Asia-Pacific Engineering Journal (Part A), Vol. 1, No. 2 (1991) 197–217
@World Scientific Publishing Company

SIGNAL PROCESSING ISSUES IN REALIZING VOICE INPUT TO COMPUTERS*

B. YEGNANARAYANA and R. SUNDAR
Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India

Received 31 October 1991

ABSTRACT

In this paper we discuss some issues in processing speech signals, especially for isolated utterances of characters of a language. For processing this speech signal we have no clues of higher level linguistic information such as prosodics, lexical, syntax, and semantics. Any representation of signals in terms of fixed parameters for each short (10–20 msec) segment is not likely to provide the distinguishing features of the sounds of the characters for recognition. Processing of speech signal based on the knowledge of acoustic-phonetics of the characters of the language will enable us to identify features for discriminating different sounds. We discuss how these features are related to the parameters of the signal. For illustration, we consider the acoustic-phonetic knowledge of the Indian language Hindi. The discussion in this paper shows the need for new methods of processing signals to realize voice input to a computer.

Keywords: speech processing; isolated characters; voice input; computer interface.

1. Introduction

The main objective of this paper is to highlight the need for new algorithms to process speech signals in order to extract features needed for recognition. The goal is to develop a recognition system for isolated utterances of characters of an Indian language with a view to provide an alternative to keyboard for entry of data/information into a computer. The sequence of acoustic events in speech production for the characters of an Indian language are well defined. The manifestation of these events in the speech signal is such that standard signal processing methods [1] are not adequate in many cases to extract the features from the signal. We discuss in this paper the nature of the problem, issues in processing speech signal for extraction of the acoustic features, and some methods to solve the problem.

Speech mode of communication with a machine involves two major components. One is the recognition of characters from speech signal and the second component

*Invited paper.

is the protocol for interaction with a machine in speech mode. Our interest is in the problem of recognition of isolated utterances of characters of an Indian language. Note that in isolated utterances of characters we cannot take advantage of the higher level knowledge sources such as lexical, syntax, semantics and prosodics. We also do not have the higher level phonetic knowledge such as coarticulation. Therefore, for recognition, it is not possible to use a simple approximate representation (in arbitrary symbols such as phones) of speech signal and use higher level knowledge sources to disambiguate the sequences of phones [2].

Recognition of the characters has to take place by processing the signal to extract the acoustic features accurately. The only knowledge we have is the description of speech production of each character in terms of the acoustic features. The features and character description in terms of these features are generally unique for a given Indian language. The rules for production of utterances of characters of a language constitute the acoustic-phonetic knowledge. It is necessary to analyse the speech signal with the help of this knowledge to derive the acoustic features first, and then the description of the character in terms of the derived acoustic features. Since the production rules are unique, there is little variability in the description from speaker to speaker. Since the utterances are for characters in isolation, there is hardly any coarticulation effects that occur due to the higher level phonetics of connected speech. Absence of coarticulation effects reduces the variability in the production of these characters from speaker to speaker and for repetitions by the same speaker. But absence of clues from higher level linguistic knowledge sources makes the task of recognition dependent solely on the feature extraction from the speech signal based on the acoustic-phonetic knowledge.

Since the purpose of speech analysis here is for recognition, and not for storage of information in a compressed form, analysis for feature extraction from signal is more important than analysis for parametric representation of speech signal. Issues of noise and distortion, though important, are not considered here, because we do not yet have reliable methods for extraction of many important features such as burst, aspiration and formant transition even from clean speech signal. It is assumed that the production rules for each character are strictly followed and hence all the acoustic features for that character are present in the specified sequence. The question is how to analyse the speech signal to extract these features. Note that while there is little speaker variability in the production of a given combination of features for a character, individual speaker characteristics are reflected in the signal due to differences in the vocal tract dimensions and in the vocal fold vibrations.

Most methods of speech analysis assume a gross source-system model for speech production. The features are interpreted from the parameters of this model. Some of the models are [1]:

(a) All-pole model or pole-zero model excited by random noise sequence or modeled glottal pulse sequence.

(b) Gross spectral shape represented in terms of filterbank spectral parameters and fine spectral structure represented in terms of unvoiced or voiced source of excitation.

These methods perform analysis of a fixed size (10–20 msecs) data window. The disadvantage of this approach is that if the feature information is not captured in these gross parameters, there is no way the information can be recovered later as the redundancy of higher level knowledge sources is absent in the signal.

The problem is that the gross model of speech production is not based on production of specific sounds of a given language. For example, the parametric representation is the same for sounds for a tonal language like Chinese, phonetic language like Hindi and nonphonetic language like English. Here the terms phonetic and nonphonetic are used to highlight the correspondence and lack of correspondence, respectively, between the text and its pronounciation. Each of these languages has unique sound units for the speech of the language. These sound units are produced by specific dynamics of the vocal tract system (which includes nasal tract) and source of excitation of the system. The production of these sounds is reflected in the acoustic features such as resonances (formants) and antiresonances of the vocal tract system, nature of excitation (voiced or unvoiced or a combination), characteristics of the released burst, formant transitions, aspiration and vowel onset instant.

The problem of feature extraction .from speech signal needs to be addressed in a way different from the conventional methods based on fixed gross model of speech production. Speech signal processing should depend on the type of features which in turn should depend on the dynamics of the source and vocal tract system for the production of utterances of legal sound units of a language. The acoustic-phonetic knowledge which describes the dynamics of the production of characters of a language is described for Hindi (an Indian language) in Sec. 2. In Sec. 3 we present methods for extracting features in terms of parameters of the speech signal for different categories of sounds. The section highlights the issues and problems to be resolved for extracting the acoustic features from speech signal corresponding to the isolated utterances of the characters of Hindi.

2. Acoustic-phonetic Knowledge of Characters of Hindi

In this section we discuss the acoustic-phonetic knowledge of Hindi which consists of organization of the character set and the rules for speech production of the characters. We also discuss how this acoustic-phonetic knowledge in the form of acoustic events is manifested as a set of parameters and features in the signal.

2.1. Characters of Hindi

Hindi language exhibits phonetic nature which is explicitly brought out in the organization of the alphabet (See Table 1). There are two broad categories in the alphabet – vowels and consonants. The vowels of Hindi include both vowel and diphthongal sounds. There is a phonetic distinction between short and long forms of some of the vowel sounds which are represented by distinct symbols.

Table 1. The Hindi alphabet

VOWELS (V)
Vowels and **Diphthongs**

अ /^/	आ /A/	इ /I/	ई /i/	उ /U/	ऊ /u/
ए /e/	ऐ /aj/	ओ /o/	औ /aw/		

CONSONANTS (C)
Stops and nasals

PLACE OF ARTICULATION	MANNER OF ARTICULATION				
	UNVOICED		*VOICED*		*NASAL*
	UNASPIRATED	*ASPIRATED*	*UNASPIRATED*	*ASPIRATED*	
VELAR	क /k/	ख् /k^h/	ग् /g/	घ् /g^h/	ङ् /g-/
PALATAL	च् /tS/	छ् /tS^h/	ज् /dZ/	झ् /dZ^h/	ञ् /G/
RETROFLEX	ट् /t/	ठ् /t^h/	ड् /d/	ढ् /d^h/	ण् /N/
DENTIALVEOLAR	त् /T/	थ् /T^h/	द् /D/	ध् /D^h/	न् /n/
BILABIAL	प् /p/	फ् /p^h/	ब् /b/	भ् /b^h/	म् /m/

Semi vowels

य् /j/	र् /r/	ल् /l/	व् /v/
PALATAL	*ALVEOLAR TRILL*	*ALVEOLAR LATERAL*	*LABIODENTAL*

Fricatives

श् /s'/	ष् /S/	स् /s/	ह् /h/
PALATAL	*RETROFLEX*	*ALVEOLAR*	*GLOTTAL*

Notes :
1) Characters in Hindi occur in the following forms : C, V, CV, CCV. CCCV
(V - stands for any vowel from the vowel set and.
C - stands for any consonant from the consonant set)
2) Symbols are indicated in both Hindi script and Computer phonetic alphabet CPA
3) Phonetic symbols used are : CPA after Lenning & Brassard [3] with modification for aspiration as suggested by Ganesen [4]

Consonants in Hindi are subdivided into three groups - stop and nasal consonants, semivowels and fricatives. The structure within each group reflects different places of articulation and manners of production of consonants. The five rows in the stop and nasal group correspond to the five different places of articulation – velar, palatal, retroflex, dentialveoloar and bilabial. The columnwise arrangement reflects grouping based on the manner of production of the consonants. The first four columns belong to the nonnasalized category. Consonants in the fifth column contain the nasal consonants. Consonants in the first and second columns belong to the unvoiced category. Glottal excitation is absent during the production of these consonants. The third and fourth columns represent the voiced consonants. In these cases glottal excitation starts prior to the consonant release. Consonants are also categorized on the basis of presence or absence of aspiration. Consonants in the first and third columns are unaspirated, whereas the consonants in the second and fourth columns are aspirated. Nasal consonants are by nature, voiced and unaspirated. Semivowels too are naturally voiced and unaspirated, and are differentiated based only on the place of articulation. The order of arrangement is from the inner to the outer point of articulatory stricture in the mouth – palatal, alveolar (trill), alveolar (lateral) and labiodental. Fricatives in Hindi are always unvoiced. The order of arrangement in the table is based on the point of stricture in the mouth – palatal, retroflex, alveolar and glottal.

Characters in Hindi are combinations of consonants (C) and vowels (V). They occur in the form C, V, CV, CCV and CCCV, where C is any consonant from the consonant group and V is any vowel from the vowel group. In this study we consider only utterances of the type V and CV.

2.2. Speech production mechanism

Phonetics deals with the systematic study of human speech sound units in order to describe and classify them. For want of a better description, phonetic features are described usually in terms of the articulatory process (see Fig. 1). As speech sounds are characterized by both dynamic and static behavior of the articulators, the description of speech sound units is based on the articulatory positions, articulatory dynamics and sequence of events.

Among the two broad classes of speech sounds (vowels and consonants), vowels are produced with a relatively free passage of air stream in the vocal tract with the vocal cords set into strong vibrations. The quality of a vowel is a function of the shape of the vocal tract which is dictated by the tongue and lips. The velum too plays a role in that it dictates whether the nasal tract is coupled to the vocal tract or not.

Consonants are produced when the air stream through the vocal tract is obstructed in some way by the articulators. These positions are followed by a release towards an open tract configuration of the following vowel. The consonants differ depending on the place where the obstruction takes place as well as the manner of articulation in the production of the consonant. The obstruction can occur in many ways. Stop consonants are produced by an abrupt release of the built-up air pressure at a point of complete closure in the oral tract. When the articulators are brought close together creating a narrow constriction, the air flow is partially obstructed leading to turbulence at the constriction. This results in fricative sounds.

The presence/absence of vocal cords vibration before the articulatory release determines whether the consonant is voiced or unvoiced. Stop consonants with aspirated manner are produced when the stop release is followed by a period wherein the glottis is constricted with a slight opening. This leads to turbulence at the glottis. The turbulence after passing through the vocal tract results in aspiration. Differentiation of consonants based on the place of articulation is determined by the position at which maximum constriction is made in the vocal tract.

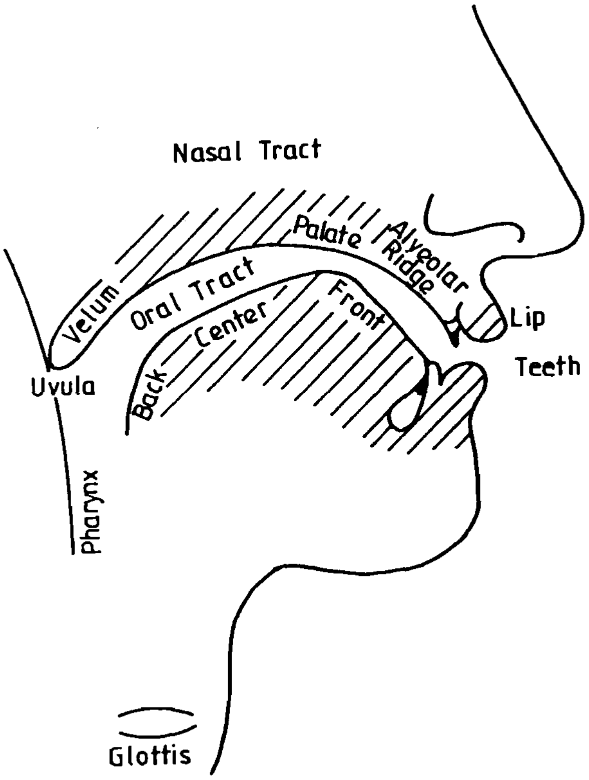


Fig. 1. Sketch of the parts of human speech production mechanism. Specific position of articulation is for the front vowel / i /.

From the above discussion it is obvious that all the meaningful sounds we produce are well defined. Sequences in which these sounds are produced are also unique. But the sounds do not conform to assumptions made in the usual models of speech production for analysis purposes. Some of the problems are the following:

Dynamic behavior of the source and system.

Coupling between the system and source and between the vocal tract and nasal tract.

Excitation at different points in the vocal tract system.

Multiple excitation.

Influence in the signal due to coarticulation.

In most cases the information is available in the signal and is visible (and perceivable) in the temporal and spectral domains as in the signal waveform and spectrogram, respectively. The problem is lack of suitable signal processing techniques to extract the relevant information.

2.3. Nature of speech signal

A typical speech waveform and its characteristics are illustrated in Fig. 2 taking an example of the case of unvoiced aspirated utterance [tS$^{h\wedge}$]. The signal shows the initial burst, frication, aspiration and voicing characteristics of the vowel. The periodicity in the vowel segment is called pitch period. The damped sinusoidal behavior within each pitch period reflects the formant structure.

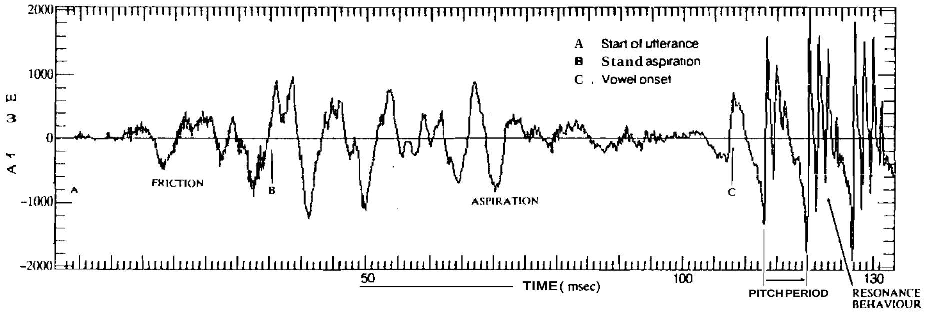


Fig. 2. Speech waveform for the utterance [tS$^{h\wedge}$].

2.4. Methods of processing speech signals

Most methods of speech analysis are based on the assumption of quasistationarity of the signal over short 10–20 msec of data and a source-system model within the selected segment [1]. Time domain parameters (like energy, amplitude, zero crossing, autocorrelation, pitch period, etc.) and frequency domain parameters (like energy in spectral bands, smoothed spectral envelope, formant frequencies, pitch frequency, etc.) are computed for each segment of data. Even when a model like AutoRegressive (AR) or AutoRegressive Moving Average (ARMA) is assumed, the parameters are still extracted from each fixed duration segment.

The main issue in the standard signal processing methods is the uniform interval (10–20 msec) analysis. Such an analysis will fail to capture the abrupt changes in the characteristics of the signal that occur in most of the consonantal sounds. This will also fail to capture the vowel onset instant after aspiration. Thus the various events that take place during production of consonants will not be represented well in the parameter contours.

3. Acoustic Feature Extraction

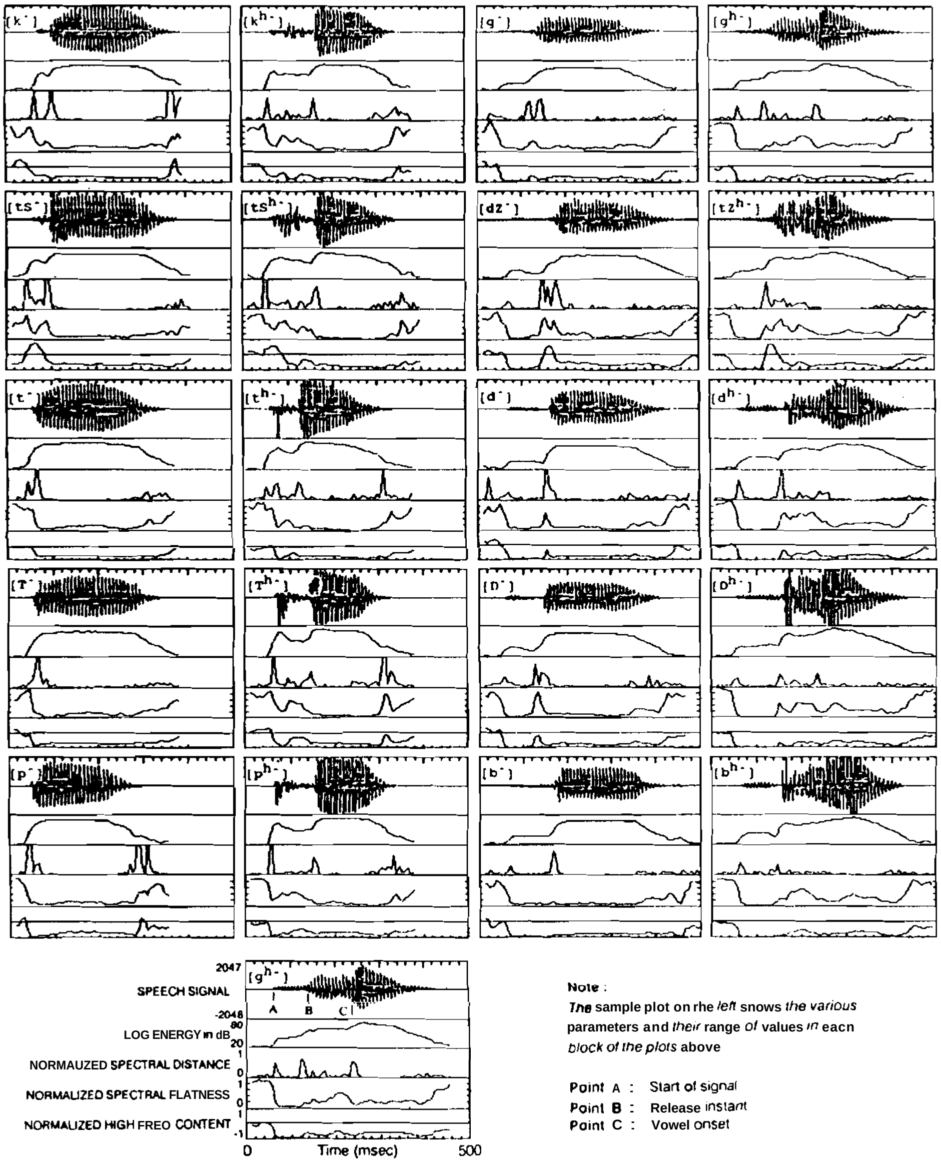
3.1. Scope of this study

In this section we discuss issues in signal processing for acoustic feature extraction and show in some cases how new algorithms are needed to solve the problems. The objective is mainly to highlight the issues, as we have no solution to many of the problems yet. We assume that all the characters are uttered strictly as per the rules of acoustic-phonetics of the language. We also manually mark the regions where the acoustic events take place. We consider the features for each category of characters as listed in alphabet. We discuss what acoustic features are needed to describe each category and the characteristics of the speech production system that needs to be considered for analysis of speech to extract the features. Note that by manual segmentation into regions of acoustic events, we have overcome the effects of choice of arbitrary placement of the analysis window. Automatic implementation of this segmentation is extremely difficult. Note that we are not talking about automatic identification of features or automatic recognition of characters from speech signal at this stage. We are mainly addressing the issue of how to obtain parameters that best describe the acoustic features. As we would notice, parameters for each feature vary over a range of values making it difficult to identify characters uniquely within each category.

3.2. Acoustic features and their manifestation in parameters

For each category of sounds we give a description in terms of acoustic features, and discuss issues involved in processing the signal. The following discussion is based on studies using several samples of speech signals corresponding to the utterances of the characters of Hindi. Signal waveforms along with some parameter contours for one set of characters are given in Figs. 3 and 4. All parameters are computed using an analysis segment of 25.6 msec with an overlap of 19.2 msec between adjacent frames. In the parameters, spectral distance is computed between adjacent frames using Itakura distance [1]. Spectral flatness is computed using the normalized error in linear prediction analysis [1]. The high frequency content refers to the ratio of energy in the high frequency (1–5 kHz) and low frequency (0–1 kHz) regions.

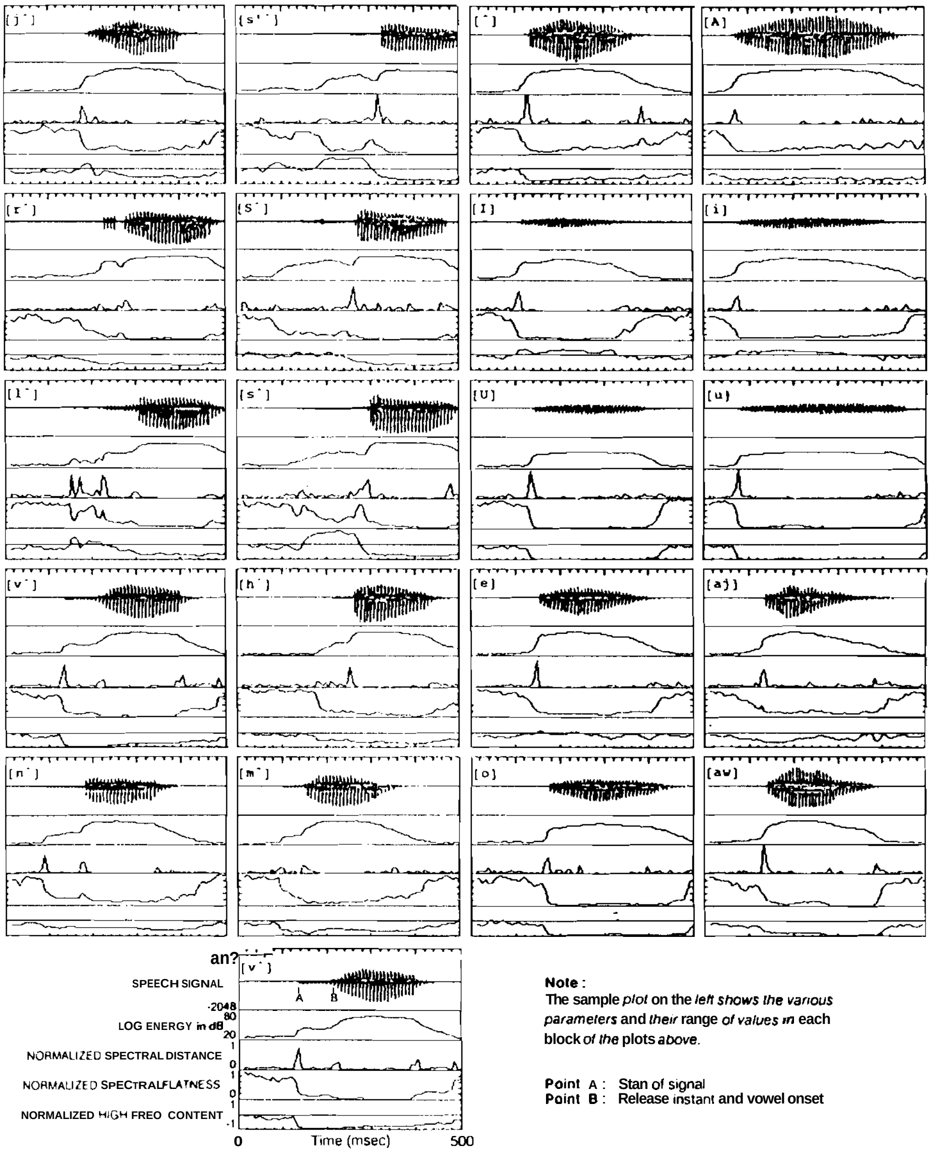
First, the alphabet is grouped into four broad categories (A, B, C and D) mainly based on the degree of stricture that occurs in the vocal tract during production of the characters in each group. The grouping of characters shown in Table 2 is similar to the grouping given in Table 1 in Sec. 2, except that nasals are also included in the semivowel category due to similarities in the signal waveforms. Each of these categories is further subdivided to form groups according to manner and place of articulation. These are indicated by categories E through S in Table 3. We shall discuss the distinguishing acoustic features for each category as well as for members within each category. We shall also indicate important parameters and their behavior for identifying the features.



Note :
The sample plot on rhe left snows the various parameters and their range of values in eacn block of the plots above

Point A : Start of signal
Point B : Release instant
Point C : Vowel onset

Fig. 3. Signal and parameter contours for utterances of Hindi Stop consonants.



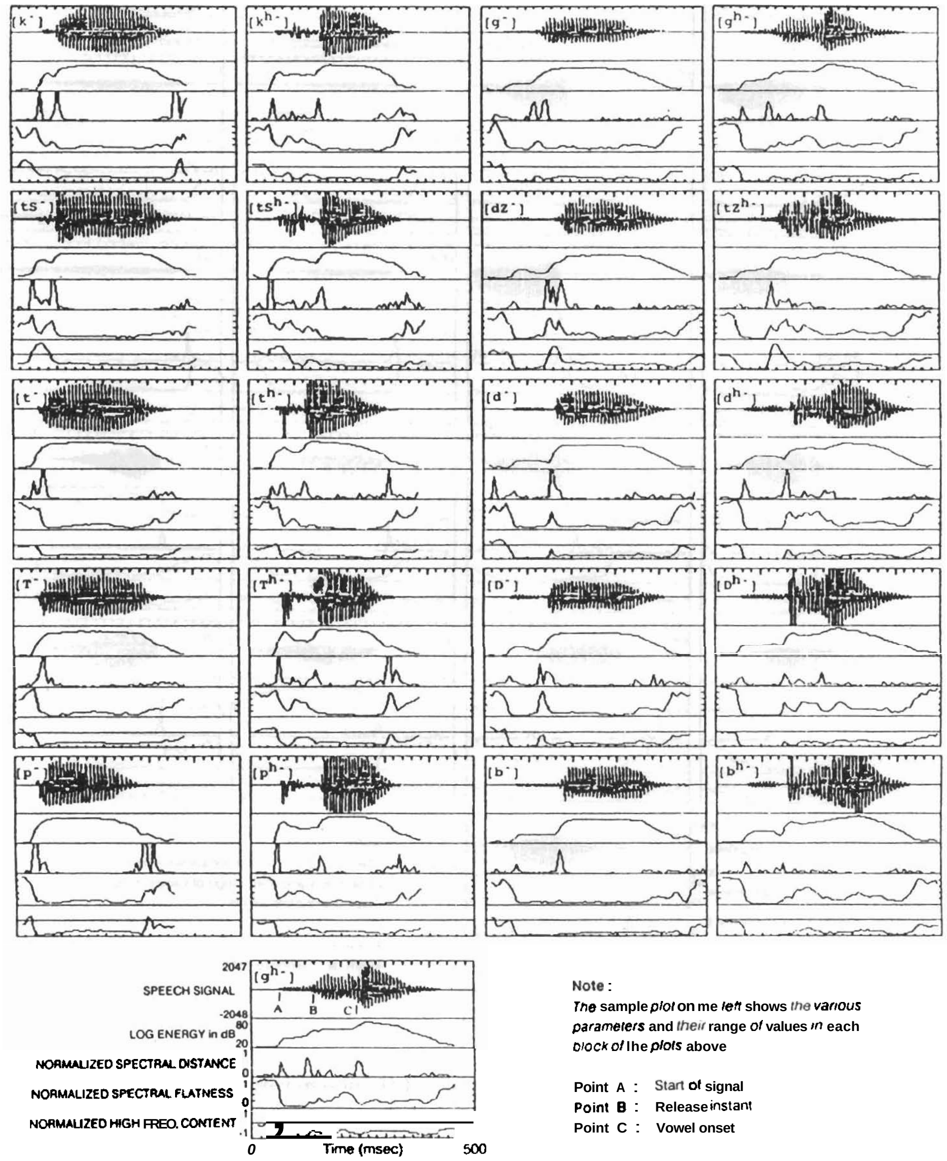
Note :
The sample *plot* on the *left shows the various parameters* and *their* range *of values in* each block *of the* plots *above.*

Point A : Stan of signal
Point B : Release instant and vowel onset

Fig. 4. Signal and parameter contours for utterances of Hindi senuvowels, fricatives, nasals, vowels and diphthongs.

Category A consists of vowels (short and long) and diphthongs. These correspond to the case where the vocal tract is open during production. We can call the tract in this configuration the resonance tract. The source of excitation of the tract

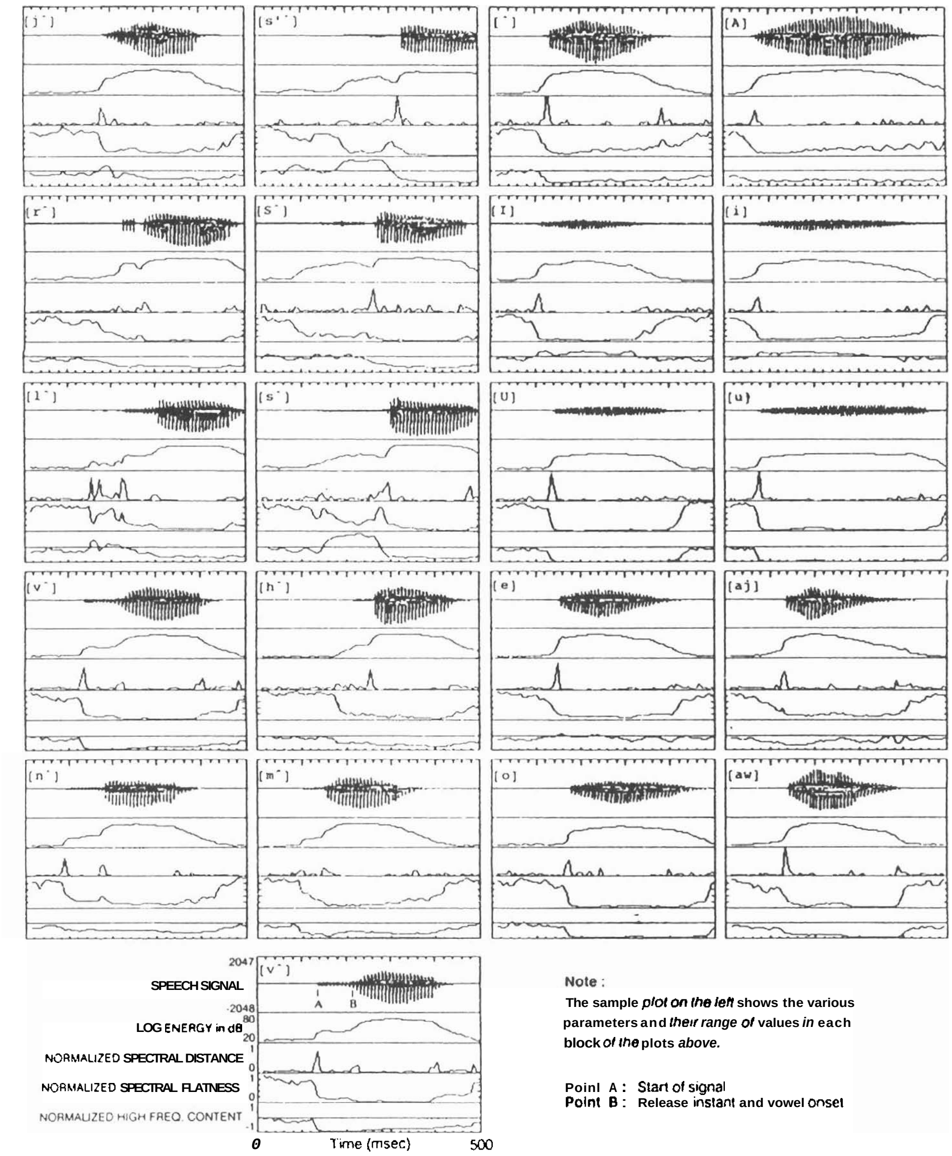
shall discuss the distinguishing acoustic features for each category as well as for members within each category. We shall also indicate important parameters and their behavior for identifying the features.



Note :
The sample *plot* on me *left* shows *the various parameters* and *their* range *of* values *in* each *block of* Ihe *plots* above

Point A : Start of signal
Point B : Release instant
Point C : Vowel onset

Fig. 3. Signal and parameter contours for utterances of Hindi Stop consonants.



Note :
The sample *plot on the left* shows the various parameters and *their range of* values *in* each block *of the* plots *above.*

Poinl A : Start of signal
Point B : Release instant and vowel onset

Fig. 4. Signal and parameter contours for utterances of Hindi semivowels, fricatives, nasals, vowels and diphthongs.

Category **A** consists of vowels (short and long) and diphthongs. These correspond to the case where the vocal tract is open during production. We can **call** the tract in this configuration the resonance tract. The source of excitation of the tract

Table 2. The Hindi alphabet rearranged to reflect classification based on features.

VOWELS AND DIPHTHONGS

Tongue hump position	Back	Back	Central/ Back	Front	Front	Central to Front	Back
Tongue hump height	Mid	Low	Low	Mid	High	Low to High	Low to Mid
Lips rounding	Yes	Yes	No	No	No	No	No to Yes
Length of vowel							
Short	/ U /	—	/^/	—	/ I /	—	—
Long	/ u /	/ o /	/ A /	/ e /	/ i /	/ aj /	/ aw /
	VOWELS					*DIPHTHONGS*	

SEMIVOWEL AND NASAL CONSONANTS

Coupling of nasal tract *by lowering of velum*	**Coupled**	**Decoupled**
Sub group	Nasal	**Semivowels**

NASALS

Velar	Palatal	**Retroflex**	**Dentialveolar**	Bilabial
/ g- /	/ G /	/ N▮	/ n /	/ m /

SEMIVOWELS

PLACE OF ARTICULATORY RELEASE			
Palatal	**Alveolar** trill	**Alveolar lateral**	**Labiodental**
/ j /	/ r /	/ l /	/ v /

FRICATIVE CONSONANTS

PLACE OF ARTICULATORY RELEASE			
Glottal	Palatal	**Retroflex**	**Alveolar**
/ h /	/ s' /	/ S▮	/ s /

STOP CONSONANTS

MANNER OF ARTICULATION		PLACE OF ARTICULATORY RELEASE				
		Velar	Palatal	**Retroflex**	**Dentialveolar**	**Bilabial**
UNVOICED	Unaspirated	▮k /	/ tS /	/ t /	/ T /	/ p /
	Aspirated	/ k^h /	/ tS^h /	/ t^h /	/ T^h /	/ p^h /
VOICED	Unaspirated	/ g /	/ dZ /	/ d /	/ D▮	▮b /
	Aspirated	/ g^h /	/ dZ^h /	/ d^h /	/ D^h /	/ b^h /

is the glottal vibration. The characteristic feature of this group is a strong periodic glottal vibration exciting the resonance tract. The signal exhibits periodicity due to voice pitch and formant structure due to vocal tract resonances. Both periodicity and formant structure are steady during the production of vowels. This also results in steady spectral characteristics throughout. For diphthongs the spectral structure varies slowly from beginning till end.

Table 3. Feature based classification of Hindi alphabet.

BROAD CATEGORIES

A.	Open tract	(vowels and dipthongs)
B.	Approximant tract	(Semivowels and nasal consonants)
C.	Narrow tract	(fricative consonants)
D.	Closed tract	(stop consonants)

VOWELS

E.	Position cl tongue hump	front - central - back
F.	Height of tongue hump	low - mid - high
G.	Shape formed by lips	unrounded - rounded
H.	Length of the vowel	short - long

DIPHTHONGS

I.	Change in position cl tongue hump	central to front (/ aj /)	——
J.	Change in height cl tongue hump	low to hlgh (/ aj /)	low to mid (aw /)
K.	Change in shape formed by lips	——	unround to round (law /)

NASAL CONSONANTS

L. velar - palatal - retroflex - dentialveolar - bilabial

SEMIVOWEL CONSONANTS

M. palatal - alveolar trill - alveolar lateral - labiodental

FRICATIVE CONSONANTS

N. glottal - palatal - retroflex - alveolar

STOP CONSONANTS

Manner:

O. unvoiced unaspirated
P. unvoiced aspirated
Q. voiced unaspirated
R. voiced aspirated

Place:

S. velar - palatal - retroflex - dentialveolar - bilabial

Category B consists of semivowels which correspond to slightly open vocal tract during production and nasals in which the nasal tract is coupled to the vocal tract. The initial part of the semivowel is produced by the periodic glottal excitation of a slightly open vocal tract. Therefore the signal prior to articulatory release exhibits periodicity and formant structure. There is change in the energy at the articulatory release, but changes in the formant structure are smooth into the following vowel region. The vowel onset takes place along with the release. This is because the vocal cords are in vibration even prior to the release, and the release which causes opening of the vocal tract sets the vowel onset too. There is significant change in the spectral characteristics before and after release. The nasals exhibit the same behavior but have the additional quality of nasalization in the consonantal as well as in the initial part of the vowel region.

All fricatives belong to category C, where the initial excitation is typically produced due to the turbulence of air at a narrow constriction in the vocal tract. Hindi fricatives are always produced as unvoiced. The vowel onset occurs immediately after the release, and the duration between release and vowel onset is nearly zero. The sailent features of this group are the presence of high frequency energy due to noise like nature in the signal and the absence of periodic glottal excitation prior to the release. The duration of the signal before release is also important in distinguishing this group from other groups. There is a significant change both in source and system characteristics after the vowel onset. The gross spectral characteristics are nearly uniform during frication followed by an abrupt change after release.

The utterances that come under category D are stop consonants. Due to the abrupt nature of the release, the signal exhibits a distinct characteristic in the form of a "burst" immediately after release. In the case of voiced stops the release is preceded by a low energy .voicing without any high frequency content. Due to the short duration of the burst signal, it is difficult to use this as a reliable feature. The unaspirated stops show a distinct increase in the pitch period immediately after vowel onset. Nonuniform spectral characteristics in the region upto vowel onset followed by rapid changes in the formant as well as pitch period are significant characteristics of this group.

The acoustic features such as glottal vibration, pitch periodicity, formant structure, steadiness in pitch and formant structure, and the like, can be derived from parameters extracted from speech signal. The relevant features and the corresponding parameters are listed in Table **4.** The significant features and parameter trends for the broad categories A, B, C and D are illustrated in Table 5. One point to note is that most of the parameters indicated are gross parameters and hence show very low dependence on speaker characteristics.

Next we consider features for discriminating among members of each of these broad categories. First we consider group A (vowels and diphthongs). There are three classes in this group - short vowels, long vowels and diphthongs. The short vowels and long vowels differ in the shape of the vocal tract during their production. The shape of the vocal tract is determined by the position of the articulators - tongue and lips. The shape of the vocal tract can be described along three dimensions - position of the tongue hump and shape at the lips, as indicated in Table **3** (categories E, F and G). The choice of this form of description permits a grada-

Table 4. Features in the signal and the corresponding parameters that bring out these features.

	Features	Parameters
1.	Strengthof signal	- Average energy
2.	Steadinessin signal strength	- Average energy diierence between neighbouring frames
3.	Change in signal strength	- Average energy difference between neighbouring frames
4.	Steadinessin spectral characteristics	- Spectral diince
5.	Change in spectral characteristics	- Spectral distance
6.	Extent of high frequency components	- High-low freq. energy ratio. first linear prediction coefficient
7.	Waveform periodicity	- Pitch
8.	Steadiness in periodicity	- Pitch difference between neighbouring frames
9.	Glottal pulse behaviour	- Glottal roll off measure
10.	Formant structure	- Spectral flatness
11.	Steadiness In formant structure	- Spectral distance and spectral flatness
12.	Continuity cl formant structure	- Spectral flatness difference between adjacent frames
13.	Formant structure – absence to presence	- Spectral distance
14.	Presenceof nasalization	- Detection of antiresonancein the gross spectrum
15.	Resonance transitions	- Formant changes
16.	Start cl signal	
	– Increasein signal strength from silence	
	– Change in spectral characteristics	
	– Changefrom flat to structuredspectrum	
17.	Release instant	
	– Increase In signal strength,	
	– Change in spectral characteristics	
	– Change In spectral flatness	
	– Change in extent cl hlgh frequencycomponent.	
18.	Vowel onset	
	– Increase in signal strength to high value	
	– Increase In strength of glottal pulses	
	– Change to presenceof distinct formant structure.	
	– Change to steadiness In spectral characteristics	

tion within each dimension independently. The diphthongs which exhibit changing articulatory position can be described in terms of changes along these dimensions. This is indicated in Table 3 (categories I, J and K).

Parameters which bring out gradation in the indicated dimensions can be derived from spectral characteristics. The first two formants F1 and F2 are good cues for this purpose [5]. The movement of the position of the tongue hump from back to front is reflected in the increase of F2. The gradation of the height together with the position of the tongue hump is reflected in the movement along a line with negative slope, starting with large F1 and small F2 for low height, to small F1 and large F2 for high height. The lip rounding feature is reflected along a line with positive slope starting with small F1 and small F2 for rounded lips, to large F1 and/or F2 for unrounded lips.

Table 5. Features and parameter trends for broad categories.

Characteristics of significant events

Start point	– energy	-very low to hgh
	– energy **difference**	- high
	– spectral distance	- high
	– **spectral** flatness	-distinct decrease
Release point	– energy **difference**	- high (increase)
	–spectral distance	- high
Vowel onset	– energy	- increase to **high** level
	– **pitch**	**- present**
	– **glottal pulse behaviour**	- stronger **glottal vibration**
	– **spectral flatness**	- decrease to **low**
	– **spectral distance**	- decrease to **low**

Significant features

1. Strength cl signal
2. Steadiness in signal **strength**
3. Steadiness in spectral characteristics
4. Extent cl **high** frequency **components**
5. Presence cl **waveform** periodicity
6. **Glottal pulse behaviour**
7. Presence cl **formant** structure
8. Steadiness cl **formant** structure
9. Occurrence cl **articulatory release**
10. Instant cl vowel onset
11. Duration **characteristics**

Behaviour of parameters

PARAMETERS	CATEGORIES			
	OPEN TRACT	***APPROXIMANT*** TRACT	***NARROW*** TRACT	CLOSED TRACT
Average energy	**high**	mid range	mid range to **high**	——
Energy difference	very low	low (except I r /)	low	**high at release**
Spectral distance	very low	low	**mid range**	**high at release**
Spectral flatness	very **low**	**low**	**low to** mid range	——
High frequency content	**mid** ***range***	mid range	**high**	——
Pitch	present	present	absent	——
Pitch difference	low	low	——	——
Glottal pulse ***behaviour***	**strong**	strong	**mid** range	——
Presence *of* release	no	Yes	**yes**	**yes**
DURATION				
Start to release	——	> minimum	**> minimum**	——
Release to vowel onset	——	~ 0	~ 0	> mlnlmum
Vowel onset to end	> minimum	——	——	——

The diphthong /aj/ can be identified from the movement of formants from large F1 and small F2 to small F1 and large F2. Similarly, the diphthong /aw/ can be identified from the movement of formants from large F1 and small F2 to small F1 and small F2. These follow trends as indicated by the movement of articulators from the initial equivalent vowel position to the final equivalent vowel position for a diphthong.

Category B consists of two classes – semivowels and nasals. As mentioned before, these were grouped together because of similarities in the gross nature of the waveforms. The nasalization in the signal before and after release is an important feature for distinguishing nasals from semivowels. This calls for deriving parameters which can detect antiresonance in the spectral characteristics. Further distinction within each class is based on the place of articulation. The nasals (category L) have five places of articulatory release similar to nonnasals. The differences among them is reflected mainly in the formant transition after release. For semivowels (category M), there are four places of articulatory release – palatal, alveolar trill, alveolar lateral and labiodental. The palatal semivowel exhibits gradual increase in signal amplitude from the start of signal. The signal prior to release is also vowel like. The alveolar trill exhibits a modulation of signal amplitude before release. The alveolar lateral exhibits spectral characteristics which characterize the "lateral" voicing before release. The labiodental also exhibits unique spectral characteristics of the region before release. The formant transitions after release are also good cues in discriminating among semivowels. Table 6 lists the significant features and parameter trends for discriminating members in this group.

There are four types in the fricative group (category N), each type is distinguished by the place of constriction – glottal, alveolar, retroflex and palatal. The spectral characteristics of these in the frication region are distinct because the energy concentration is different in different frequency bands for each of these cases. In addition, the glottal frication exhibits formant structure depending on the following vowel. In this case there are no formant changes after release since the glottal release does not change the vocal tract shape. The other three fricatives have distinct formant transitions depending on the vowel following release. The significant features and the parameter trends for discriminating the members in this group are listed in Table 7.

Stop consonants are grouped into four categories (O, P, Q and R) depending on the manner of production – unvoiced unaspirated, unvoiced aspirated, voiced unaspirated and voiced aspirated. The unvoiced/voiced feature is reflected in the absence/presence of weak voicing before the release of the stop. In the unvoiced case the signal starts only after the release instant. The voiced stop is characterized by the presence of voicing without formant structure and hence has energy only in the region of the pitch frequency. The unaspirated stop shows significant energy in the signal immediately after release in the form of a burst. In the case of aspirated stops there is signal for a significant duration after the release upto vowel onset. This signal is characterized by the presence of formant structure similar to that of the following vowel, but without any periodicity. Even in the case of velar and palatal unaspirated release there is significant duration between the release and vowel onset, but the absence of formant structure distinguishes them from the aspirated stops.

Table 6. **Features and parameters trends** lor nasals and **semivowels.**

Significant **features**

Before ***release***

1. Nasal characteristics

   Separates the nasals from the semivowels

   (Calls for detection of antiresonance in the spectral characterstics)

2. **Energy contour**

After release

3. **Transitions in** higher formants

   Significant characteristics

   (Differentiation between alveolar trill and alveolar lateral)

4. **Formant transitions**

   Robust characteristics to distinguish among the **semivowels** and nasals

   (Calls for algorithms to **obtain** formant tracks)

Behaviour of Parameters

PARAMETERS	MANNER OF ARTICULATION				
	NASALS	***SEMIVOWELS***			
		PALATAL	***ALVEOLAR TRILL***	***ALVEOLAR LATERAL***	***LABIODENTAL***
BEFORE RELEASE					
Antiresonance in spectral characteristics	**present**	**absent**	**absent**	**absent**	**absent**
Average energy	**steady**	**gradual increase**	**significant modulation**	**steady**	**steady**
AFTER RELEASE					
Transition in higher formants	—	—	**change from low to high across release**	**absence of change**	—

For each category of the stop consonants (O, P, Q and R) there are five different consonants in Hindi depending on the place of articulation (category S) – velar, palatal, retroflex, dentialveolar and bilabial. The distinguishing characteristics among different places of articulation are in the release waveform and post release transitions. Determining the characteristics of the release waveform (burst) is difficult due to the short duration of the signal. Nevertheless we still consider the burst features also in the following discussion, because at some later point in time one may evolve signal processing algorithms to characterize it.

The spectral characteristics of the velar burst waveform show consistent peak frequency in the smooth spectrum. The duration of the release waveform is also a

Table 7. Features and parameter trends for Fricatives.

Significant features

Before release

1. Spectral characteristics
2. Steadinessof spectral characteristics
3. Presence of formant structure

After release

4. Formant transitions
5. Steadiness of formant structure

Behaviour of parameters

PARAMETERS	PLACE OF ARTICULATORY RELEASE			
	GLOTTAL	*PALATAL*	*RETROFLEX*	ALVEOLAR
BEFORE RELEASE				
Extent of high frequency components	low	high	significant	high
Region of spectral peak	very low	mid range	low	high
Spectral flatness	low	low	mid range	high
Change in spectral flatness	low	high	significant	high
AFTER RELEASE				
Change in spectral flatness across release	very low	high	high	significant

significant feature. In the case of voiced stop, the release characteristics is embedded in the voicing waveform.

The palatal stop exhibits fricative signal of significant duration at the release due to the large contact area at the closure point as well as the articulatory dynamics in this position. The fricative nature implies the presence of high frequency components. The spectral distance over adjoining frames is not low indicating randomness in the signal. The energy at the release point is abrupt which differentiates this from the fricative class.

The retroflex, dentialveolar and bilabial stops have very short duration release waveform characteristics compared to velar and palatal stops. The release waveforms show distinct and consistent behavior but are difficult to quantify. The retroflex release has energy concentration in the mid-frequency range. The dentialveolar release exhibits distribution of energy in the spectrum. The waveform has slight fricative nature. This is attributable to the large area of closure contact. The bilabial release rvaveform has clear low frequency characteristics for one or two cycles. The spectral characteristics reflect this by a significant peak at the low frequency end. Table 8 shows the significant features and Table 9 the parameters which elucidate the features for various manners and places of articulation.

Table 8. Features lor stop consonants.

Significant features

Before release

1. Presence/absence of voicing
2. Silence if voicing absent

Between release burst and vowel onset

3. Duration between release and vowel onset
4. Presence/absense of aspiration

At *relase burst*

5. Duration of burst
6. Spectral characteristics of bunt
7. Amplitude of burst
8. Extent of high frequency components

After vowel onset

9. Formant transitions – good cues for determining place of articulatory release.

4. Summary and Conclusions

So far we have considered description of sound units of a language in terms of acoustic features and the manifestation of these features in parameters of the signal. Since the features are based on the speech production mechanism, any speech recognition system based on these features is likely to be speaker independent and also robust. At present, design of voice input systems for isolated utterances of words adopt one of the following two approaclies – template matching and feature-based.

In template matching, utterance of each word is represented as a sequence of spectral parameter vectors and the test and reference utterances of words are matched using dynamic time warping algorithm [2]. Limitations of this approach are well known in the recognition studies for alpha digit tasks [6,7,8]. Feature-based approach appears promising [9], but at present the choice of features is dictated more by our ability to derive them from speech signal, rather than from the requirement of deriving the corresponding speech production features. Our discussion in the previous section shows that features for distinguishing several sound units show almost similar characteristics in the signal. While many of the parameters listed can be derived from the signal with known methods of processing, there are several critical features which are difficult to extract reliably. Some of these features are – point of

Table 9. Parameter trends for stop consonants.

Behaviour of parameters

Before release

PARAMETERS	MANNER OF ARTICULATION	
	UNVOICED	*VOICED*
Duration	0	> minimum
Pitch	absent	present
Spectral flatness	high	very low
Spectral distance	——	low
Average Energy difference	——	very low
Extent of high frequency components	——	very low

Between release burst and vowel onset

PARAMETERS	MANNER OF ARTICULATION	
	UNASPIRATED	*ASPIRATED*
Duration	short	long
Spectral flatness	——	low
Energy change	low	initially high. decaying towards vowel onset

At release burst

PARAMETERS	PLACE OF ARTICULATORY RELEASE				
	VELAR	*PALATAL*	*RETROFLEX*	*DENTIALVEOLAR*	*BILABIAL*
Duration of burst	short	long	very short	very short	very *short*
Peak frequency of smooth spectrum	mid range	high	high	mid range (spread)	very low
Extent of high frequency component	mid range	high	high	high	low

release, vowel onset instant, detection of presence of glottal vibration, formant transition after release, spectral features (resonance and antiresonance) of nasals and nasalized vowels, characteristics of burst, and characteristics of fricatives. Many of these features need analysis of short data records to capture rapidly changing features like formant transition and burst. New signal processing algorithms would be needed to extract this information from speech signal. But it is interesting to note that even with the existing method of signal processing, it is possible to develop a feature-based recognition system for isolated utterances of characters whose performance degrades gracefully depending on our ability to process the signal [10,11].

In conclusion, in this paper we have addressed the problem of processing speech signals for developing recognition systems in this paper. We have taken isolated utterances of characters of an Indian language to highlight the issues in processing. In particular, we have shown that the knowledge of acoustic-phonetics enables us to analyze the signal to extract the relevant features for recognition. We have discussed various categories of sounds and described parameters and features needed to distinguish them. For implementing an automatic recognition system for these sounds it is necessary to extract these features. We have also pointed out that existing signal processing methods are not adequate to process the signal for extracting these features. The discussion in this paper demonstrates the limitation of the existing speech recognition systems based on uniform parametric representation of fixed size sgements of speech.

References

[1] L. R. Rabiner and R. W. Schafer, ***Digital Processing of Speech Signals*** (Prentice Hall, New Jersey, 1978).

[2] W. A. Lea, ***Trends in Speech Recognition*** (Prentice Hall, Englewood Cliffs, New York, 1980).

[3] M. Lennig and P. Brassard, "Machine-readable phonetic alphabet for English and French," ***Speech Communication*** 3, 166 (1984).

[4] S. N. Ganesan, A ***Contrastive Grammar of Hindi and Tamil,*** University of Madras, Madras, 1975.

[5] B. Ladefoged, A ***Course in Phonetics*** (Harcourt, Brace Jovanovich, New York, 1975).

[6] R. Sundar, S. Raman, and B. Yegnanarayana, "Studies on speech recognition of Hindi stop consonants," ***Proc. Euro. Conf. Speech Technology,*** Edinburgh, Sept. 1987, pp. 95–98.

[7] B. Yegnanarayana, S. Raman, and R. Sundar, "Signal dependent analysis for speech recognition," ***Proc. IEEE Int. Conf. Speech Input/Output Techniques and Applications,*** London, March 1986, pp. 31–36.

[8] B. Yegnanarayana and T. Sreekumar, "Signal-dependent matching for isolated word speech recognition system," ***Signal Process.*** 7, 161–173 (1984).

[9] R. A. Cole ***et al.,*** "Speaker-independent recognition of spoken English letters," ***Proc. Int. Joint Conf. Neural Networks,*** San Diego, June 1990.

[10] J. Harrington, "Automatic recognition of English consonants," ***Aspects of Speech Technology,*** Edinburgh University Press, Edinburgh, 1988, pp. 69–143.

[11] R. Sundar, "Recognition of isolated utterances of Hindi characters based on acoustic-phonetic knowledge," Ph.D. Thesis (in preparation), IIT Madras, India.