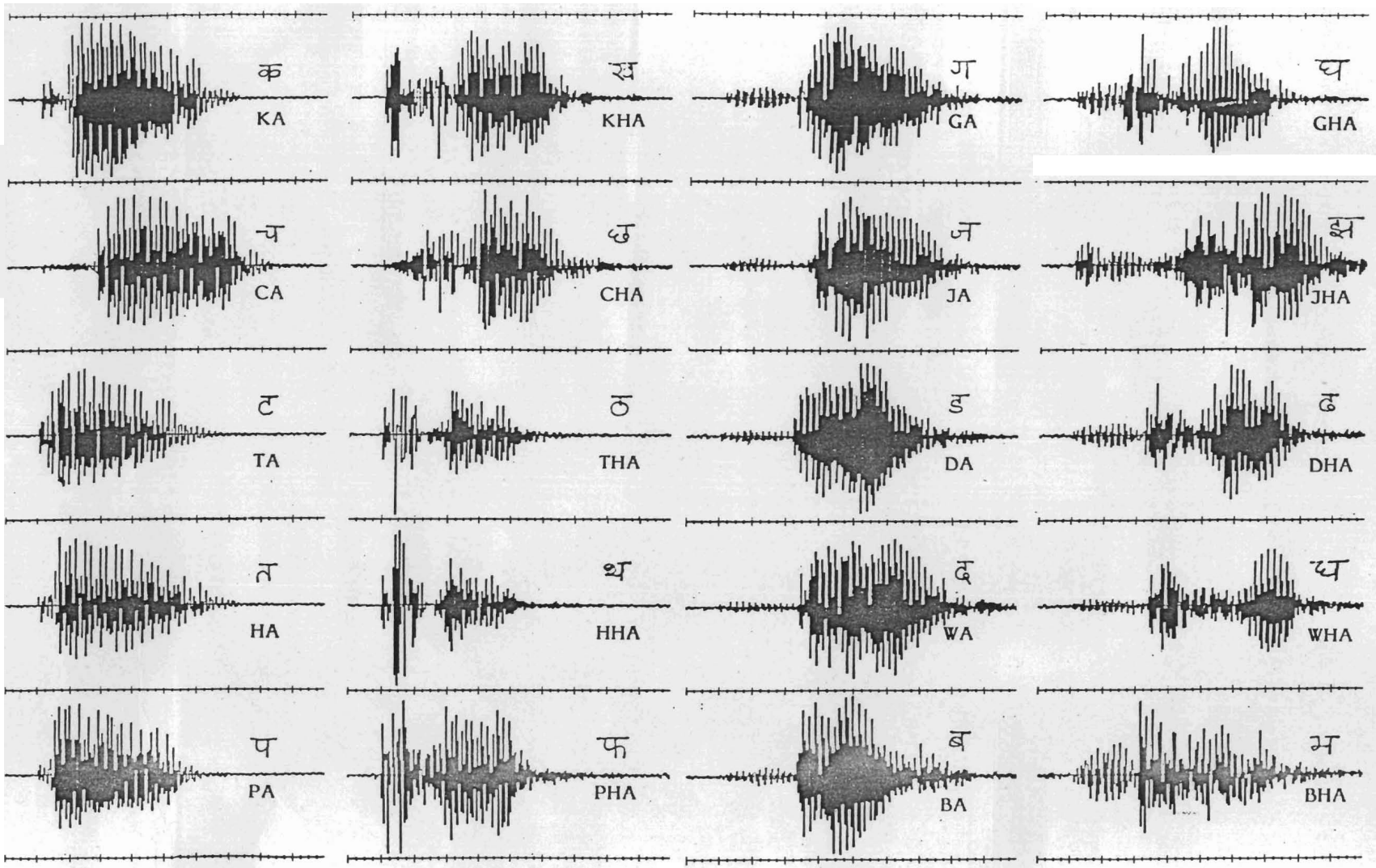


FIG. 1 THE WAVEFORMS OF DIFFERENT UTTERANCES OF HINDI STOP CONSONANTS

(for pronunciation refer to Table - 1.2)

The characteristics of the signal for the consonants due to different manners of production are relatively well distinguishable. In contrast, the features associated with different places of articulation are generally weak. This makes the speech recognition of the consonants in a given column more difficult than those in a given row. Hence the major portion of this thesis is devoted to the discussion of the methods for the recognition of a set of stop consonants characterized by different places of articulation. The consonants belonging to the unvoiced and unaspirated group are considered for detailed investigation.

Refinements in the form of associating weight-vectors appropriate for different segments of the speech signal for matching purposes are shown to be inadequate. In view of the nature of production of these signals, portions of utterances having predominantly consonantal characteristics are identifiable. It is shown that the signal corresponding to these portions is transient in nature and hence representation of those parts with enhanced temporal resolution results in better recognition performance. Further refinements such as addition of prosodic information are shown to improve the performance. After developing the algorithm for the recognition of this subset of consonants with the vowel-ending 3T [ΔI, the study is extended to four other vowel-endings. Problems associated with the recognition of consonants belonging to other groups are also discussed.

The improvemets in the system performance have been monitored using a new measure of performance, called Performance Index (PIX). PIX enables one to evaluate the performance of a

recognition system, using small data-sets. The statistical measure of percentage recognition score (PRECS), which requires the use of large data sets, is also computed. It is shown that PIX gives a good indication of the system performance with small data-sets just as well as PRECS does it with large data sets.

We conclude that strategies appropriate to the respective subsets of consonants result in better recognition performance than those based on a uniform and fixed approach. Based on these studies a scheme for recognition of Hindi stop consonants is proposed.

~~~~~~~~~~
~~~~~~~~~~

PROLOGUE

Panini, the fifth-century Sanskrit grammarian and one of the earliest phoneticians the present day literature [1] can trace, describes the rules of speech, grammar, etc., for the Sanskrit language. In his treatise, Astadhyayi [2,3], he refers to aphorisms as 'Shiva-sutras'. According to Indian mythology, sound was originally produced by the beating of the hand-held percussion instrument of Lord Shiva. It could be that the natural and systematic way in which speech sounds are produced for the alphabet of Sanskrit suggested the work of divine forces to the ancients. Even if such a line of thinking is not acceptable to the current scientific world, certain pointers in that treatise are worth studying. A cursory study of the vowels, semivowels, consonants, etc., in almost all the Indian languages, does show some order in their arrangement. This has relevance to the way sound is produced by human beings. Utterances of different letters of the alphabet belonging to well-defined classes and subclasses are marked by small changes in the characteristics of speech-producing mechanisms. These points have been brought out by Panini. Similar works, such as Tolkappiyam 141, can be found in Dravidian literature, and for other Indian languages as well.

Examination and exploitation of the orderliness in which the Hindi alphabet is arranged form the basis of the work reported in this thesis. The pattern of arrangement of the letters of the alphabet, and the nature of the utterances suggest certain speech-signal processing techniques, which have not been studied hitherto.

This work is a logical extension of our work on the development of input/output peripherals suitable for Indian languages. Specifically, this effort was motivated by our goal to develop working models for speech-to-text transcription, for speech in Indian languages. Taking advantage of the phonetic nature of Indian languages, a subgoal of transcribing speech was defined. In its simplest form, speech-to-text transcription consists of reading out each character of a written text in isolation. Then the character-wise utterance becomes synonymous with reading out the alphabet. Thus speech recognition in this form reduces to the recognition of alphabetical letters. Hence, in this thesis, 'word' is synonymously used with 'character', 'letter' or 'utterance'.

As noted by Allen [1] in some other context, it is to be borne in mind that "this technique is a means to a limited practical end, and by no means the ultimate analysis". For, as Bhartrhari points out in Vakyapadiya 151, "within the sound-unit the component features have no independent existence, nor the sound-units within the word; nor have the words any separate existence apart from the sentence". Hence the study of isolated characters is only the beginning. Whitney 161 echos similar thought in his commentary on Atharva-Pratisakhya when he observes that, "without studying the word-isolates one might make errors in the continuous text; it is for this reason that the study of the isolates is necessary".

Commenting on reinterpretation of ancient Indian texts, Allen [13] observes that "it implies that they display a level of phonetic discourse beyond the full comprehension", of commentators of a single period, "such as only the advances of the late nineteenth and twentieth centuries enable us to appreciate today". He adds that "the recognition that analyses so advanced in their technique should have been evolved at so early a date may well inspire a salutary scientific humility".

It is with this humility that the work reported here was carried out, with the help of modern facilities. It is hoped that in its small way it will contribute something in this area.

1. DESCRIPTION OF THE SPEECH RECOGNITION PROBLEM

"A goal, whatever it may be, is what gives meaning to our existence. It is the carrot on the stick that keeps us striving - that keeps us interested - that gives us a reason for getting out of bed in the morning."

EARL NIGHTINGALE

1.1 Notivationfor this work

With the advent of microprocessors, enormous processing capability at low hardware prices has been realised. This has steadily influenced applications ranging from data processing to dedicated control. Development of more and more intelligent and user-oriented systems include those that are menu-driven and which can converse or usefully interact with the operator. This implies accepting input or generating output and, if possible, even processing in the operator's natural language. This trend has influenced users and manufacturers of computer systems also in developing countries like India.

In November 1978, the Electronics Commission of the Government of India organised a symposium on "Linguistic Implications of Computer based Information Systems" [7]. Surveying the fifty-odd papers presented in that symposium and the bibliography compiled by Om Vikas [8], one sees how processing and input/output in Indian languages has started taking wide and deep roots in this country. That work in this area flourished in

the following five years was amply demonstrated by the exhibits of commercial and prototype systems at the Technology Exhibition [9]. The exhibition was conducted by the Electronics Commission in connection, with the Third World Hindi Conference, held in New Delhi in October 1983.

In this background of computerization in the Indian context, Mahabala and this writer [10] discussed how dot matrix technology could be adapted for generating computer outputs in Tamil, a language of southern India. The fonts in Indian languages do not lend themselves easily to impact printing, without sacrifices on the presentability. This writer and others C113 describe how outputs with enhanced presentability can be generated, in multiple Indian languages using a stencil cutter. The system was demonstrated at the Technology Exhibition when computer outputs were generated simultaneously in Tamil, Hindi and English [12].

Whereas these constitute efforts towards generation of outputs in Indian languages, input in these languages mainly consisted of entry through keyboard. The input is through relabelling of the keyboard with the inclusion of a software interface. The software will map the different ASCII codes, generated by the different key strokes, to the various characters to be printed out [10,13,]. The phonetic nature of Indian languages suggests certain novel designs of keyboard. Mudur and Wakankar [15] have shown that, with the phonetic nature of the Devanagiri script forming the logical basis for the letter forms and for the keyboard design, further processing of Devanagiri by the computer can be made as natural as in the case of the European languages. The basic alphabets of Indian languages

consist of vowels, consonants and diacritical marks. In addition to these, there are composite characters having separate patterns, which are formed by a combination of vowels and consonants. Because of the phonetic structure of Indian languages of the Brahmi family, the sequence of vowels and consonants that form a given composite character is essentially same in different languages though the resultant patterns are different. Taking these into account, Narasimham and Sinha [16] have proposed the design of a keyboard consisting of keys for vowels, consonants and diacritical marks, and have prescribed the rules of keyboarding based on the phonetic structure of Indian language scripts.

As evident from the works referred to above, the advantages of the phonetic nature were quickly realised. But the advances in the state of the art of speech recognition and in the technology of implementation were not yet mature enough to consider exploiting these phonetic features so as to view speech itself as the medium of input. The current developments have not yet resulted in' voice–input systems which permit any speaker to input words of unlimited vocabulary for any type of environment. However limited successes in speech recognition, speaker recognition, spotting of words, etc., have been achieved. Different types of speech recognition systems for speech in English and Japanese have been built. Most of these are speaker dependent systems. This implies that the user has to train the system prior to its actual use, by uttering the words of vocabulary for which the system is designed. In addition the vocabulary is also restricted. The performance of some of these systems is satisfactory mainly because the vocabulary consists of

non-confusable words [17-19]. The type of input utterances which they accept is another major specification. The input could be the normal continuous speech, or discrete utterances (also known as isolated words), or connected speech.

Isolated word speech recognizers accept words of a given vocabulary, uttered with a typical pause of about 200 msecs between them. A sample utterance would be ROMEO .... AND .... JULIET, with the periods of silence as indicated. As a development over these discrete speech recognizers there are connected-speech recognizers which do not require the pause between the uttered words. In the case of connected speech, for instance, an N-digit number will be uttered digit by digit but without the pause as in the case of discrete utterances, However these recognizers perform better for discrete utterances than for continuous speech, since the acoustic variations of words spoken without pause are greater than those with pause. This is due to the coarticulation effect between the adjacent sounds. The physiological basis for this is that the positions of the tongue, jaw and lips for one speech-sound are influenced by those for the previous and following sounds. In the case of isolated words, the intervening silence portions help define the initial and final positions of the articulatory organs fairly uniquely for the respective sound seqments. Thus the coarticulatory effects are not present. Also, in the case of discrete utterances, the endpoints of the words can be fairly accurate1y determined. Recognizers for normal continuous speech are still a far-off dream. Most of the commercially successful systems are speaker dependent, isolated word, selected (small) vocabulary systems.

The work reported in this thesis is concerned with development of speech recognition methods suitable for speech in Indian languages. In developing these methods the phonetic nature of these languages is fully exploited. Advances in technology should permit implementation of these methods or variations of these, in real time.

Integrating an output unit which generates output in Indian languages [11], with a voice-input system, helps in developing working models for speech-to-text transcription. Noting that all aspects of speech recognition have not been fully understood and that progress in this area has been achieved by restrictive specifications and short-term goals, we have defined a subgoal. There is one written character in script which represents one basic unit of speech-sound, in the sense that it can be independently uttered. Hence the subgoal consists of transcribing speech in its crudest and simplest form : reading out each character of a written text in isolation. There will be exceptions to such naive assumptions, with composite sounds which will have to be mapped to composite characters, using special rules. Barring such exceptions, the characterwise utterance becomes synonymous with reading out the letters. Thus speech recognition in this phase reduces to the recognition of letters uttered in isolation.

The stop consonants pose major challenges to machine recognition, since the different consonants are produced by small changes in the characteristics of speech-producing mechanisms. In contrast with the widely known sets of confusable vocabulary,

such as the E-set { B,C,D,E,G,P,T,V,Z }, the Hindi stop consonants appear to be more confusable. Therefore a different approach to the processing of speech signals for these words is needed. In this thesis we report the development of knowledge-based signal processing for recognition of Hindi stop consonants. Our studies show that employment of speech-specific knowledge is the way to handle confusable vocabulary. Based on these studies, a scheme for recognition of Hindi stop consonants is proposed. We believe that this may form the first step in the development of methods for continuous-speech recognition since stop consonants carry significant information in continuous speech.

1.2 The Hindi alphabet

The Hindi alphabet consists of 10 vowels (including 2 diphthongs), 4 semivowels, 4 fricatives and 25 stop consonants (including 5 nasals). The stop consonants are ordered in a systematic manner in most of the Indian languages and this order may suggest ideas for developing a recognition system. We choose a subset of Hindi stop consonants and describe their characteristics in this section. Table-1.1 shows the arrangement of the Hindi stop consonants.

The stop consonants are produced by complete1y closing the oral cavity and then releasing the built-up air pressure. The different consonants have different points of closure in the oral cavity. The innermost point of closure in the mouth is at the glottis. The other points of closure are where the back, middle and front parts of the tongue press against the appropriate

regions of the upper palate. Finally, beyond the teeth there is the closure of the lips. Thus the five rows in Table-1.1 represent five different classes of consonants. These correspond to the five places of articulation.

Table-1.1 : Hindi **stop** consonants and their characteristics

Classification based on articulation (Row-wise)	Classification based on excitation (column-wise)				
	Unvoiced		Voiced		
	Unaspirated	Aspirated	Unaspirated	Aspirated	Nasals
	K-P set	KH-PH set	G-B set	GH-BH set	
Velar (K-set)	क (K) k	ख (KH) kʰ	ग (G) g	घ (GH) gʰ gʱ	ङ (NG) ŋ
Palatal (C-set)	च (C) tʃ ʧ	छ (CH) tʃʰ ʧʰ	ज (J) dʒ ʤ	झ (JH) dʒʰ ʤʰ ʤʱ	ञ (GN) ɲ
Retroflex (T-set)	ट T ʈ	ठ (TH) ʈʰ	ड (D) ɖ	(DH) ɖʰ ɖʱ	ण (NN) ɳ
Dental (H-set)	त H t̪ t	थ (HH) t̪ʰ tʰ	द (W) d̪ d	ध (WH) d̪ʰ dʰ d̪ʱ dʱ	न (N) n n̪
Bilabial (P-set)	प (P) p	फ (PH) pʰ	ब (B) b	भ (BH) bʰ bʱ	म (M) m

□ (as in S.N. Ganesan [20])

■ (deviations from S.N. Ganesan, as in Peter Ladefoged [21])

The first four consonants in each row of Table-1.1 belong to the non-nasalised category. The first two are unvoiced and are **referred** to a5 surd consonants [22]. The next two are voiced and are called **sonorant** members. The fifth one, **which** is a nasal, is

also of the sonorant type. These members can also be grouped on the basis of aspiration. The first and the third are of the unaspirated type whereas the second and the fourth are aspirated. With row-wise and the column-wise arrangement of the consonants as in Table-1.1 the pattern of classification is according to the place of articulation and manner of production.

We have used some English letters to represent each of the consonants. For want of appropriate single letters in English, the choice of these is not based on any phonetic consideration. For instance, the letter 'H' is used to denote the sound 'th', and 'W' is used to denote its voiced version, namely 'dh'. The corresponding symbol of International Phonetic Alphabet (IPA), if used, is given within square brackets. The consonants are usually uttered in the context of a vowel or another consonant. The Hindi letters in Table-1.1 are consonants having the vowel-ending अ (A) [ə]. Some examples of the unaspirated consonants with other vowel-endings are given in Table-1.2.

Although Table-1.1 gives the traditional way of classification, Peter Ladefoged [21] does not consider the classification of the fourth group iie. voiced and aspirated) as reflecting the actual condition. According to him these sounds "are neither voiced (in the sense of having regular vibrations of the vocal cords) nor aspirated (in the sense of having a period of voicelessness during and after the release of the closure) ..... The state of the glottis during the closure and the release is neither voiced nor voiceless but murmured". Since there is no IPA diacritic for murmur he uses (..) underneath the symbol for

the voiced sound. Although this thesis is primarily concerned with the recognition of stop consonants, vowels are also considered since during production there will be significant influence of the vowel-part on the consonantal part and vice versa. From the point of view of the organs involved in the production of vowels, one may consider the three short vowels अ (A) . . (E) [I] and उ (U) [U] as basic. The long versions of these are आ [a] , ई [i] and ऊ [u] . Certain combinations of these result in the diphthongs ऐ (AI) [aj] and औ (O) [aw]. The two other

Table-1.2 : Examples of certain monosyllabic utterances

Utterance (Consonant-vowel	Pronounced as in
KA	cut
CA	chubby
TU	to
HE	thin
PI	pilot
KO	cow
GA	gum
JE	jig
DI	die
WO	thou art
BE	be

vowels ए [ɛ] and ओ [o] have been considered in different periods of history by different writers as long vowels or as diphthongs [1].

The four semivowels are ज [j], र [r], ल [l], व [ʋ] and the four fricatives are झ (ɦ), ष (ʂ), स [s], ह [h]. These are also characterised by different places of articulation, in addition to other features such as roll, taps, frills etc.

1.3 Approach to the speech recognition of Hindi stop consonants

The input to the recognition system consists of assorted types of sounds, some of which form distinct groups. The fact that they form groups could prove to be either an advantage or a disadvantage. When distinguishing between the groups it is an advantage, whereas for discriminating between the members within a group it is a disadvantage as it demands very refined processing. These two steps may require different types of strategies for recognition.

Different segments of speech signal within a syllable or word may require different methods of processing. Uniform and fixed strategies may not be adequate. Vowels are basically produced by "the laryngial voice-source, in which the quasiperiodic modulation of the breath stream by the vibrations of the true vocal folds produces the buzzing sound we call voicing" [23]. "Noise sources, from turbulence in the breath stream resulting from flow through a constriction or past an obstruction" [23] generate fricatives. "Burst or transient sources which result from the release of a pressure built up against a closure in the vocal tract" [23] generate stop

consonants. In addition to these physical processes by which different sounds are produced, intervals of silence between segments of sounds also seem to cue perception of certain of these sounds [24]. Figure 1.1 shows waveforms of some of these sound-patterns. Thus different classes of sounds are produced by different ,mechanisms. Hence it is appropriate to look for different methods of solutions to handle these.

Considering the recognition of stop consonants, listed in Table-1.1, there could be one method to distinguish the set of consonants in a row and another to distinguish those forming a column. Since there are distinct system-excitation features distinguishing the members in a row and distinct articulatory processes distinguishing the members in a column, the different methods for recognition will have to take into consideration the appropriate information during recognition.

"Stop-sounds can be considered to be made up of at least two typical sound segments, the occlusion and the burst, and the latter phase may in some instances be split up into three successive and partly overlapping phases, the explosion-transient, a short fricative and a h-sound" [25]. The burst may last for about 50-100 msec only. This is the most important speech segment of the utterance containing the primary features of the consonants. The waveform corresponding to this leading region essentially consists of transients. There will not be any periodic components and hence one cannot extract any vocal tract resonant frequency information. But the following vowel-part will display well-formed resonant characteristics and hence the resonant frequency components, called formants, could be estimated for that

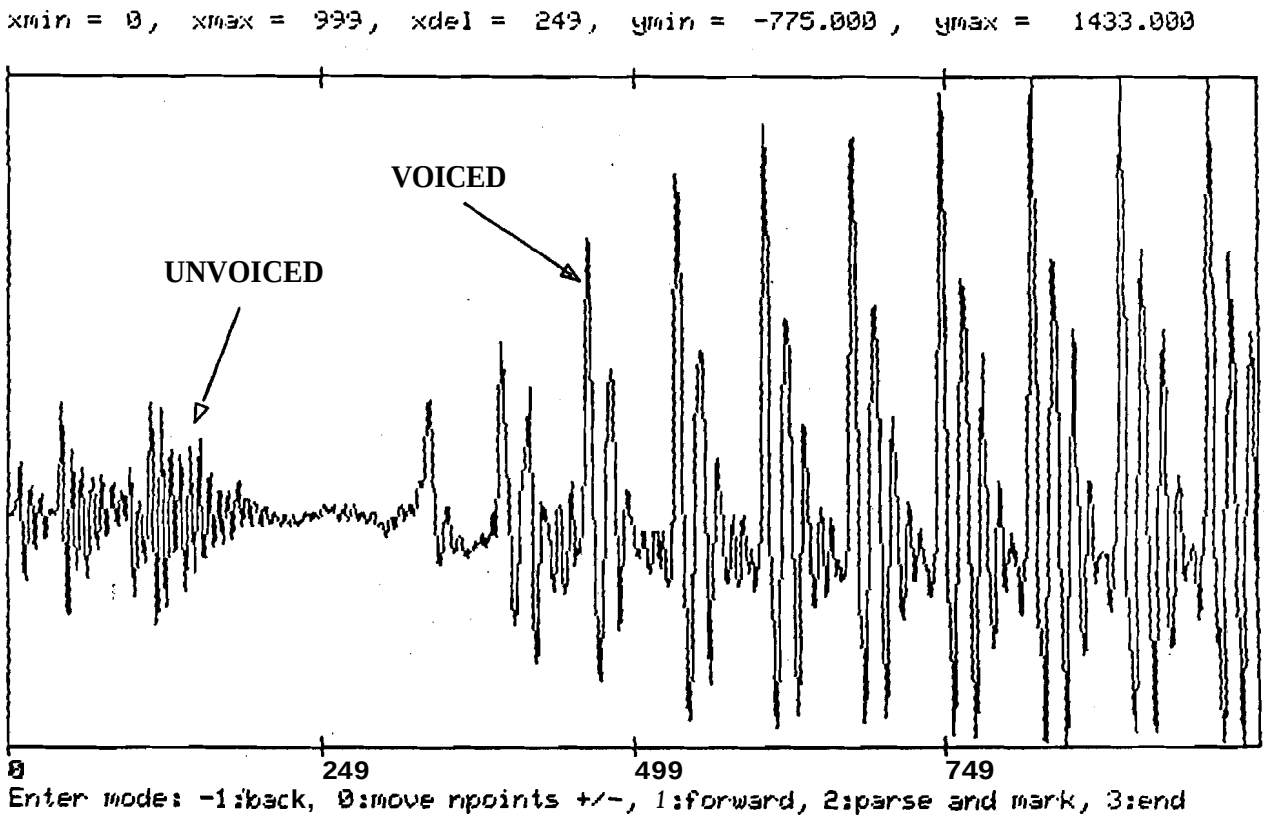


FIG 1.1(a) WAVEFORM OF UTTERANCE 'KA' (as in CUT)

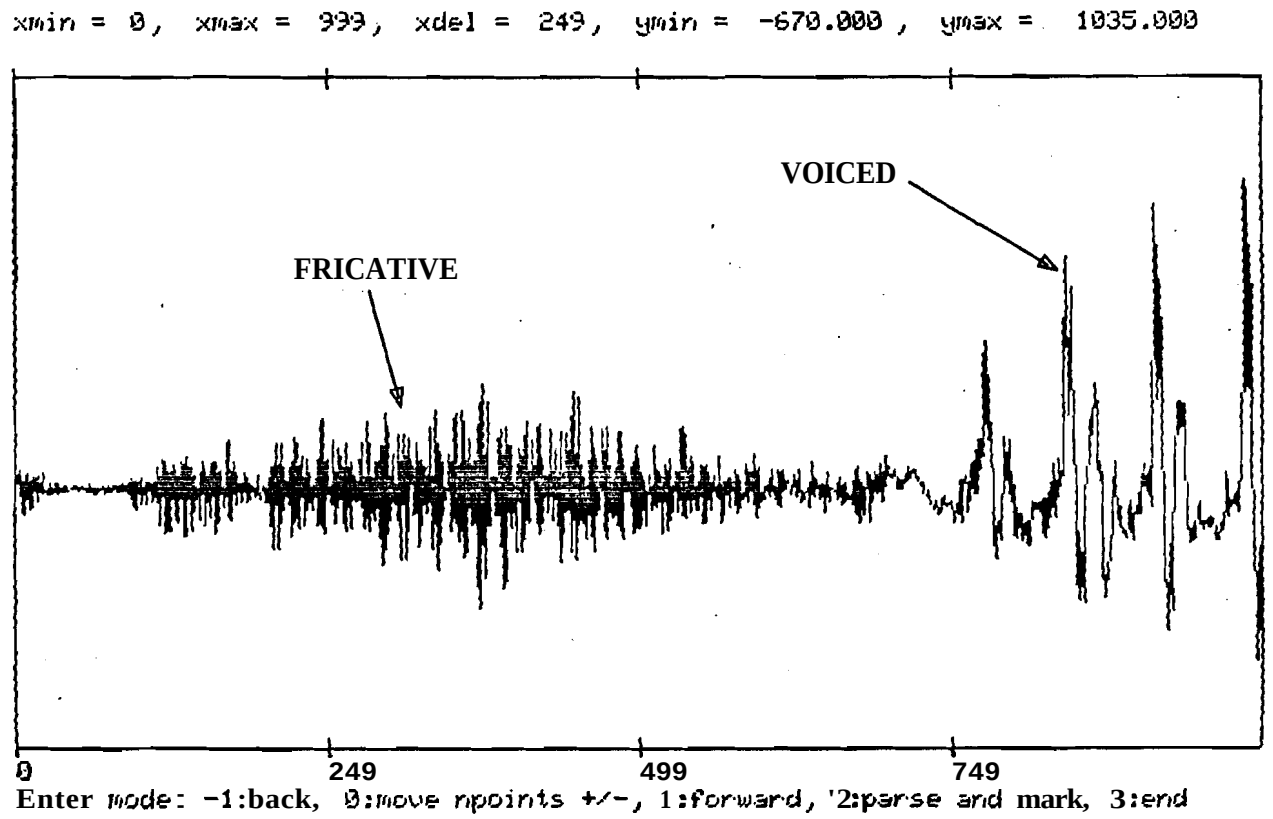


FIG 1.1(b) WAVEFORM OF UTTERANCE 'CA' (as in CHUBBY)

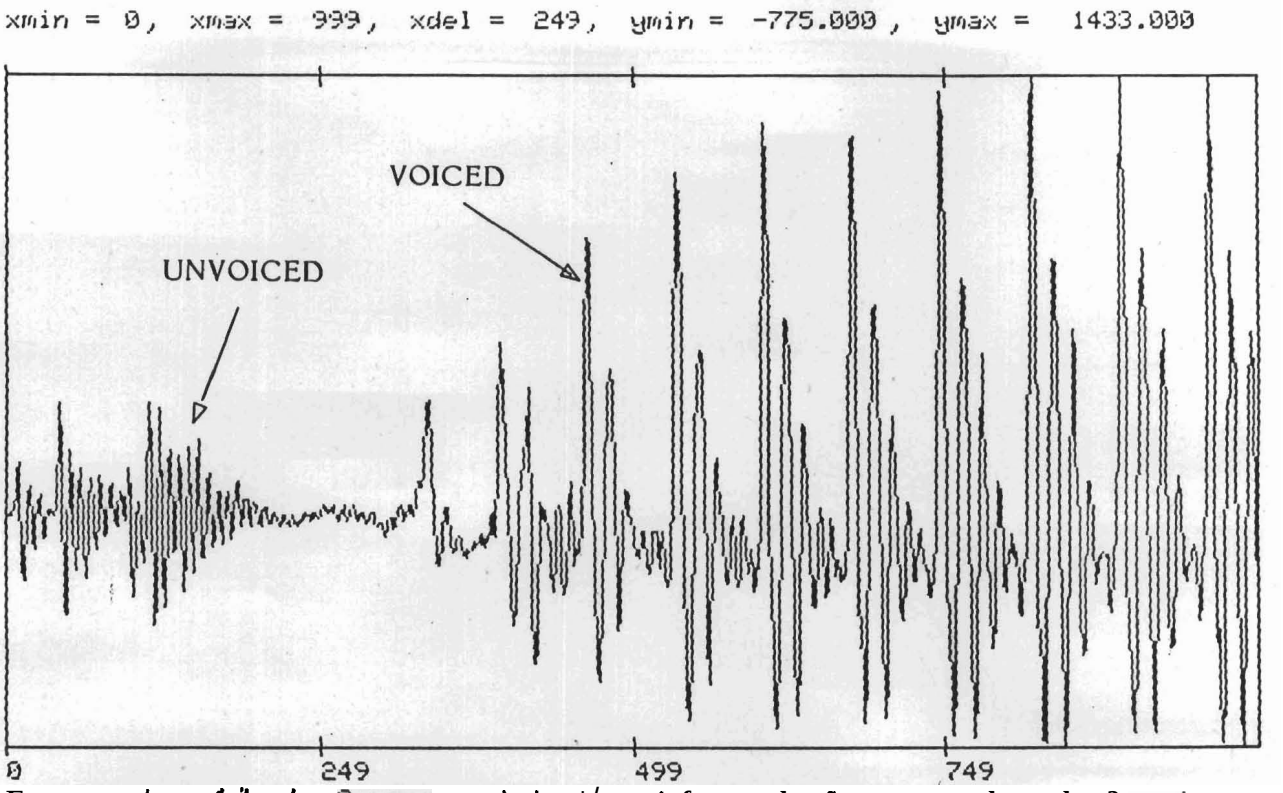


FIG 1.1(a) WAVEFORM OF UTTERANCE 'KA' (as in CUT)

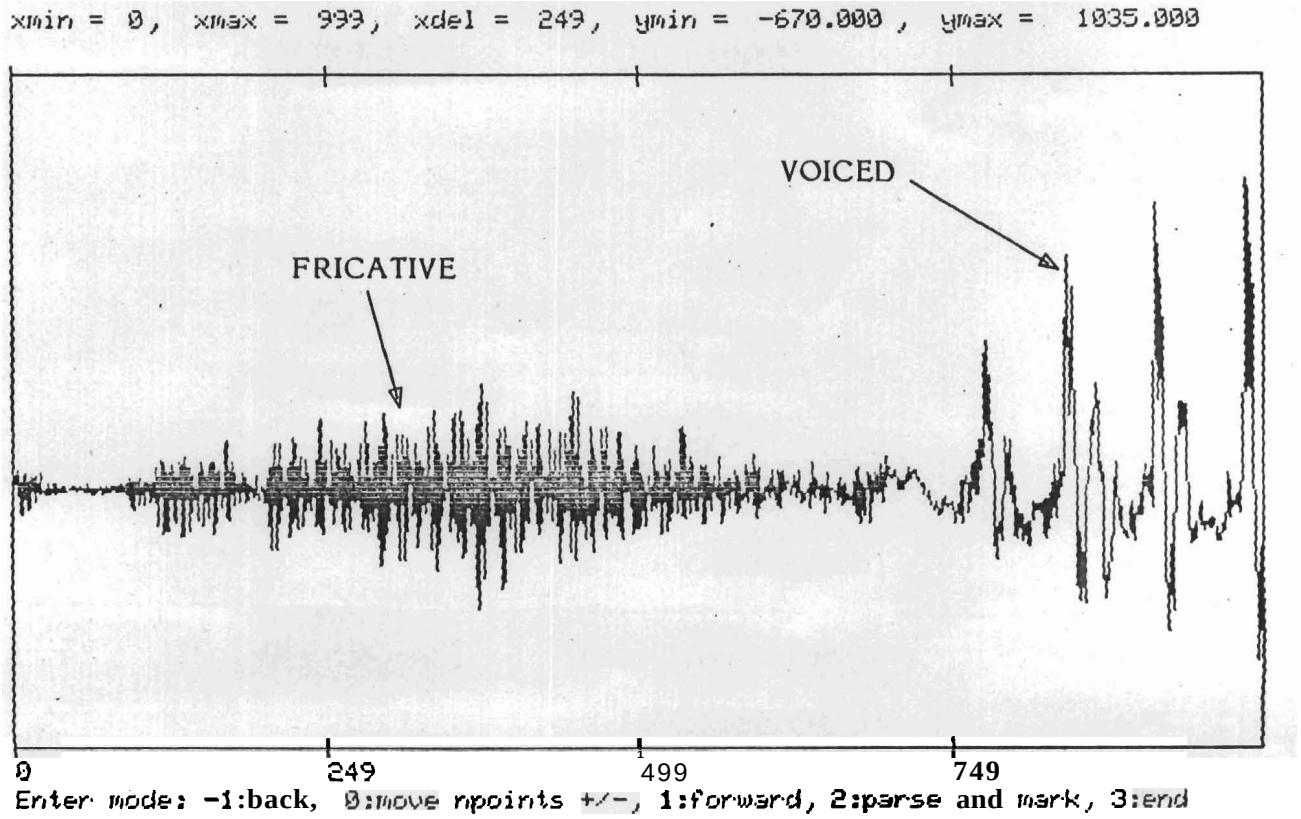


FIG 1.1(b) WAVEFORM OF UTTERANCE 'CA' (as in CHUBBY)

part. Each position of the articulatory organ has its specific formant pattern (F-pattern). "The time variations of the F-pattern across one or several adjacent sound segments, which may be referred to as the F-formant transitions, are often important auditory cues for the identification of a consonant, supplementing the cues inherent in the composition of the sound segments traditionally assigned to the consonant" [25]. Thus there are primary characterising features of these consonants and secondary features such as formant transitions. The methods we develop must be mainly based on t'he primary characteristics.

"The identification of a phoneme from the physical data contained in successive sound segments involves first a phonetical categorization, essentially with regard to 'manner of production', and then within each category, a choice related to 'position of articulators", for instance the choice of one of /b/,/d/,/g/ when the phoneme has been identified as a voiced stop" [25]. In the context of identification of the consonants in Table-1.1 this would mean the identification of the column first and then the row to which the consonant belongs. Figure 1.2 shows a possible hierarchical approach to the speech recognition of Hindi stop consonants.

A hierarchical approach has the inherent danger of leading the recognizer onto wrong paths since each node in the decision tree represents a point of no-return. Therefore, risk factors should be assigned to each stage of decision. Notwithstanding such refinements, the classification and arrangement of consonants according to the manner of production and place of

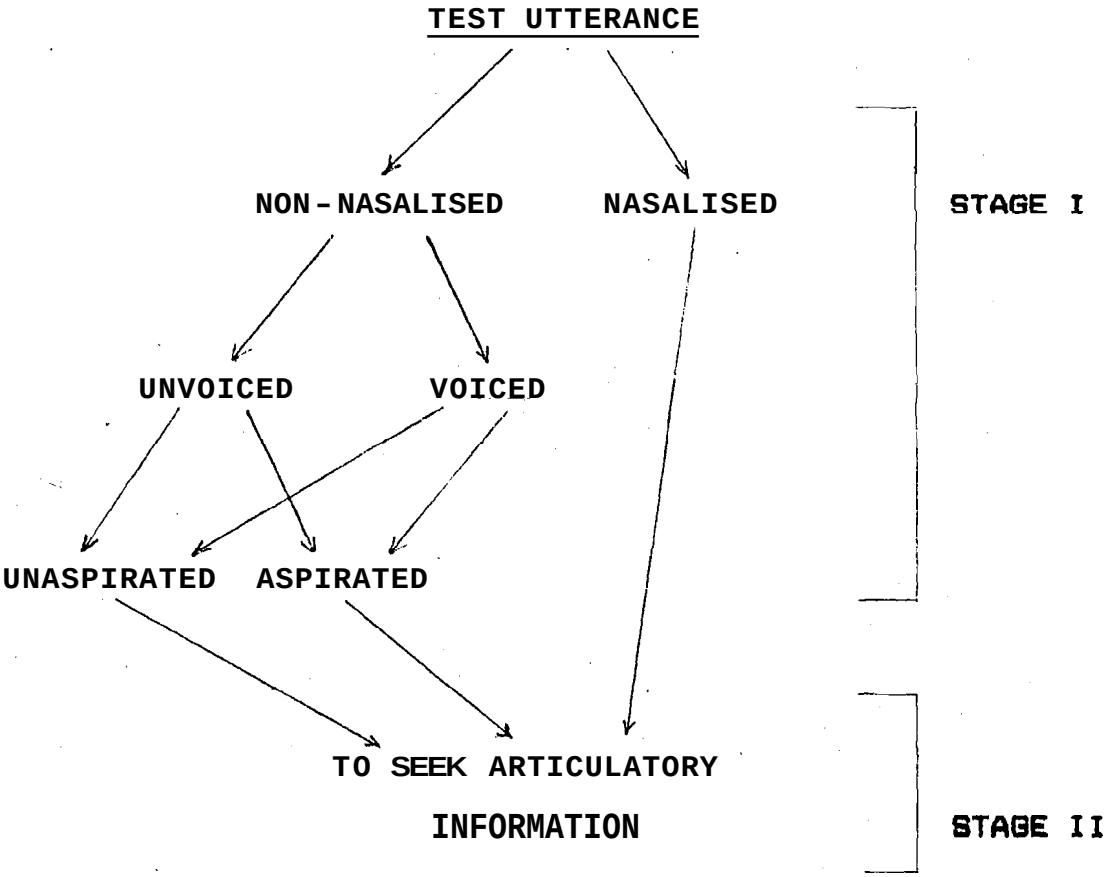


Figure. 1.2 Hierarchical approach to the recognition of consonants

articulation suggest recognition approaches which exploit the distinguishing features of these classes. The determination of place of articulation by directly estimating the articulatory parameters has been reported in literature [26]. Parameters estimated from the speech signal have been used in the identification of vowels [27], semivowels, and consonants [28], in continuous speech. Estimation of articulatory parameters and extrapolation of these to articulatory movements have been applied for removal of coarticulation effects and for improving the vowel discrimination [29]. Processes for obtaining the articulatory information are different from those for getting the excitation-information.

The characteristics to be looked for in each of these stages being different, it follows naturally that different processes of recognition are involved. Also, in the short speech segment under study, there is a combined effect of these characteristics. Some of these characteristic features might be distinguishable in the time domain, some in the frequency domain, some in both and some in neither. No uniform method of treating the speech segments can be expected to serve the purpose well. Thus the crucial stage in this work would be identifying different strategies for isolated word speech recognition for different subsets of vocabulary.

1.4 Organization of *the* thesis

We have shown that transcription of speech in Indian languages to text can make use of the phonetic nature of these languages. The main reason is the correspondence between a unit of independently utterable speech sound and a written character. The most difficult thing in this transcription task is the recognition of stop consonants. A cursory study of these consonants has led us to infer that it is necessary to evolve different strategies of isolated word recognition for different subsets of the vocabulary. The objective of this thesis is to develop the relevant strategies for recognition of isolated utterances of Hindi stop consonants.

In Chapter 2 the state of the art of isolated word speech recognition is described. The limitations of the existing methods for dealing with confusable vocabulary are highlighted. The need for a different approach for the recognition of Hindi stop

consonants is discussed. During development of the recognition methods for this vocabulary one has to evaluate the influence of the changes in the design parameters on the performance of the recognition system.

A new measure, called Performance Index [30], is used to capture the effects of these changes. The measure reflects the characteristics of the distance matrix formed for a pair of test and reference data sets. Chapter 3 deals with a comparative study of Performance Index and the widely used measure of Percentage Recognition Score. The limitations of the conventional approach to the recognition of Hindi stop consonants is demonstrated through the use of Performance Index.

Some limitations of the conventional methods can be overcome using a signal-dependent approach for recognition. In this approach adaptive strategies for parameter extraction and matching are used based on the information derived from the signal [30] [31]. Chapter 4 shows that although the signal-dependent approach is satisfactory for alpha digit vocabulary, the system performance is still poor for the confusable vocabulary of Hindi stop consonants [31]. This shows that other sources of knowledge also have to be used for improving the recognition performance.

As the recognition performance of humans in respect of this vocabulary is fairly good, it is hypothesized that more knowledge about this vocabulary should be used in addition to the information from the signal. Methods by which humans acquire and use this knowledge are not known. But "a system can have and use

knowledge without its designers having or being able to use that knowledge" [32]. Hence fragments of knowledge of the signal and the vocabulary are brought to bear on the recognition problem. Systems developed under ARPA SUR project such as HARPY and HEARSAY II used knowledge at higher levels. We show that, for dealing with the confusable vocabulary, use of appropriate knowledge at lower levels, from the points of view of speech production and speech perception, helps in improving recognition performance. This will aid application of higher level knowledge for subsequent processing as may be required for dealing with continuous speech.

Since speech can be divided into a sequence of sound segments we may consider each segment as constituting a stationary part or a transient part or as a boundary between such parts. It was hypothesized that the characteristics of speech could be altered in the stationary parts by voluntary control of articulators and modulation of excitation characteristics, and that similar exercise of control is not easy in the short transient portions. Thus one could characterize the signal based on features around these transient portions C331 [34]. Chapter 5 discusses these points and shows that combinations of such innovations hold the key to improved recognition for the confusable vocabulary. Different experiments carried out on this basis are discussed.

Chapter 6 deals with experiments on large sets of data so as to resolve certain issues related to choice of reference, considerations in the determination of the warping path, conditions of recording of speech data and effects of different vowel-endings. Addition of prosodic features to the knowledge-base is also discussed.

The overall scheme for recognition of Hindi stop consonants is outlined in chapter 7. The recognition results show that processing of different segments using knowledge at different levels is the way to deal with natural signals such as speech.

In Chapter 8 we conclude that, for the performance of a machine-based recognition to approach that of humans, the recognition methodology should exhibit the trends usually employed by humans, ,namely, reasoning from knowledge in each domain [35].

2. REVIEW OF ISOLATED WORD RECOGNITION STRATEGIES

Minister to visiting foreign dignitary:
"Since this meeting is just for ten minutes, . shall we take the world issues, arms race, Afghanistan, north-south dialogue, Lebanon, Iraq-Iran, trade, peace etc., as discussed?"

from R.K.LAXMAN's cartoon
The Times of India.

2.1 State of the art of speech recognition

"Speech recognizers commercially available today are effective only within narrow limits. They have relatively small vocabularies and frequently confuse words. Users must develop the skill to talk to the recognizer, and the machine's performance often varies widely from speaker to speaker" [36]. These limitations in performance can be overcome only through a better understanding of the speech recognition process employed by humans. According to Doddington and Schalk [36], the reasons for the present-day predicament are: "Speech perception at an acoustic level is critical to advanced recognition capability. Commercial word recognizers do not 'hear' the way human beings do, and this lack of adequate acoustic-level perception is the source of most of the shortcomings of these machines. Phoneticians already know a great deal about how speech is recognized acoustically, yet there has so far been little activity in transferring this knowledge to machines". This

indicates the paradigm-shift in the field of speech recognition from the traditional pattern matching-based algorithms to utilization of speech-specific knowledge starting from the basic acoustic-phonetic level.

Traditionally, speech recognition has been treated as a classical pattern recognition problem. The motivation for this model is that the information in the spoken sound can be extracted from the distribution of energy with frequency, and that repetitive utterances of the same word are likely to have similar distributions. Implied in this is the presumption that different words will have different distributions. Hence the recognizers look for the temporal variation of the short-time amplitude spectrum of the speech input.

Such a scheme is referred to by Newell [32] as that of generalised input-output functions. Treating the speech signal as just another waveform, these schemes apply the widely known signal processing techniques to parametrize the input. They do not consider how the signal is produced or how the speech is perceived by a human listener. The shortcoming in the traditional approach will become apparent when one considers "speech as a complex code in which the intelligence of a message is represented both locally by spectral features and globally by a hierarchy of structural features" [37]. These refer to the higher levels of knowledge which are employed by humans.

In contrast to the "generalised input-output functions" there are "knowledge-source driven representations" [32]. They assert that information from the signal alone is not enough to determine fully the message of the signal, and that knowledge

from other levels of human-processing should be used. These levels are from the viewpoints of speech production, sensory reception and speech perception [38]. From the speech production point of view there is information in the vocal tract resonances, manner and place of articulation, coarticulatory features, etc. Analyses based on the response of the human ear, auditory nerves, etc., are from the viewpoint of sensory reception. Use of voice onset times, formant transitions etc., is from the speech perception point of view. In the case of speech understanding the linguistic implications, like syntax, semantics and pragmatics, can be added to the sources of knowledge.

Levinson and Shipley [39] demonstrate the level of sophistication of the conversations possible with an airline flight-reservation system, the design of which incorporates information at the syntactic and semantic levels also. The system accepts words with pauses and responds similarly. A recognition accuracy of more than 98% has been reported. However, incorporation of higher levels of knowledge will not be possible in all cases. For instance, if a machine has to recognize isolated English letters, there is no way one can derive higher level knowledge from those isolated utterances. Such systems have to make do with whatever can be gathered at the acoustic and phonetic levels only. Notwithstanding the refinements through inclusion of higher levels of knowledge, the analysis starts at the acoustic level. Hence great care should be taken in handling the signal at that level. Any loss of information or misinterpretation at that level cannot be compensated for by refined processing at later stages.

Thus the observation made by Fant [40] in the mid-70'5, more or less holds good even to this day. According to Fant, "several groups have gone ahead and developed syntactical and semantic analysis schemes. This is a most interesting development from the applied linguistics point of view and is needed for the realization of task-oriented applications, but people have perhaps been a bit too eager to exploit these domains while neglecting the primary phonetic element recognition." Hence there is a need to consider the details of even a traditional model, study the issues involved, and innovatively improve upon the existing methodologies to overcome their limitations.

2.2 Review of isolated word recognition methods

2.2.1 The pattern recognition model

Figure 2.1 shows the four blocks of the conventional model of speech recognition. In this thesis, all the discussions are based on the assumption that the speech signal is in digital form. The analog speech signal is sampled at the rate of N samples/ second, which represents a rate greater than the Nyquist rate for the highest frequency of interest. Typically the sampling rate is 10 kHz/sec. The incoming speech signal is conveniently divided into time-sequential frames, each consisting of a specific number of samples.

In the training phase some parameters or features will be extracted frame-by-frame, and stored as reference templates. In the recognition phase the input speech signal will be subjected to the same step of parameter extraction. Basically, this is a

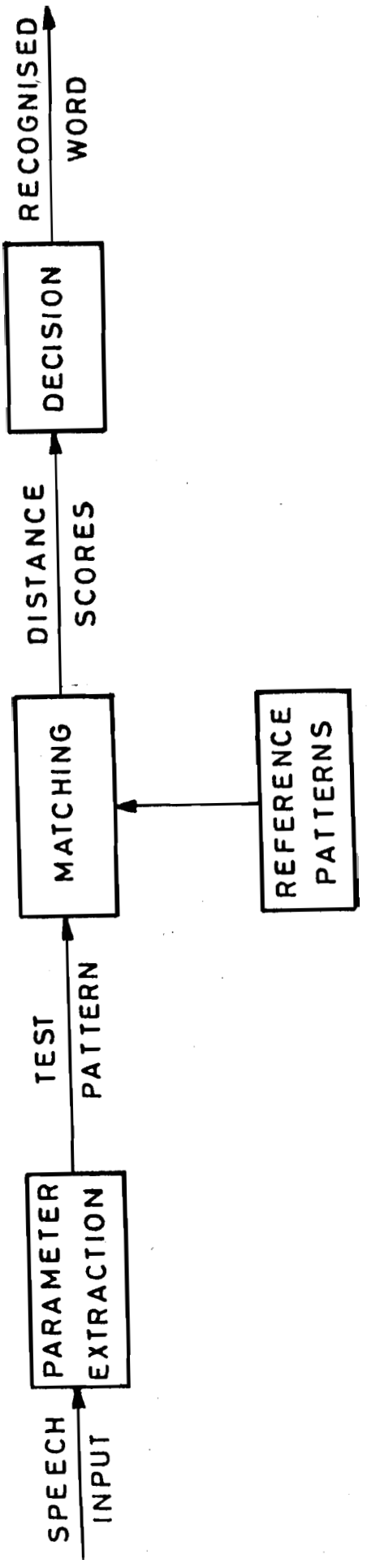


Fig. 2.1 The conventional PR model for speech recognition

data-reduction technique so that subsequent processing needs to deal only with the extracted parameters representing the speech. Matching consists of comparing a test utterance with all the reference templates and computing the distance between them. In the decision stage, the pattern of the reference template which matches most closely with that of the input test utterance is identified and the corresponding reference word is the recognized word.

The success of these models mainly depends on how faithfully the speech signal is represented. The entire information about the utterance, whether used as test input or as reference word, is soaked in those parameters which represent the speech signal. The information available to a recognizer is only at the acoustic and phonetic levels, when isolated utterances are inputs. Hence there is a great need for choosing the appropriate parametric representation and distance-metric.

2.2.2 Parameter extraction

Information in a speech signal can be represented either in the time domain or in the frequency domain. A third way of representation is based on modelling and uses the parameters of a filter whose spectrum fits best with that of the input speech signal.

Almost all recognizers depend by far on some sort of analysis done in the frequency domain. Speech is a nonstationary signal, but, with the assumption that it is stationary over short time-intervals, the usual mathematical techniques meant for processing stationary signals are applied. To compensate for the approximation due to this assumption, two factors are usually

considered. The first ope is concerning the time-interval of the signal under study, and the second has to do with the selective weighting of the time-samples, using the Hamming window and the like functions. The time-interval should reflect the desired spectral resolution.

An FFT algorithm can be used for short-time spectral analysis. It can be used to compute the short-time Fourier transform,

$$X_i(w) = \sum_{n=0}^{N-1} x_i(n) \cdot w(n)\, e^{-j\omega n},$$

where $x_i(n) = x(n+i), \quad n = 0,1,.....N-1$

$$i = 0,L,2L,....,$$

at equally spaced frequencies,

$$w_k = (k\, 2\pi/M) \text{ for } k=0,1,...M-1.$$

This facilitates computation of the outputs, of a uniformly spaced bank of filters, and hence the determination of the distribution of energy over the desired frequency bands.

In the filter-bank analysis model the range of frequency of interest (upto 5 kHz) will be conveniently divided into Q pass-bands. Q usually ranges from 5 to 32. Among the various ways in which the pass-bands are divided, the one corresponding to critical band is the most common, since it has been shown to resemble that of the auditory peripheral system [41, 42]. These pass-bands are spaced linearly upto 1 kHz and logarithmically above that. The output of each of the band-pass filters is generally passed through a nonlinearity, which is usually in the form of a detector. Then it is low-pass filtered to give a signal

which is proportional to the energy of the speech signal in the band. Thus a Q-dimensional vector is computed for each frame. The time-sequence of these vectors represents the speech pattern. The parameter-smoothing is achieved partly by the filter bandwidths and mostly by the integrator. Thus one can obtain a smooth spectrum which approximates the transfer function of the vocal tract. It includes the general characteristics due to the glottal excitation and radiation, but excludes the effects of the periodicity of the excitation.

Assuming that speech production on a short-time basis can be viewed as the convolution of an excitation function (either random noise or a quasi-periadic pulse train) with the vocal tract impulse response, there are methods for separating the components of a convolution. Homomorphic processing belongs to this class [43]. In this approach, the filtering is performed on the logarithm of the spectrum. By taking the logarithm of the magnitude of the Fourier transform, the sum of the logarithms of the transforms of the excitation and the vocal tract impulse response will be obtained. By performing the inverse discrete Fourier transform, which is a linear operation, an additive combination of the cepstra of the excitation and vocal tract components will result. In the case of voiced speech this means that there is a slowly varying component due to the vocal tract transmission and a rapidly varying periodic component due to the periodic excitation. Applying a window to this cepstrum, analysis of low-pitched voices can be performed by eliminating the effects of periodicity. This is called cepstral smoothing.

Methods belonging to this class of analysis are referred to as spectral-smoothing methods [44]. In contrast to these there are methods based on spectral modelling, by which a complete representation is given by the parameter values alone. A model for the spectrum is assumed, usually in terms of poles and zeros, and the parameters are computed on the basis of minimizing the error between the original and the model-spectra. This means that the parameters of the model should fit best with those of the input speech signal spectrum. The method based on Linear Predictive Coding (LPC) belongs to this class.

The underlying principle of LPC is that a speech sample can be approyimated as a linear combination of past samples. By minimizing the squared difference between the actual and predicted values of the samples over a finite interval, a set of predictor-coefficients is computed.

From direct measurements on the speech waveform, or from the parameter's of a model, speech is represented in parametric form. Typically, these include determination of the first few formant-frequencies, their bandwidths and amplitudes, their rate of change in specific segments, distribution of energy over selected frequency bands, autocorrelation coefficients, zero-crossing rates, the ratio of peak to rms value of the signal, the ratio of energy in low to that in high frequency bands, the ratio of either of these to overall energy etc. Different recognizers have employed different variations of these in the various stages of recognition such as for endpoint detection, discrimination among voiced/unvoiced/silence segments of speech, distance computation

and determination of discriminant function in the final stage of decision. In this work, we have chosen to represent speech using log melspectral parameters and LPC's, the actual choice of which depends on the segments.

2.2.3 Matching of test and reference patterns

When a test utterance is compared with the reference words there are two points to be considered. One is the time-normalization of the patterns and the other is the measure of similarity. Since no two repetitions of a word, even by the same speaker, will be of the same duration the test utterance has to be normalized with the reference word. A distance measure indicates the similarity or otherwise of the two patterns being compared.

In general, creation of reference templates consists of recording of the words in the vocabulary spoken by the person concerned. These are stored as reference templates. Refinements of this step consist of the speaker giving more samples of each. of the words. Then the system either computes the average of these samples or uses them as a set of multiple-reference tokens. In the experimental stage, a few of these samples will be used as reference and the remaining as the test data. In one method of selection of reference each of the samples is used as the reference and the remaining as the test data. Then a reference set is created by selecting those members which have performed well as reference. Several methods of reference creation have been reported in literature [45, 461. The problem of reference selection is a crucial one since the performance of a system greatly depends on it.

The most successful and the widely used method of time-normalization is based on dynamic programming. The algorithm is called Dynamic Time Warping (DTW) algorithm [47]. Other methods based on, linear time-normalization, matching for significant events in the utterance, etc., have been in use.

Time-normalization basically consists of finding a warping function w(n) which will register the two patterns in time, as shown in Fig.2.2. If the test and reference patterns consist of N and M frames respectively then,

$$T = \{T1, T2, \ldots, TN\} , \text{ and },$$

$$H = \{R1, R2, \ldots, RM\}$$

If the endpoints of the test and reference words match, the values of the warping function w(n) for the end-frames are,

$$w(T1) = R1 \quad \text{and} \quad w(TN) = HM$$

The algorithm for the matching process will determine the intermediate values of w(n). The optimum warping path w(n) corresponds to the path which results in minimum cumulative distance between the initial (T1,R1) and the final (TN,RM) points.

The DTW method is inextricably linked to the chosen similarity measure in the form of frame-by-frame distance computation. A simple recursive technique is used to determine the warping path, in which the accumulated distance from the initial point (T1,R1) to the final point (TN,RM), is minimised [47]. If n is the independent grid search variable, and if all the valid paths through the grid correspond to monotonically increasing time-indices, the recursive relation is given by:

$$D(n,m) = d(Tn, Rm) + \min_{x <= m} D[(n-1),x] .$$

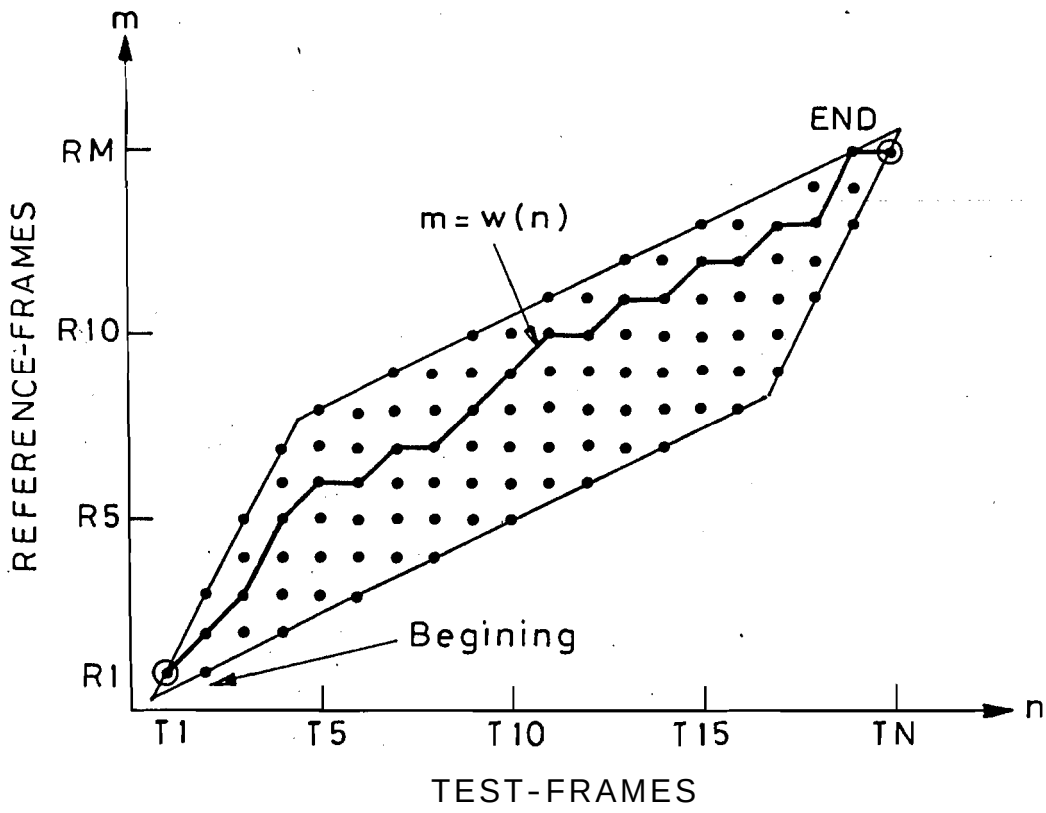


Fig. 2·2 An example of time-warping function. The parallelogram shows the possible domain of (n, m) coordinates.

In this equation, D(n,m) is the total cumulative distance. It is obtained by adding the local distance d, between the test frame Tn and the reference frame Rm being matched, to the minimum of the cumulative sums of distances obtained until the previous match. The global and the local constraints specify valid paths between the grid points (n,m) and n – 1 , For instance, the constraint that no two repetitions of the same word can differ in time–duration by more than 100% limits the slope of the warping path function w(n) to 1:2 or 2: 1, as shown in Fig.2.3.

The local constraints take the form of specifying valid neighbouring grid–points, as shown in Fig.2.4. Three valid paths, are shown to the point (n,m), with the slopes being 0, 1 and 2. The path with slope 0 is not permitted if a path of slope 0 has already resulted, for the matching of the previous frame, as shown between the grid–points (n-2,m) and (n-1,m). This implies that not more than two adjacent test frames will be matched with the same reference frame.

The warping path w(n) is constrained to within the parallelogram of side–slopes 1/2 and 2, as shown in Fig.2.3. In the discussion on DTW, two assumptions were made. One is regarding the endpoint conditions and the other is concerning the distance measure.

In an implementation, which allows the endpoints to be determined based on DTW results, a range of begin and final endpoints may be permitted, thereby deferring the choice of the right ones to after the computation of cumulative distance. Again, the implementation may specify further constraints leading

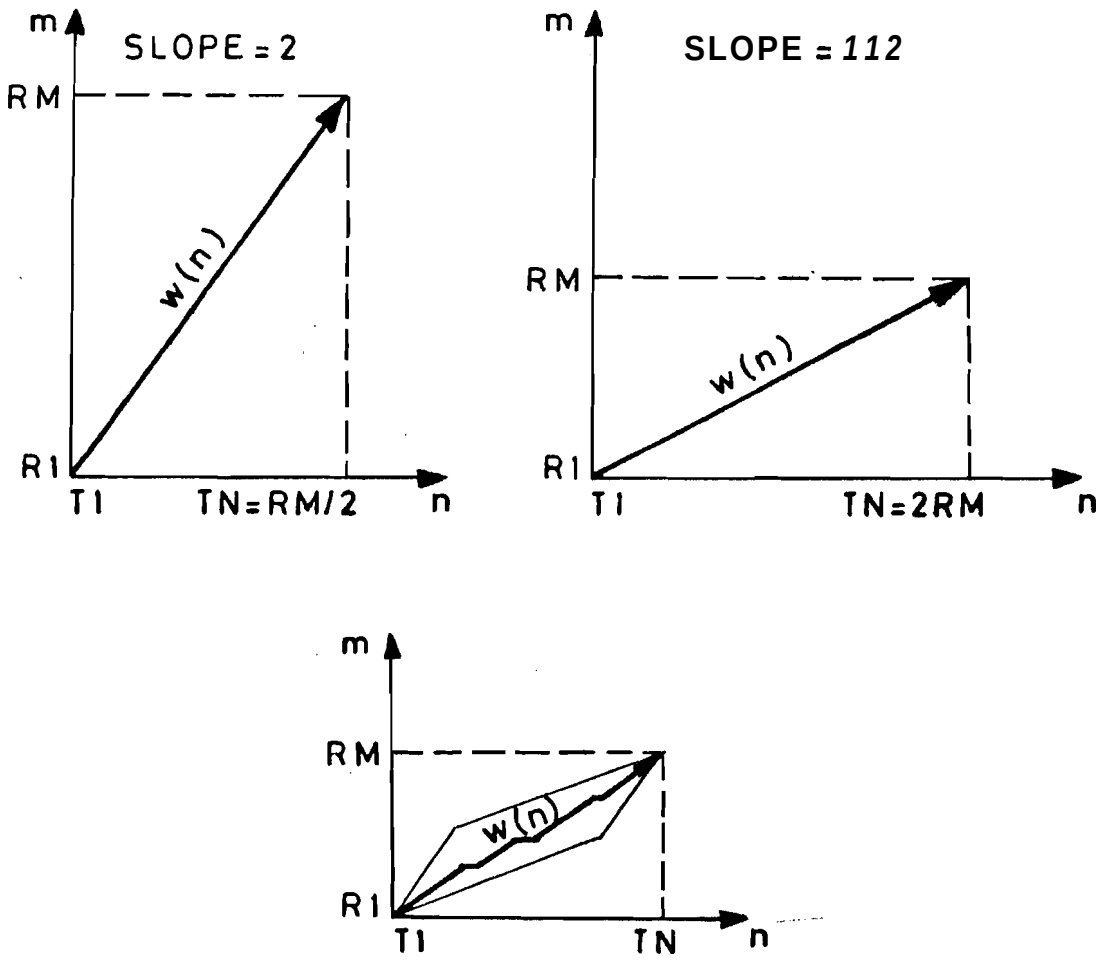


Fig.2·3 Global constraints restricting the slope of the warping-path function

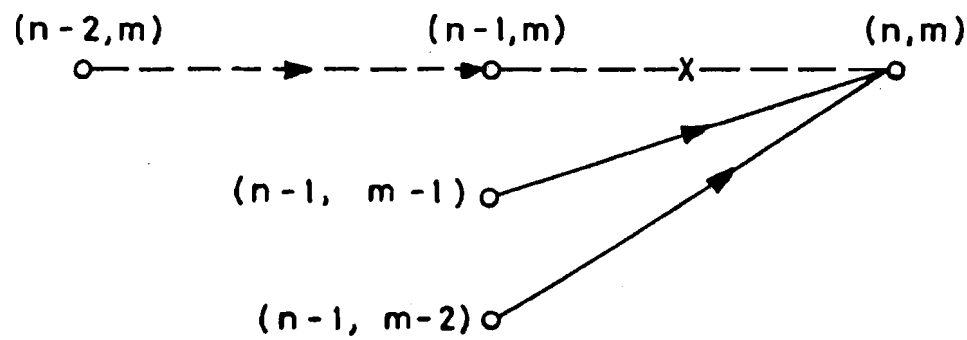


Fig.2·4 Local path-constraints specifying valid paths of w(n)

to no frame in the test utterance being kept out of consideration. Variations such as these result in variations in the region of the possible paths.

The two main factors in the choice of a distance measure are the physical significance of the measure with reference to the parameters and the computational complexity. Considering each frame as a point in space, the straightforward measure that one thinks of is the Euclidean distance, given by,

$$d(T,R) = \sum_{i=0}^{p-1} (Ti-Ri)^2 , \text{ where}$$

Ti and Ri are the test and reference vectors in the p-dimensional space.

In order to reduce the computation involved, absolute values of the differences are considered in the Chebyshev distance measure, which is given by,

$$d(T,R) = \sum_{i=0}^{p-1} |(Ti-Ri)| .$$

Silverman and Dixon [48] suggest distance measures based on log power-spectra for classification of speech sounds. But measures based on the smoothed spectral behaviour have been extensively used. Using an rms log spectral measure, which is the root-mean-square measure between the test and reference log spectra obtained from LP analysis, Pfeifer [49] reports good results in a speaker identification test. For a speaker verification test, Atal [50] has shown that a weighted Euclidean

distance measure based on cepstral coefficients resulted in the highest scores among several parameters derived from LPC. The unweighted Euclidean distance based on cepstral coefficients is referred to as cepstral distance measure. Gray and Markel [51] propose the use of rms distance D between the log spectra after each log spectrum has been cepstrally smoothed to L coefficients. Thus,

$$D^2 = (C_0 - C_0')^2 + 2 \sum_{k=1}^{L} (C_k - C_k')^2 .$$

The popularity of the measures based on a smoothed spectrum is mainly due to the easy computation of parameters through LP analysis or cepstrum analysis. For example, the rms log spectral measure can be computed efficiently using cepstral coefficients [51]. Also, the smoothed spectrum contains most of the significant features like formants of the vocal tract system.

2.2.4 Decision rule for recognition

The distance scores are used to arrive at the decision on recognition. The simplest method would be to choose the reference word which is at the least distance from the test utterance. In other words, the nearest neighbour is chosen. Variations of this are in terms of specifying that the chosen candidate should be at a distance less than a specified threshold value or that, relatively, the distance value should be lower by a certain amount than that of other candidates, and so on.

Another widely used decision rule is the K-nearest neighbour rule, when there are multiple reference patterns for a given word. The distance scores of the input utterance, with all of the multiple references for that word, are averaged. The word, for which this average score is the Least, is recognized as the one closest to the test input.

If the matching stage permits different distance measures then the decision rule may specify a discriminant function with appropriate coefficients, for each of those measures. Generally, such complex computation is resorted to, only if time for computation and costs are not major considerations as, for instance, during experimentation.

2.2.5 Performance of isolated word speech recognition systems

The performance of even the so-called successful isolated word recognizers is limited by the issues of speaker dependence, the nature and size of vocabulary and the environmental conditions.

The problems due to speaker dependence stem from the fact that the recognizers are sensitive to the characteristics of the speaker. The Texas Instruments' study of different commercially available isolated word recognizers shows that the performance is better for male speakers than for female speakers [36]. This confirms the speaker dependence from the point of view of sex. Table-2.1 lists the results of that study,

Table-2.1 : Texas Instruments' study of system performance for male and female speakers

Manufacturer	Model	Error (%) for	
		Hen	Women
Threshold Technology	T-500	1.2	1.7
Interstate Electronics	VRM	2.0	3.7
Heuristics	7000	4.4	7.3
Centigram	MIKE 4725	6.3	8.0
Scott Instruments	VET/1	11.2	14.0

The size and nature of the vocabulary are the two factors that determine the performance of a system. The number of comparisons during matching influences the time and cost of computation. They increase linearly with the size of the vocabulary. Also, the user has the additional burden of having to remember the words included in the vocabulary. These factors do not directly affect the system performance. More than the size, the nature of the vocabulary will be the factor affecting the system performance. If acoustical1y-similar words are included in the vocabulary the resulting increase in the confusability has a disastrous effect on the system performance. Compared with a vocabulary of multisyllabic words the performance for monosyllabic words is always poorer. Proper nouns like the names of cities, people, places, etc., are multisyllabic and are less confusable. But English letters and most of the digits are

monosyllabic and hence are more confusable. Table-2.2 lists the results of recognition scores obtained by Itakura C471 and Rosenberg et al., [52,53] for experiments with certain vocabularies.

Table-2.2 : Comparison of recognition scores for different vocabularies

Reference	Vocabulary	% Recognition
Itakura [47]	Names of 200 Japanese Cities	77.3
	English Letters (A-Z) and Digits 0-9	88.6
Rosenberg [52,53]	Cities, Airports, Days, and Numbers	71.6
	English Letters (A-Z)	79.5

As examples of confusable vocabulary one finds, in the literature, reference to the subsets of English letters, such as the E-set consisting of CB, C, D,E,G,P, T,V,Z}. Speech recognition systems may have only a 50% chance of success in discriminating between them. A 1983 U.S. patent [54] for a word recognizer claims an improvement to 75% in this performance. Experimental systems which can recognize voiceless stop consonants such as {P,T,K}, with 89% recognition accuracy, have been reported [55].

The environment, in which these recognizers are used, is another major factor affecting the system performance. In ideal laboratory-conditions, very high scores of even upto 99.9% recognition may be possible. But in actual field-tests the score drops to 97% or lower. The reason is that the inevitable noise makes it difficult to detect the endpoints of an utterance. Even in laboratory-conditions the speaker himself may contribute to noise by way of lip-pops, quick-inhalation before speaking, aspiration after speaking, etc. Also, during the training phase, the speaker may take utmost caution and carefully articulate the words for purposes of creation of reference templates, but during the recognition phase, he may be casually uttering the words. The preferred speech level is that of authoritative, confident, across-the-desk conversation [36].

Thus speech recognition is a multi-faceted problem. The advances achieved to this day are based on limiting the problem in terms of specifying the environment, the manner of uttering the words, careful selection of vocabulary, etc. Thus isolated word recognition has come to stay and has been exploited commercially.

Isolated word speech recognition in Indian languages has one point of advantage, namely the correspondence between the letters and the units of speech-sound. It is physically possible to utter each unit in isolation in most cases, such that concatenation of these units constitutes the speech. In other words, speech consists of independent1y utterable letters. Even if 100% recognition is ultimately achieved for the E-set and similar

vocabularies in English, it will not help in the development of systems for continuous speech, since English is a non-phonetic language. On the other hand a cent-percent sucessful recognition scheme for discrete utterances of Hindi letters will be a forerunner to continuous speech recognizers. This way one can process meaningful speech and still avail of the know-how already developed for isolated word recognition.

2.3 Need for a different approach for Hindi consonants

In our study of recognition of speech in Indian languages, recognition of consonants assumes great importance since they carry significant information.

As seen from Table-1.1, the two ways of classifying the stop consonants are by 'place and manner' information. The different manners of production are associated with distinctive characteristics such as voicing, aspirations and their combinations. But the utterances according to different places of articulation will display different characteristics only to the extent that is possible because of,

i) the different initial positions of the articulators, and

ii) the movement of the articulators to their steady state positions, as determined by the succeeding vowel.

In contrast with the different excitation features associated with the 'manner' those associated with the different 'places' are weak features. Hence discrimination among those consonants distinguished by different places of articulation, is difficult.

In order to design our system for a typical confusable vocabulary the voiceless stop consonants, given in the first column of Table-1.1, are considered first. The members of the set are {KA, CA, TA, HA, PA}.

Acoustically they differ among themselves in their short leading consonantal part and at the interface with the succeeding vowel-part. If the distinguishing primary features are not captured in these short durations, there is almost no chance of recovery from error in recognition based on the remaining portions of the utterance. In this aspect, this set differs from the widely known E-set and hence is a more confusable set.

The approach to recognition should be based on the nature of the signal in the respective segments of the speech pattern. One can extend this logic to another set of consonants produced by the same place of articulation but characterized by different manners of production. The processes for these two sets of consonants will have to be based, in one case on place-information and, in the other case on manner-information. Thus, when conventional methods follow fixed and uniform strategies in signal representation and matching, they result in degraded performance for monosyllabic words. It is worth noting that refinements such as DTW are for the whole-word approach and do not assume "a priori knowledge about which parts of speech patterns contain linguistically important information" [56], while it is these parts which primarily distinguish the similar-sounding members in a given vocabulary. Therefore, at signal level, different segments such as voiced, unvoiced, silence need to be represented with

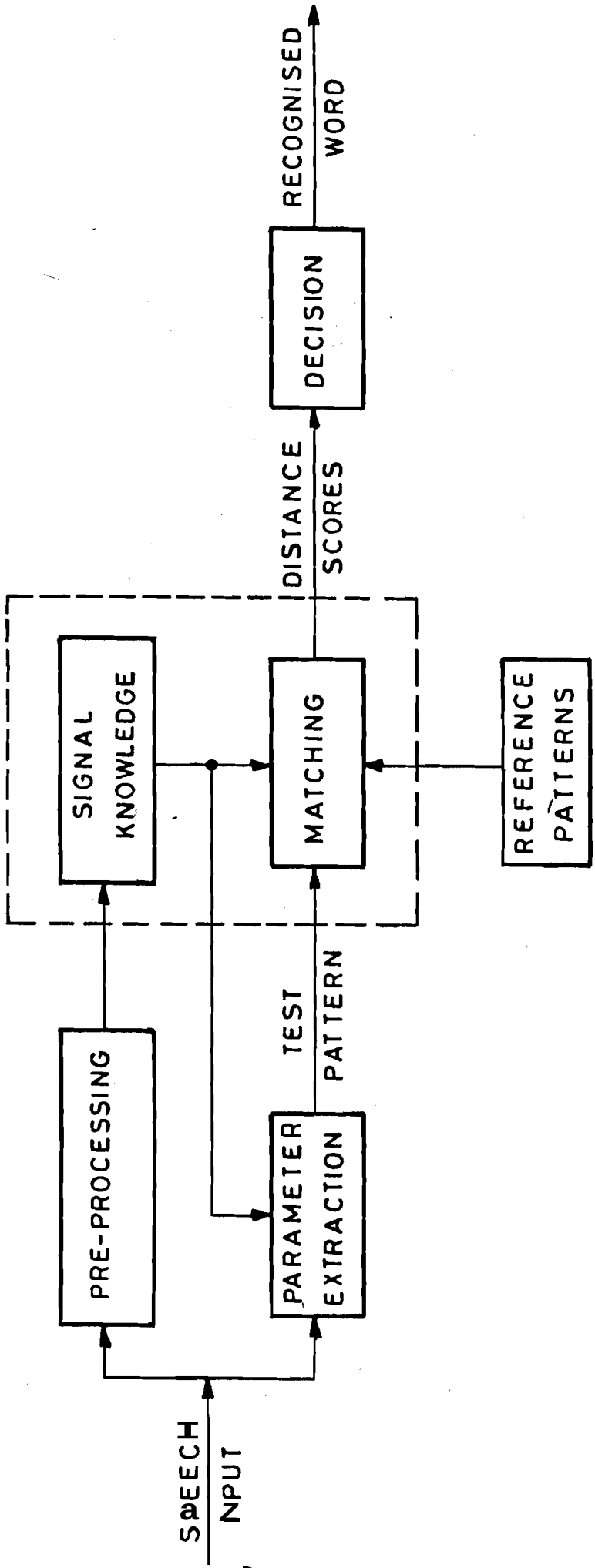


Fig. 2·5 The envisaged IWSR system based on signal knowledge

appropriate spectral and temporal resolution. Decision on these choices is to be based on information derived from the signal through certain additional pre-processing, as shown in Fig.2.5.

A different approach to recognition of confusable words consists of identifying and isolating those parts of utterance which primarily characterize the consonantal parts [33, 571.

To develop a recognition scheme for the confusable word set, we manually isolated the leading consonantal part by observing the speech waveform and by listening to different portions of the speech utterance, Through listening-tests we verified that the leading consonantal part can be isolated. Without the leading part the listener has difficulty in recognizing the consonant when the succeeding vowel-part alone is played to him. This implies that the isolated consonantal part does carry the discriminating information required for recognition. In view of this, to carry out appropriate processing on respective parts, we need to determine the regions of the consonantal and vowel segments.

Once the approach to recognition is finalised through more refinements, experiments are carried out with more data sets and for different vowel-endings. The results suggest that the approach to recognition should be based on the knowledge of the signal as well as on the knowledge of the vocabulary. In the process of experimentation, small changes are made in the system design at different stages, and the performance of the system is evaluated. In order to monitor conveniently the improvements or otherwise in the system performance, a new measure of performance evaluation is used.

3. SIGNIFICANCE OF PERFORMANCE INDEX IN THE EVALUATION OF SPEECH RECOGNITION SYSTEMS

"We have a tendency to fell a tree when we need a toothpick."

ROBERT LEMBKE in Die Weltwoche, Switzerland

3.1 Need for a new measure of performance evaluation.

At present the widely used measure of performance evaluation of speech recognition systems is the Percentage Recognition Score. In the following discussion it will be referred to as PRECS. Another form of this measure is the error rate, Computation of these measures requires a large amount of test data. Hence this measure is useful for the final evaluation of a developed system. During the initial stages of the system development, when innovative changes are frquently made, there may be only marginal improvements in the system performance. Measures such as PRECS are not convenient to study these marginal improvements. The following observation of Woodard and Lea [58] summarizes the present situation: "The procedure for evaluating the performance5 of speech recognizers is an important but often overlooked topic of research. Yet there can hardly be any more important problem in speech recognition than determining how well algorithms or devices work, Researchers need to be able to determine and predict the impact of design changes on performance."

When a recognition system is under development, a measure which indicates the system performance with fewer tests and one which responds to the manipulation of the system design parameters is desirable. It will help defer exhaustive testing with large sets of data to the final stages of system development. Performance Index [30,59] is one such useful measure. In the discussion to follow it will be referred to as PIX. It is derived from the distance matrix, formed for a pair of test and reference data sets, and hence reflects the characteristics of the distance matrix. Using this measure we discuss the performance of the system in handling the confusable vocabulary in its various stages of development.

Computation of PRECS is based on the yes-no outcome of the tests to verify whether the system recognized the input utterance correctly or not. It does not give information on how closely the test utterance resembles the correct reference word. Also there is no indication on how dissimilar the test utterance is to the other reference words. This information is not very important for the evaluation of a system, the design of which has been finalised. At that stage the system is expected to handle data with all the variations possible or permissible. Hence the relevance of these statistical measures of performance evaluation.

Since PIX is computed from a distance matrix it reflects the characteristics of the latter. The distance matrix contains information on the distances between a test utterance and the reference words. This is indicative of the degree by which the

chosen reference word stands against the rejected ones. Since PIX can be computed with only one pair of test and reference sets it directly brings out the effect of the innovative changes in the system design. However, for a reliable estimation of the system performance, tests may be conducted with more data sets and the average value of PIX computed. At a stage when the performance of the system has been improved to some acceptable level, statistical measures such as PRECS may be computed.

A measure based on the distance matrix is likely to be more consistent than percentage recognition. It is because of the following reason. For small test data such as a pair of test and reference sets PKECS assumes values in discrete steps, say, multiples of 10% for a ten-word vocabulary. On the other hand, a measure like FIX which is based on distance matrix can assume a continuum of values over the range 0 to 100%. The nature of PIX and its utility in the development of a speech recognition system are studied in this chapter.

3.2 Definition of Performance Index

In a distance matrix the diagonal elements are the distances between two repetitions of the same word and the nondiagonal elements are the distances between two different words. If, as in Tables 3.1 and 3.4, the diagonal element in every row is smaller in value than the corresponding nondiagonal elements, the contribution of the test to PRECS is 100%. It means that there is no confusion between the reference words and these test utterances. However PRECS does not reflect the deviation of the diagonal elements from the nondiagonal elements, which alone will indicate the degree by which the chosen reference stands against the rejected references.

The two matrices shown in Tables 3.1 and 3.4 are for two different systems when tested with the same pair of test and reference data sets. Both record 100% successful recognition. Just a visual inspection would reveal that the distance matrix for system 2 is better than that for system 1. The reason for this is the better discriminability between the diagonal and the nondiagonal elements in a qiven row of the second matrix. PIX quantifies this information and thus reflects the 'goodness' of system 2 over system 1, for the qiven data. It is computed as illustrated in Tables 3.1 to 3.6.

The nondiagonal elements of each row in the distance matrix are normalized with respect to the corresponding diagonal element and expressed in percentage. The resulting matrix is called the normalized distance matrix. The next step is the derivation of an index matrix from the normalized matrix using the piecewise-linear mapping function illustrated in Fig.3.1. The basis for this mapping function is that the elements which are well discriminated and those which are not are given weights based on some nonlinear weighting function. The nonlinear weighting ensures the following in the computation of system performance.

(i) The elements which are discriminated very well should not contribute too much to PIX.

(ii) The elements which are marginally discriminated are not completely kept out of consideration while computing PIX.

(iii) The elements which are discriminated in varying degrees between these two extremes are given appropriate weightages.

(iv) The elements which are discriminated very poorly are reckoned as those which contribute the most to the confusion and so their contribution to PIX will be set to zero.

Table 3.1 : Distance matrix for system 1

Test Utterance \ Reference Word	Five	Six	Seven	Eight	Nine
Five	1526	3620	2508	4115	1856
Six	4706	1600	2748	3136	6374
Seven	2565	2260	1475	4294	3836
Eight	3984	1534	3005	1297	4540
Nine	1416	3394	2255	3559	1355

Table 3.2 : Normalized distance matrix for system 1

Test Utterance \ Reference Word	Five	Six	Seven	Eight	Nine
Five	100	237	164	269	121
Six	294	100	172	196	398
Seven	174	153	100	291	260
Eight	307	118	231	100	350
Nine	104	250	166	262	100

Table 3.3 : Index matrix for system 1

Test Utterance \ Reference Word	Five	Six	Seven	Eight	Nine
Five	---	100	99	100	42
Six	100	---	100	100	100
Seven	100	88	---	100	100
Eight	100	36	100	---	100
Nine	14	100	100	100	---

PIX = 88.95

Table **3.4** : Distance matrix for system 2

Test Utterance \ Reference Word	Five	Six	Seven	Eight	Nine
Five	565	1833	1427	2120	968
Six	1738	475	1149	1812	2542
Seven	1142	1011	486	1758	1827
Eight	1101	589	904	306	1179
Nine	635	1517	1196	1829	580

Table **3.5** : Normalized distance matrix for system 2

Test Utterance \ Reference Word	Five	Six	Seven	Eight	Nine
Five	100	324	252	375	171
Six	**366**	**100**	**242**	**381**	**535**
Seven	**235**	**208**	**100**	**362**	**376**
Eight	**360**	**192**	**295**	**100**	**385**
Nine	**109**	**261**	**206**	**315**	**100**

Table **3.6** a Index matrix for system 2

Test Utterance \ Reference Word	Five	Six	Seven	Eight	Nine
Five	---	**100**	**100**	**100**	**100**
Six	**100**	---	**100**	**100**	**100**
Seven	100	100	---	100	100
Eight	100	100	100	---	100
Nine	19	100	100	100	---

PIX = **95.95**

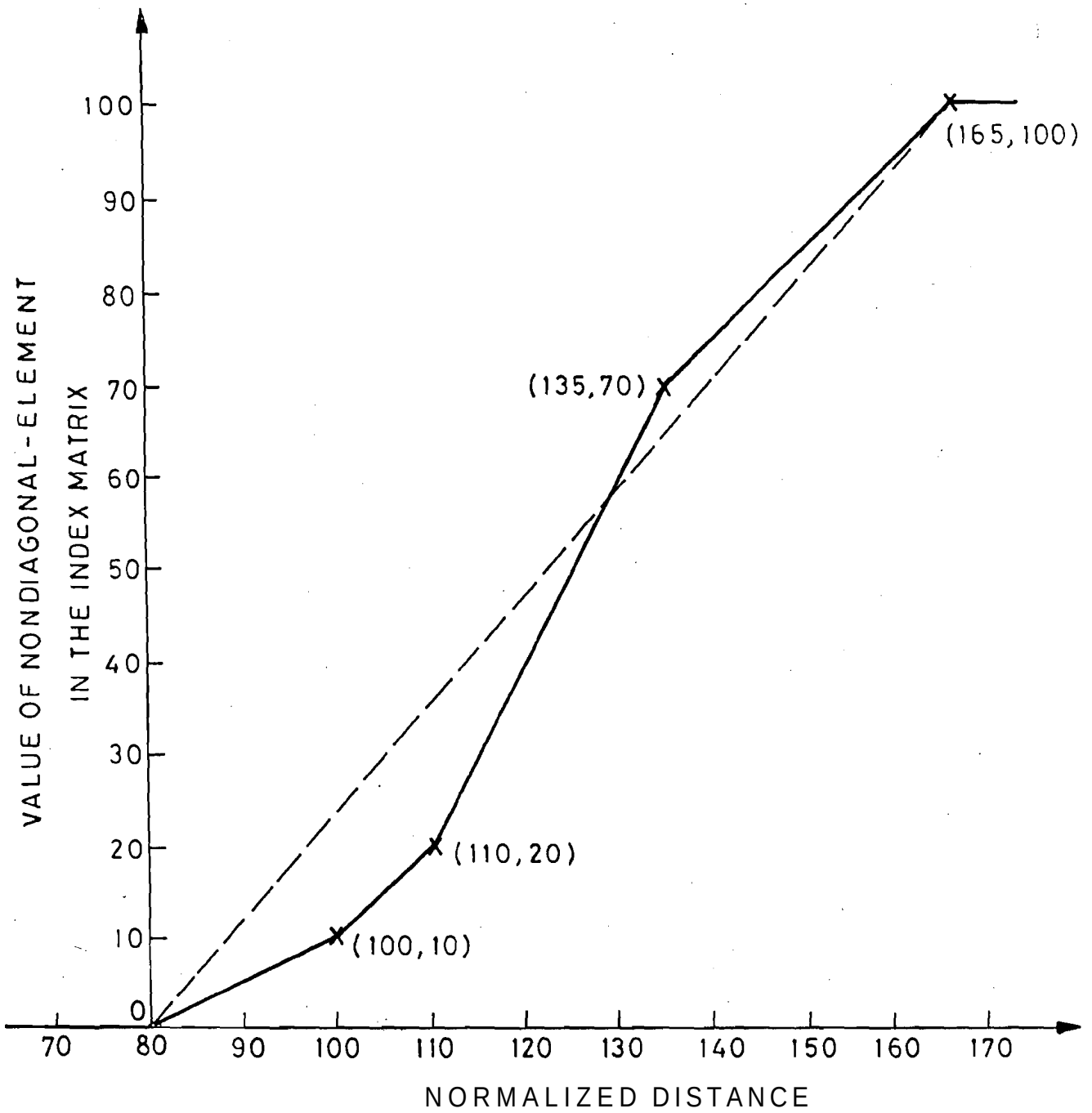


Fig.3·1 Mapping functions for computation of PIX.

based on these considerations, the index matrix is derived as follows:

(i) Any off-diagonal element the value of which is larger by 65% than that of the diagonal element is considered as non-confusing and its contribution to the index is taken as 100.

(ii) Any off-diagonal element the value of which is 80% or less than that of the diagonal element is considered as totally confusing and its contribution to the index matrix is taken as zero.

(iii) Contributions to the index matrix by the off-diagonal elements whose values lie between these two extremes are obtained using the mapping function shown by the solid lines in Fig.3.1.

The next step in the computation of PIX is to find the average of all the nondiagonal elements. Thus PIX is defined as the average of all the nondiagonal elements of an index matrix, derived from the normalized distance matrix. Computing PIX in this manner one can observe that the performance of system 2 is better than that of system 1.

The breakpoints on the piecewise-linear mapping function curve were empirically determined and were found to be satisfactory. For the range of values of the elements in the distance matrices, these breakpoints yield values of PIX which highlight the differences in the perfomances of the systems under comparison.

Other breakpoints could as well be chosen, but may not result in PIX values very much different from those obtained with this mapping function. With a different mapping function the degree by which the system performances are differentiated would vary. For instance, as an extreme case, if a linear mapping function is assumed, it will be as shown by dotted' lines in Fig. 3.1. For the illustrative two systems considered, the index matrix will be as shown in Table-3.7. It can be seen that the PIX values still show that the performance of system 2 is better. But the choice of a linear mapping function has slightly reduced the performance differences. The breakpoints chosen for the piecewise-linear mapping function yielded PIX values for a better comparative study.

Table 3.7 : Index matrices with a linear mapping function

3.7(a) : For System 1

---	100	100	100	51	
100	---	100	100	100	
100	91	---	100	100	
100	48	100	---	100	
30	100	100	100	---	PIX = 91.0

3.7(b) : For System 2

---	100	100	100	100	
100	---	100	100	100	
100	100	---	100	100	
100	100	100	---	100	
36	100	100	100	---	PIX = 96.8

PIX is a measure of the performance of the system for a given pair of reference and test data sets. For the same sets of data, if innovative changes are made in the system, PIX would reveal whether it is for the better or for the worse. With a small number of tests the trend could be confirmed. If necessary, the recognition experiments could be performed with a large number of tests, and statistical measures of performance computed. In the case of numerous design parameters, instead of fixing their values arbitrarily and thus running blind, one could shortlist the parameters and vary them for a few runs. After computing PIX for those cases, further extensive experimentation with more data can follow. Thus PIX enables one to evaluate the changes in the performance of a recognition system using small test data on modest computational facilities.

3.3 Comparative study of Performance Index (PIX) and Percentage Recognition Score (PRECS)

3.3.1 Methodology of the study

The objective of this study is to determine how well PIX and PRECS compare with each other. The performance of a recognition system is evaluated in terms of PIX and PRECS for two types of vocabulary, One is the set of digits (zero, one, ... nine), the other is a subset of Hindi stop consonants, namely (KA, CA, TA, HA, PA). The system is tested for speaker dependent condition.

75 sets of test data for digit-task and 64 test sets for stop consonants form the total population of data considered for analysis. The performance of the system is evaluated based on studies with total population as well as with samples of size of about 15 test sets drawn from the population. At this stage one may recall that, for reliable computation of statistical measures such as PRECS, the size of the sample should be large enough. Notwithstanding this fact, PRECS is computed for the sample sets as well as for the total population. Eased on the PIX for each of the test sets, the values of mean, median and standard deviation of PIX are computed for the sample as well as for the total population. Table-3.8 shows the computed values of PIX and PHECS for the digits vocabulary.

Table-3.8 : PRECS, PIX, etc., for digits set

Measure	Of the Total Population	Of the samples drawn : Set no.				
		1	2	3	4	5
Size	75	15	15	15	15	15
FHECS	94.13	80.00	94.66	98.75	100.00	98.33
Median of PIX	94.90	86.51	93.61	96.40	97.30	95.90
Mean of PIX	92.92	87.14	91.47	95.98	94.94	95.66
S.D. of PIX	7.28	6.91	7.10	5.13	7.52	4.63

3.3.2 Relevant statistics for small sampling

In statistics, samples of size $N > 30$ are called large samples and those of size $N < 30$ are called small samples [60]. The sampling distributions of most of the commonly considered statistics are approximately normal for large samples. This approximation becomes better with increasing N. For small samples this approximation is not good and becomes worse with decreasing N. This calls for appropriate modifications. Student's t-distribution [60] is one such modification and is suitable for evaluation of results about averages based on small samples. It is defined as,

$$t = \frac{\bar{x} - \mu}{s} \sqrt{N-1} ,$$

for samples of size N drawn from a normal (or approximately normal) population with mean μ. For each sample one computes t, using the sample mean $\bar{x}$ and the sample standard deviation s.

One can define 95%, 99% or other confidence intervals by using appropriate standard tables of the t-distribution [60]. In this manner, one can estimate the population mean μ within specified limits of confidence.

For example, if $-t_{0.975}$ and $t_{0.975}$ are the values of t for which 2.5% of the area lies in each "tail" of the t-distribution, a 95% confidence interval for t is

$$-t_{0.975} < \frac{\bar{x} - \mu}{s} \sqrt{N-1} < t_{0.975}$$

From this the estimated interval that contains μ is

$$\bar{X} \pm t_{0.975} \frac{s}{\sqrt{N-1}} .$$

In this expression $t_{0.975}$ represents the 97.5-percentile value, while $t_{0.025} = -t_{0.975}$ represents the 2.5-percentile value. When the population mean μ is known one can determine the confidence interval (computed using the sample set) within which μ lies. This gives an indication of whether the sample set is representative of the aggregate.

For the purpose of cross-verification the 95% and 99% confidence intervals have been computed by another method, called Bootstrap [61]. In this, the sample evaluates itself by pulling on its own bootstraps. A large number of copies, running to a million or so, are made from a sample set. If the sample set consists bf, say, 15 members, the copies are shuffled and 15-member sample sets are extracted. The mean of each of these hundred-odd sets is computed. Cutting out the high and low values in the range of the 'means' obtained thus, the appropriate 95% and 99% confidence intervals are computed.

To test the significance of the standard deviation at a given confidence interval the χ^2-distribution can be used. The statistic χ^2 is defined as

$$\chi^2 = \frac{N s^2}{\sigma^2} ,$$

where σ is the population standard deviation. When the standard deviations of the sample sets and the population are known the statistic χ^2 can be computed. For specified limits of confidence

(CF), F can be obtained for these sample sets using appropriate Tables [60]. Comparison of χ^2 and χ^2_{CF} will reveal the extent to which the standard deviation of the sample set conforms to that of the population.

3.3.3 'statistical significance of PIX for Number set

Since we know the mean and standard deviation of the PIX for the population as well as for sample sets we can apply these formulae and check the extent to which the samples are representatives of the aggregate. For the different sample sets the confidence intervals of the mean PIX at 95% and 9 9 levels and the statistic χ^2 for the respective set are computed. For each sample set ^{2} at 95% confidence level is obtained from the standard tables [60]. These are shown in Table-3.9.

Table-3.9 : Appropriate statistics of PIX for digits set

Population Mean = 92.92

Sample Set-No	Sample Mean	Confidence intervals 95%	Confidence intervals 99%	$\chi^2 = \frac{N s^2}{\sigma^2}$	$\chi^2_{0.95}$
1	87.14	83.19 - 91.09	81.64 - 92.64	13.51	23.7
2	91.47	87.41 - 95.53	85.82 - 97.12	14.27	23.7
3	95.98	93.05 - 98.91	91.90 - 100.0	7.45	23.7
4	94.94	90.64 - 99.24	88.95 - 100.0	16.00	23.7
5	95.66	93.01 - 98.31	91.97 - 99.35	6.07	23.7

The 95% and 99% confidence intervals based on the Bootstrap method for the digits set have been computed and are shown in Table-3.10. It may be seen from Tables-3.9 and 3.10 that the confidence intervals may be computed by either method.

Table-3.10 : Confidence'intervale based on the Bootstrap method

Population Mean = 92.92

Sample Set-No	Sample Mean	Confidence intervals 95%	99%
1	87.14	83.42 - 90.54	81.51 - 91.30
2	91.47	87.50 - 94.75	86.63 - 95.33
3	95.98	93.19 - 98.03	92.21 - 98.36
4	94.94	91.02 - 98.14	89.09 - 98.61
5	95.66	93.06 - 98.15	90.93 - 98.51

On studying ~ a b l e 3.8 and 3.9 we make the following observations:

1. With reference to Table-3.8, both PRECS and PIX show that, except for the first set, the remaining sets give better indication of the aggregate.

2. With reference to Table-3.9, the 99% confidence interval of PIX includes the population mean for all the sets except for the first. set. Only for two sets do the 95% confidence intervals include the population mean.

3. In Table-3.8, PHECS computed with sample sets of data give good indication of PHECS obtainable with large sets of data. The reason for this can be that the values of PRECS are high. In addition, it also identifies the outlying sample set, which is

the first set. Thus PRECS, based on sample sets themselves, show that four sample sets are representatives of the aggregate.

4. With reference to Table-3.9, using the χ^2-distribution for testing the significance of the standard deviations of PIX at 95% confidence interval, we note that the χ^2 values are less than those for 95% level. Thus the sample standard deviations of PIX conform to the population standard deviation. In other words, the variability indicated by the respective sample test is comparable to the variability indicated by the large number of tests.

3.3.4 Statistical significance of PIX for Hindi stop consonants *set*

The second vocabulary considered for these cornparative studies was that of Hindi stop consonants. Table-3.11 lists the PRECS, median, mean and standard deviation of PIX for the sample as well as for the total population of the test sets. Table-3.12 shows the appropriate statistics computed for the sets of this vocabulary.

The 95% and 99% confidence intervals based on the bootstrap method for the Hindi stop consonants set have been computed and are shown in Table-3.13. It may be seen from Tables-3.12 and 3.13 that the confidence intervals may be computed by either method.

Tables 3.11 and 3.12 lead to the following observations:

1. For any of the sample sets the 95% as well as the 99% confidence intervals of PIX includes the population mean of PIX. It implies that the samples are true representatives of the aggregate, and the value of PIX determined from sample data is comparable to the value of PIX for large data.

Table-3.11 : PRECS, PIX, etc., for the Hindi stop **consonants** set

Measure	Of the Total Population	Of the samples drawn: Set No. 1	2	3	4
Size	64	16	16	16	16
PRECS	63.75	55.00	66.25	70.00	63.75
Median of PIX	50.55	47.90	49.71	51.90	54.21
Mean of PIX	49.63	48.21	46.73	50.32	53.27
S.D. of PIX	10.03	9.33	9.90	9.77	9.88

Table-3.12 : Appropriate statistics of PIX for the Hindi stop consonants set

Population Mean = 49.63

Sample set No.	Sample Mean	Confidence Interval 95%	99%	$\chi^2 = \frac{N s^2}{\sigma^2}$	$\chi^2_{0.95}$
1	48.21	43.08 - 53.34	41.11 - 55.31	13.84	25.0
2	46.73	41.29 - 52.17	39.19 - 54.27	15.59	25.0
3	50.32	44.95 - 55.69	42.88 - 57.76	15.18	25.0
4	53.27	47.84 - 58.70	45.74 - 60.80	15.53	25.0

Table-3.13 : Confidence intervals based on the Bootstrap method

Population Mean = 49.63

Sample Set-No	Sample Mean	Confidence intervals 95%	99%
1	48.21	43.73 - 52.79	40.95 - 54.08
2	46.73	41.96 - 51.66	39.89 - 52.58
3	50.32	45.60 - 54.75	43.31 - 56.03
4	53.27	48.59 - 58.16	45.29 - 59.48

2. As expected, the PRECS computed with small sets are not reliable.

3. Compared with the previous case of digits set, the values of PIX in this case are low and it reflects the poor values of PRECS. Once this aspect is known from tests with sample sets of data, PRECS need not be computed at all until the system performance a5 measured by PIX improves.

4. For digits vocabulary the values of PIX and PRECS are high and are nearly equal. For stop consonants the values of PIX are uniform but low, and the values of PRECS are widely varyinq. These imply that the system does not perform well for the second case. Specifically, the system does not distinguish the members of the vocabulary very well. This is indicative of the confusability of the stop consonants vocabulary since the system used is the same in both the cases.

5. As in the previous case, the χ^2-values indicate that sample standard deviations of PIX conform to the population standard deviation and hence the variability indicated by the respective sample test is comparable to the variability of the large number of tests.

3.4 Conclusions

Based on the studies on a system with two vocabularies of different nature we arrive at the following conclusions.

1. When PRECS is low the performance of the system will be poor. Thus, depending on the characteristics of small test data, there will be fluctuations in the values of PRECS. This will be indicated by low values of PIX.

2. Low values for PIX are obtained, irrespective of whether PIX is computed using small or large sets of data.

3. The low values of PIX for sample sets indicate that the system is not capable of recognizing the vocabulary under consideration.

4. The low values of PIX are pointers to the fact that PRECS would be low and hence the system design must be improved so as to enhance the system performance. When PIX has improved, PRECS can be computed reliably with large data sets.

5. Since PIX reflects the characteristics of the distance matrix, larqe values of PIX for sample sets indicate that the corresponding value of PRECS will be high.

6. Thus the computation of PIX obviates the need for exhaustive testing of a system with larqe test data at each and every stage of the system development. In other words, PIX is a cost-effective measure of evaluation in the early stages of system development.

7. PIX could also help in the comparison of two systems with nearly identical PRECS, when the test is carried out with small but carefully chosen typical data sets. By definition, PRECS is a reliable measure only for large amounts of test data. The nearly identical values of PRECS of two systems need not necessarily be due to the differences in the characteristics of the systems. It could as well be due to the variability in the test data. Thus PIX is a better measure of the system performance when sample data is carefully chosen.

8. Since a system is supposed to handle large amounts of data, the characteristics of which are expected to have some variability, PRECS is the the ultimate measure. But PIX helps the deferral of the computation of PRECS until the system has a certain minimum acceptable level of performance.

Our studies in this chapter indicate that PIX is an effective measure of the performance of the systems as well as vocabularies. However, both PIX and PRECS are useful in practice because PIX reflects the characteristics of the distance matrix whereas PRECS gives the overall recognition accuracy. Thus PIX fulfils the long-felt need for a measure "to determine and predict the impact of design changes on performance". Our task is to develop a system which can handle confusable vocabulary. This resulted in a series of experiments leading to changes in the system design, and PIX is used to discuss the results of all of these.

4. SIGNAL-DEPENDENT MATCHING FOR SPEECH RECOGNITION

"Women have a passion for mathematics.
They divide their age5 by two, double
the price of their dresses, treble their
husband's salaries and add five years to
the ages of their best friends."

- MARCEL ACHAHD
quoted in The Daily Express

4.1 Introduction

In this chapter we discuss certain limitations which generally impair the performance of a speech recognition system. Most of the common approaches consist of tackling these issues individually. The method based on signal-dependent matching is discussed as a possible integrated approach for tackling these issues. Study of this approach in the context of the recognition of Hindi stop consonants is relevant, since handling of this vocabulary calls +or refinements in the conventional methods of processing.

The recognition reliability and computational complexity of an IWSR system are greatly affected by the accuracy of detection of the endpoints of an utterance. In some cases there may not be any unique endpoints, and hence even manual identification of these points from the speech waveform may not be possible. The alternative is to accept the ambiguity in the endpoints and try to design algorithms for matching, to take care of the ambiguity.

Generally, random fluctuations can be observed in any parameter-contour of the representation of speech data. These are due to natural variations occurring at the speech signal level itself, or introduced during various stages of speech signal processing. These fluctuations cause unpredictable distances during matching. Methods to reduce the effects of fluctuations are parameter smoothing [62], parameter reduction and parameter averaging [63]. These techniques do not consider the reasons for fluctuations. As a consequence it is possible that some of these techniques may average out relevant information contributing to the discriminability. One method of overcoming this problem is to represent the speech utterance as a sequence of symbols. The matching of test and reference utterances then consists of matching the corresponding symbol strings [64]. Another approach is to make use of information' from the signal. and decide the strategy for compensating the fluctuations adaptively.

A speech signal consists of different types of segments and these do not require the same kind of matching. For example, a silence frame when matched with another silence frame yields sometimes arbitrarily large distance due to noise in the data. When it is known that a frame belongs to the silence category, its parameters do not contain more information than can be used for classifying the frame as a silence frame. Hence a silence frame does not require the same procedure for distance computation as may be needed for, say, the voiced category. But, the different versions of the widely used DTW algorithm do not permit any change in the matching strategy during comparison.

In order to take care of these problems as part of the processing, signal-dependent matching was originally proposed [30, 313. It is based on making use of information derived from the signal to influence certain decisions in the stages of parameter extraction and matching.

4.2 An algorithm for signal-dependent matching

The process of making the matching algorithm adaptive, consists of two distinct steps. The first step is to derive the information and the second is to use it in the matching algorithm., The main difficulty in developing the matching strategy is to know which information is to be used. Based on the signal characteristics of each frame, the parameters to represent the frame and a measure for dissimilarity between the frames have to be chosen.

From the point of view of implementation, changes in the parameter set and in the matching strategy are ultimately intended to give due weightage to the distance computed for each test frame. This can be implemented by creating a weight-function for each utterance. The weight-function, consisting of tuples for each frame, indicates the parameter set to be used for each and the weighting-factor for the distance computed for that frame. The values of each tuple should be determined based on information from the signal.

It is to be noted that the weight-function should depend only on the test utterance and not on the reference template. If the weight-function depends on the reference data also, the

distances between a given test utterance and different reference words do not have any common basis for comparison. For the same reason, DTW algorithms, which use symmetric comparison or which skip test frames based on reference data, are not suitable for signal-dependent matching.

The signal-dependent matching is implemented in the following manner. Since there is no significant difference among the warping paths obtained for different parameter sets [30], the warping path is determined using two parameters per frame. One corresponds to the low-frequency energy-level and the other to the high-frequency energy-level. The interframe distances along the path are computed according to the weight-function. In other words, different parameter sets, distance measures and weight-factors ran be used for each test frame, based on the information from the signal. Thus the distance computations for each test frame could be very complex without increasing the overall complexity in the procedure of matching.

4.3 Speech material and representation

Speech data consists of a minimum of two repetitions of the word set for the given vocabulary. One of the repetitions is used as reference and the other as test data. The speech signal is sampled at 10 kHz and digitized and stored as 16-bit numbers. Each utterance is divided into nonoverlapping frames of size 256 samples. Each frame is multiplied by a Hamming window-function and its discrete Fourier transform is computed. The spectrum is reduced to 16 log-spectral values on an approximate melfrequency-scale as given in [63]. To compensate for signal level differences in the speech data, normalization is achieved using the propor-

tional normalization method of Das [65]. This is carried out on spectral time-samples. While compensating for inter-utterance signal level differences it also preserves the intra-utterance amplitude relationships. Each test utterance is compared with each of the reference words using the DTW algorithm. The distance between a test frame and a reference frame is computed by summing the absolute difference between the corresponding log-me1spectral values.

During the course of this study there had been changes in the analog front-end used. Details of the latest version of it are given in Appendix-I. Brief description of the different program modules developed for this study is given in Appendix-11. Typical print-outs of the results are included in Appendices III to VI. The PIX-values are shown at the right-bottom corners of the index matrices. Of the two columns of letters shown at the extreme right end of the print-outs, the left column corresponds to the test inputs, and the right column to the recognized outputs.

The DTW algorithm used is similar to the asymmetric algorithm proposed by Itakura [47], except that the search region is modified to incorporate some flexibility for skipping a maximum of four reference frames at the begining. This will to some extent take care of the ambiguities in the detection of the begin-frame. In isolated utterances the end is fairly well defined by the falling energy-contour. It is to be noted that this implementation of DTW does not permit skipping of any test frame. This is important from the signal-dependent matching point of view.

As stated earlier, the weight-vector for each test frame is derived on the basis of the signal characteristics of each. The first component of the weight-vector indicates the parameter set to be used for that frame. Usually this parameter set automatically decides the 'type of distance computation to be adopted. The second component of the vector is a scale-factor for the distance computed for the frame. To determine these components a frame is classified into one of voiced/ unvoiced/silence categories using the relative energy levels of different frames.

There is no need for this classification to be very accurate. Simple algorithms serve this purpose. The method is based on identifying the frame corresponding to the maximum energy, and classifying other frames with reference to that. The frames in the end-portions with less than 80%of the peak energy level are classified as silence frames. The frame showing a 30% increase of energy level as compared with the average energy level of all the previous frames, is taken as the boundary-frame of the nonsilence portion. Among the nonsilence frames those with a higher value of the low frequency components of energy are considered as voiced frames and the others as unvoiced ones.

For nonvoiced frames all the 16 log-melspectral parameters are used, For voiced frames the deviation of the log-melspectral values from the mean value for the frame is used as the parameter. These zero-mean parameters provide some kind,of energy normalization, which is generally required for voiced frames, owing to significant fluctuations in energy from one repetition of the utterance to another.

The scale-factors for the respective frames are determined based on the energy level relative to the maximum, as shown in Table-4.1. These weightages are somewhat arbitrary and could be refined. The weightage is reduced for the transition frames between the three classes, as it is difficult to reproduce the characteristics of the transition frames for different repetitions.

Table-4.1 : Weightage assignment for different test frames

Frames	Weight
(a) With Energy level < 0.8 Maximum	0.2
(b) 0.8 Maximum <= Energy level <= Maximum	1.0
(c) In the transition region	0.5

4.4 Experiments with other vocabularies

For a comparative evaluation, three subsets of alphadigit vocabulary have been cansidered first. They are the E-set consisting of letters {B,C,D,E,G,P,T,V,Z}, the A-set consisting of letters {A,J,K} and the digits set consisting of numbers {0,1,..9}. The performances of the two systems, one based on the conventional approach following fixed strategies and the other on the signal-dependent approach outlined above, have been evaluated. Table-4.2 shows the improvement in the values of PIX for the signal-dependent approach over those for the conventional approach.

Table-4.2 : Performance Index for three sets of vocabulary by conventional and signal-dependent approaches

Approach	{B, C, D, E, G, P, T, V, Z}	{A, J, K}	{0, 1, 2, ... 8, 9}
Conventional	70.25	85.63	84-50
Signal-dependent	84.26	97.33	93.47

4.5 Experiments with Hindi stop consonants

Since the signal-dependent approach has resulted in better system performance for alphadigit vocabulary, it is adopted for the development of our recognition system for Hindi stop consonants. The set of voiceless stop consonants, consisting of {KA, CA, TA, HA, PA}, is considered first.

Under identical conditions of frame size, parametric representations etc., as have been used in the case of alphadigit vocabulary, the PIX-value obtained for Hindi stop consonants is only 25.8. This clearly shows that the consonants set is more confusable than the widely known E-set and that our approach should be further refined to handle this confusable vocabulary.

Acoustically each member of this word set consists of two parts: the short, leading consonantal part and the following dominant vowel-part. As these members have the same vowel-ending, the distinguishing primary features in the consonant-vowel transition part should be captured. Otherwise, there is almost no chance of recovery from error in recognition based on the remaining portions of the utterance. Further refinements to the process of recognition should include this knowledge also, in order to enhance the performance of the system.

5. KNOWLEDGE-BASED PROCESSING FOR CONFUSABLE VOCABULARY

> "Learning is acquired by reading books, but the much more necessary learning, the knowledge of the world, is only to be acquired by reading men and studying all the various editions of them."
>
> - LORD CHESTERFIELD

5.1 Introduction

Humans do not have much difficulty in recognizing isolated utterances of Hindi stop consonants. It may be possible to achieve similar performance by machines if some knowledge about the vocabulary is incorporated into the system. In this chapter we show that knowledge-based processing improves the system performance significantly.

The knowledge related to speech production and speech perception of these utterances is made use of. Specifically, we show that the confusability of the vocabulary requires the use of appropriate knowledge. A strategy of confining the attention to those parts, which primarily contain the information distinguishing the different utterances, has been evolved for improved recognition. The experiments show that the discriminating parts should be appropriately represented so as to reflect the characteristics of the consonant. Specifically, the representation of the predominantly consonant part requires good temporal resolution to capture the articulatory dynamic characteristics of the particular consonant. Table-5.1 summarizes

the enhancement of the system performance from a mere 25.8 for the whole utterance to 74.6 for the consonantal part, in response to the different refinements discussed in this chapter,

Table-5.1 : System performance for different refinements discussed in this chapter

Vocabulary : { KA,CA,TA,HA,PA }

Expt. No.	Method	PIX
1.	Fixed strategy : Uniform frame site and resolution	25.8
2.	Improved temporal resolution	50.5
3.	More temporal but less spectral resolution	68.0
4.	Performance evaluation for consonantal part alone	74.6

5.2 Speech-specific knowledge

Knowledge-based methods of speech processing employ speech-specific knowledge to solve complex problems [66]. The knowledge could vary from the lower levels of acoustics and phonetics to the higher levels of syntax, semantics and pragmatics. In the development of speech-understanding methods there have been successful attempts to employ knowledge at higher levels [38, 39].

Continuous speech recognition systems developed under the ARPA-SUR project, such as HARPY and HEARSAY-II, employ techniques based on knowledge at higher levels, in order to improve the system performance. Incorporation of higher levels of knowledge

will not be possible in all cases. For instance, if a machine has to recognize isolated English letters, there is no way one can derive higher level knowledge from those isolated utterances. Such systems have to make do with whatever can be gathered at the acoustic and phonetic levels only.

Application of knowledge at lower levels such as the signal level had not been considered in literature for a long time. Hence, while referring to higher levels of processing, Fant [40] observed that "people have perhaps been a bit too eager to exploit these domains while neglecting the primary phonetic element recognition". Since then the scenario has been changing. The use of a variety of acoustic properties has gained ground as it seems to hold some promise for solutions in complex tasks [67-69]..

The two categories of knowledge that we use at signal level are the acoustic-phonetic knowledge (knowledge of speech production) and the knowledge of speech perception. The phonetic nature of the Hindi language and the symmetry of the structure of the stop consonants are exploited in evolving a hierarchical approach to .recognition. The part of the knowledge that influences the decisions at signal processing level is considered, and accordingly selective processing of the different segments of the signal is carried out.

5.3 Effect of knowledge on recognition results

Each member of the set of voiceless stop consonants {KA,CA,TA,HA,PA} consists of a short, leading consonantal part followed by a dominant vowel part. The consonantal part is characterized by a signal which is largely transient in nature.

During the production of the transients the articulators are in rapid mot'ion and do not attain any steady-state positions. The representation of this part of the signal should be such as to reflect the dynamics of the articulators. On the other hand, during production of the succeeding vowel or diphthong-part the articulators remain in some steady-state positions or move slowly. This part requires a representation which captures information on the the shape of the articulators. Thus the two main parts of these utterances are to be represented appropriately. The transition region between these two parts exhibits the characteristics of both the parts. Hence the dividing line between these parts cannot really be precisely identified.

Information such as this determines the temporal and spectral resolution required and also the type of parameters to be used for signal representation. In order to capture the articulatory dynamic characteristics of the consonantal part, the representation requires good, temporal resolution. by allowing the frames to overlap, the inter-frame interval is reduced, thereby achieving higher temporal resolution. Table-5.2 shows that by enhancing the temporal resolution in the representation of the Hindi stop consonants the system performance has been increased from 25.8 to 50.5.

Noting that there is further scope in representing the different parts with appropriate resolution, the next step is to decide on the number of parameters for the respective parts. The vowel-part requires relatively high sepectral resolution compared to what is required for the consonantal part. This way the

spectral shape representing the steady vocal tract system may be preserved well. For representation of the consonantal parts a high resolution spectrum is not required since the articulators during the production of these parts do not remain in steady-state positions. A gross shape of the spectrum is sufficient. In accordance with this observation, when the number of parameters for the consonantal part is reduced to 4, the PIX is found to increase to 68.0.

In order to perform the above-mentioned selective processing based on the knowledge of the vowel and consonant production, it is necessary to segment the utterance into consonantal and vowel-parts. One method of segmentation is based on the variations of the energy per frame. When the overall energy of a frame reaches about 85% of the peak for that utterance, followed by three frames of no significant dip, the frame is marked as the beginning of the vowel. Manual inspection of some typical spectrograms led to adoption of this empirical method of determining the consonant-vowel interface.

Table-5.2 : Performance of the system for varying the spectral and temporal resolutions

Expt. No.	Frame size (No. of samples/frame)		No. of log-mel spectral parameters		PIX
	Consonant	Vowel	Consonant	Vowel	
1.	256	256	16	16	25.8
2.	256 (with overlap of 192)	256 (with overlap of 192)	16	16	50.5
3.	256 (with overlap of 192)	256	4	16	68.0

Although there is improvement in the performance of the system it is not as high as that obtained for the alphadigit vocabulary. This shows that there is further scope for improvement.

5.4 Speech recognition method for confusable vocabulary

5.4.1 Segmentation based on transient analysis

The frames corresponding to the consonant and vowel-parts have been identified in the previous experiment on the basis of energy level differences. The consonantal part is characterized by a signal which is essentially transient in nature and the vowel-part corresponds to a relatively steady sound-pattern. Therefore, there must be more variations in the values of the parameters representing the consonantal part. Relatively the parameters for the vowel-part will not fluctuate much over groups of successive frames. Hence the segmentation of the utterance based on transient analysis would be a more appropriate method.

The method of Watari et al., [33], of detecting the transient parts of the speech signal has been adopted for our analysis. In its original form this method was used in a scheme for developing an algorithm for speaker independent recognition, in which only the transient parts of the speech are used as recognition elements. This method makes use of the spectral distribution of energy. The transient detection parameter is the deviation from the mean of parameter values in the adjacent frames. Two sets of speech waveforms and the transient plots for

a few utterances are shown in Figures 5.1 through 5.4. Peaks in the transient plots correspond to the maximum amount of variations in the parameter values and the valleys correspond to the minimum variations. Thus the first valley denotes the transition from silence to consonant region. The valley following a prominent peak corresponds to the beginning of the steady parts of the vowel segment, in the case of voiceless stop consonants. The segment lying between these two valley points is the consonantal part, and what follows is the vowel-part until the next prominent peak. The procedure for computation of the transient detection parameter is as follows.

1. Let $B_j(n)$ be the j th spectral value for a frame n. Then $\bar{B}(n)$ for that frame is computed as,

$$\bar{B}(n) = \mathrm{Log}\left(\sum_{j=1}^{J} B_j(n) + C\right),$$

where C is a constant introduced to avoid spurious transient detections in the silence portions. The value of C has been empirically determined as 100, for our purpose- Whereas $B_j(n)$ gives the spectral distribution of energy for a given frame, $\bar{B}(n)$ gives the overall energy for that frame.

2. The parameter value +or each spectral band is normalized with reference to the overall energy. Its inverse, $\hat{B}_j(n)$, is computed as follows.

For $\bar{B}_j(n) = \mathrm{Log}\,(B_j(n) + C)$,

$\hat{B}_j(n) = \bar{B}(n) - \bar{B}_j(n)$.

3. Taking M frames on either side of the frame under consideration the values of the normalized parameters are averaged. For frames at the extreme ends appropriate correction, such as considering only those frames that are actually in existence, has been incorporated. This yields an averaged parameter for each band.

$$\text{i.e. } BM_{j,n} = \frac{1}{2M+1} \sum_{m=-M}^{M} \hat{B}_j(m+n)$$

The value of M in our case has been empirically determined to be 6.

4. For each band, the absolute sum of the differences between the averaged parameter $BM_{j,n}$ and the normalized parameter is computed, for each of the 2M+1 frames. The sum of the resulting components of a j-tuple is the transient detection parameter D(n).

$$\text{i.e. } D(n) = \sum_{j=1}^{J} \sum_{m=-M}^{M} | \hat{B}_j(n+m) - BM_{j,n} |$$

The transient detection parameter will point to the fluctuations in the parameteric values only if there are variations,

i) with reference to an average computed over a few frames, and,

ii) by different amounts in the different frequency bands.

Any sporadic fluctuations get smoothed out in the processes of summing, averaging, thresholding and normalization. For the steady-state portions of the utterance the spectral distribution of energy does not essentially vary over adjacent frames. Hence

the value of the transient detection parameter will be low. On the other hand, it will be high for the transient portions because of the variations in the spectral distribution of energy over the adjacent frames. Thus a transient plot for a given utterance shows the peaks and valleys which help in the segmentation of the utterance into appropriate parts. This method yields one parameter per frame. which is convenient for the generation of the transient plots. However, since there is smearing over adjacent frames, the points on the plot are to be carefully interpreted. Further, for weak fricatives, aspirations, and similar sound segments the process of choosing the parameters empirically will have to be studied in detail. However, in our studies of Hindi stop consonants, the plots obtained for different utterances have been found to be consistent. The segmentation based on transient plot has been manual1y checked with the respective speech waveform for large sets of data, before adoption of this method for purpose of segmentation.

Figures 5.1 through 5.4 show transient plots for some of the utterances. To ensure that we do not lose any possible clues about- the consonant in the transition region, the end of the consonantal region is fixed at a point half—way between the last valley, prior to the low transition region, and the peak preceding that. With reference to these two sets of figures it may be noted that, although the energy content in the speech waveforms of the two sets differs considerably, there is consistency between the two sets of tranient plots.

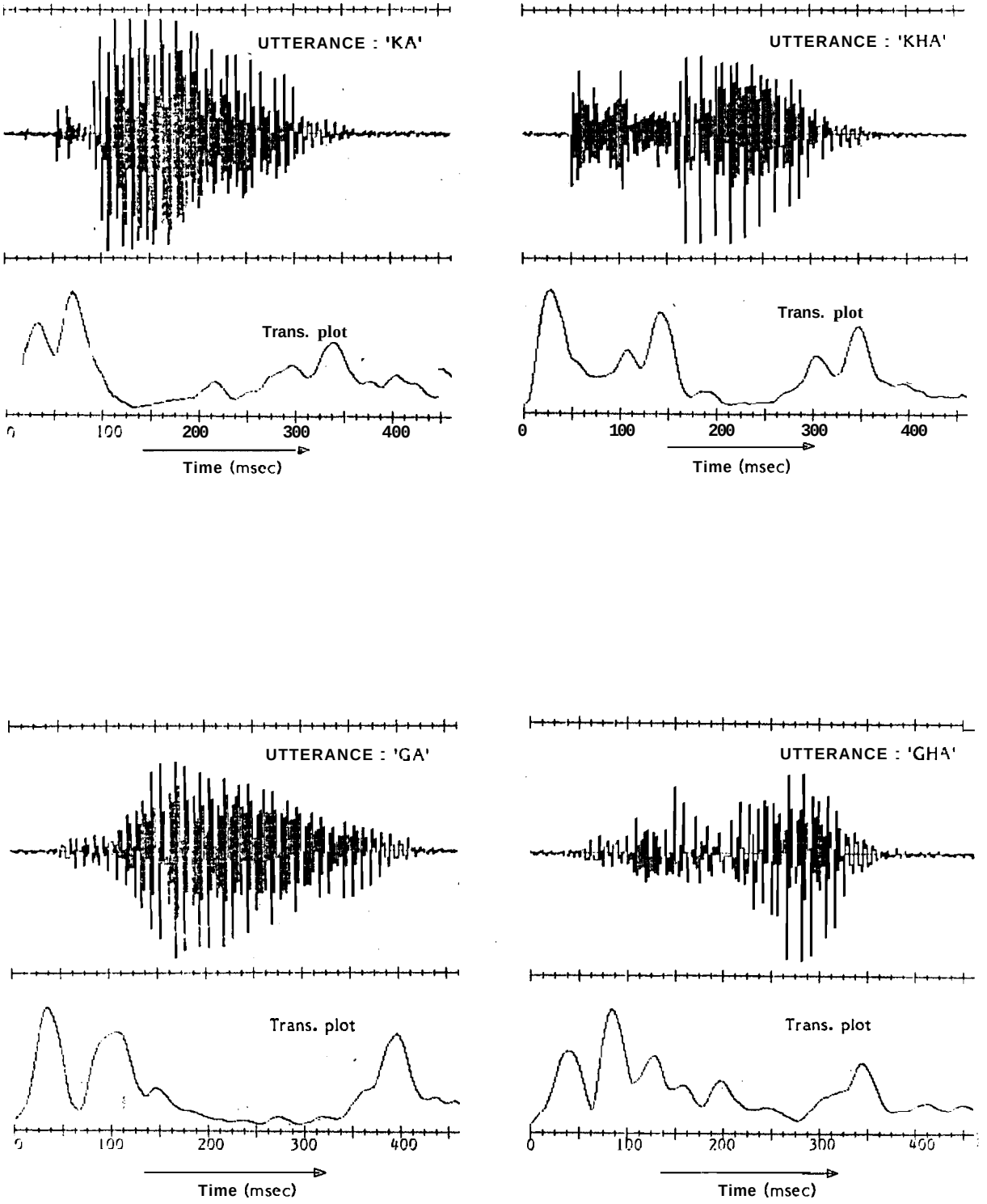
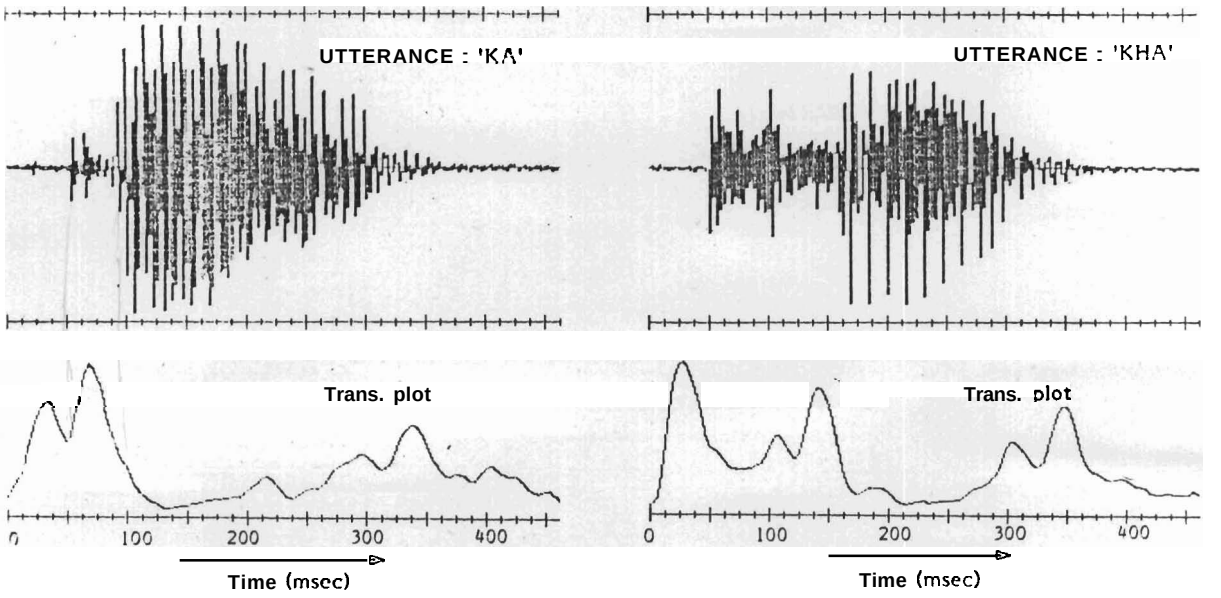


FIG 5.1 SPEECH-WAVEFORMS AND THE RESPECTIVE TRANSIENT-PLOTS FOR THC UTTERANCES (KA, KHA, GA, GHA) : Set 1



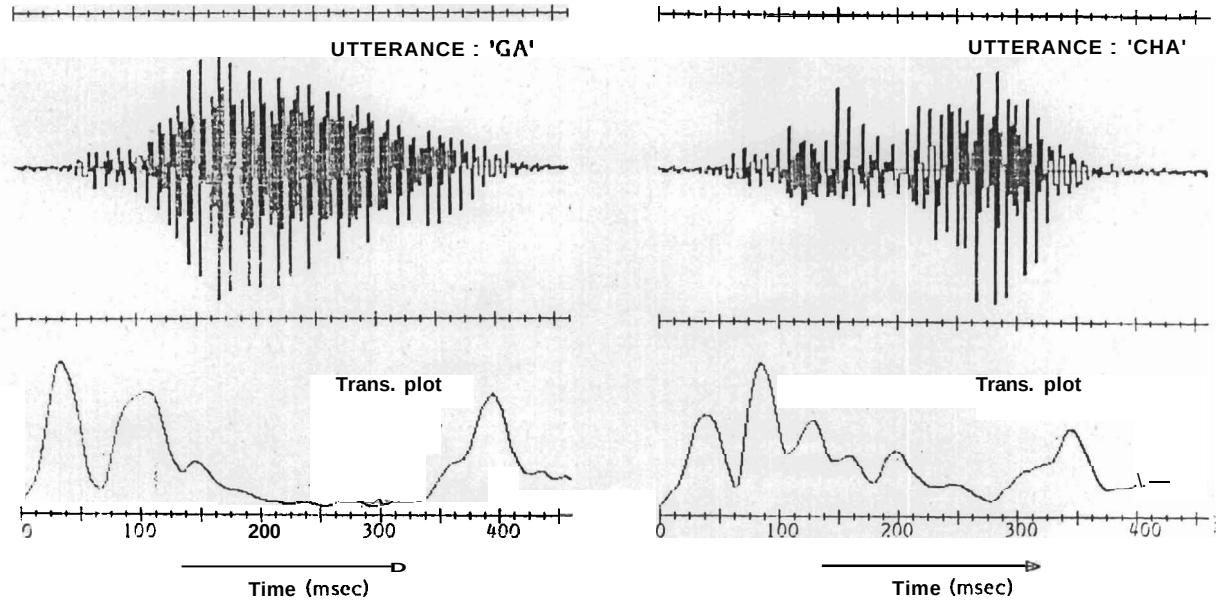


FIG 5.1 SPEECH-WAVEFORMS AND THE RESPECTIVE TRANSIENT-PLOTS FOR THE UTTERANCES (KA, KHA, GA, GHA) : Set 1

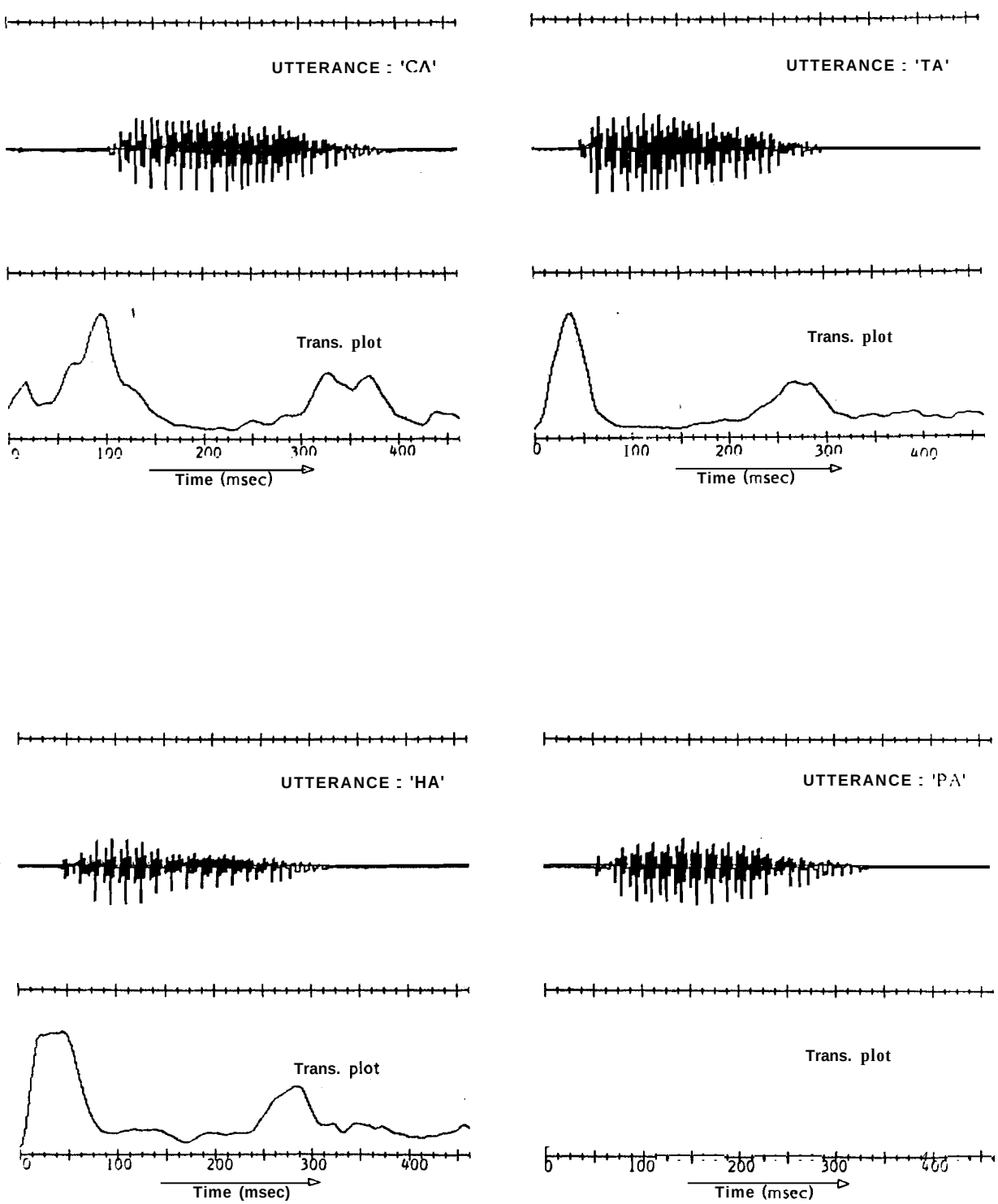


FIG 5.4 SPEECH-WAVEFORMS AND THE RESPECTIVE TRANSIENT-PLOTS FOR THE UTTERANCES (CA, TA, HA, PA) : Set 2

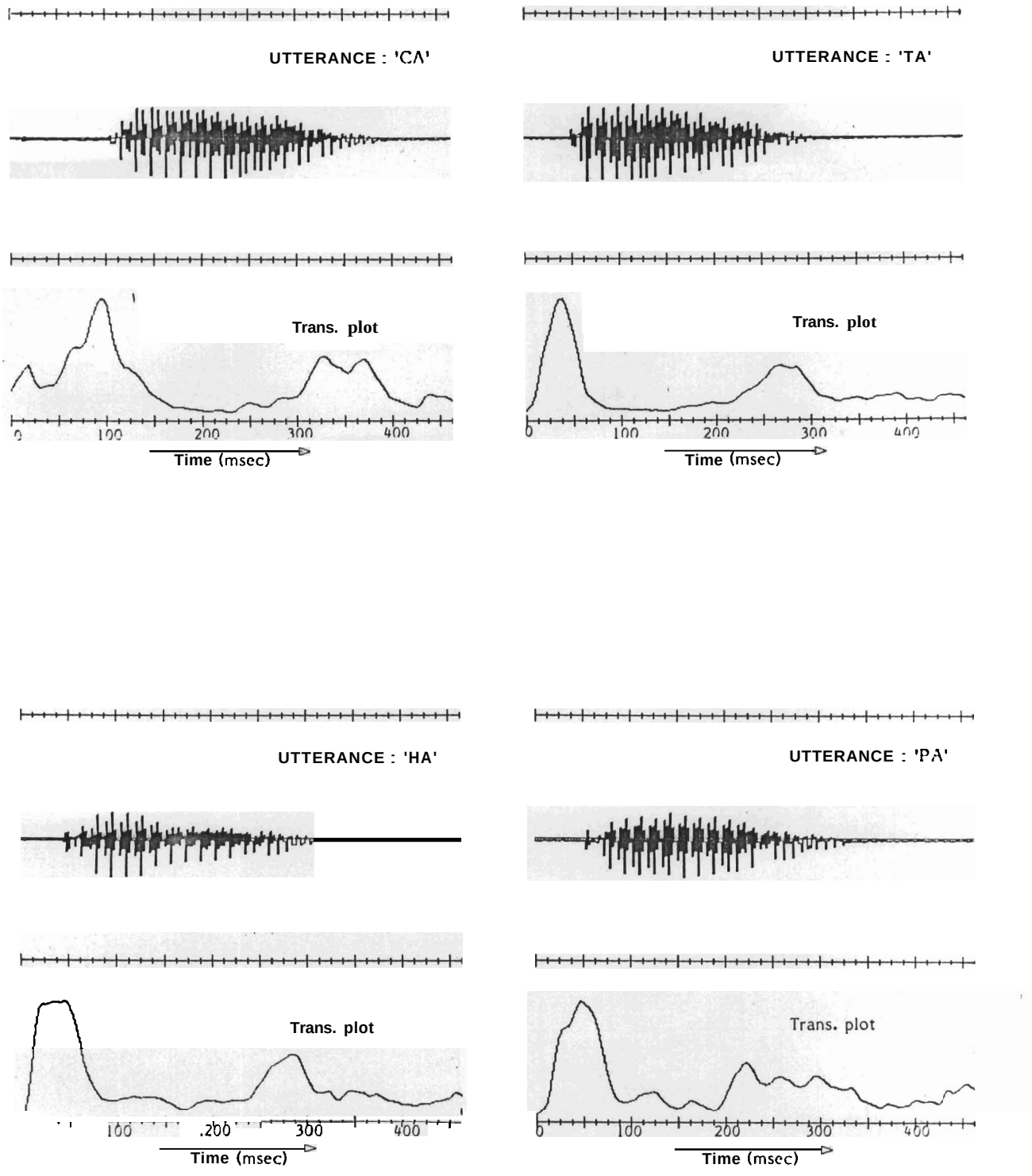


FIG 5.4 SPEECH-WAVEFORMS AND THE RESPECTIVE TRANSIENT-PLOTS FOR THE UTTERANCES (CA, TA, HA, PA) : Set 2

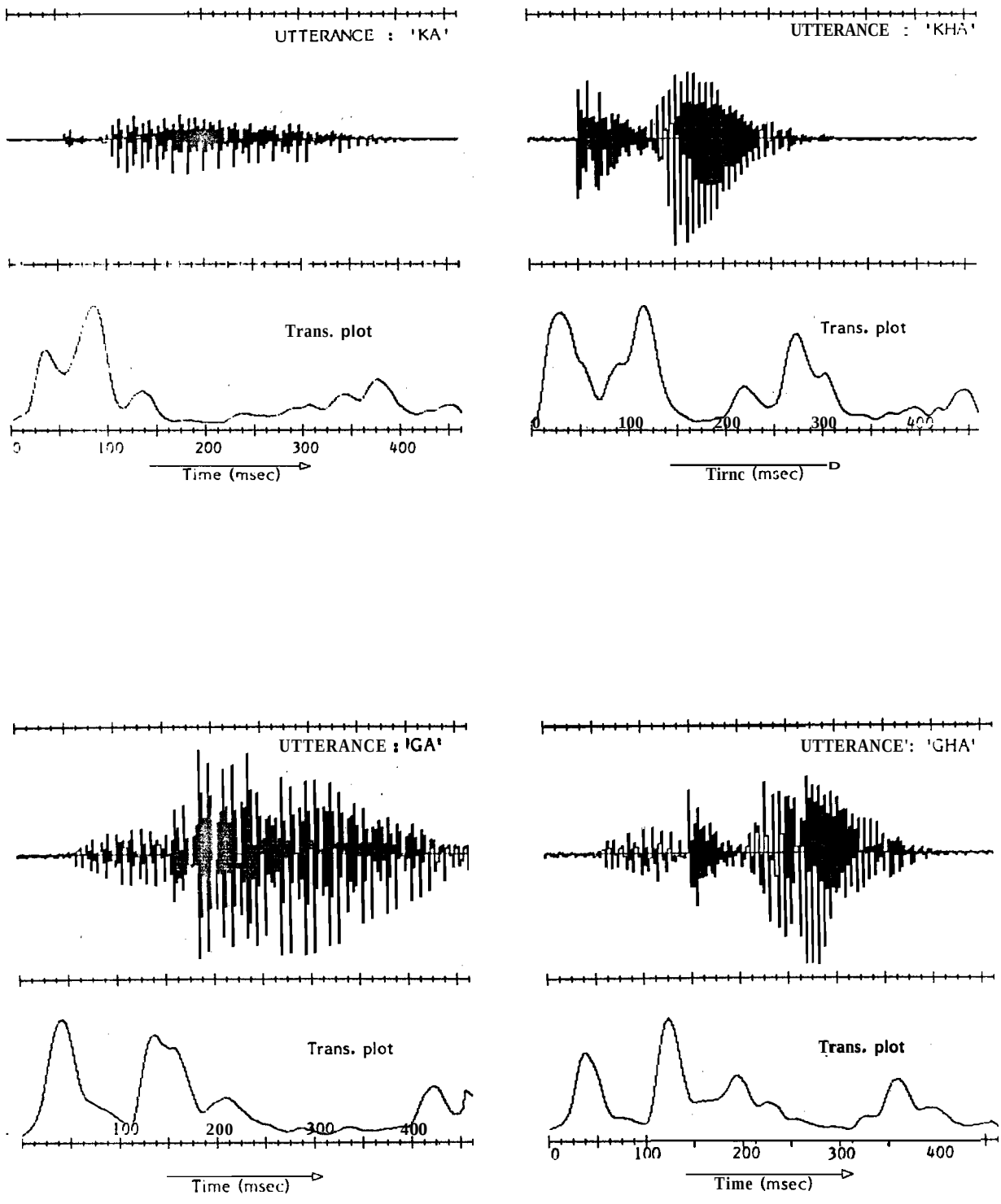


FIG 5.3 SPEECH-WAVEFOKMS AND THE RESPECTIVE TRANSIENT-PLOTS FOR THE UTTERANCES (KA, KHA, GA, GHA) : Set 2

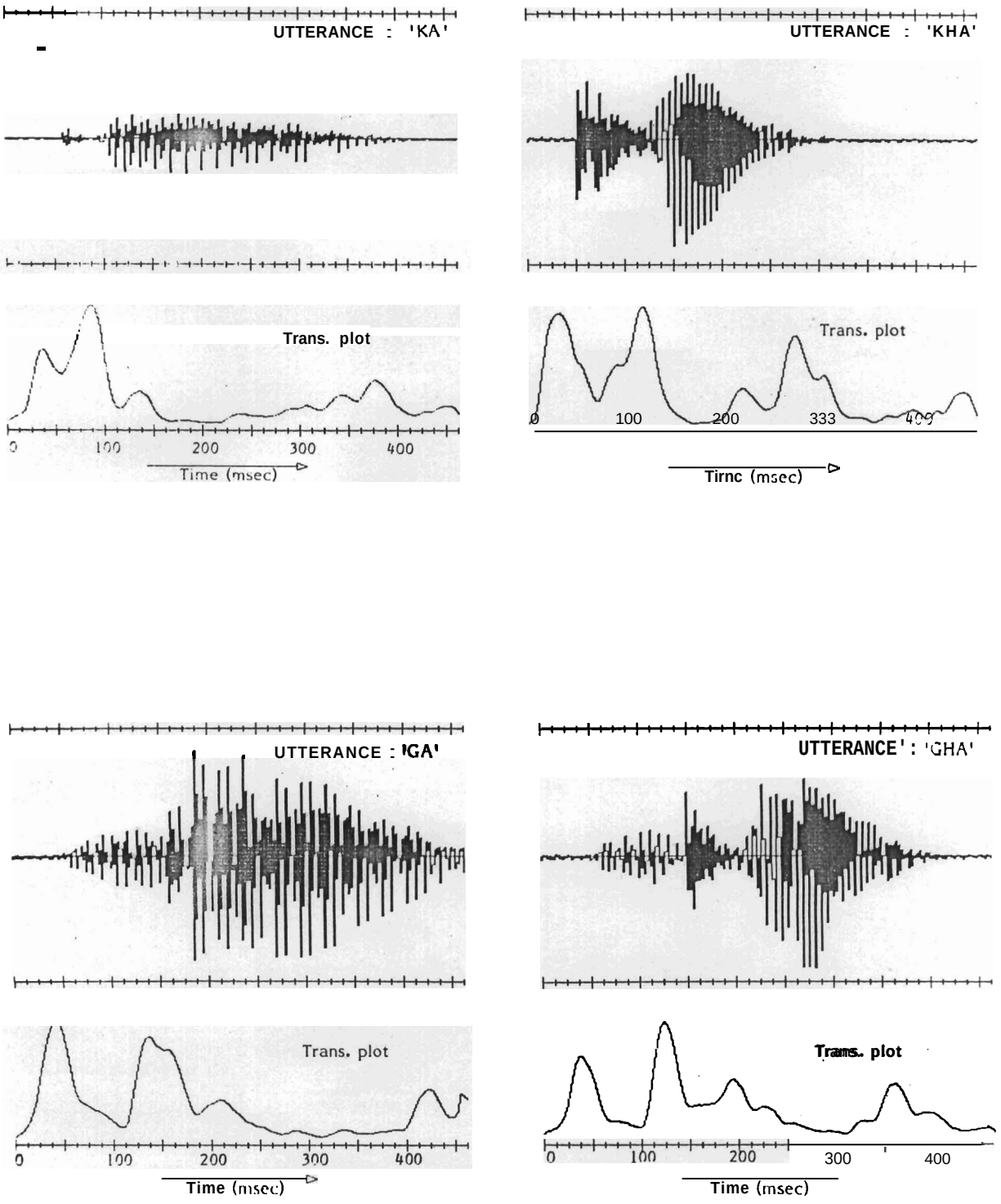
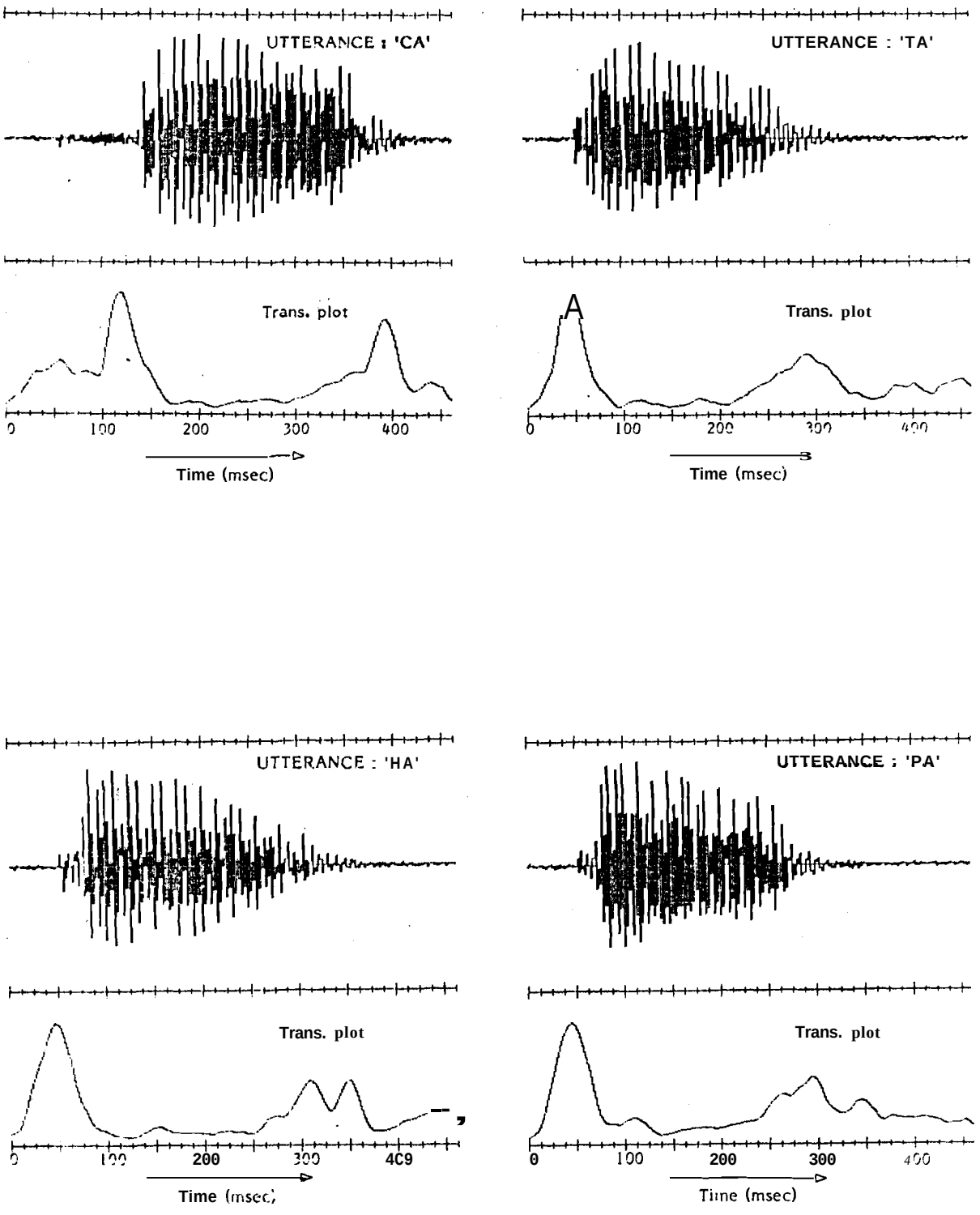


FIG 5.3 SPEECH-WAVEFORMS AND THE RESPECTIVE TRANSIENT-PLOTS FOR THE UTTERANCES (KA, KHA, GA, GHA) : Set 2



PIC 5.2 SPEECH-WAVEFORMS AND THE RESPECTIVE TRANSIENT-PLOTS FOR THE UTTERANCES (CA, TA, HA, PA) : Set 1

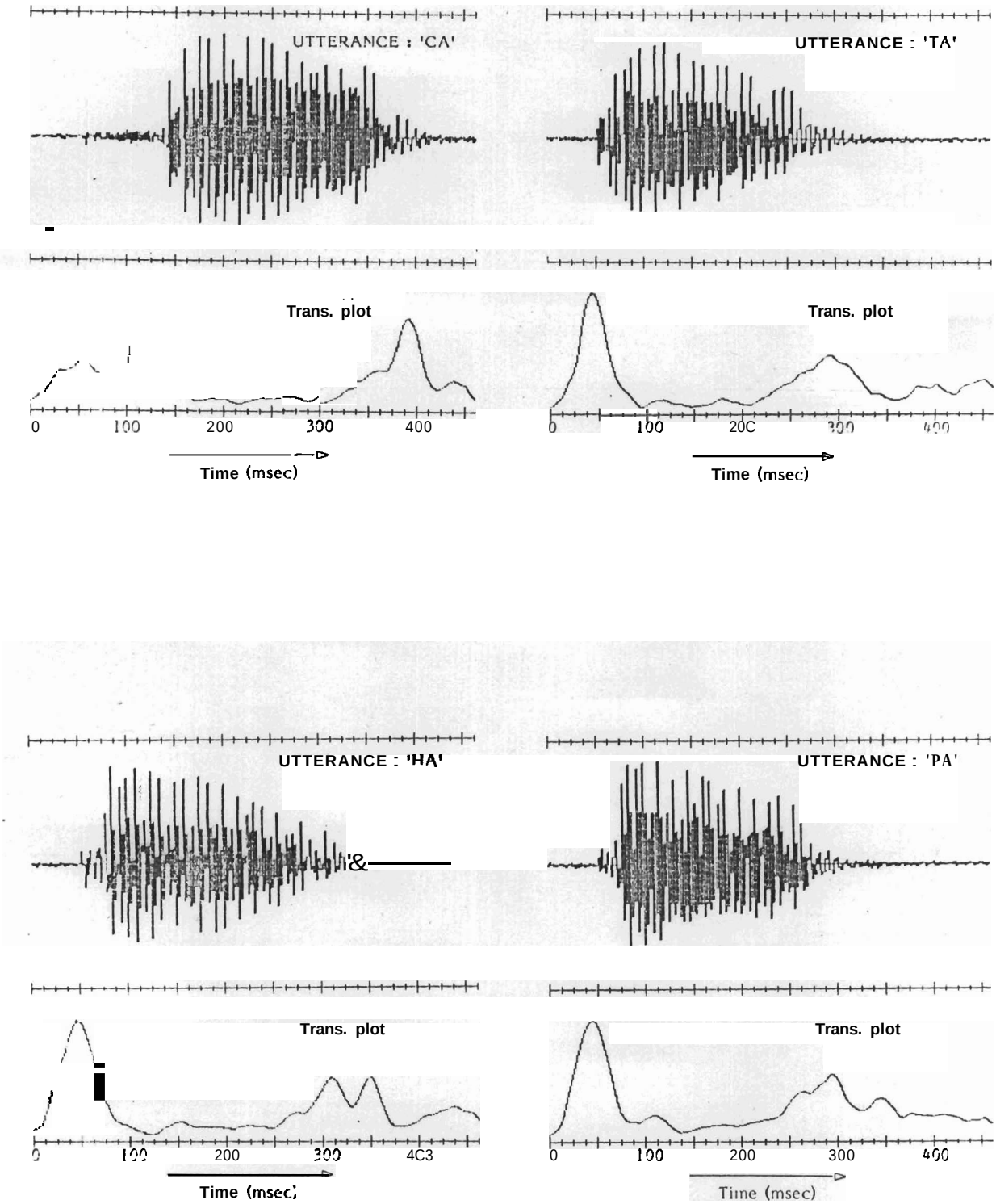


FIG 52 SPEECH-WAVEFORMS AND THE RESPECTIVE TRANSIENT-PLOTS 1'011 THE UTTERANCES (CA, TA, HA, PA) : Set I

Results reported in this and the following sections are based on experiments carried out on 8 data sets, for a given reference set. Table-5.3 shows the PIX-values for the two methods of the consonant/vowel segmentation. PIX-values are slightly higher for the energy-based method, presumably because there is less averaging of the parameters over adjacent frames. Our studies on a large number of data sets show that both the methods perform equally well. Our vocabulary consists of members which are primarily characterized by a signal which is largely transient in nature. Hence the method based on transient analysis is used for further studies.

Table-5.3 : Performance of the system for segmentation based on Energy and Transient detection

No. of samples/frame = 128
Overlapping = 0
Parameters = 8 log-melspectral values

PIX for part	Segmentation basis	
	Energy	Transient
Overall	43.70	43.70
Consonant	65.60	63.87
Vowel	34.64	32.96

5.4.2 Processing of the utterance by parts

The results shown in Table-5.2 are for appropriate representation of the consonantal and vowel-parts. However, the value of PIX computed is for the whole utterance. This is referred to as the overall PIX. The PIX values in Table-5.3, obtained as part of the studies on segmentation, throw light on another aspect of the vocabulary. That is concerning the parts of the utterance which contribute to the confusability of the vocabulary. The overall PIX shows that the performance of the system for handling the whole utterance is influenced by the low PIX for the vowel-part. (See Appendix-III for a sample output). The higher PIX obtained for the consonantal part is an indication of better discrimination amonq these parts. Thus the low overall PIX is due to the dominant influence of the vowel-part. This gives us the clue that if the influence of the dominating part is de-emphasized, or possibly eliminated, the performance of the system would then depend only on the parts which are discriminable. Thus the confusability of the vocabulary can be overcome.

That the dominant vowel-part influences the overall PIX is confirmed in the next experiment. The vocabulary for this experiment is a word set consisting of the same consonantal part but with different vowel-endings. The members of this set are {KA,KE,KU,KI,KO}. Table-5.4 shows that the overall PIX as well as the PIX for the vowel-part is high, thereby confirming the dominant influence of the vowel-part.

Table-5.4 : PIX for the word set having the same consonantal part but different vowel-endings.

Part of the utterance	PIX
Overall	96.45
Consonant	59.25
Vowel	98.03

That the PIX for the consonantal part is not very low for this word set leads us to conclude the following.

i) The consonantal segment is so short that it cannot be identified accurate1y.

ii) Any segmentation process can only identify different regions as predominantly vowel part or as predominantly consonantal part.

iii) In the method adopted for segmentation, the consonantal part includes the interface regions which will exhibit the characteristics of both the parts.

5.4.3 . Choice of appropriate parameters for the respective parts

The objective of this set of experiments is to determine suitable parameters and resolution required for the representation of different segments. This phase of experimentation consists of two parts. In the first part the word set consists of different consonants but with the same vowel-ending; 'A'. The second part deals with word sets of same consonant' but with different vowel-endings.

5.4.3.1 Experiments with same vowel-ending

The consonantal part is represented by 8 spectral parameters only. Two representations of the vowel part, one using 8 spectral parameters and the other using 8 LPC parameters, have been followed. The distance between two linear prediction spectra is computed using cepstral distance [51]. Table-5.5 shows the results of this experiment. In this Table, the second and the fourth column5 correspond to mixed representation: the consonantal part by spectral parameters and the vowel-part by LPC parameters.

Table-5.5 : PIX showing the effect of appropriate spectral and temporal resolutions

No. of samples/frame = 128

No. of parameters = 8

PIX for part	Overlappinq = 0		Overlapping = 64	
	Spec.	Spec.& LPC	Spec..	Spec.& LPC
Overall	43.70	47.51	47.91	54.27
Consonant	65.60	63.45	74.59	74.25
Vowel	34.64	42.00	29.97	41.94

These results confirm our earlier finding that the system performance for the consonantal part improves with increased temporal resolution due to the overlapping frames. Compared with the choice of spectral parameters for the entire utterance, the choice of spectral parameters for the consonantal part and LPC

for the vowel-part results in a slight improvement in the system performance. The reason is that the increase in the PIX for the vowel-part is reflected in the overall PIX. This implies that representation of the vowel-part by LPC parameters yields better results. The reason for this improvement is that LPC parameters better represent the formant-structure associated with vowel production. Even though the vowel-part is same, what is it that has led to the increase in the PIX, implying enhancement in the discriminability among the vowel-parts ? The answer lies in the coarticulation effect of the consonant on the vowel part. The coarticulation effect of the different consonants on the succeeding dominant vowel affects the characteristics of the vowel to some extent. This variation is represented better while using LPC for the vowel-part. This confirms once again our earlier observation: the nature of the speech signal is such that there are only regions which are predominantly of one or of the other type with no sharp boundaries. Here it may be worthwhile to recall Bhartrhari [5] that, "within the sound-unit the component features have no independent existence".

The results shown in Table-5.5 offer one other useful piece of information. Comparing the PIX for the consonantal and vowel-parts we observe the effect of varying the resolution in their representation. The increased temporal resolution has not very much affected the PIX for the vowel-part. Similarly the choice of different parameters for the vowel-part has not affected the PIX for the consonantal part, These two observations suggest that appropriate decisions for the respective segments can be taken without the one affecting the other.

5.4.3.2 Experiments with different vowel-endings

The ,justification for the use of **LPC** for the vowel-part is confirmed through a set of experiments on word sets with different vowel-endings. If one sample of the word set {KA, KE, KU, KI, KO} is the reference set, another sample set with the same consonant forms the test set. Similarly, word sets are formed for other consonants also. For representing the vowel part using **LPC** parameters, the mean PIX is 96.36 whereas it is 89.80 for representation with spectral parameters. Thus this finding also confirms that suitable parameters should be used for different segments.

5.4.4 Evaluation of segmentation based on different strategies

The objective of the next set of experiments is to study the effect of varying the width of the consonantal part. Simultaneously, the effect of the number of spectral parameters for representing the consonantal part is also studied. The data sets used for'these experiments are different from those used for the previous ones. Results of these experiments are shown in Table-5.6.

Table-5.6 : Effect of the width and spectral resolution of the.consonantal part

Number of spectral parameters	PIX for consonant width				
	knowledge-based	5	7	9	11
8	72.29	69.72	69.29	69.53	69.75
4	74.46	71.65	70.64	70.60	70.26
2	77.64	73.00	72.55	73.45	73.25

The results confirm that determination of the consonantal segment based on the analysis of the signal leads to better system performance. Also the system performance is better when fewer spectral parameters are used for representing the consonantal part. The unvoiced frames in these parts require lesser spectral resolution, since the relative energy-distribution in the low and high frequency bands is all that may be required to characterize these segments. (See a sample output in Appendix-111).

The results of these experiments show that for the handling of confusable vocabulary a totally different approach is needed. It consists of isolating the character-distinguishing consonantal part and handling it separate1y for recognition. before evolving such a radical scheme we show that tests on, more data have to be performed for resolving some issues like selection of a good reference set, inclusion of prosodic features, etc.

6. REFINEMENTS BASED ON EXPERIMENTS WITH LARGE DATA SETS

" Today-- stress is tomorrow-- good old days."

- FRANKFURTER ALLGEMEINE ZEITUNG
Blick durch die Wirtschaft

6.1 Introduction

That the system performance can be enhanced by methods of representation and processing appropriate to the different parts of the signal, was seen in the last chapter. The PIX-values were computed based on 8 sets of test data for a particular reference set. In this chapter, we discuss the choice of the reference set, introduction of additional knowledge in the form of prosodic features, the influence of the recording environment, the effect of different vowel-endings and the influence of each of these on the performance of the system.

Processing confined to the consonantal part has been justified as the method for handling confusable vocabulary. This approach is found to hold good for different vowel-endings also. Addition of prosodic features to the knowledge-base is shown to help in pruning the search among alternatives and also to result in the enhancement of system performance to some extent. Further improvement in the system performance, due to the refinements discussed in this chapter, is summarized in Table-6.1.

Table-6.1 : System performance for further refinements discussed in this chapter

Vocabulary : { KA,CA,TA,HA,PA }

Expt. No.	Method	PIX
1.	After choosing a good reference	76.2
2.	Warping within the consonantal part	80.8

6.2 Choice of reference set

Using 9 sets of data, with each word set as the reference and the remaining 8 sets as the test data, exhaustive testing was carried out. Table-6.2 shows the PIX-values for different reference sets. They are the mean-values of PIX obtained for each test set, for the overall and for the consonantal part. The third row from the bottom shows the average of the PIX-values due to all the tests. The last two rows show the PIX of the system for two different good reference sets chosen as follows.

A good reference set should include members which are representatives of the respective speech-sounds. In our case this is determined by studying how each utterance performs as a member of the reference set. All speech-sounds may appear to be equally good to a human listener. But the same cannot be said of a machine. Because of certain characteristics of the recognition system, determined by the methods adopted for representing and processing the signal, certain data sets are better choices as

Table–6.2 : PIX for different reference sets

Frame size : 128 samples

Overlap : 64 samples

No. of Parameters : 8 log-mel spectral

Reference Set	PIX for part	
	Overall	Consonant
1	51.12	74.17
2	45.50	74.77
3	36.40	60.63
4	44.87	69.89
5	48.60	76·55
6	46.85	76.70
7	53.79	75.58
8	50.12	66.97
9	56.43	73.77
ALL	48.16	72.00
GOOD 1	48.33	73.30
GOOD 2	48.85	76·21

reference than others. From the exhaustive tests conducted we have information on how each utterance has performed as a member of a reference or of a test set. These are tabulated as shown in Tables 6.3 and 6.4.

Table–6.3 shows the number of times the different instances of an utterance have been correctly identified for a given reference word. The identity of the reference word is given in this Table by the number of the set to which it belongs. For

Table-6.3: Number of successful recognitions with the respective members as the reference data

Member \ Set No.	1	2	3	4	5	6	7	8	9
KA	7	7	2	8	7	8	7	4	7
CA	8	8	8	8	8	8	8	8	8
TA	6	6	0	8	7	5	5	6	5
HA	5	3	1	4	3	5	4	0	6
PA	8	5	8	0	8	7	7	6	5

instance, KA1 is the first sample of the utterance A The entry of 7 in the first row and the first column is to be interpreted as follows. With KA1 as a member of the reference set, 7 out of the remaining 8 samples (KA2 through KA9) were correctly identified. One test utterance was wrongly identified as one among the words CA1 through PA1. When KA3 was the reference, only 2 out of 8 samples were correctly identified. Thus we may say that KA1 is a better reference than KA3.

Before we arrive at this conclusion there is a point worth noting. When a test utterance is identified as closer to one of the different reference words, such a decision is arrived at in a relative sense. It may be due to the fact that the test and the chosen reference word are similar utterances, which is what we expect the system to recognize. But it can also be due to the +act that the system has been forced to identify the test with the chosen reference, since the other members in the reference set are very dissimilar to the test utterance. That is, in

reality, these two words are not samples of the same utterance. This implies that a successful recognition has information about other members of the reference set in addition to the chosen one. When a priori knowledge about the distances between various members of the vocabulary is available, one can set a threshold and overcome such problems. Determination of a threshold depends on various factors such as the type of vocabulary, the recording environment, and other decisions taken in the various stages of processing,. It is a crucial decision since its effect will be cumulative.

We have adopted an approach which uses additional information on how an utterance has performed as a member of a test set. It is based on the hypothesis that if an utterance performs well as a reference it should also be identifiable as a good test input. This need not be true always. Further, recognition results are not absolute but relative. Therefore information about a sample being a good test input is used as additional information. Table-6.4 shows the number of times a specific test utterance has been recognized correctly with different samples of the word as reference data.

On inspection of Tables, 6.3 and 6.4, one can see that different choices of reference set are possible. Two of these considered for subsequent experimentation are:

{ KA6, CA1, TA1, HA6, PA5 } and

{ KA4, CA4, TA5, HA1, PA1 }

Table-6.4 : Number of succeesful recognitions with the respective members as the test data

Member \ Set No.	1	2	3	4	5	6	7	8	9
KA	6	6	2	6	7	8	6	8	7
CA	8	8	8	8	8	8	8	8	8
TA	6	7	6	1	6	7	5	4	6
HA	5	4	1	1	5	6	6	0	4
PA	4	6	5	6	7	7	6	7	6

With each of these as the reference set and the rest of the data grouped into 8 test sets, PIX-values are computed. Mean PIX for these two cases are the entries in the last two rows of Table-6.2. They show that with the choice of a good referece set the average behaviour of the system can be estimated well with fewer sets.

6.3 Considerations in the determination of warping path

6.3.1 Influence of the dominating vowel-part

In this section we study the influence of the dominant vowel-part in the determination of warping path. The warping path is determined using one set of parameters for the whole utterance. The distances along this path are computed separately for the consonantal and vowel-parts. In this connection the following questions arise.

1. Does the warping path go through the distinguishing regions of interest in the consonantal part ?
2. Does the warping path go through the point corresponding to the consonant-vowel interfaces of the test and the reference data ? Or does it get so influenced by the dominant vowel-part that it misses the anchor point, as in the case of warping path # 1, shown in Fig.6.1 ?

When it is difficult to identify categorically the regions of distinguishing characteristics within the consonantal part. the next best thing would be to ensure that the warping path does go through the anchor region corresponding to the consonant vowel interfaces. An illustration of this is the warping path # 2, in Fig. 6.1. Based on the hypothesis that the distinguishing features in that region would have influenced the decisions in fixing the warping path, one can hope that the warping path goes through the regions of interest in the consonantal segment. This is the basis for our studies on confining the process of warping and matching to within the consonantal region. The DTW algorithm fixes the endpoint for the warping path as (N,M), which implies that for warping within the consonantal region the point (N,M) will coincide with the anchor point.

Since warping is confined to the short consonantal region there will be problems due to the global constraints of 1:2 or 2:1, imposed by the DTW algorithm. Hence suitable modifications in the warping algorithm have been made. Details of these modifications, which are based on the speech production viewpoint, are discussed in the following section.

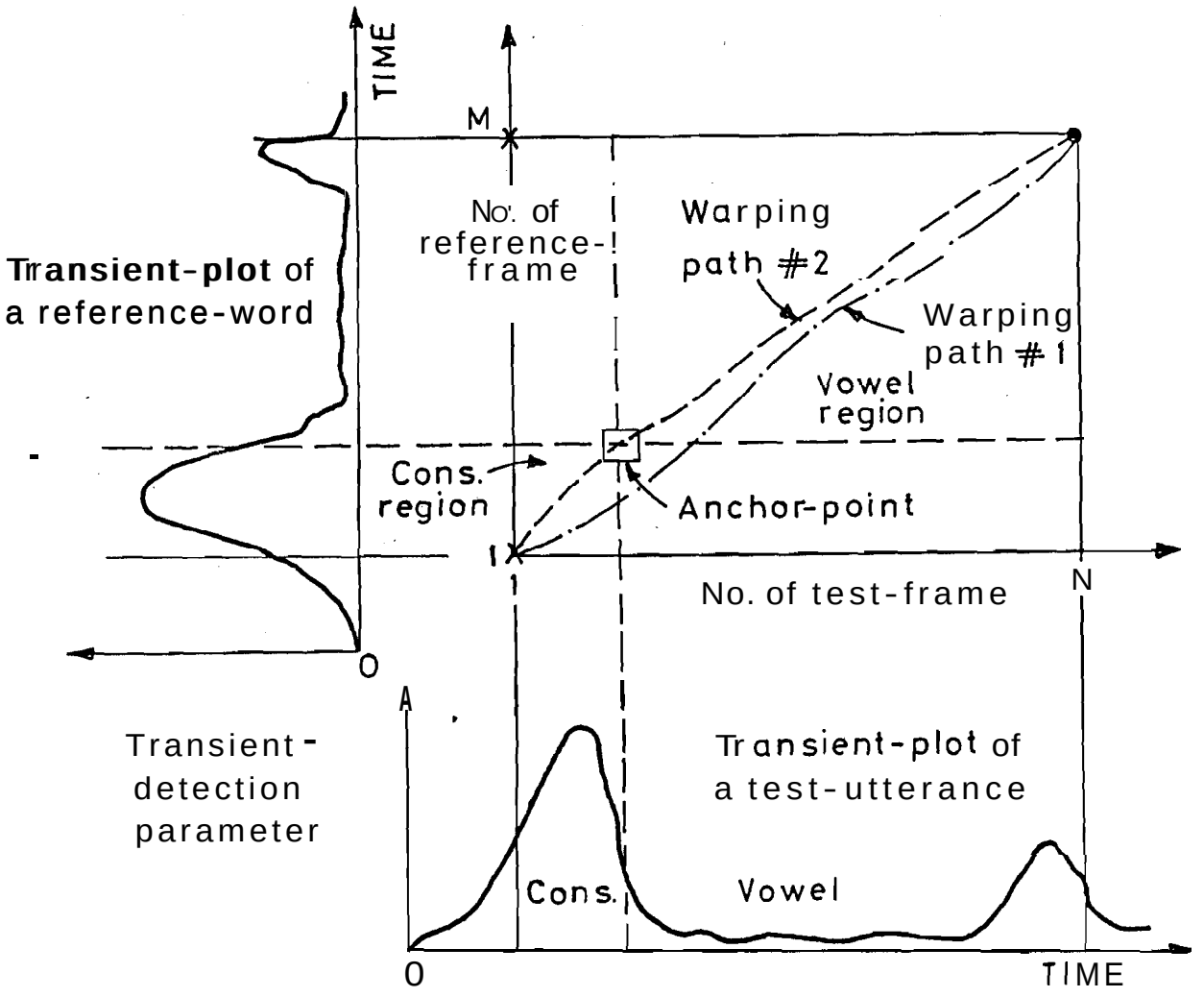


Fig.6·1 Warping paths :

#1 - Influenced by the dominant vowel-part

#2 - Going through the anchor-point.

6.3.2 Addition of prosodic features to the knowledge-base

Different consonants have different places of articulation. The fist one 'KA' is a velar consonant and the last one 'PA' is a bilabial consonant. During the production of these consonants, after the release of air from the site of constriction the articulators move towards the steady-state positions which correspond to the following vowel-part. Since the site of constriction is different for the different consonants, the time taken for the movement of the articulators will vary from consonant to consonant. This duration must be greater for constrictions at the back of the vocal tract than for those at the front. This is roughly indicated by the number of frames for the consonantal part. It is only a rough estimate since features such as frication may show the consonantal part more clearly than others. To a lesser degree this duration depends on the vowel. Table-6.5 shows the average length of frames for the consonantal part, for consonants with different vowel-endings.

The duration of the consonant is a prosodic feature and is added to the knowledge-base. Since the duration of the consonantal part cannot be voluntarily extended we hypothesize that any two consonants, which do not come within the 1:2 or 2:1 constraint, may be classified as belonging to different groups.

For instance, with reference to Table-6.5, the test utterance CA will be compared only with the reference words of KA and CA. Thus the choice of TA, HA and PA are eliminated straightaway. The spectral characteristics of the test utterance will decide whether it is KA or CA. In this manner the prosodic

Table-6.5 : Average length of consonantal segment for utterances with different vowel-endings

Frame size = 128 samples/frame

Overlapping = 64 samples

Vowel-ending	Consonant (Number of frames) K	C	T	H	P
A	10	14	4	5	5
E	14	21	4	7	6
U	13	17	5	7	8
I	11	14	3	4	3
O	10	15	3	5	4

feature is used to prune the search among the possible alternatives. In the signal-dependent approach, the number of consonantal frames in the test utterance is compared with those in each of the reference words and appropriate action is taken. Based on this refinement, the matching algorithm is modified such that whenever this constraint (1:2 or 2:1) is not met the corresponding entry in the distance matrix is made an unusually large number. For the purpose of finding out which of the two constraints is not met two different (large) numbers are entered. For the two chosen reference sets, the system performance is shown in Table-6.5.

Table-6.5 : System performance for warping and matching within the consonantal part

Frame size = 128 samples/frame

Reference set	Overlap = 64 samples		Overlap = 96 samples
	Overall	Consonant	Consonant
1	48.33	73.30	80.77
2	48.85	76.21	80.86

6.4 Effect of condition of recording on the system performance

The speech data for the experiments described were recorded in a noise-free room on a hand-held cassette tape recorder, using the built-in condenser microphone. In a similar environment, using a sophisticated tape-deck and a close-talking noise-cancelling microphone, another set of speech data was recorded. In both the cases, the speaker and the analog front-end for digitizing the speech were the same. The vocabulary of the set of consonants with the vowel-ending 'A' was also same. Thus only the recording environment was different. Table-6.7 compares the PIX of the system for the two sets of speech material.

The figures in Table-6.7 show that the system performance can be improved if proper care is taken while recording- However, the experiment with the new set of data also confirms the general trend of improvement in system performance with the refinements in the recognition methodology outlined earlier.

Table-6.7 : System performance for speech material recorded in different environments

Frame size = 128 samples/frame

Data set	Overlap = 64 samples		Overlap = 96 samples
	Overall	Consonant	Consonant
Old	48.85	76.21	80.86
New	50.95	80.59	82.78

With the reference chosen from the old set and the test data drawn from the new set, the performance of the system was found to reduce drastically. The resulting PIX of 17.36 for overall and 44.78 for consonant parts show that, without the use of more sophisticated normalization techniques, the speech material cannot be drawn from different sources. As our main focus is on arriving at a scheme for recognition of the confusable vocabulary, these points have been noted but not pursued further. A practical recognition system must take note of these problems and adopt suitable corrective measures. This will enhance the performance of the system still further.

6.5 Performance for vocabulary having other vowel-endings

The stage is now set to carry on studies with other word sets, consisting of the same set of consonants studied so far but with vowel-endings other than 'A'. 'A', 'E' and 'U' are the three basic vowels. There are short and long versions of these. As these two versions could be distinguished by prosodic features such as duration of the utterance, the short versions alone are

considered for our study. In addition, two of the six diphthongs described in Chapter 1 are considered. They are 'I' and 'O'. Thus there are five consonants K,C,T,H and P, with five vowel-endings A,E,U,I, and 0. The five sets of vocabulary considered for our experimentation are :

{KA, CA, TA, HA, PA3 {KE, CE, TE, HE, PE}

{KU, CU, TU, HU, PU} CKI, CI, TI, HI, PI3

CKO, CO, TO, HO, PO}

The generally uniform characteristics of these five sets of consonants can be. inferred from Table-6.5, which shows the average length of the consonantal segments for the respective utterances. All the refinements carried out on the system for the recognition of the first set of consonants have been performed while dealing with the other four sets also. Table-6.8 shows the performance of the system for these vocabularies.

From Table-6.8, different vowels are seen to influence the overall recognition to different degrees. The general trend of improvement in system performance with the refinements discussed thus far, is confirmed for the vowel-endings 'E' and 'U' also. However, the results for diphthong-endings show a different trend. As the results indicate, the warping path through the consonantal part is not influenced by what follows in the case of diphthongs. This difference in performance for the endings of vowel and diphthong may be attributed to the different dynamic characteristics of the articulators, during the production of the respective sounds.

Table–6.8 : Performance Index for the sets of consonants with five different vowel–endings

Frame size = 128 samples/frame

Vowel ending	Overlap = 64 samples		Overlap = 96 samples
	Overall	Consonant	Consonant
A	48.85	76.21	80.86
E	65.83	85.52	90.02
U	67.47	89.76	90.09
I	70.09	87.23	85.23
O	65.03	81.56	80.74
All	60.90	83.53	85.08
Standard Deviation	12.75	8.33	7.46

The entries in the last two rows of Table–6.8 correspond to the values of mean and standard deviation of PIX, for the entire data. Reducing values of the standard deviations corresponding to the increasing values of the mean–PIX show that the the results obtained have generally been consistent.

Now that the results have been computed for large data it may be appropriate to study the corresponding Percentage Recognition Score (PRECS). Although PRECS is not a reliable measure for small sets of data, Table–6.9 lists the values of PRECS separately for the five sets of consonants also.

Table-6.9 : Percentage Recognition Score for the sets of consonants referred to in Table-6.8

Vowel ending	Overall	Consonant	Consonant (anchored)
A	56.50	80.00	84.70
E	62.20	88.88	93.33
U	67.50	92.50	97.50
I	74.30	88.50	88.50
O	77.10	80.00	82.80
All	65.00	85.00	88.75

The trend of improvement in system performance, as shown by the PIX values in Table-6.8, is generally confirmed by the PRECS-values in Table-6.9. Even though the series of refinements have resulted in the increase of PIX from 25.8 to 80.86, the figures of performance indicate that there is scope for further improvement. Further studies will have to be conducted to see whether the system performance can be improved upon by innovations in the approach at signal levels, or whether it calls for knowledge from higher levels. If it is the latter, it would imply that a certain amount of fuzziness at lower levels can be tolerated, with the hope of resolving the issues at higher levels of processing. Instead of going into these issues we proceed to develop a scheme for recognition of consonants, based on the studies conducted thus far.

7. A SCHEME FOR RECOGNITION OF HINDI STOP CONSONANTS

"It is always easier to travel than to stop.
As long as one travels towards a promised land
the dream is there; to stop means to face reality."

- LOUIS L'AMOUR
The Ferguson Hifle

7.1 Introduction

We have seen that appropriate processing on specific parts of the signal having high discriminability holds the key for recognition of confusable vocabulary. We have arrived at this conclusion on the basis of our experiments with a vocabulary of Hindi stop consonants belonging to the unvoiced and unaspirated category. Since other categories of consonants are characterized by different excitation features some modifications in the method of recognizing them may be needed. At this point of development we have a scheme for recognition of consonants belonging to the unvoiced and unaspirated category, if the vowel-ending is known. Determination of the vowel-ending and the category/manner of production is necessary before the overall scheme for recognition of consonants can be finalised.

In continuous speech, if the vowels can be spotted, a method for recognition of consonants is all that is required to recognize the speech input. From this viewpoint the results of this work assume great significance. The features due to the steady-state patterns of a vowel facilitate its easy recognition.

The phase of consonant recognition may be considered as consisting of two parts. The first one is concerned with the determination of the manner of production and the second with the place of articulation, in that order.

In this chapter we study the issues related to these phases of recognition. Based on these studies we propose an overall scheme for recognition of Hindi stop consonants and project that improvement to an order of about 80% recognition accuracy at the phonetic level is possible. It will lead to the development of recognition systems with a near-100% successful performance, if knowledge at higher levels of syntax and semantics is also invoked in the implementation of the recognition methodology.

7.2 Determination of the vowel-ending

The following discussion shows how the vowel-ending can be determined for a given consonant belonging to the unvoiced and unaspirated category.

The first phase of experimentation deals with the recognition of vowels. The word sets consist of the vowels {A,E,U,I,O}. The mean-PIX for this vocabulary is 99.98 and the PRECS is 100%. These figures indicate that vowels can be recognized accurately.

The next set of experiments evaluates the performance of the system when these vowels are uttered in the context of the different consonants. Since the vowel-part has been found to be the dominant one in an utterance, we verified the possibility of using a vowel set as the reference with which the incoming test

utterance is compared. The set of vowels {A,E,U,I,O} is the reference set. The test sets are { KA,KE,KU,KI,KO }, { CA,CE,CU,CI,CO 3, etc. PIX is computed for each utterance, namely, a specific consonant with a specific vowel-endinq. The results are shown in Table-7.1. The mean-PIX is 86.43 and PRECS is 90.73% .

Table-7.1 : PIX for different stop consonants with a vowel set as the reference

Reference : { A,E,U,I,O }

Consonant \ Vowel ending	A	E	U	I	O
K	90.39	85.68	83.68	97.14	83.87
C	74.29	50.14	81.25	100.00	74.42
T	88.18	77.04	90.60	100.00	92.43
H	88.25	71.89	92.33	99.25	96.64
P	94.96	81.75	84.21	94.04	88.32

Mean-PIX : 86.43

PRECS : 90.73%

The value of PRECS was computed. For all the vowel-endings except 'E', PRECS is in the range 92%-100%. For 'E' as the vowel-ending PRECS is 72%. This brought down the overall PRECS to 90.73%. However, the values of PIX ranging from about 50 to 100 show that this is not a very reliable method of determination of vowel-ending. Hence for the next set of experiments the reference set includes the set of vowels uttered in the context of the specific consonants.

The data sets for these experiments consist of the consonants with different vowel-endings. separate experiments are carried out for each of the five consonants such that the reference and the test data will have the same consonant but different vowel-endings. For instance, if the reference set is CKA, KE, KU, KI, KO}, the test set also consists of utterances with these vowel-endings and in the context of the consonant 'K'. Thus data sets were formed for other consonants also. For all of these data sets the mean-PIX is 96.36 and PRECS is 100% .

The results of all these experiments demonstrate the easy recognizability of the vowel-part. Once the vowel-ending is known, the place of articulation can be determined using a reference set appropriate to the category/manner of production. We have already seen the recognition methodology to be adopted for the stop consonants belonging to the unvoiced and unaspirated category.

7.3 Determination of the manner of production

Given an utterance with a known vowel-ending, we discuss the method of determining the manner of production. We consider the case of the vowel-ending 'A', and proceed to discuss the experiments conducted to determine the manner.

For this set of experiments, the data sets consist of stop consonants due to different manners of production. The results are shown in Table-7.2. The PIX values are in the range of 90's, and the PRECS is 100%. (See Appendix-IV for a sample output).

This confirms our earlier observation that the consonants due to the different 'manners" have distinctive features and hence determination of the manner of production will not be as difficult as that of the place of articulation.

Table-7.2 : PIX for the vocabulary of stop consonants due to different manners of production

Ref./Test Word set	PIX
{ KA, KHA, GA, GHA }	97.85
{ CA, CHA, JA, JHA }	91.25
{ TA, THA, DA, DHA }	92.50
{ HA, HHA, WA, WHA }	96.63
{ PA, PHA, BA, BHA }	95.33

Since the 'place' is a weak feature and the 'manner' is a dominant one, it appears that instead of having multiple word sets (one for each consonant) as reference sets only one set may suffice. The next set of experiments verifies this point. The reference set is { KA,KHA,GA,GHA } and the test data sets consist of { KA,KHA,GA,GHA }, { CA,CHA,JA,JHA }, etc. (See Appendix-V for a sample output.) The PIX of the system is computed individually for each of the consonants in these sets, and is shown in Table-7.3. The mean-PIX is 88.99 and PRECS is 98.21%.

Table-7.3 : PIX for stop consonants with a single reference-set for determination of manner of production

Reference : { KA, KHA, GA, GHA }

Consonant (Test Data)	Unvoiced-Unaspirated	Unvoiced-Aspirated	Voiced-Unaspirated	Voiced-Aspirated
K	100.00	100.00	92.66	85.55
C	100.00	69.00	83.11	82.33
T	100.00	88.78	93.00	76.67
H	100.00	89.56	85.11	77.67
P	100.00	90.00	96.34	70.00

Mean-PIX : 88.99

PRECS : 98.21

The variation of PIX over a wide range from 69 to 100 reduces the mean-PIX to 88.99. The wide variation of PIX-values suggest that the choice of a single reference for the determination of manner of production needs to be carefully studied. However, if this test reveals the manner, further testing for place determination can be taken up. The experiments discussed in this section demonstrate that the determination of the 'manner' will not pose much of a problem, although difficulties may arise in respect of some consonants.

7.4 An overall scheme of recognition

The overall scheme of recognition consists of the following steps. First comes the determination of the vowel-ending. It has not been considered in detail here, as it has been shown to be a relatively simple task. In the second step, the unknown utterance

is compared with five sets of reference words, one set for each of the five stop consonants. Each set consists of four utterances belonging to the four different categories. For instance, if the unknown utterance has the vowel-ending 'A", the five reference sets are:

{ KA, KHA, GA, GHA } { CA, CHA, JA, JHA 3

{ TA, THA, DA, DHA } { HA, HHA, DA, DHA }

{ PA, PHA, BA, BHA }

For each of these five comparisons the respective reference word which is at the least distance from the unknown utterance is identified. Applying the majority rule of 3-out-of-5, the 'manner" of the unknown utterance is determined. With the vowel-ending and the 'manner' established, the third step consists of determining the place of articulation. Our work has, been mainly concerned with the details of this third step for the category of the unvoiced and unaspirated consonants. For the consonants belonging to other categories, further refinements may be called for.

With *300* sets of test-utterances for each of the five different vowel-endings, the manner of production as outlined above. was determined. The overall PRECS is around *80%.* Table-7.4 shows the number of successful determinations for different vowel-endings and places of articulation.

It may be seen that PRECS for different vowel-endings is generally in the range *80%-90%* except for 'U', for which it is only 66%. This implies that the details of this scheme need further refinements, such as an additional support based on studies from other aspects of the input signal. It may be noted

Table-7.4 : Number of successful cases of determination of 'manner'

(each entry is out of **60**)

'Place'	Vowel ending	A	E	U	I	O
Velar	(K)	**54**	**50**	**38**	**52**	**52**
Palatal	(C)	**54**	**51**	**38**	**47**	**37**
Retroflex	(T)	52	**51**	**36**	**55**	**57**
Dental	(H)	**56**	**48**	50	**50**	**53**
Bilabial	(P)	55	52	**37**	**44**	**44**
PRECS :		90%	**84%**	**66%**	**82%**	**81%**

that the different excitation-features associated with different categories of consonants have not been explicitly utilized in the determination of the manner of an unknown utterance. The reason is that simultaneous determination of the 'manner' and the 'place' of a given utterance is not possible in a hierarchical approach. Hence the more dominant feature of 'manner' is sought to be established first, by the method described above. When the 'manner' is known, the place of articulation will be established, using tests which make use of distinctive features associated with the particular 'manner'. Table-7.5 shows the values of PIX for vocabularies characterized by different manners of production, and having the vowel-ending ' A The consonants in each word set are characterized by different places of articulation. Adopting the same approach as was done for the category of unvoiced and unaspirated consonants, PIX was computed for other categories of consonants also.

Table-7.5 : PIX for vocabularies characterized by different manners of production

(Consonantal segment is the one prior to the steady-state region)

Frame size : 128 samples/frame

Overlapping : 96 samples

Vowel-ending : 'A'

SET	Unvoiced-Unaspirated (KA-PA set)	Unvoiced-Aspirated (KHA-PHA set)	Voiced-Unaspirated (GA-BA set)	Voiced-Aspirated (GHA-BHA set)
PIX	80.86	65.99	40.80	45.33

From Table-7.5 one can see that the method which has worked well for the unvoiced and unaspirated category does not hold good for other categories. As the features vary from category to category, recognition of consonants belonging to different categories calls for appropriate definitions of the consonantal parts in the respective speech signals.

The transient plots of the speech waveforms may be used to identify regions which predominantly characterize the consonants belonging to different categories. The excitation characteristics being different for each of these four sets of consonants, processing confined to different regions may be a solution to enhance the system performance. With this in mind the following sets of experiments have been performed. The frame size, the temporal resolution, and the definition of the consonantal region have been varied to determine the appropriate choice of the parameters for the various categories of consonants. Since the

consonants belonging to the third and fourth categories are voiced, a larger frame site (256 samples/frame) was chosen to represent the formant information accurately. For the same reason, LPC's were used to represent the steady-state parts of the signal.

Figure 7.1 shows the different regions of a typical transient plot. A11 the utterances do not have the characteristics shown in the plot. For instance, region A-C is the only one which includes a peak in the case of consonants belonging to the unvoiced and unaspirated category. Similarly, region D-E, with its small peaks following a prominent second peak, is typical of consonants helonging to the voiced and aspirated category. Knowledge such as this will be useful in the identification of the appropriate consonantal regions.

On careful study of several transient plots, the different regions which are likely to predominantly characterize the consonants have been figured out. The samples of the transient plots which correspond to those of four stop consonants (C,T,H,P) for the four different manners of production, are shown in Figs. 7.2 to 7.5. The regions A-E in the transient plots of the various consonants appear to contain the information we are looking for in the identification of the category of these consonants. Hence subsequent experiments were confined to processing in these regions. Experiments carried out with different parameters show that. the system performance can be sought to be enhanced through appropriate choice of these. The results obtained are based on tests with 100 utterances for each run of the experiment.

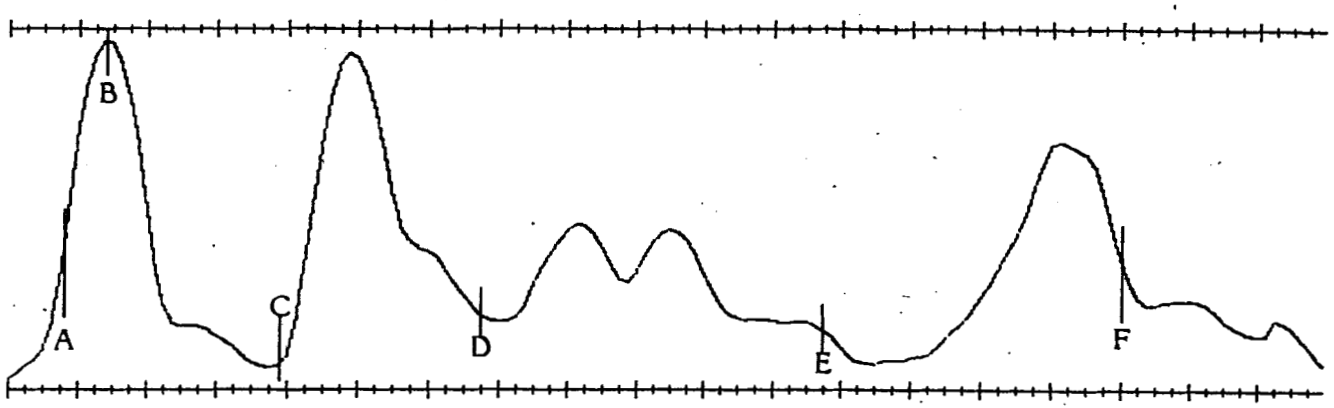


FIG 7.1 DIFFERENT SIGNIFICANT REGIONS IN A TRANSIENT-PLOT

NOTE : All the utterances do not have characteristics which result in a transient-plot shown above.

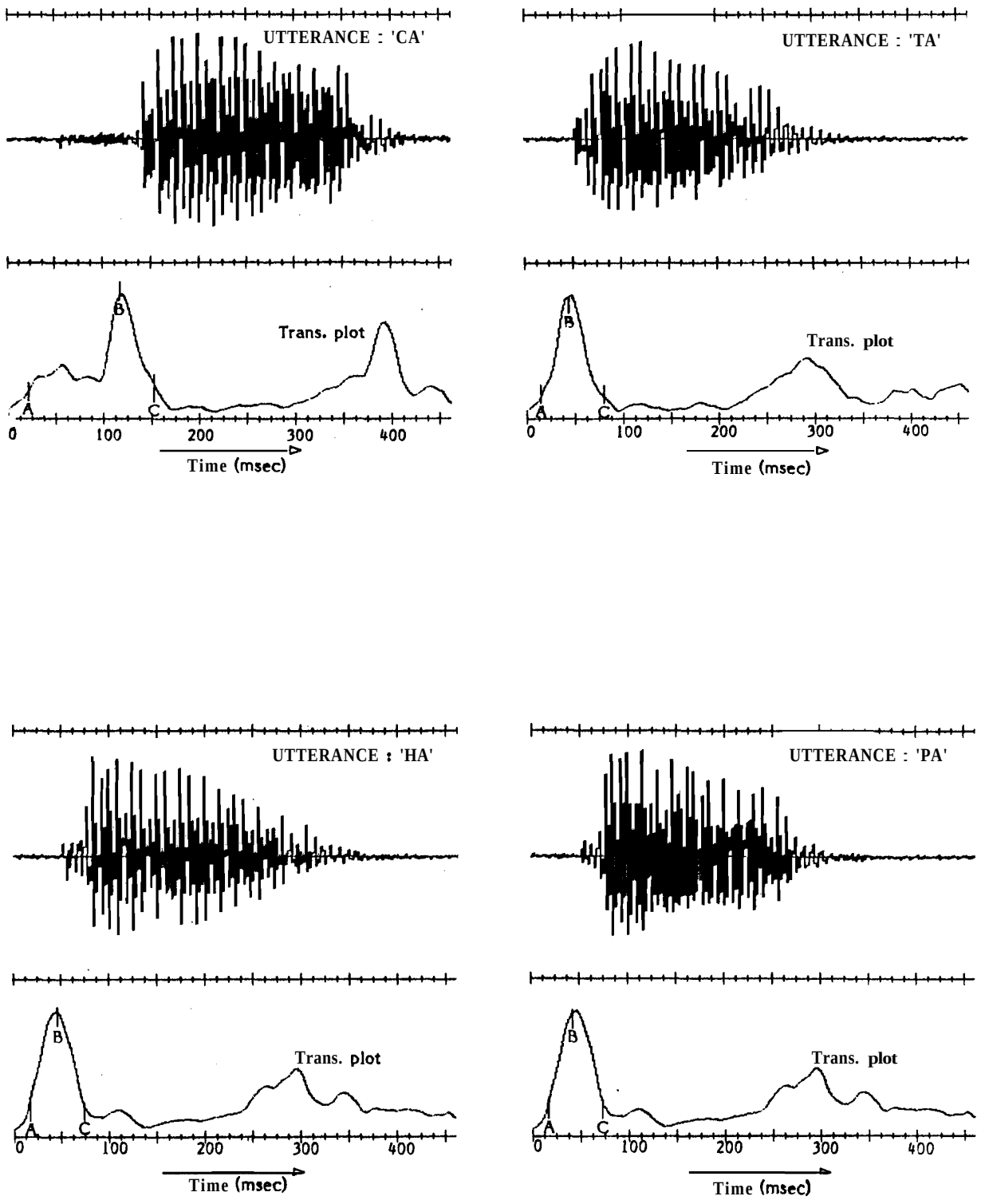


FIG 7.2 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (CA, TA, HA, PA) OF UNVOICED-AND-UNASPIRATED CATEGORY

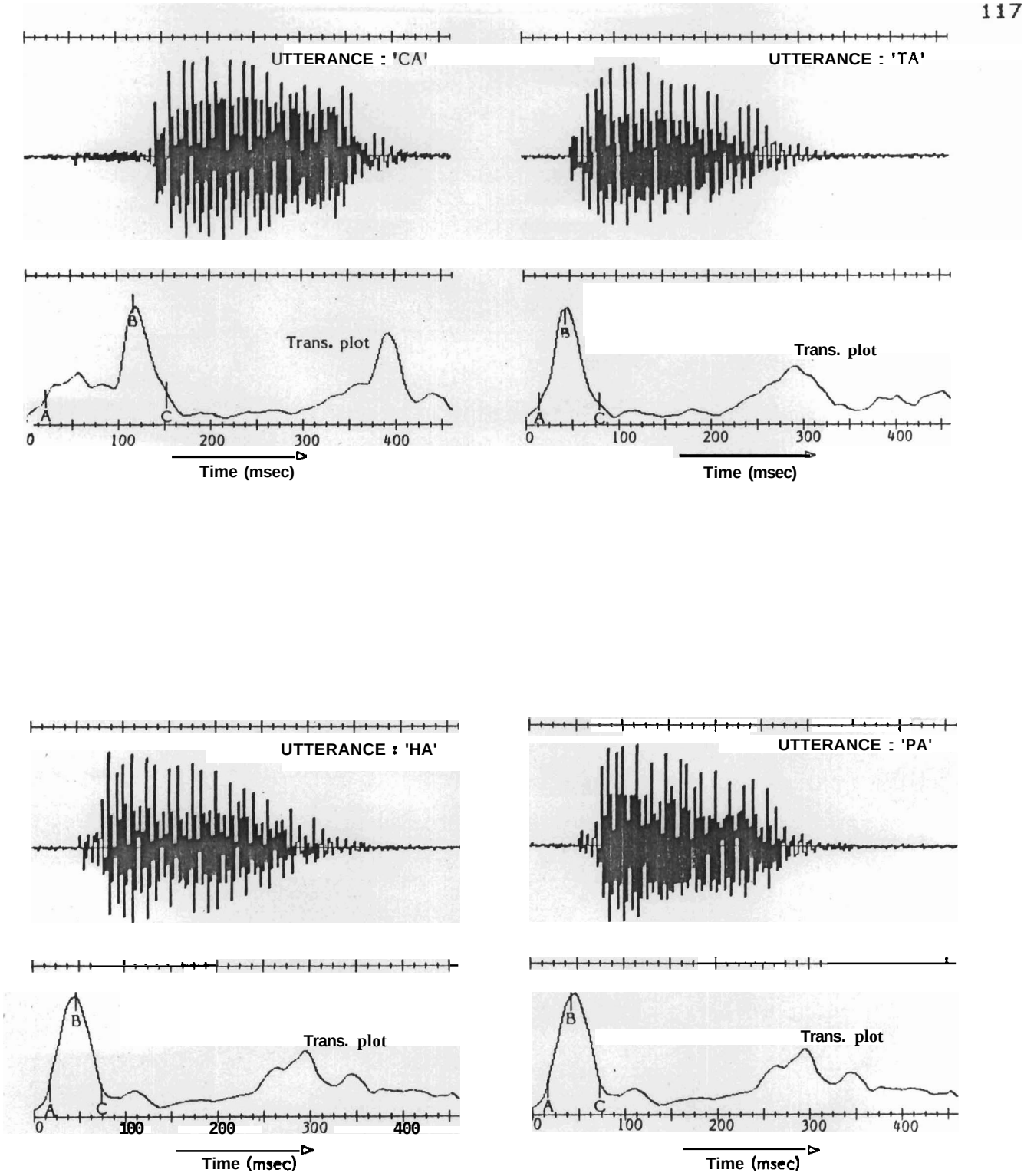


FIG 72 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (CA, TA, HA, PA) OF UNVOICED-AND-UNASPIRATED CATEGORY

Table-7.6 shows the variations in the system performance for the vocabulary of consonants belonging to the unvoiced and aspirated category. Specifically, the members of the vocabulary are {KHA, CHA, THA, HHA, PHA}.

Table-7.6 : PIX for the vocabulary of unvoiced and aspirated consonants

Vocabulary : { KHA, CHA, THA, HHA, PHA }

Consonantal segment	Warping region	Frame size	Over-lapping	PIX
A - D	A - E	128	96	65.99
"	"	256	128	70.41
"	"	256	224	73.49
B - D	B - D	128	96	67.94
"	"	256	128	67.27
"	"	256	224	70.86
B - C	B - D	128	96	71.67
"	"	256	128	70.44
"	"	256	224	73.18
C - D	C - D	128	96	48.89
"	"	256	128	52.04
"	"	256	224	54.43

The PIX-value of 65.99 obtained for the fist set of conditions, is shown to improve to 73.49 when a larger frame size (256 samples/frame) is chosen. On reducing the temporal resolution by cutting down the interframe overlap to 128 samples only, PIX is found to reduce to 70.41. This indicates that adequate temporal resolution in the representation of the speech

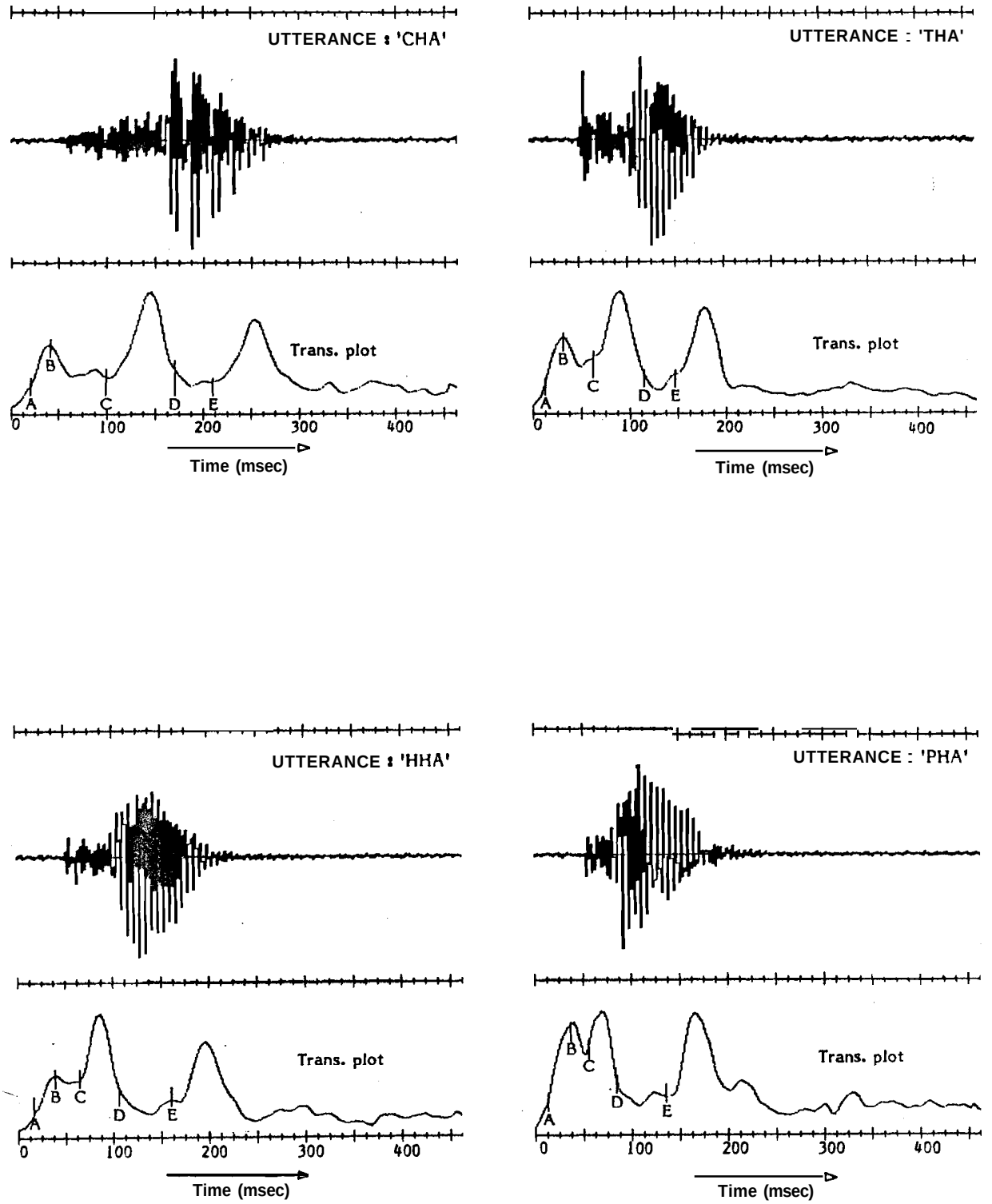
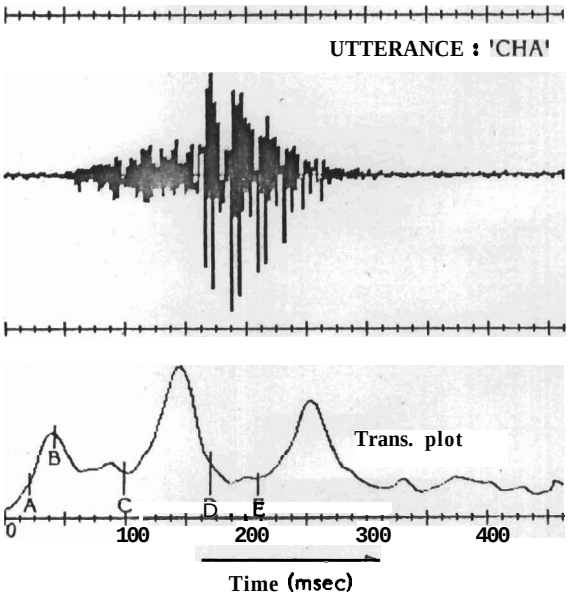
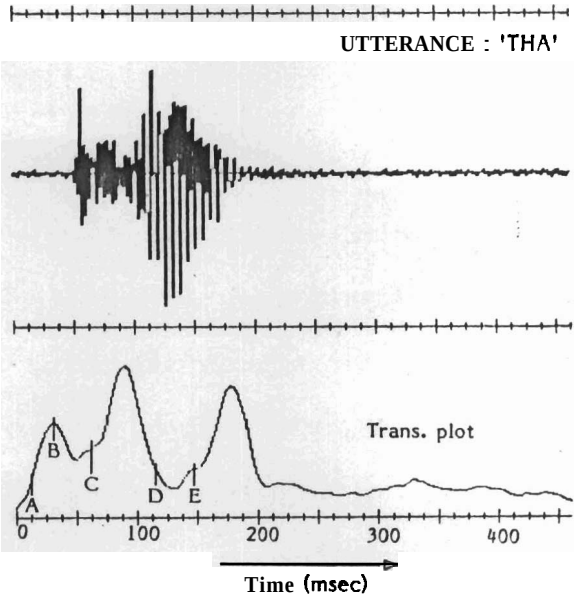
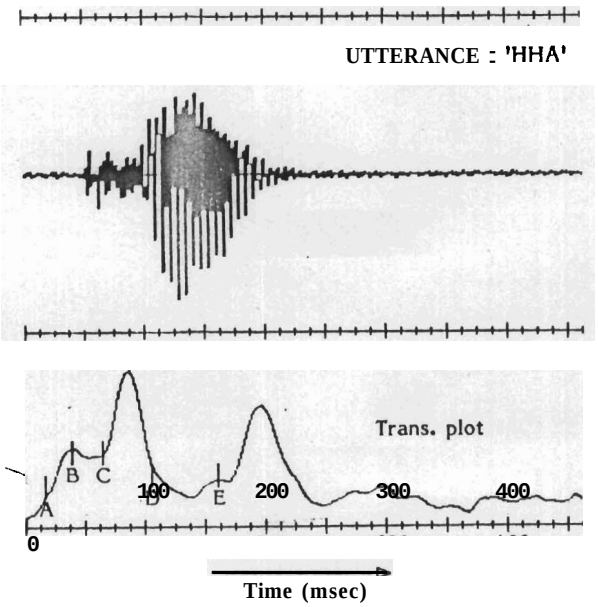


FIG 7.3 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (CHA, THA, HHA, PHA) OF UNVOICED-AND-ASPIRATED CATEGORY







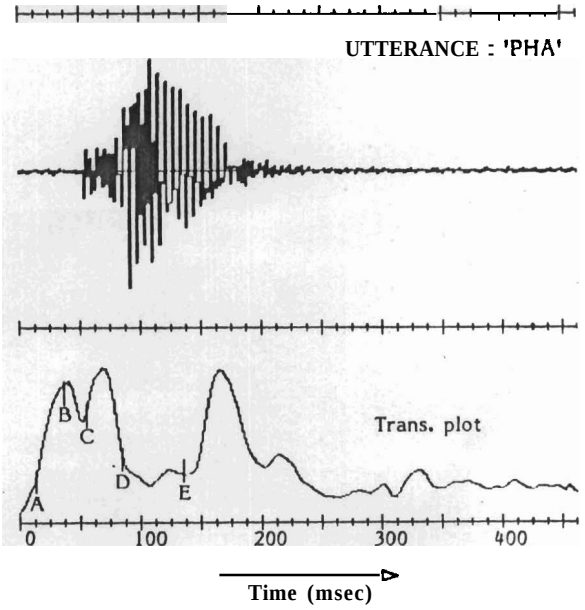


FIG 7.3 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (CHA, THA, HHA, PHA) OF UNVOICED-AND-ASPIRATED CATEGORY

data is needed. At the same time, since there are other features such as aspiration and voicing due to the succeeding vowel-part, a larger frame size is required.

The generally uniform PIX-values in the range of 70's obtained for processing over the region B-C show that this region is a prominent one. However, since the same degree of performance is obtained for processing over the region A-E it may be said that the effect of the prominent features of this class of consonants is spread over a wider region. If minimization of the processing time is the criterion, one has the option to confine attention to region B-C.

Table-7.7 shows the variations in the system performance for the vocabulary of consonants belonging to the voiced and unaspirated category. Specifically, the members of the vocabulary are {GA, JA, DA, WA, BA}.

The PIX-values in the Table indicate that confining the processing to region C-D, and choice of large frame size and high temporal resolution, enhance the system performance. (See Appendix-VI for a sample output.) The large frame size helps since the consonants are voiced. Since the utterances are consonants, and hence characterized by transient features, the high temporal resolution is also justified. The region C-D corresponds to the voice-onset region of the succeeding vowel-part and also exhibits the essential characteristics of the voiced consonant. Processing confined to this portion of the speech signal has resulted in a PIX-value of 75.12 compared to that of 40.80 obtained for processing almost the whole utterance.

Table–7.7 : PIX for the vocabulary of voiced and unaspirated consonants

Vocabulary : { GA, JA, DA, WA, BA }

Consonantal segment	Warping region	Frame size	Over–lapping	PIX
A – D	A – E	128	96	40.80
"	"	256	128	54.17
"	"	256	224	56.99
B – D	B – D	128	96	56.51
"	"	256	128	57.16
"	"	256	224	60.86
B – C	B – D	128	96	33.65
"	"	256	128	39.16
"	"	256	224	38.79
C – D	C – D	128	96	70.62
"	"	256	128	64.58
"	"	256	224	75.12

The high value of PIX when the essential voiced part of the signal is processed suggests that representation of this part using LPCs must result in enhanced system performance. The PIX–values in Table–7.8 confirm this observation. It may further be noted that the PIX–values are nearly the same for this type of representation, irrespective of the frame size and the degree of temporal resolution. The reason for this may be that both the consonant and the succeeding vowel–part are voiced and that LPCs better represent the formant information accurately.

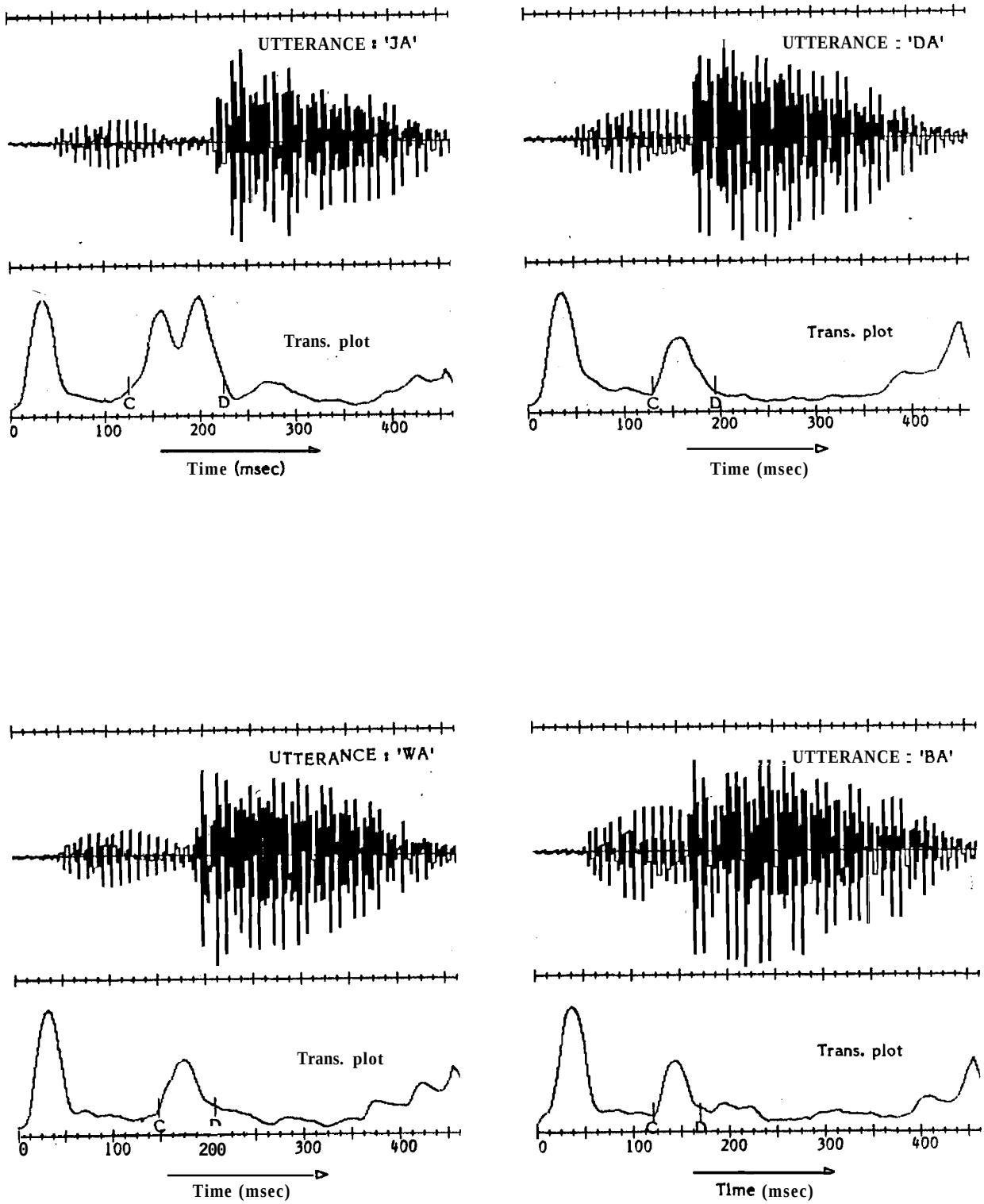


FIG 7.4 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (JA, DA, WA, BA) OF VOICED-AND-UNASPIRATED CATEGORY

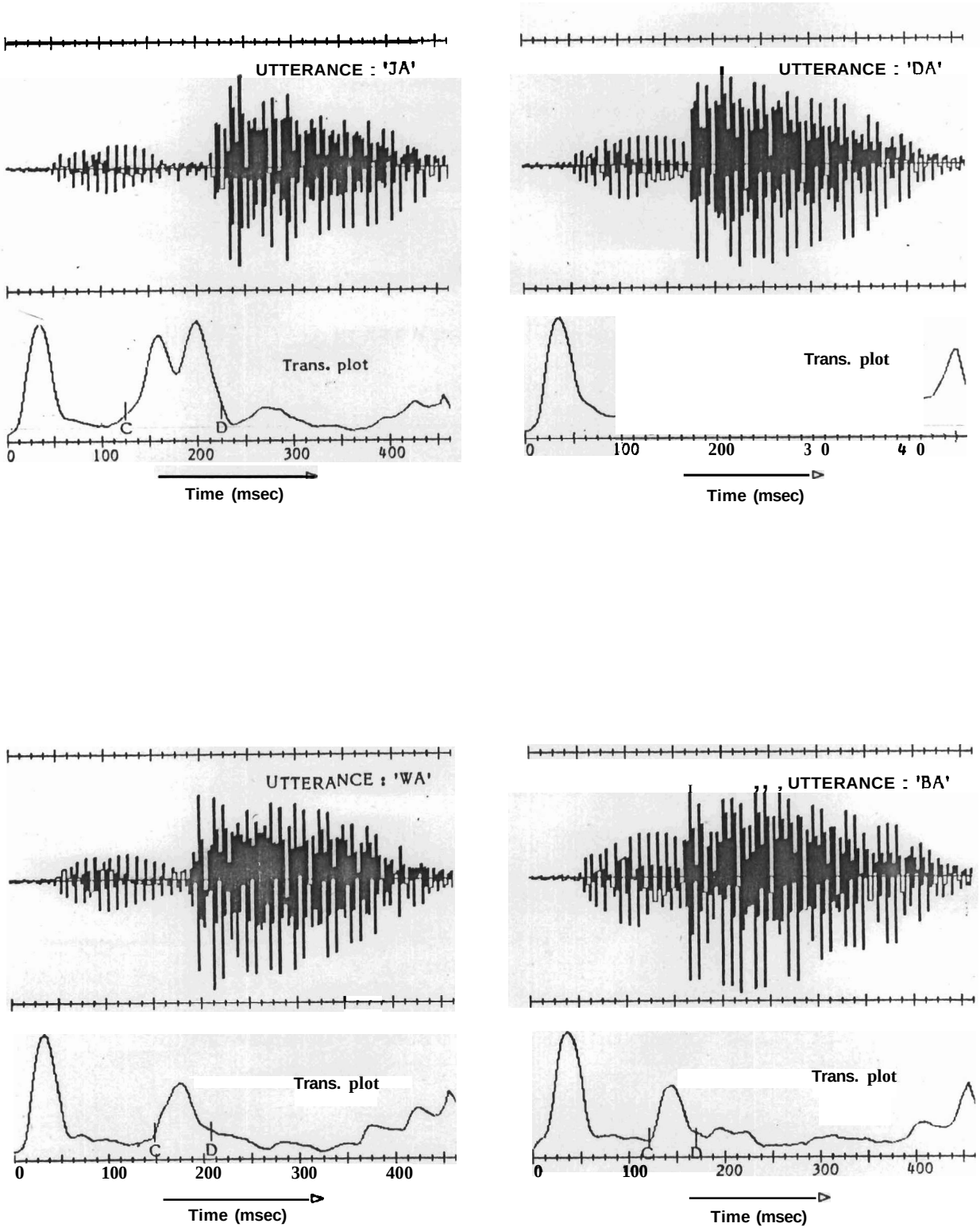


FIG 7.4 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (JA, DA, WA, BA) OF VOICEDANDUNASPIRATED CATEGORY

able-7.8 : PIX for different representations of voi'ced and unarpirated conronantr

Vocabulary : { GA, JA, DA, WA, BA }

Consonantal segment : Region C–D

Frame size	Overlapping	PIX	
		Spectral parameters	LPC-parameters
128	96	70.62	79.10
256	128	64.58	78.25
256	224	75.12	79.75

Table–7.9 shows the variations in the system performance for the vocabulary of consonants belonging to the voiced and aspirated category. Specifically, the members of the vocabulary are { GHA, JHA, DHA, WHA, BHA }. The transient plots of the consonants belonging to this category show some prominent variations in the region D–E following the more prominent second peak. This may be due to the combination of voicing and aspiration–effects in the production of these consonants.

The PIX–values in Table–7.9 show that the system performance can be enhanced by using large frame size, high temporal resolution, and by confining the warping to region C–E and distance computation to region C–D. The reason for this is more or less the same as for the previous case, except that the warping is extended to region C–E thereby showing the importance of that region.

Table-7.9 : PIX for the vocabulary of voiced and aspirated consonants

Vocabulary : { GHA, JHA, DHA, WHA, BHA }

Consonantal segment	Warping region	Frame size	Over-lapping	PIX
A - D	A - E	128	96	45.33
"	"	256	128	48.79
"	"	256	224	57.98
B - D	B - D	128	96	47.25
"	"	256	128	52.24
"	"	256	224	55.59
B - C	B - D	128	96	42.49
"	"	256	128	43.63
"	"	256	224	44.26
C - E	C - E	128	96	57.32
"	"	256	128	59.51
"	"	256	224	63.02
C - D	C - D	256	224	64.23
C - D	C - E	128	96	57.22
"	"	256	128	62.41
"	"	256	224	70.42

Since this set of consonants is also voiced, representation of the consonantal segment using LPCs results in better system performance. However, in addition to voicing, there are other features due to aspiration. Hence large frame size and high temporal resolution in combination with better representation using LPCs, result in a high PIX-value of 81.29, as seen from Table-7.10.

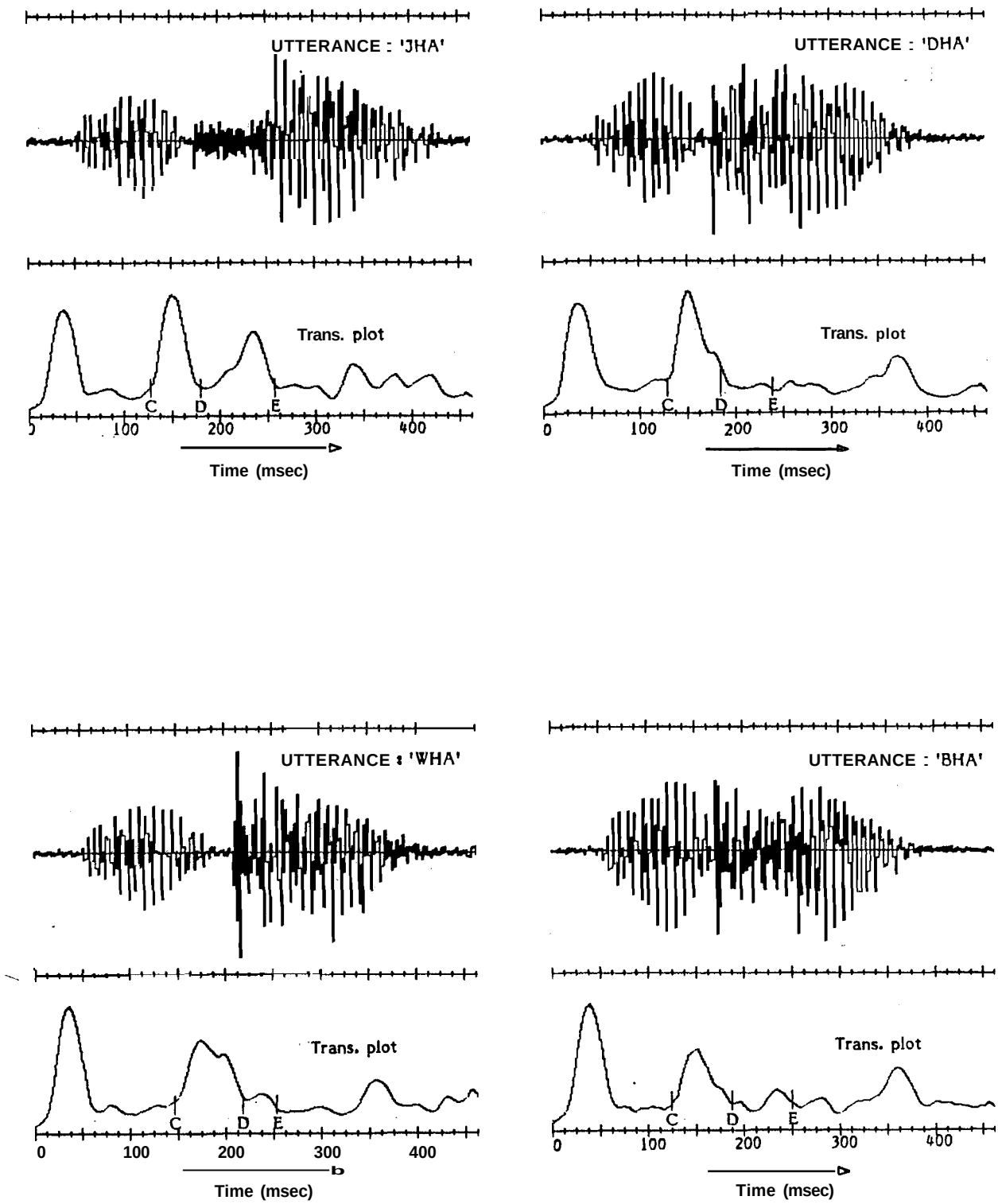


FIG 7.5 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (JHA, DHA, WHA, BHA) OF VOICED-AND-ASPIRATED CATEGORY

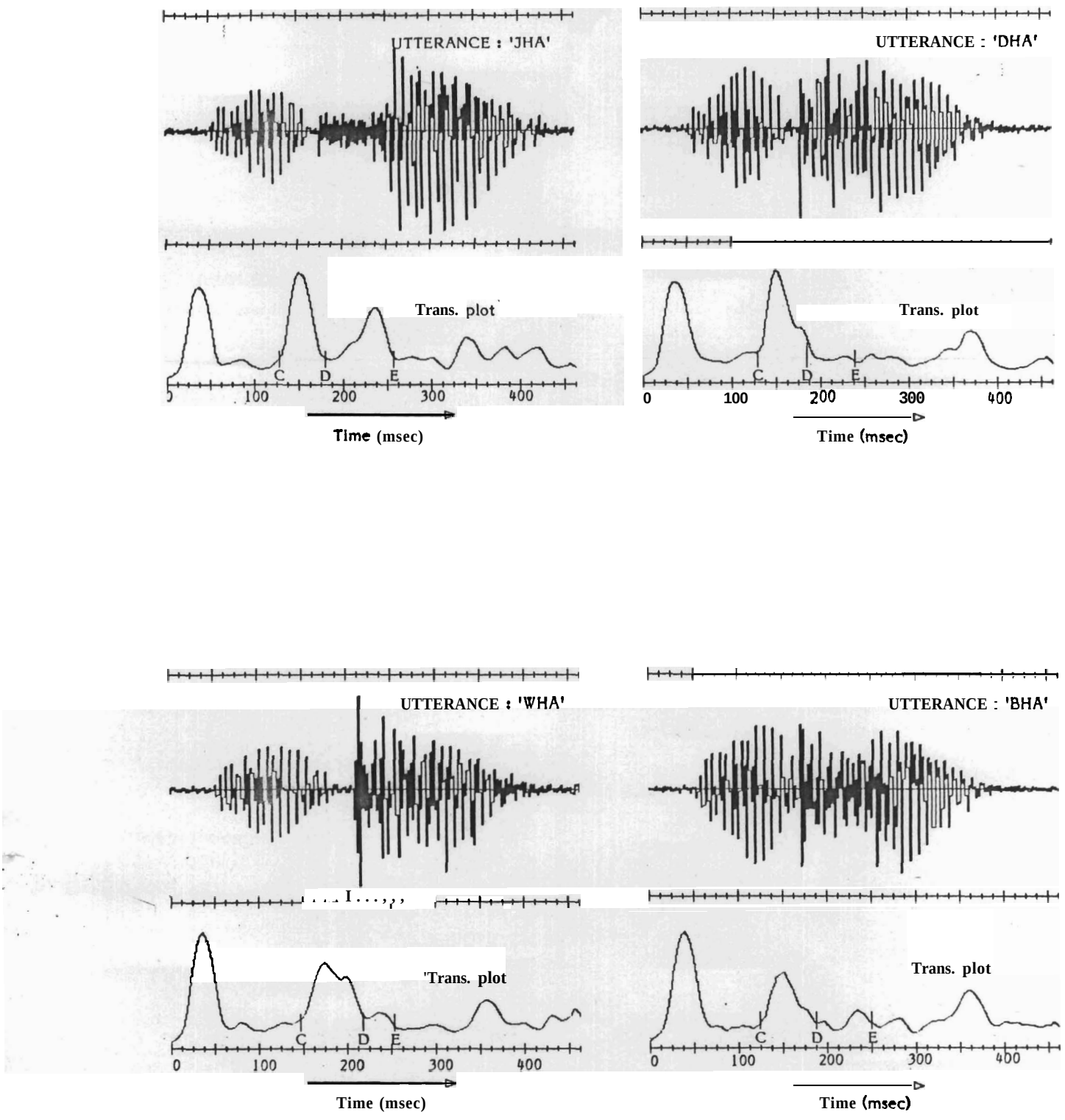


FIG 7.5 TRANSIENT-PLOTS SHOWING THE SIGNIFICANT-REGIONS FOR THE UTTERANCES (JHA, DHA, WHA, BHA) OF VOICED-AND-ASPIRATED CATEGORY

Table-7.10 : PIX for different representations of voiced and aspirated consonants

Vocabulary : { GHA, JHA, DHA, WHA, BHA }

Consonantal segment : Region C-D

Frame size	Overlapping	PIX	
		Spectral parameters	LPC-parameters
128	96	57.22	67.56
256	128	62.41	77.65
256	224	70.42	81.29

The results of these different sets of experiments show that the discriminability among the consonants can be taken advantage of, if they are appropriately represented and processed. It leads to enhanced system performance. The different refinements carried out resulted in PIX-values in the range 70-80 only. This implies that there is scope for further improvement. We nevertheless now have an approach to deal with a confusable vocabulary. The hierarchical approach as suggested by the classification and arrangement of consonants according to the manner of production and place of articulation is seen as a possible approach. The motivation for this is a study of the speech production and the speech perception of these consonants.

In combination with about 80% recognition accuracy in the determination of 'manner', it will appear that the overall recognition accuracy which can be achieved by the methods

proposed in this thesis, is only about 65%. In this work. the focus has been on the determination of the place of articulation, which is a weak feature contributing to the confusability of the vocabulary. In contrast, a near-100% accurate determination of the different manners of production is possible, because of the dominant excitation features characterizing them. In this case, the overall accuracy approaches that of the place-determination achieved so far namely 70-80%. Consonants are the crucial elements, but difficult to recognize in continuous speech. If they can be recognized with about 80% accuracy, with the application of additional knowledge at higher levels, this scheme can be used for connected/continuous speech recognition. Earlier attempts at continuous speech recognition resulted in greater than 90% sentence-recognition with much less than 80% recognition accuracy at lower levels of processing.

Erman et.al. [70] refer to a system's ability to exploit selectively its best data and most promising methods, as opportunistic problem-solving. While analyzing the performance of the HEARSAY-II system, they observe that "opportunistic processing at the lower levels was ineffective and harmful because the credibility ratings of hypotheses were insufficiently accurate to form hypothesis-islands capable of focussing the search effectively.. .. It is only with the word-sequence hypotheses that the reliability of the ratings is high enough to allow active search". Based on this approach, the C2-configuration of the system "achieved correct interpretation of 90% of a test set of utterances (with 73% of the utterances being recognized word-for-word correctly)" [71]. The performance of the systemc at the phonetic level must

have been still lower, since Klatt [721] reports that "the contribution of the ARPA project to improved phonetic recognition strategies is not in proportion to the level of effort expended". Further he notes that Schokey and Reddy [733] "discovered that linguists are actually not very consistent at phonetic transcription if the language is unfamiliar. From their data, they speculate that machines should not be expected to be better than 60 to 70 percent accurate in producing a phonetic transcription". In the case of a phonetic language, problems such as these must be less severe.

In spite of inaccurate phonetic transcription in English, a 90% recognition accuracy was achieved. With improved phonetic strategies resulting in about 80%-recognition of Hindi consonants, one hopes to achieve a near-100% recognition of connected speech in Hindi. The results of this work lend support to such hopeful thinking.

8. CONCLUSIONS

> Minister to partymen: "Yes, thank you - my mission was successful. I found out there is no solution to world problems."
>
> From R.E. LAXMAN's cartoon
> The Times of India

8.1 Summary

The phonetic aspect of Indian languages could be the corner-stone in the development of a system which recognizes speech in these languages. In its simplified form, speech may be considered as reading out letters of the alphabet. In which case, the input to the speech recognition system consists of discrete utterances of letters. Of all the letters, recognition of consonants is the most difficult, and a challenging problem. But it is crucial for continuous speech recognition. Therefore recognition of Hindi stop consonants was the focus of this work. Study of these conso-nants suggested certain signal processing techniques for the recognition of non-nasalised consonants. Because of small changes in the vocal tract system in the production of these consonants, certain sub-classes of these consonants prove to be more confus-able than the sets of vocabulary hitherto widely known. In this thesis, we have reported the various studies conducted during our efforts to develop a recognition system which can handle the confusable vocabulary of Hindi stop consonants.

The development of the approach to recognition of the Hindi stop consonants was based on a series of experiments. In the process of experimentation, small changes were made in the system design at different stages and the performance of the system was evaluated. In order to conveniently study the response of the system to these changes there was a need for a measure of evaluation which would permit testing of the system with fewer data sets. The new measure, called Performance Index (PIX), was used for this purpose. It reflects the characteristics of the distance matrix formed for a pair of reference and test data sets. Comparative study of PIX and the widely used Percentage Recognition Score (PRECS) showed that PIX is a cost-effective measure of evaluation of a recognition system under development. Using PIX, the results of our experiments were discussed.

A signal-dependent approach, which uses adaptive strategies for parameter extraction and matching, was considered first as the basis for our recognition system. Experiments with different vocabularies showed that, whereas for alphadigit vocabulary this approach proved to be a promising one, for a subset of Hindi stop consonants it was found to be wanting. The results showed that information derived from the signal alone was not sufficient and that some knowledge about the speech production of these consonants should be utilized. These consonants are distinguishable among themselves only by the weak characteristics associated with the different places of articulators in their production. This knowledge was sought to be utilized in the next step of refinement.

Since a speech signal could be considered as a sequence of sound segments which are definable by articulatory events, the consonantal and vowel parts of these consonants were identified. Appropriate representation of these parts, in accordance with the nature of the signal characterizing them, resulted in improvement of the system performance. Specifically, the consonantal part, characterized by a signal which is largely transient in nature, needed better temporal resolution, so that the dynamics of the articulators may be represented faithfully. Likewise, voiced segments needed better spectral resolution. Regarding the choice of parameters, representation of the vowel-part by LPCs was found to be better. The reason is that LPCs are more suitable for representing the formant-structure associated with vowel production. It was noted that the dominant vowel-ending was the main reason for the confusability of this vocabulary and that the short leading consonantal part had the features necessary for distinguishing the members of the vocabulary. It was concluded that appropriate temporal and spectral resolution of the different parts, and confining the processing to those parts which carry the crucial information, were two major steps towards successful handling of the confusable vocabulary.

Experiments were continued with a large number of data sets to study the influence of a good recording environment, good reference set, determination of the warping path, and different vowel-endings. At this stage, PRECS also was computed so that it might be compared with PIX. The following observations were made, based on the results of these experiments.

Data sets created in different recording environments cannot be arbitrarily mixed. Appropriate normalization techniques should be applied before handling such data sets. A good reference-set reliably indicates the average behaviour of the system. The warping path within the consonantal part is marginally influenced by the dominant vowel-part and it varies depending on whether it is a basic vowel or a diphthong. The trend of improvement in system performance due to the innovative changes is similar for data sets with different vowel-endings. In the course of these experiments it was also found that the knowledge-base of the system can be widened by including prosodic features. The other important observation was concerning the PIX. The computation of PRECS with large data conclusively showed that PIX, computed with fewer tests, was indicating the trend in the system performance.

The studies conducted led to the evolution of a scheme for recognition of Hindi stop consonants. Recognition of the vowel ending and the manner of production of a consonant does not pose major problems. Given a test utterance the method of recognition consists of fist identifying the vowel-ending and then the manner of production. Thereupon, the place of articulation is to be determined using selective processing, and with an appropriate reference set. Recognition is thus achieved through the choice of appropriate techniques in three passes.

These studies showed that based on knowledge in a limited domain, and by breaking up the problem into a number of smaller search spaces which can be tackled separately, machine performance can be improved so as to approach that of humans. Thus knowledge-based processing appears to be the way to deal with natural signals such as speech signals.

8.2 Conclusions

Our approach to the problem of speech recognition was developed experimentally. Hence many heuristics figured in it and played crucial roles. These could possibly be supported by appropriate quantitative analysis and validated through more experiments with large amounts of data. Against this background of what was reported we now proceed to critically review the various features of this work.

1. Since the measure of evaluation used for the studies was PIX, we may first consider it for this discussion. PIX fulfilled the need for a measure "to determine and predict the impact of design changes on performance". We also saw how it compared well with the widely used Percentage Recognition Score. Specifically, the two points in favour of PIX are that it enables performance evaluation with relatively few data sets and that it reflects the characteristics of the distance matrix.

The mapping function and the values associated with it are however based on heuristics such as the following :

i) "we do not allow a well-distinguished member to dominate the performance score",

ii) "we give the marginally distinguishing members lesser weight as they may jump the fence one way or the other".

A quantitative analysis, possibly based on the use of fuzzy logic, will have to be undertaken to study the effects of such heuristic decisions in the present definition of PIX. PIX should

also be studied for much larger data, and for varying sizes of reference and test data sets, which would result in different sizes of distance matrices.

2. The approach to recognition and subsequent refinements were based on the motivation that knowledge at signal level should be fully utilized. Subsequently this was refined to include the knowledge of speech production and speech perception of the consonants.

Perhaps there is scope for refinement of the siqnal-dependent approach. It might hold solutions for related problems at signal level. In another situation it was experimentally verified. Refinement to the signal-dependent approach, in the form of weighting the parameters within a frame so that those with high S/N ratio values in the spectra are emphasized more than those with low values, has led to improved system performance for distorted speech [31,74]. Hence refinements to the siqnal-dependent approach, in a direction different from what has been discussed in this thesis, could be pursued.

3. Reduced spectral resolution by using fewer parameters, and high temporal resolution by overlapping the frames was advocated for the representation of the consonantal part. As may be seen from Table-5.6, reduced sepctral resolution does not result in a significant improvement in the system performance. Nevertheless, the results were consistent in the sense that there were no cases in which PIX-values were widely varying for different spectral resolutions. It may be that for a different type of parameters this improvement in the performance could be

studied more reliably. Thus this aspect also calls for a more detailed study. That the effect of high temporal resolution is better system performance, has however been conclusively proved.

4. Representation of the vowel-part by LPCs was always found to be better than the representation by the spectral parameters. Since it is not meaningful to use lesser spectral resolution for vowels, the number of parameters used was always 8 or more. However, studies for representation of a vowel by parameters with less resolution would show deterioration in the system performance. This, along with the studies for consonants mentioned above would conclusively prove that different segments need to be dealt with appropriately not only from the point of view of processing but also from that of representation.

5. Techniques based on vocabulary-specific knowledge were discussed. But the degree to which these would be helpful in enhancing the system performance will have to be studied when other factors such as recording environment, refined pre-processing, etc., have been taken care of. For instance, with reference to Table-6.7, the PIX-value of 80.59 was achieved for confining to consonantal part alone. The reason might be that the recording environment was better. For data created in a different environment the PIX was only 80.86 even after the combined refinements of enhanced temporal resolution and restriction of the warping to the consonantal part. Hence the system performance needs to be evaluated after taking a few more such precautions during data-creation and pre-processing. However, the vocabulary-specific knowledge which led to the conclusion that the consonantal regions had more discriminative information has been proved to be reliable

by the results obtained at different stages of experimentation. Any process that ignores this vital point is bound to result in degraded system performance.

6. The combined effect of providing high temporal resolution, addition of prosodic feature related to the consonantal duration, and the restriction of warping to within the consonantal region, led to improved system performance. The individual effects of these are not known and so deserve further study. Apparently this step has not resulted in significant improvement. But, for a correct evaluation of these steps taken on a system which has already been performing well, one has to measure the magnitude of reduction in the error rate, rather than the improvement in the recognition score. This implies that testing should be carried out with rather large data. The values of PIX in Table-6.8 show that these steps are in the right direction. But their real efficacy 'is better evaluated with the measure of error rate.

7. The segmentation of the utterance into consonantal and vowel-parts was determined by a method of transient analysis. It involved computation over a band of adjacent frames so as to avoid parametric fluctuations of data in steady-state portions. This results in a fuzzy estimation of the consonantal part. Subsequent analysis of results for the two. parts showed that there cannot be any sharp boundary between these parts and that there are only segments which are predominantly of one or the other type. Possibly, a quantitative analysis based on fuzzy logic might help in determining the degree of predominance of a segment, and this could be utilized to define a more effective

boundary from the point of view of recognition. For evolving any such measure, more tests under strictly observed conditions should be done so that other factors do not affect the estimation of predominance of a feature.

8. Studies on the transient plots of the consonants due to different manners of production showed that they contain more information useful for selective processing on the input signal. For this work, transient plots for discrete utterances alone were studied. Their usefulness in continuous speech recognition is to be studied. For instance, transient plots may help in spotting the vowel-parts in the continuous speech signal. The extent to which these plots will be useful, in the presence of co-articulatory features characterising continuous speech, is to be studied. If successful, the scheme of vowel spotting, manner determination, followed by place determination at lower levels of processing and application of knowledge at higher levels such as syntax, semantics, etc., appears to be a promising approach to continuous speech recognition.

9. By considering five different vowel-endings the trend of improvement in system performance, in response to the innovative changes in the system design, was generally confirmed. But, from Tables 6.8 and 6.9 one can see that the degree of the improvement varies for different vowel-endings. These tables also show that the behaviour of the system for vowel-endings and diphthong-endings is different. The degree of influence of these parts on the consonantal segment appears to be another interesting study. A thorough study of these two aspects based on tests with large amounts of data can throw more light on the

relative effects of the various types of vowels. Such fragments of knowledge may be useful in enriching the knowledge-base of a recognition system.

10. A detailed study for recognition of the vowels and different manners of production was not undertaken. The main reason was that they were not likely to be as confusable as the word-sets, whose members are based on different places of articulation. However the studies conducted confirmed our basic philosophy of selective processing for different types of consonants. But a thorough study of these, especially in the context of consonants with different vowel-endings will help to refine the methodologies for a knowledge-based approach. Such a study is useful for the development of practical recognition systems.

11. A three-pass hierarchical scheme has been proposed for the recognition of stop consonants. The main drawback of any hierarchical scheme is the reaching of a point of no-return to other alternatives, on leaving the nodes of decision at a given level. A better scheme might be one based on the 'beam-search technique', in which decisions are always deferred to a later stage of processing. At each stage, all 'meaningful' alternatives will be tried and the results accumulated. Only when the results run counter to some very elementary expectations, or when the computation becomes complex, will further search along certain lines be abandoned. To implement such a scheme, tests on large data, more information about the vocabulary, careful preparation of the data sets, etc., are necessary. Based on apriori knowledge, appropriate decisions on 'thresholds' and

'discriminant functions' can be reliably made. These decisions are crucial since pruning of search would essentially depend on them. The use of beam-search techniques implies more computation. But if the decisions are not reliably made this method is not an effective alternative. Ultimately it is the cost-effective solution that will be widely adopted.

12. This study has as its focus the solution of the confusability of the stop consonants, for a chosen set of vowel-endings. It does not include some of the vowels, diphthongs, semi-vowels, and also the nasalised consonants. In addition to these there are also composite letters such as the stop consonants with semivowel-ending. Some of our studies may be extended to these vocabularies. It is also possible that, in the study of semi-vowels, more understanding of stop consonants could emerge.

13. We started with the assumption that we would utilize the phonetic nature of the alphabet of Indian languages. On that basis speech, in its simplest form, was treated as a string of discrete utterances. However, this restricted specification of slowed-down speech is not a natural one. Hence the aspects of developing a system which would accept continuous speech should be looked into.

14. Since consonants are the crucial elements in continuous speech, our effort to improve the recognition of stop consonants is to be seen as providing some inputs to the development of methods for continuous speech recognition. There are successful attempts to develop speech-understanding systems even though these do not rely much on accurate description of the speech

signal at phonetic levels. There is scope for applying the results of a work which focusses on improving system performance by refining the processing at signal levels. Thus the studies reported in this thesis should be considered as throwing some light in that direction and are "by no means an ultimate analysis" [1].

~~~~~~~~~~~~
~~~~~~~~~~~~

REFERENCES

1. Allen W.S., "Phonetics in Ancient India, Geoffrey Cumlerlege, Oxford University Press, London, 1953.

2. Ballantyne J. R., "Varadaraja's Laqhu Siddhanta Kaumudi", Motilal banarsidass, Varnasi, 1961.

3. Iyer K.A.S., "On the fourteen Mahesvara Sutras", Proc. of the Fourth All Indian Oriental Conference, 1928.

4. Tolkappiyar, "Tolkappiyam Eluttatikaram", Nachinarkkiniyar's commentary - Ilampuranar's commentary- The South India Saiva Siddhanta Works Publishing Society, Tinnevelly, 1955.

5. Coward H.G., "Bhartrhari", Twayne Publishers, Boston, 1976.

6. Whitney W. D., "Saunakiya Caturadhyayaka (Atharva-Pratisakhya)", Ed. & transl., JAOS (Vii) 333#

7. Electronics Commission, Government of India, "Proceedings of Symposium on Linquistic Implications of Computer-based Information Systems", New Delhi, 10-12 November 1978.

8. Om Vikas, "Use of Non-English Lanugages in Computer - A Selected bibliography", Electronics Commission, Information, Planning & Analysis Group, October 1978.

9. Electronics Commission, Government of India, "Electroniki Bharati", (in Hindi), A special issue in connection with the Third World Hindi Conference, October 1983.

10. Mahabala H.N. and Raman S. "Determination of Minimum Matrix Size of Tamil Characters for Optimal Presentability using On-line Teletype as Tamil Typewriter", Proc. of Computer Society of India Convention, CSI-78, February 1978.

11. Raman S., Sundar R. and Hahabala H.N. "Microprocessor-based Stencil-Cutter as Computer-Output for Text in Indian Languages", Journal of IETE, New Delhi, Special Issue on Computer Applications in Processing of Indian Languages and Scripts, Vol. 30, No. 6, 1984, pp. 234-237.

12. Mahabala H.N., Raman S., Sundar R. and Prasad R.N., "Computer Output in Indian Scripts on Stencils" (in Hindi), Electroniki Bharathi, Special Issue of Electronics Commission, October 1983, p. 16.

13. Vishwanath G., "Computer Input and Output in Indian Languages", Proceedings of Symposium on Linguistic Implications of Computer-based Information Systems, New Delhi, November 1978.

14. Narasimham P.V. H.M.L., Prasad B. and Rajaraman V. "Code-Based Keyboard for Indian Languages", J. Comp. Soc. of India, 2,2; December 1971, pp. 33-37.

15. Mudur, S.P., and Wakankar, L.S., "Computer Input/output in Devanagiri", Proc. of Symposium on Linguistic Implications of Computer-based Information systems, New Delhi, November 1978.

16. Narasimham, P.V.H.M.L, and Sinha R.M.K., "Phonetically Coded Key-boarding in Indian Languages", Proc. of Symposium on Linguistic Implications of Computer-based Information Systems, New Delhi, November 1978.

17. Yeunung Chen, "Vocabulary Selection for High Performance Speech Recognition", Proc. Intl. Conference on Acoustics, Speech and Signal Processing, 1983, ICASSP 83, Vol.2, pp. 757-760.

18. baker J.M. et al., "Speech Recognition Performance Assessments and Available Databases", Proc. Intl. Conference on Acoustics, Speech and Sig. Proc., 1983, ICASSP 83, Vol.2, pp. 527-530.

1 Lea W.A., "What Causes Speech Recognisers to Make Mistakes?", Proc. ICASSP-82, 1982, Vol.3, pp. 2030-2033.

20. Ganesan S.N., "A Contrastive Grammar of Hindi and Tamil", University of Madras, 1975.

21. Peter Ladefoged, "A Course in Phonetics", Harcourt Brace-Jovanovich Inc., 1975.

22. Vidhata Mishra, "A Critical Study of Sanskrit Phonetics", The Chowkkamba Sanskrit Studies Vol. LXXXIII, Varnasi, India, 1972.

23. Broad D.J. and Shoup J.E., "Concepts for Acoustic Phonetic Recognition", 'Speech Recognition', Ed. Raj Reddy, Academic Press, 1975, pp. 243-274.

24. Liberman. A.M., "The Sounds of Silence", Proc. Intnl. Symp. on Speech Processing, 1980, Ed. P.V. S. Rao, pp. 5.1 - 7.17.

25. Fant G., "Speech, Sounds, and Features", The MIT Press, 1973, pp. 5–6, 25.

26. Shirai K. and Honda H., "Estimation of Articulatory Parameters from Speech Waves", Trans. IECE Japan, Vol. J61-A, No.5, 1978, pp. 409-416.

27. Shirai K., "Vowel Identification in Continuous Speech Using Articulatory Parameters", Proc. ICASSP–81, 1981, pp. 1172-1175.

28. Shirai K. and Kobayashi T., "Recognition of Semivowels and Consonants in Continuous Speech Using Articulatory Parameters", Proc. ICASSP-82, 1982, Vol.3, pp. 2004–2007.

29. Ishizaki S., "Dynamic Speech Discrimination Using an Articulatory Model", 6th IJCAI, Vol. 1, 1979, pp. 422–424.

30. Yegnanarayana B. and Sreekumar T. "Signal–dependent Matching for Isolated Word Speech Recognition Systems", Signal Processing, Vol.7, No. 2, Oct. 1984, pp. 161–173.

31. Yeqnanarayana B. Sreekumar T. and Raman S., "Signal–dependent Matching for Speech Recognition", Proc. of IEEE Intl. conference on Systems, Man and Cybernetics, 1983, Vol. 1, pp. 150-154.

32. Allen Newell, "A Tutorial on Speech Understanding Systems", Speech Recognition: Invited Papers presented at the 1974 IEEE Symposium, Ed. D. R. Reddy, Academic Press, 1975, pp. 3–54.

33. Watari M., Akabane M. and Sako Y., "A Speaker Independent Word Recognition based on Transient Matching", Proc. of International Conference on Acoustics, Speech and Signal Proc.1983, ICASSP-83, Vol.2, pp. 715-718.

34. Raman S. And Yeqnanarayana B., "Performance of Isolated Word Recognition System for Confusable Vocabulary", Proc.of IEEE Intl. Conference on Acoustics, Speech and Signal proc., San Diego, Mar. 1984, ICASSP84, Vol. 2, pp. 17.5.1 - 17.5.4.

35. Gevarter W.b., "Expert Systems: Limited but Powerful", IEEE Spectrum, Vol.20, No. 8, August 1983, pp. 39-45.

36. Doddinqton G.R. and Schalk T.B., "Speech Recognition: Turning Theory to Practice," IEEE Spectrum, 18(9), September 1981, pp. 26-32.

37. Rabiner L.R. and Levinson S.E., "Isolated and Connected Word Recognition - Theory and Selected Applications", IEEE Trans on Commn., Vol. Com-29, No.5, May 1981, pp. 621-659.

38. Lea W.A., "Speech Recognition: Past, Present and Future", Trends in speech Recognition, Ed. W.A.Lea, Prentice Hall Inc., Prentice Hall Inc., 1980, pp. 39-98.

39. Levinson S.E., and Shipley K.L., "A Conversational Mode Airline Information and Reservation System Using Speech Input and Output", Bell Syst. Tech. Jounl., Vol.59, Jan 1980, pp. 119-137.

40. Fant G. in his key-note address in the IEEE Symposium on Speech Recognition, 1974, 'Speech Recognition', (Ed) D.R.Reddy, Academic press, 1975.

41. Zwicker E., "Subdivision of the Audible Frequency Range into Critical Bands, (Frequenzgruppen)", J. Acous Soc. Am., Vol.23, 1961, p. 248.

42. Schroeder M.R., "Recognition of Complex Acoustic Signals", Life Sc. Res. Rep., (Ed) T. H. Bullock, Vol.55, 1977, pp. 323-328

43. Oppenheim A.V. and Schafer R.W., "Homomorphic Analysis of Speech", IEEE Trans. Audio and El. Acous., Vol. AU-16, No. 2, June 1968, pp. 221-226.

44. John Makhoul, " Linear Prediction in Automatic Speech Recognition", Speech Recognition, (Ed.) D.R. Reddy, Academic Press, 1975, pp. 183-220.

45. Scott P.B., "VICI- A Speaker Independent Word Recognition System", Conf. Rec., 1976-ICASSP, Philadelphia, April 1976, pp. 210-213.

46. Rabiner L. R., "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words", IEEE Trans. Acous., Speech & Sig. Proc., Vol. ASSP-26, No.1, Feb.1978, pp. 34-42.

47. Itakura F., "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing. Vol. ASSP-23, February 1975, pp. 67-72.

48. Silverman H.F. and Dixon N.R., "Comparison of Several Speech-Spectra Classification Methods", IEEE Trans. Acous. Speech and Signal Proc., Vol. ASSP-24, August 1976, pp. 289-295.

49. Pfeifer L. L., "Inverse Filter for Speaker Identification", Speech Communications Research Lab., Santa Barbara, CA, Final Report RADC-TR-74-214, 1974.

50. Atal B.S., "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", J. Acous. Soc. Am., Vol.55, June 1974, pp. 1304-1312.

51. Gray A.H. and Markel J.D., " Distance Measures for Speech Processing", IEEE Trans. Acous. Speech and Signal Proc., Vol. ASSP-24, No.5, October 1976, pp. 380-391.

52. Rosenberg A.E., and Itakura F., "Evaluation of an Automatic Word Recognition System Over Dialed-up Telephone Lines", Journal of Acoustical Society of America., Vol.60, suppl.1, 1976, p. S12.

53. Rosenberg A.E and Schmidt C.E., "Automatic Recognition of Spoken Spelled Names for Obtaining Directory Listings", Bell Systems Tech. Journal, Vol.58, October 1979, pp. 1797-1823.

54. "Recognizing Confusable Words", A note on U.S. Patent No. 4 400 828, for 'Word Recognizer', awarded to Frank C. Pirz, et. al., IEEE Spectrum, Feb 1984, p. 18.

55. De Mori R., et. al., "Integration of Acoustic, Phonetic. Prosodic, and Lexical Knowledge in an Expert System for Speech Understanding", Proc. of Intl. Conf. on Acous. Speech and Sig. Processing, San Diego, Mar. 1984, ICASSP84, Vol.3, pp. 42.9.1-42.9.4

56. Sakoe. H and Chiba. S. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. Acous. Speech. & Sig. Proc., Vol. ASSP-26, Feb. 1978, pp. 43-49.

57. Raman. S, et al., "Knowledge Based Signal Processing for Recognition of Hindi Stop Consonants", Proc. IEEE Intl. Conf. on Comp. Sys. & Sig. Proc., India, Dec. 1984. Vol. 2, pp. 896-900.

58. Woodard J.P., and Lea W.A., "New Measure of Performance for Speech Recognition Systems", Proc. Intl. Conf. ASSP, ICASSP-84, Vol. 1, p. 9.6.1.

59. Raman. S. and Algazi V.R., "Performance Index as a Measure of Performance Evaluation of a Speech Recognition System", Report No. SIPL-84-7, University of California, Davis, June 1984.

60. Spiegel M. R., "Theory and Problems of Statistics", Schaum's outline series-Mc Graw Hill, 1981, pp. 188-191; pp. 344-345.

61. Efron B. and Gong G., "A Leisurely Look at the bootstrap, the Jackknife and Cross-Validation", The American Statistician, Feb. 1983, Vol. 37, No. 1, pp. 36-48.

62. Rabiner. L.R., et.al., " Applications of a Non-linear Smoothing Algorithm to Speech Processing", IEEE Trans. Acous. Speech. & Sig.Proc., Vol. ASSP-23, No.6, Dec. 1975, pp. 552-557.

63. Davis S.B and Mermelstein P., "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acous. Speech & Sig. Proc., Vol. ASSP-28, No.4, Aug. 1980, pp. 357-366.

64. Yegnanarayana B., Subramani G., and Raman S., "Signal to Symbol Transformation for Isolated Word Speech Recognition Systems", Proc. of International Conference on Systems, Man and Cybernetics, 1983, Vol.I, pp. 345-349.

65. Das S. K., "Some Expirements in Discrete Utterance Recognition", IEEE Trans. on Acous., Speech, and Signal Processing, Vol. ASSP-30, No. 5, Oct. 1982.

66. Zue V.W., and Kopec G.E, "Trends in Speech Recognition", Trends & Perspectives in Signal Processing, Sep. 1983, pp. 12-15.

67. Cole, R.A., et al., "Feature-based Speaker-independent Recognition of Isolated English Letters", Proc. Intl. Conf. ASSP, ICASSP-83, Vol.2, March 83, pp. 731-733.

68. Huttenlocher D.P. and Zue V.W., "A Model of Lexical Access Eased on Partial Phonetic Information", IEEE Intl. Conf. on ASSP, ICASSP-84, San Diego, 1984, pp. 26.4.1 - 26.4.4.

69. De Mori R., et. al., "Parallel Algorithms for Syllable Recognition in Continuous Speech", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. PAMI-7, No. 1, Jan. 1985, pp. 56-69.

70. Erman L. D., et. al., "The Hearsay-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty", Computing Surveys, Vol. 12, No. 2, June 1980, pp. 213-253.

71. Erman L.D. and Lesser V.R., "The Hearsay-I1 Speech Understanding System: A Tutorial", Trends in Speech Recognition, (Ed) W.A. Lea, Prentice Hall Inc., 1980, pp. 361-381.

72. Klatt D. H., "Overview of the ARPA Speech Understanding Project(1)", Trends in Speech Recognition, Ed. W.A. Lea, Prentice Hall Inc., 1980, pp. 249-271.

73. Shockey L. and Reddy D.R., "Quantitative Analysis of Speech Perception", paper presented at the Stockholm Speech Commn. Seminar, 1775.

74. Yegnanarayana B. and Sarat Chandran, "Performance of Isolated Word Recognition System for Degraded Speech", Proc. of Intl. Conf. on Acous., Speech and Sig. Proc., 1984, ICASSP-84, pp. 9.9.1 - 9.9.4.

~~~~~~~~~~~~~
~~~~~~~~~~~~~

LIST OF AUTHOR'S PUBLICATIONS RELATED TO THIS WORK

1. Mahabala H.N. and Raman S. "Determination of Minimum Matrix Size of Tamil Characters for Optimal Presentability using On-line Teletype as Tamil Typewriter", Proc. of Computer Society of India Convention, CSI-78, February 1978.

2. Raman S., Sundar R. and Mahabala H.N. "Microprocessor-based Stencil-Cutter as Computer-Output for Text in Indian Languages", Journal of IETE, New Delhi. Special Issue on Computer Applications in Processing of Indian Languages and Scripts, Vol. 30, No. 6, 1984, pp. 234-237.

3. Yeqnanarayana B., Sreekumar T. and Raman S. "Signal-dependent Matching for Speech Recognition", Proc. of IEEE Intl. conference on Systems, Man and Cybernetics, 1983, Vol.1, pp. 150-154.

4. Raman S. and Yeqnanarayana B. "Performance of Isolated Word Recognition System for Confusable Vocabulary", Proc. of IEEE Intl. Conference on Acoustics, Speech and Signal Proc., San Diego, Mar. 1984, ICASSP84, Vol.2, pp. 17.5.1 - 17.5.4.

5. Raman. S and Alqazi V.R. "Performance Index as a Measure of Performance Evaluation of a Speech Recoqnition System", Report No. SIPL-84-7, University of California, Davis, June 1984.

6. Raman. S, et.al., "Knowledge Based Signal Processing for Recoqnition of Hindi Stop Consonants", Proc. IEEE Intl. Conf. on Comp. Syst. & Sig. Proc., India, Dec. 1984, Vol. 2, pp. 896 - 900.

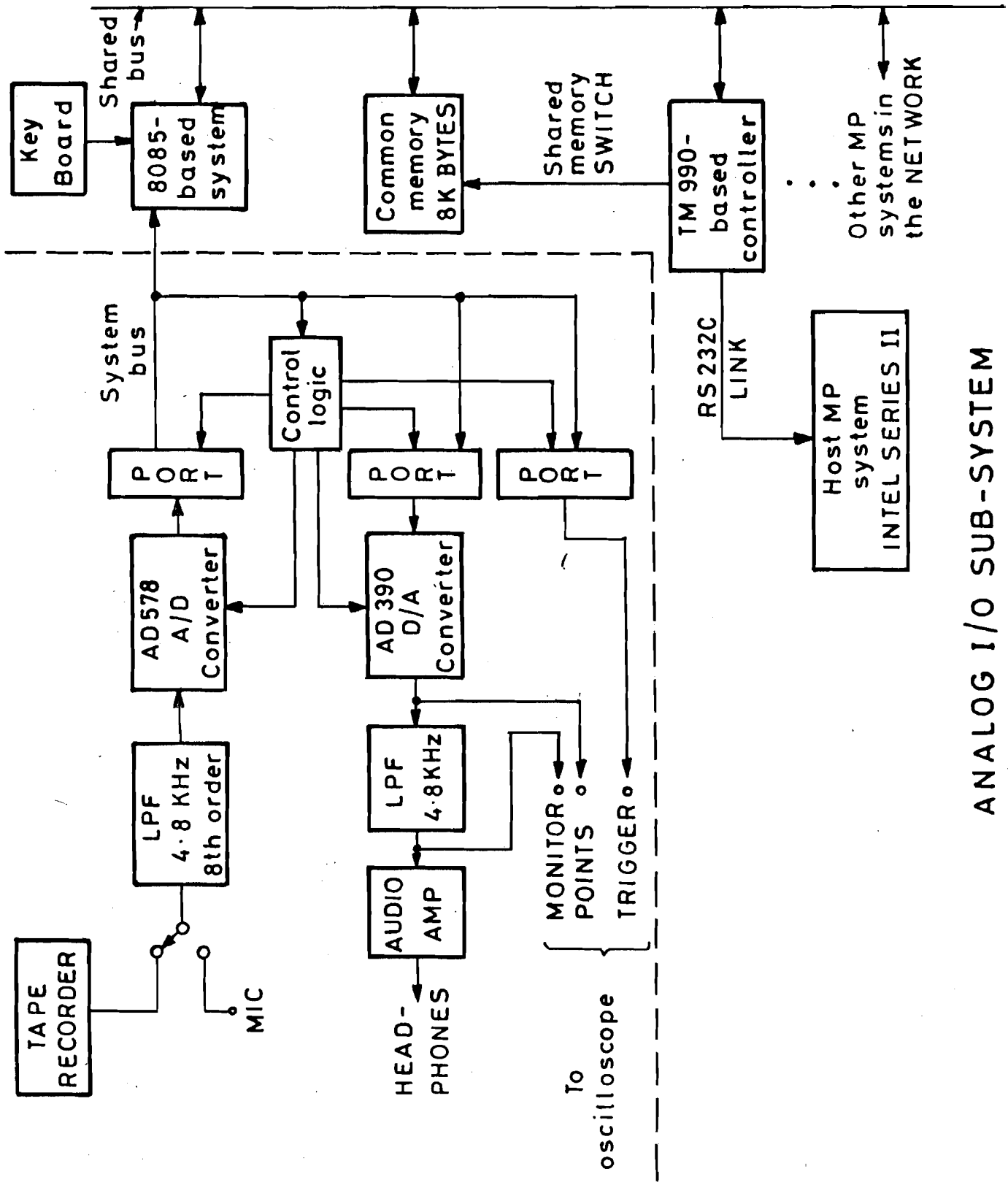
~~~~~~~~~~~~~
~~~~~~~~~~~~~

APPENDIX - I : Analog 1/0 sub-system

The analog 1/0 sub-system is built around an 8085 single-board microcomputer which is connected to a network of micro-computers, sharing a common-memory. The network is linked to an Intel series-II development system through an RS232C serial-communication interface.

The analog input-section consists of an eighth-order low-pass filter with a cut-off frequency of 4.8 kHz followed by a 12-bit successive-approximation Analog-to-Digital Converter AD578. The low-pass filter prevents aliasing effects which may arise due to under-sampling. The analog output is implemented using an AD390 12-bit Digital-to-Analog Converter (DAC), followed by a simple R-C low-pass filter and an audio amplifier. The output of the audio amplifier feeds the head-phones. The analog output of the DAC is brought out to facilitate monitoring of the digitized waveform on an oscilloscope. A block-diagram of the system is shown overleaf.

The analog signal is sampled and digitized at a rate of 10 kHz under control of the micro-computer. The digitized data is stored in the local memory. The user interacts with the system through keyboard-commands. Facilities exist in the system for editing the digitized speech-signal to retain selected portions. Audio output is provided during the sampling and editing process for useful feedback. This also aids in the study of selected regions of the digitized signal. The sampled-and-edited data is uploaded to the Intel development system, and stored as a file on the floppy-disk for subsequent processing. The system includes facilities for down-loading and listening to data-files stored in the floppy-disk.



ANALOG I/O SUB-SYSTEM

APPENDIX - II : Description of Program modules

The different Pascal modules used in the recognition process are for:

1. Parameter extraction
2. Transient detection
3. Amplitude normalization, and,
4. Dynamic programming algorithm for matching and recognition.

Parameter extraction : The digitized speech signal is converted into parametric form for better representation. The parameter extraction is performed by the SPLP module. SPLP contains programs for computing the spectral parameters, the Linear predictor coefficients (LPC), and the cepstral parameters derived from LPC. The digitized samples are windowed using the Hamming window. The spectral parameters are obtained by performing either 128 or 256-point Fast Fourier Transform (FFT), on the windowed data. The FFT results are reduced to either 8 or 16 spectral parameters by summing them up on a log-melspectral scale which is linear up to 1kHz and logarithmic above 1kHz. The spectral parameters are obtained with desired time resolution by performing overlapped FFT with a shift of 32, 64 or 128 samples. The LPCs are computed for the required time-resolution after performing high frequency pre-emphasis and windowing on the digitized data. Eight LPCs are computed, apart from energy, (the first component). The LPCs cannot be used directly in the recognition stage as it is not amenable to distance computation and hence the eighth-order cepstrum of the LPCs are derived and used.

Transient Detection : This program determines the regions of transients in the digitized waveform. The input to this program is the parameter file resulting from the parameter extraction module. The transient parameter is computed using the spectral time-samples in the parameter file. The output of this program is a Transient plot. Peaks in this plot identify regions of maximum variations while valleys indicate regions of steady-state in the input signal. The transient plot is used to identify relevant regions for subsequent processing.

Amplitude Normalization : Amplitude normalization is necessary to overcome mis-recognition due to differences in amplitudes of the utterances. The algorithm used is 'Proportional normalization' which is performed on the spectral parameters. The normalized parameter file is used in the recognition algorithm.

Dynamic Programming algorithm for recognition : Itakura's dynamic time warping algorithm is used to perform time-alignment of the test and reference utterances as well as for distance computation. The recognition program takes in a reference set, consisting of up to five utterances, and a similar test set. The distance is computed for each combination of test and reference words. The reference and the test sets are ordered in such a way that the Performance Index (PIX) is evaluated for each pair of test and reference sets. The program outputs the distance matrix. the recognition score and the PIX-value. The user can select the type of parameters : spectral or LPCs, to be used in matching.

APPENDIX -III: An output showing the PIX-values for overall and consonantal part

REFERENCE SET - R : { KA6, CA1, TA1, HA6, PA5 }
TEST SET - S : { KA5, CA5, TA5, HA5, PA6 }

FRAME-SIZE : 128 Samples / frame
OVERLAPPING : 64 Samples

RSkp16.t Date:C-V interface: Transient. Begin-end: Energy. Cepstral: non-weighted.
Reference set : 16K6 (47) 16C1 (50) 16T1 (50) 16H6 (43) 16P5 (45)
Test set : 16k5 (50) 16c5 (48) 16t5 (48) 16h5 (47) 16p6 (47)

Overall .. with spectral parameters :

995	1589	1513	1327	1473	100	160	152	133	148	10	95	87	66	83	KK	
1188	1023	1383	1154	1276	116	100	135	113	125	32	10	70	26	50	CC	
1180	1580	9e0	952	892	122	163	190	98	92	44	98	15	9	6	TP	
1098	1368	1106	929	925	118	147	119	103	100	36	82	38	10	10	HP	
849	1413	1154	856	788	108	179	146	109	100	18	100	31	19	10	PP	52.50

Number of frames of consonant in test set - 1: 11 2: 9 . 3: 4 4: 7 5: 7

Consonant .. with 8 parameters/frame :

363	591	783	652	799	100	1o3	216	180	220	10	98	100	100	100	KK	
409	243	547	424	446	168	100	225	174	184	100	10	100	100	100	CC	
252	261	92	159	161	274	284	100	173	175	100	100	10	100	100	TT	
265	324	297	194	219	137	167	153	100	113	72	100	88	10	26	HH	
223	267	335	178	160	139	167	209	111	100	74	100	100	22	10	PP	89.00

Consonant .. with 4 parameters/frame :

341	557	767	620	767	100	163	225	182	225	10	98	100	100	100	KK	
313	209	495	328	350	150	100	237	157	167	85	10	100	92	100	CC	
234	245	60	143	149	390	408	100	238	248	100	100	10	100	100	TT	
233	300	269	170	201	137	176	158	100	118	72	100	93	10	36	HH	
213	223	313	150	122	175	183	257	123	100	100	100	100	46	10	PP	91.01

Consonant .. with 2 parameters/frame :

283	407	705	546	669	100	144	249	193	236	10	79	100	100	100	KK	
227	181	429	282	312	125	LOG	265	156	172	50	16	100	91	100	CC	
232	227	54	89	139	430	420	100	165	257	130	100	10	100	100	TT	
199	268	257	118	161	169	227	218	100	136	100	150	100	10	71	HH	
185	189	301	122	72	257	263	418	169	100	100	100	100	100	10	PP	94.55

Vowel .. with Spectral parameters :

632	998	730	675	674	100	158	116	107	107	10	93	32	17	17	KK	
779	780	836	730	830	100	100	107	94	106	10	10	17	7	16	CH	
928	1319	876	793	731	106	151	100	91	83	16	86	10	5	1	TP	
833	1044	809	735	706	113	142	110	100	96	26	77	20	10	8	HP	
626	1146	819	678	628	100	182	130	108	100	10	100	60	18	10	PK	31.71

Vouel .. with 8 Cepstral parameters :

767	979	597	798	732	100	128	78	104	95	10	56	0	14	7	KT	
687	674	684	921	1040	102	100	101	137	154	12	10	11	72	89	CC	
987	1248	760	845	699	130	164	100	111	92	60	99	10	22	6	TP	
974	1327	700	958	766	102	139	73	100	80	12	74	0	10	0	HT	
835	1689	1038	811	827	101	204	126	98	100	11	100	52	9	10	PH	35.21

Overall .. 8 Spec. in consonant and 8 Ceps. in vouel :

1130	1570	1380	1450	1531	100	139	122	128	135	10	74	44	56	70	KK	
1096	917	1231	1345	1486	120	100	134	147	162	40	10	68	82	97	CC	
1239	1509	852	1004	860	145	177	100	118	101	80	100	10	36	11	TT	
1239	1651	997	1152	985	108	143	87	100	86	18	78	3	10	3	HP	
1056	1956	1373	989	987	107	198	139	100	100	17	100	74	10	10	PP	53.05

APPENDIX - IV: An output demonstrating the easy recognition of consonants due to different manners of production

REFERENCE SET - 2 : { KA2, KHA2, GA2, GHA2 }
TEST SET - 1 : { KA1, KHAI, GA1, GHAI }

FRAME-SIZE : 128 Samples / frame
OVERLAPPING : 64 Samples

21KGAE.6T Date:C-V interface: Transient. Begin-end: Energy. Cepstral: non-weighted.
Reference set : 16KA2 (50)16KHA2 (50) 16GA2 (50)16GHA2 (50)
Test set : 16KA1 (50)16KHA1 (50) 16GA1 (50)16GHA1 (50)

Overall .. uith spectral parameters :

1001	3196	3267	3413	100	319	326	341	10	100	100	100	KK
3988	1249	2521	1908	319	100	202	153	100	10	100	88	CC
3403	3260	1148	1642	296	284	100	143	100	100	10	78	TT
3176	3323	1886	1391	228	239	136	100	100	100	71	10	HH 94.75

Number of frames of consonant in test set - 1: 10 2: 18 3: 22 4: 36

Consonant .. uith 8 parameters/frame :

253	1294	577	635	100	511	228	251	10	100	100	100	KK
1561	550	1603	1092	284	100	291	199	100	10	100	100	CC
887	2265	401	740	221	565	100	185	100	100	10	100	TT
1682	2552	1371	922	182	277	149	100	100	100	84	10	HH 98.67

Consonant .. uith 4 parameters/frame :

225	1276	545	597	100	567	242	265	10	100	100	100	KK
1509	466	1511	1010	324	100	324	217	100	10	100	100	CC
809	2243	351	688	230	639	100	196	100	100	10	100	TT
1490	2468	1259	788	189	313	160	100	100	100	95	10	HH 99.58

Consonant .. uith 2 parameters/frame :

157	1274	431	449	100	811	275	286	10	100	100	100	KK
1423	396	1343	860	359	100	339	217	100	10	100	100	CC
757	2227	285	662	266	781	100	232	100	100	10	100	TT
1388	2426	1149	648	214	374	177	100	100	100	100	10	HH 100.00

Vouel .. uith Spectral parameters :

748	1902	2690	2778	100	254	360	371	10	100	100	100	KK
2427	699	918	816	347	100	131	117	100	10	62	34	CC
2516	995	747	902	337	133	100	121	100	66	10	42	TT
1494	771	515	469	319	164	110	100	100	99	20	10	HH 76.92

Vouel .. uith 8 Cepstral parameters :

766	1686	2512	3137	100	220	328	410	10	100	100	100	KK
1760	613	815	731	287	100	133	119	100	10	66	38	CC
1829	882	671	1222	273	131	100	182	100	62	10	100	TT
1143	573	316	419	273	137	75	100	100	72	0	10	HT 78.17

Overall .. 8 Spec, in consonant and 8 Ceps. in vouel :

1019	2980	3089	3772	100	292	303	370	10	100	100	100	KK
3321	1163	2418	1823	286	100	208	157	100	10	100	92	CC
2716	3147	1072	1962	253	294	100	183	100	100	10	100	TT
2825	3125	1687	1341	211	233	126	100	100	100	52	10	HH 95.33

APPENDIX - V : An output demonstrating the dominant features due to different manners of production

REFERENCE SET - I : { KAI, KHAI, GAI, GHAI }
TEST SET - I : { TAI, THAI, DAI, DHAI }

FRAME-SIZE : 128 Samples / frame
OVERLAPPING : 64 Samples

K1TDAE.6T Date:C-V interface: Transient. Begin-end: Energy. Cepstral: non-weighted.
Reference set : 16KA1 (50)16KHA1 (50) 16GA1 (50)16GHA1 (50)
Test set : 16TA1 (50)16THA1 (50) 16DA1 (50)16DHA1 (50) .

Overall .. uith spectral parameters :

1142	2941	3736	3351	100	258	327	293	10	100	100	100	KK
2966	1358	2805	3568	218	100	207	263	100	10	100	100	CC
2500	2376	1399	1992	179	170	100	142	100	100	10	77	TT
3269	3485	1773	1393	235	250	127	100	100	100	54	10	HH 94.25

Number of frames of consonant in test set - 1: 4 2: 16 3: 19 4: 27

Consonant .. uith 8 parameters/frame :

243	360	212	183	100	148	'87	75	10	83	3		OKH
1187	517	1207	1084	230	100	233	210	100	10	100	100	CC
919	1722	334	381	275	516	100	114	100	100	10	28	TT
1363	2478	895	736	1B5	337	122	100	100	100	44	10	HH 71.50

Consonant ,. uith 4 parameters/frame :

215	356	202	175	100	166	94	81	10100		7		OKH
1089	381	1081	998	286	100	284	262	100	10	100	100	CC
833	1694	276	331	302	614	100	120	100	100	10	40	TT
1261	2448	831	652	193	375	127	100	100	100	54	10	HH 75.08

Consonant .. uith 2 parameters/frame :

195	296	200	173	100	152	103	89	10	87	13	4	KH
1033	325	931	828	318	100	286	255	100	10	100	100	CC
697	1550	248	259	281	625	100	104	100	100	10	14	TT
947	2176	773	500	189	435	155	100	100	100	90	10	HH 75.67

Uouel .. uith Spectral parameters :

899	2581	3524	3168	100	287	392	352	10	100	100	100	KK
1779	841	1598	2484	212	100	190	295	100	10	100	100	CC
1581	654	1065	1611	148	61	100	151	83	0	10	86	TC
1906	1007	878	657	290	153	134	100	100	88	68	10	HH 85.42

Vowel .. uith 8 Cepstral parameters :

712	2158	3269	4468	100	303	459	628	10	100	100	100	KK
1325	673	1265	2828	197	100	188	420	100	10	100	100	CC
1009	551	721	1734	140	76	100	240	75	0	10	100	TC
2501	1139	844	758	330	150	111	100	100	85	22	10	HH 81.83

Overall .. 8 Spec. in consonant and 8 Ceps. in vowel :

955	2518	3481	4651	100	264	365	487	10	100	100	100	KK
2512	1190	2472	3912	211	100	208	329	100	10	100	100	CC
1928	2273	1055	2115	183	215	100	200	100	100	10	100	TT
3864	3617	1739	1494	259	242	116	100	100	100	32	10	HH 94.33

APPENDIX -VI: Outputs showing the improvements in the system-performance for selective processing on **CA-BA** set

REFERENCE SET - 2 : { GA2, JA2, DA2, WA2, BA2 }
TEST SET - 3 : { GA3, JA3, DA3, WA3, BA3 }

FRAME-SIZE : 256 Samples / frame,
OVERLAPPING : 224 Samples

OVERALL REGION : **A - E**
CONSONANT REGION : **A - D**

23GBA.23K Date:C-V interface: Transient, Begin-end: Transient.Cepstral: non-weighted.
Reference set : 23GA2 (50) 23JA2 (50) 23DA2 (50) 23WA2 (50) 23BA2 (50)
Test set : 23GA3 (50) 23JA3 (50) 23DA3 (50) 23WA3 (50) 23BA3 (50)

Overall .. with spectral parameters :

1362	2076	1282	1189	1651	100	152	94	87	121	10	87	7	3	42	KH	
1808	1388	1436	1752	1875	130	100	103	126	135	60	10	13	52	70	CC	
1136	1318	914	884	1180	124	144	100	97	129	48	79	10	8	58	TH	
1488	1227	1252	1433	1444	104	86	87	100	101	14	3	3	10	11	HC	
1088	1864	1307	933	849	128	220	154	110	100	56	100	89	20	10	PP	41.15

Number of frames of consonant in test set - 1: 32 2: 50 3: 50 4: 50 5: 48

Consonant .. with 8 parameters/frame :

805	1006	774	891	1197	100	125	96	111	149	10	50	8	22	84	KT	
1808	1388	1436	1752	1875	130	100	103	126	135	60	10	13	52	70	CC	
1136	1318	914	884	1180	124	144	100	97	129	48	79	10	8	58	TH	
1488	1227	1252	1433	1444	104	86	87	100	101	14	3	3	10	11	HC	
1008	1733	1242	880	790	128	219	157	111	100	56	100	92	22	10	PP	42.65

OVERALL / CONSONANT REGION : C - **D**

23GBAT.23M Date:C-V interface: Transient, Begin-end: Transient.Cepstral: non-weighted.
Reference set : 23GA2 (22) 23JA2 (26) 23DA2 (21) 23WA2 (19) 23BA2 (20)
Test set : 23GA3 (19) 23JA3 (33) 23DA3 (18) 23WA3 (18) 23BA3 (18)

Overall ,. with,spectral parameters :

410	778	651	556	635	100	190	159	136	155	10	100	94	71	90	KK	
1991	904	1810	1327	1849	220	100	200	147	205	100	10	100	82	100	CC	
535	785	375	417	449	143	209	100	111	120	78	100	10	22	40	TT	
535	654	436	410	538	130	160	106	100	131	60	95	16	10	62	HH	
507	970	547	488	354	143	274	155	138	100	78	100	90	73	10	PP	77.55

Number of frames of consonant in test set - 1 19 2: 33 3: 18 4: 18 5: 18

Consonant ,. uith 8 parameters/frame :

410	778	651	556	635	100	190	159	136	155	10	100	94	71	90	KK	
1991	904	1810	1327	1849	220	100	200	147	205	100	10	100	82	100	CC	
535	785	375	417	449	143	209	100	111	120	78	100	10	22	40	TT	
535	654	436	410	538	130	160	106	100	131	60	95	16	10	62	HH	
507	970	547	488	354	143	274	155	138	100	78	100	90	73	10	PP	77.55

BIO-DATA

NAME : S.RAMAN.

DATE OF BIRTH : January 1,1944.

QUALIFICATIONS : 1. B.Sc., (Physics)-1962-University of Madras.
2. B.E., (Elecl. & Commn. Engineering) - 1965 - I.I.Sc., - Bangalore.
3. M.S., (Design of a demonstration digital computer) - I. I.T., Madras-1976.

ADDRESS FOR COMMUNICATION : Senior Design Engineer,
Department of Electrical Engineering,
Indian Institute of Technology,
Madras-600 036. India.

ACADEMIC BACKGROUND:

CERTIFICATE/ DEGREE	INSTITUTION	PERIOD	SUBJECTS
B.Sc.,	St-Joseph's college, Tiruchirapalli.	1959-1962	Physics, Maths
B.E.,	Indian Institute of Science, Bangalore.	1962-1965	Electronics & Communication Enginnering
M.S.,	(Part-time)-Indian Inst. of Technology, Madras.	1972-1976	Digital Computer Design

POSITIONS HELD:

DESIGNATION	ESTABLISHMENT	PERIOD
Senior Technical Assistant	Dept. of Elec.Communication Engg., I. I. Sc., Bangalore.	1966-1968
Associate Lecturer	Dept. of Electrical' Engg., I. I.T., Madras.	1969-Aug 1976
Senior Design Engineer	-do-	Since Aug 1976
Visiting Scholar	University of California, Davis, USA.	Spring Quarter 1984

AREAS OF INTEREST: Development of microprocessor-based systems, Digital system design, Speech recognition, Character recognition.