Fuzzy Sets and Systems 46 (1992) 1–9
North-Holland

Use of fuzzy mathematical concepts in character spotting for automatic recognition of continuous speech in Hindi

P. Eswar, C. Chandra Sekhar and B. Yegnanarayana

Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India

Received March 1989
Revised May 1990

Abstract: In this paper we describe the use of fuzzy mathematical concepts in spotting characters from continuous speech in the Indian language Hindi. The research effort reported here is concerned with identifying the phonetic features from the acoustic parameters of a speech signal and combining the features to spot a character in continuous speech. These functions are performed by using expert systems that use confidence values for grading the conclusions arrived at each stage of processing. This paper discusses how this grading is obtained by using numerical translation of the description of a character obtained from an expert phonetician. Confidence values are assigned to the premises of the rules obtained from an expert phonetician using fuzzy membership functions. Fuzzy relations are used to grade the conclusions derived from these premises. Results of applying these techniques to character spotting are given.

Keywords: Membership functions; confidence values; expert systems; continuous speech recognition; character spotting.

1. Introduction

Automatic speech recognition in its first stage involves interpretation of speech data and its representation in terms of symbols. This speech signal-to-symbol transformation is best performed using a knowledge-based approach. The performance of any speech recognition system can be improved by choosing proper symbols for representation. Characters of the language are chosen as symbols for the signal-to-symbol transformation module of our speech-to-text system being developed for the Indian language Hindi [7, 1]. The aim here is to emulate human processes as much as possible at the signal-to-symbol transformation stage itself. In this case, the expert systems approach permits a clear distinction between the domain knowledge and the control structure needed to manipulate the knowledge. A number of speech recognition systems for continuous speech have been reported [3, 4] with varied success. The main drawback in these systems is that they use a simple approach for signal-to-symbol transformation with some abstract units as symbols, thereby increasing the complexity at higher levels of processing. Recent efforts [2, 5] try to improve the performance of signal-to-symbol transformation using speech specific knowledge. Phonemes or syllables were proposed as symbols in these systems, but it is difficult to locate their boundaries in continuous speech.

We have chosen characters of the language as symbols and the knowledge domain is primarily acoustic-phonetic knowledge [1]. The acoustic-phonetic knowledge is first acquired with the help of a human acoustic-phonetic expert who is asked to express it while interpreting speech patterns. The expert phonetician makes a decision based on data which are to some extent incomplete and ambiguous, and also based on his acquired knowledge which is dynamic in nature. When the acoustic-phonetic knowledge is expressed in the form of rules, the premises (such as facts or derived facts) in a rule are uncertain, or the reasoning is uncertain, or both are uncertain. Hence the knowledge obtained from the expert is tentative. A system should provide its decision with an indication of this

0165-0114/92/$05.00 © 1992—Elsevier Science Publishers B.V. All rights reserved

tentative nature ranging from total certainty to utter disbelief. A system which takes care of this range of uncertainties is called a graded system. We describe an expert system which grades its decisions at every stage of processing while spotting a character in a given utterance.

This paper explains how to provide numerical values in the description of the knowledge for the character spotting expert system. The acoustic-phonetic knowledge consists of descriptions of characters in terms of phonetic features and the relation between the phonetic features and the corresponding measurable acoustic parameters. This knowledge can be built gradually using the expert's knowledge and several experimental and theoretical studies. The knowledge is incomplete because it is difficult to capture the reasoning techniques used by an expert. It is ambiguous because of imprecision in the extraction of parameters from a speech signal. In order to take care of the incomplete and ambiguous nature of the knowledge, confidence values are assigned to: (1) premises of a rule, (2) conclusions of a rule based on how various premises in a rule are combined and (3) hypotheses based on how a set of rules are combined. Confidence values are assigned to the premises by measuring the closeness of the actual data to that specified by an expert using fuzzy membership functions. Fuzzy relations are used to combine different premises or different sets of rules. That is, the rules of the expert system for character spotting assign confidence values to the conclusions of the rules, and sometimes to a rule itself. Fuzzy membership functions and fuzzy relations based on Zadeh's [8] theory of possibility are used to represent the expert's knowledge as faithfully as possible. This approach is used in building character spotting expert systems in which the knowledge base is represented in the form of IF . . . THEN rules with appropriate fuzzy membership functions.

Section 2 describes the basic concepts of fuzzy set theory, fuzzy mathematical restrictions and membership functions. The section also describes how these concepts are used to assign confidence values to the premises and conclusions of the rules in the knowledge base of an expert system. Section 3 describes application of the fuzzy mathematical concepts for character spotting in continuous speech through an illustration of a character expert. Section 4 discusses the results of applying these concepts to several character spotting experts.

2. Fuzzy set theory, fuzzy restrictions and membership functions

Fuzzy restriction [8] is a relation which acts as an elastic constraint on the values that may be assigned to a variable. Such restrictions play an important role in speech recognition, and particularly in signal-to-symbol transformation, because the environment happens to be uncertain and hence fuzzy.

Let $\boldsymbol{U}$ denote the universe of discourse representing a set of elements. A fuzzy subset $\boldsymbol{A}$ of $\boldsymbol{U}$ is characterized by a membership function or compatibility function $\mu_a(u)$, which associates a real value in the range 0 to 1 with every member u in $\boldsymbol{A}$. The value of $\mu_a(u)$ represents the grade of membership of u in $\boldsymbol{A}$. The points in $\boldsymbol{U}$ at which $\mu_a(u) > 0$ constitute the support of $\boldsymbol{A}$. The points at which $\mu_a(u) = 0.5$ are called the crossover points of $\boldsymbol{A}$. Formally, a fuzzy set $\boldsymbol{A}$ with its finite number of supports $u_1, u_2, \ldots, u_n$ can be represented as ordered pairs

$$\boldsymbol{A} = \{\langle \mu_a(u_i), u_i \rangle,\ i = 1, 2, \ldots, n\}. \tag{1}$$

A finite fuzzy subset $\boldsymbol{A}$ of $\boldsymbol{U}$ is also expressed as [8]

$$\boldsymbol{A} = \mu_1/u_1 + \mu_2/u_2 + \cdots + \mu_n/u_n \tag{2}$$

where $\mu_i = \mu_a(u_i)$.

The value of μ_i represents the degree to which u_i can be a member of $\boldsymbol{A}$. In particular, $\mu_i = 1$ denotes strictly the containment of u_i in $\boldsymbol{A}$, whereas $\mu_i = 0$ denotes that u_i does not belong to $\boldsymbol{A}$. If only a single member u_1 of $\boldsymbol{U}$ is contained in $\boldsymbol{A}$, then we have a fuzzy singleton $\boldsymbol{A} = \mu_1/u_1$. If $\mu_1 = 1$, then $\boldsymbol{A} = 1/u_1$ and in this case $\boldsymbol{A}$ is called a nonfuzzy singleton.

If the universe of discourse is a continuous domain, then the fuzzy set $\boldsymbol{F}$ may be expressed as

$$\boldsymbol{F} = \int_{\boldsymbol{U}} \mu_f(u)/u \tag{3}$$

where $\mu_f(u)$ is a membership or the compatibility function of $\boldsymbol{F}$ and the integral $\int_{\boldsymbol{U}}$

denotes the union of fuzzy singletons $\mu_f(u)/u$ over the universe of discourse U.

In many cases it is advantageous to compute the membership function of a fuzzy subset in terms of a standard function whose parameters can be adjusted to fit a specified membership function in an approximate fashion. Three such functions (shown in Figure 1) are defined as follows [8]:

S curve

$$S(u:x,y,z) = \begin{cases} 0 & \text{for } u \leq x, \\ 2[(u-x)/(z-x)]^2 & \text{for } x \leq u \leq y, \\ 1-2[(u-z)/(z-x)]^2 & \text{for } y \leq u \leq z, \\ 1 & \text{for } u \geq z, \end{cases} \quad (4)$$

where $y = \frac{1}{2}(x+z)$ is the crossover point.

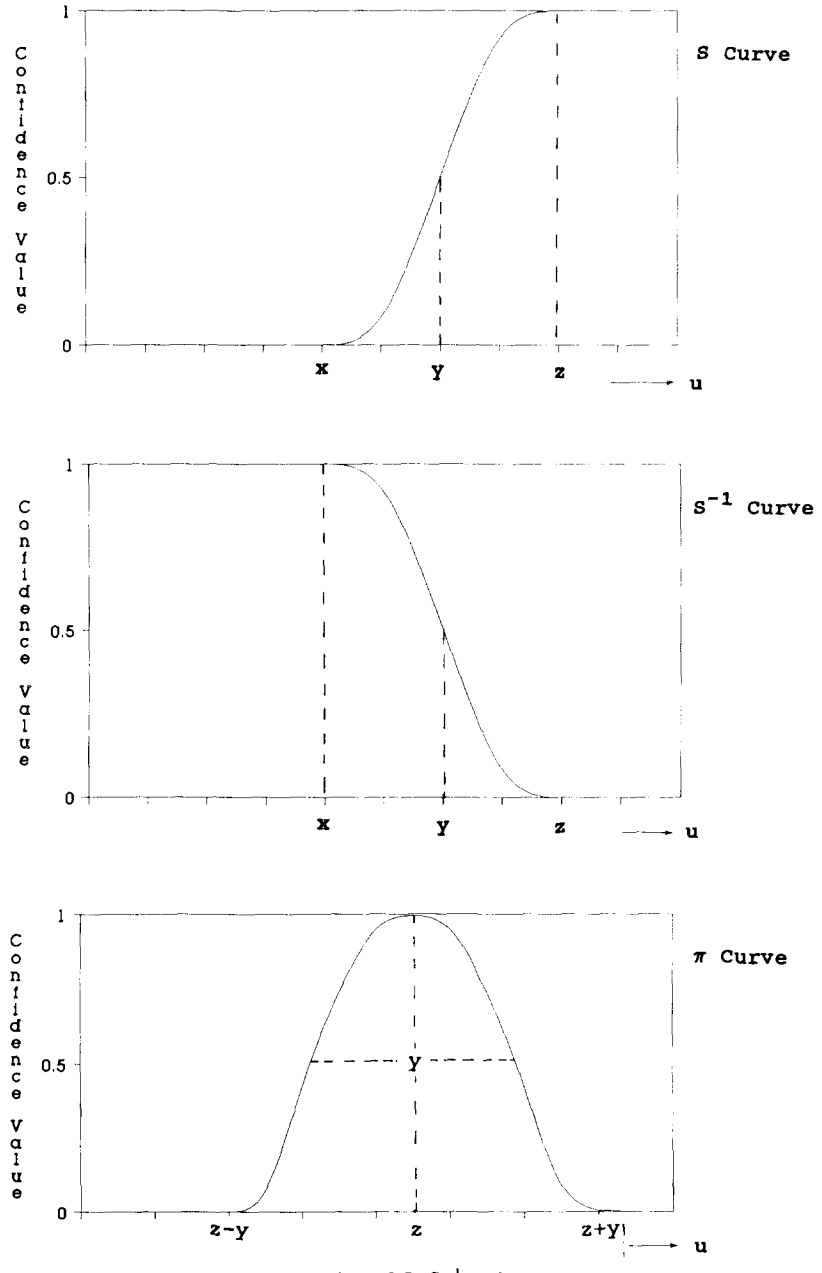

Fig. 1. Illustration of S, S^{-1} and π curves.

S^{-1} curve

$$S^{-1}(u:x, y, z) = \begin{cases} 1 & \text{for } u \leq x, \\ 1-2[(u-x)/(z-x)]^2 & \text{for } x \leq u \leq y, \\ 2[(u-z)/(z-x)]^2 & \text{for } y \leq u \leq z, \\ 0 & \text{for } u \geq z, \end{cases} \quad (5)$$

where $y = \frac{1}{2}(x+z)$ is the crossover point.

π curve

$$\pi(u:y, z) = \begin{cases} S(u:z-y, z-y/2, z) & \text{for } u \leq z, \\ S^{-1}(u:z, z+y/2, z+y) & \text{for } u \geq z, \end{cases} \quad (6)$$

where y is the bandwidth, that is, the separation between the crossover points of π, and z is the point at which π is unity.

We illustrate the concepts of fuzzy set and membership functions in the context of speech recognition. In this regard it is necessary to explain some of the basic processes in the production of speech sounds.

A cross-sectional view of the vocal tract system is shown in Figure 2 [6]. The basic source of power in the production of speech sounds generally is the respiratory system pushing air out of the lungs. The air is pushed through the trachea and it passed through the vocal cords. Sounds produced when vocal cords are vibrating are called voiced sounds and the sounds produced when the vocal cords are far apart are called voiceless or unvoiced sounds. The passage above the larynx is called the vocal tract system. There are two different kinds of speech sounds produced depending on the path through which the air moves in the vocal tract system. Nasal sounds are produced when the air passes through the nasal tract and oral sounds are produced when the air passes through the oral tract. The parts of the vocal tract system that are used to produce different sounds are called articulators (Figure 2 and 3). Articulators in the lower surface of the vocal tract (for example, tongue) move towards the upper surface (for example teeth, hard palate, soft palate). Consonants are described in terms of the place and manner of articulation. Vowel sounds are described by the

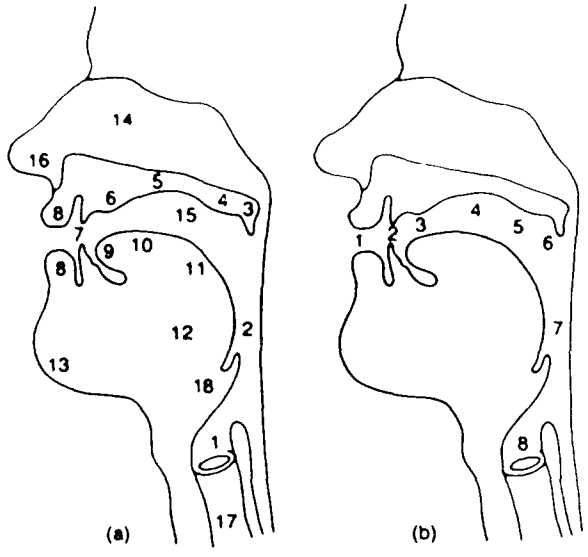

Fig. 2. A cross sectional view of the vocal tract. (a) *Speech articulators*: (1) vocal folds; (2) pharynx; (3) velum; (4) soft palate; (5) hard palate; (6) alveolar ridge; (7) teeth; (8) lips; (9) tongue tip; (10) blade; (11) dorsum; (12) root; (13) mandible; (14) nasal cavity; (15) oral cavity; (16) nostrills; (17) trachea; (18) epiglottis. (b) *Places of articulation*: (1) labial; (2) dental; (3) alveolar; (4) palatal; (5) velar; (6) uvular; (7) pharyngeal; (8) glottal (taken from [6, pp. 51]).

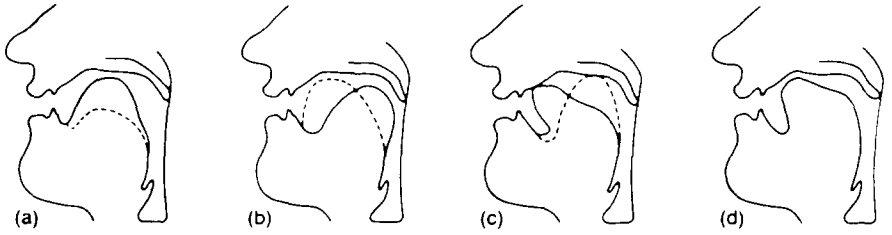

Fig. 3. Typical articulatory positions for (a) a vowel, showing two tongue height positions, (b) a high vowel, showing front and back positions, (c) a stop, showing alveolar and velar places articulation (for a nasal, the only difference lowered velum), (d) an alveolar fricative (taken from [6, pp. 53]).

position of the tongue and the lips. *Front* vowels are produced when the highest point of the tongue is in front, i.e., near the teeth. *Back* vowels are produced when the highest point is at the back, i.e., near the velum. In between these two, there are *central* vowels. Vowels are described as *high, low* and *mid* based on the height of the tongue. *Rounded* and *unrounded* vowels are described depending on the shaping of the lips. Thus any vowel can be described by using phonetic features like *front, back, central, high, low, mid, rounded* and *unrounded.* Place and manner of articulation features are used to describe a consonant sound. An important acoustic cue for the determination of the position of the tongue during the production of a vowel sound is the second formant corresponding to the second resonant frequency in the vocal tract. The range of second formant frequency is 900–2700 Hz. The i-th formant is represented as F_i.

To illustrate the use of fuzzy concepts, consider the universe of discourse consisting of all the second formant frequencies in the range 900–2700 Hz. For *back* vowels the possible range of F_2 is 900–1200 Hz. Then a nonfuzzy subset corresponding to *back* vowels is

$$back = \{900 \cdots 1200\}. \tag{7}$$

The values of F_2 may fall outside the range 900 to 1200 Hz due to some contextual effects. Therefore the fuzzy subset $back_{\text{fuzzy}}$ may be defined as

$$back_{\text{fuzzy}} = 1.0/\{900 \cdots 1200\} + 0.5/\{1400\} + 0.2/\{1600\}. \tag{8}$$

This means that the degree to which F_2 in the range 900–1200 Hz is compatible with the *back* feature is with a confidence of 1, whereas the confidence values are 0.5 and 0.2 respectively, when F_2 has values at 1400 Hz and 1600 Hz.

An example of a fuzzy set representation of an acoustic parameter of the speech signal is given in Table 1. The table shows the grade membership of the second formant frequency $F_2 = 1200$ Hz for three different phonetic features (*front, back* and *central*). The grade membership of $F_2 = 1200$ Hz for the *back* feature is 1.0. To obtain this grade membership value the function $S^{-1}(1200:1200, 1400, 1700)$ is used

Table 1. Confidence values for a second formant frequency (F_2) of 1200 Hz for different compatibility functions

Fuzzy membership	Type of fuzzy curve used
$\mu_{\text{back}} = 1.0$	$S^{-1}(1200:1200, 1400, 1700)$
$\mu_{\text{front}} = 0.0$	$S(1200:1700, 1750, 1800)$
$\mu_{\text{central}} = 0.2$	$\pi(1200:400, 1550)$

because this gives a high confidence value below a threshold. For the *back* feature the second formant frequency is less than 1200 Hz. For F_2 in the range 1200–1700 Hz, the *back* feature has a grade membership decreasing in the range from 1.0 to 0.0. Any value of F_2 greater than 1700 Hz yields zero confidence for the *back* feature. To get the grade membership for the *front* feature, the S-curve is chosen because it gives a grade membership of 1.0 for all frequencies above a threshold. The *front* feature is indicated when F_2 is greater than a threshold. Similarly the *central* feature is indicated when F_2 lies in the specified range. Hence a π-curve is appropriate to represent this behaviour.

For propagation of uncertainty, the concept that the degree of uncertainty of a combined proposition is a function of the degree of uncertainty of the component propositions is used. For this the following logical connectives of fuzzy relations are used:

$X1$ OR $X2$ = maximum ($X1$, $X2$),

$X1$ AND $X2$ = minimum ($X1$, $X2$),

not $X1$ = (max − $X1$),

where $X1$ and $X2$ are fuzzy variables and max is the maximum value of the variables.

In order to prevent an exponential growth of the number of inferences at various stages in the signal-to-symbol transformation, the inferences are pruned using only those whose confidence values exceed a preset threshold. For example, consider the rule for the *vocalic* (vowel-like) feature:

IF
(1) Low frequency energy is high OR
(2) Total energy is high
THEN
vocalic.

Here the *vocalic* decision will be assigned a

confidence value which is the maximum of the confidence values obtained from the premises (1) and (2), but the *vocalic* decision will completely fail when the confidence values obtained from both (1) and (2) fall below a certain preset threshold.

3. Use of fuzzy mathematical concepts in character spotting

An expert systems approach for character spotting in continuous speech in Hindi is proposed in [7]. In this section we explain the use of fuzzy mathematical concepts to grade the conclusions derived at each stage of processing and to integrate the results at each stage to spot the character in the given speech signal (Table 2). Interpretation of a speech signal involves generation of hypotheses from the numerical values of the acoustic parameters used to describe a character. This process consists of two steps, namely, (1) description of the phonetic features in terms of acoustic parameters and (2) description of a character in terms of the phonetic features. Interpretation of a speech signal becomes difficult because errors in the acoustic parameters may affect the features identified. Inaccuracy in feature identification may affect character spotting.

For example, normally the acoustic description of the long vowel /a:/ occurring as a separate character in continuous speech is different from that of /a:/ occurring in the character /ka:/. Though the description of the sound /a:/ in both cases is a *long unrounded open back vowel*, there is still a difference as to how the *back* feature of /a:/ in /ka:/ and the *back* feature of /a:/ occurring in isolation are obtained from the acoustic parameters. Similarly even though the consonant portion /k/ of /ka:/ is defined as unaspirated, in actual speech there is a certain amount of aspiration present. Thus the descriptions of many of the characters in terms of the phonetic features and the descriptions of the features in terms of acoustic parameters are imprecise and vague. In order to take care of such conditions, the expert system used for spotting a character in an utterance does not give any binary decision but assigns confidence values to the conclusions based on fuzzy algorithms.

Table 2. List of Hindi characters

Sign	Character
/a:/	आ
/ka:/	का
/ca:/	चा
/ta:/	ता
/ṭa:/	टा
/pa:/	पा

Important steps in spotting the character/ka:/ in an utterance using an expert systems approach are as follows:

Spotting_Character /ka:/
- Locate vocalic regions
- Locate closure regions
- Locate burst regions
- Locate aspirated regions
- Hypothesize the possible presence of character by checking whether combinations of these features are according to the description of the character
- Use intrinsic cues to identify segments of the character in the regions where its presence is hypothesized
- Use coarticulation cues to identify segments in the character, if necessary
- Use context-dependent cues to identify the segments in the character, if necessary
- Combine the above results to identify the character with appropriate confidence value

The expert system processes the speech signal to spot the character and gives an appropriate confidence value. At the lower levels of processing, similar fuzzy reasoning is applied for locating different features using various acoustic parameters. For example, the acoustic cues to locate *closure* regions are: (1) total energy should be low, (2) ratio of low frequency energy to high frequency should be low and (3) duration of the closure region should be small. The vagueness inherent in such terms as low and small has to be expressed numerically. The boundaries obtained for various features may

not be precise. A method based on fuzzy relations is used to combine these features to obtain a confidence value for possible presence of the character.

After obtaining the region where the character is possibly present, the intrinsic cues, the coarticulation cues and the context-dependent cues are used to modify the confidence value with which the segments of the character are hypothesized. Finally, the character is identified with a confidence value based on the confidence values obtained at each stage of analysis. The knowledge base of the expert system for spotting the character consists of production rules or IF . . . THEN rules derived from the description of the character. The thresholds for the acoustic parameters used to spot the phonetic features are provided in a fuzzy table. The input data to be analyzed are processed by invoking the expert system. Appropriate confidence values are assigned to the conclusions of each rule when it is fired. The assignment is made by comparing the actual input values with the thresholds in the fuzzy table. This process of assigning confidence values is continued until the rule base is completely exhausted. The result of this processing gives the confidence value with which the character is spotted. Thus fuzzy mathematical concepts help in modeling the imprecision and vagueness in the knowledge base that relates the acoustic parameters, phonetic features and the characters.

4. Results and conclusions

Character spotting expert systems were tested on 120 utterances in Hindi spoken by two male speakers. Speech data was sampled at 10 kHz and digitized with 12 bit resolution. It was observed that with just two parameters (total energy and first linear prediction coefficient) along with their fuzzy thresholds, *vocalic* regions were located with more than 95% accuracy. The other features like *closure* or *silence* needed different thresholds for *voiced* and *voiceless*

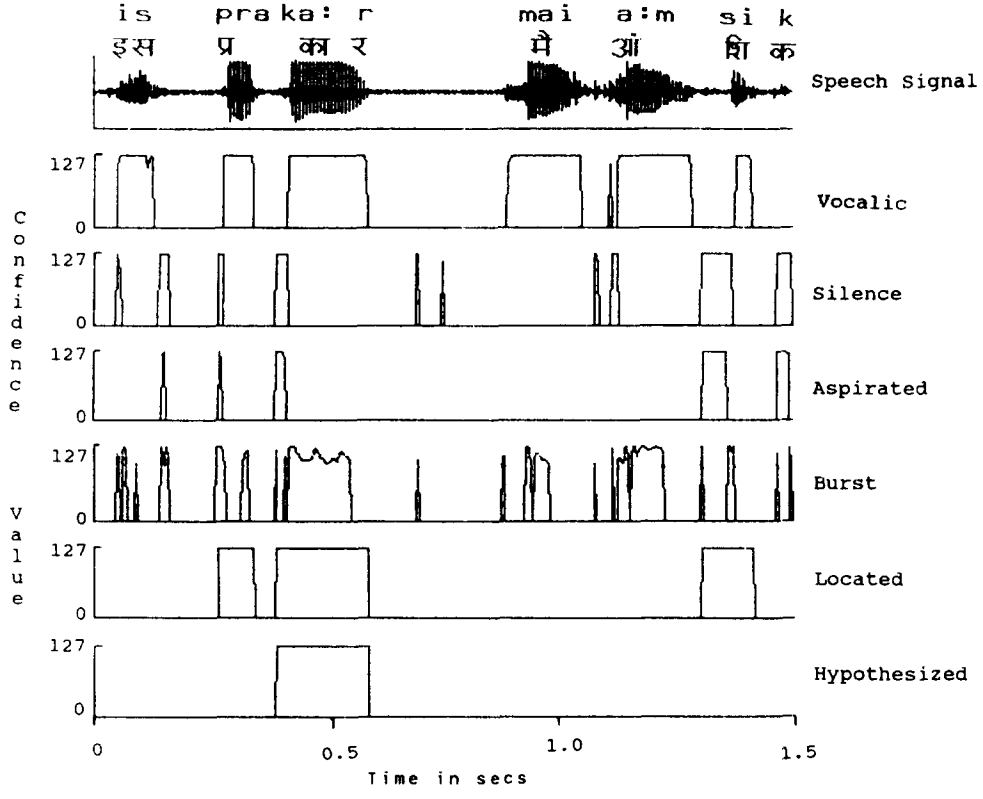

Fig. 4. Confidence value plots for the gross features, located regions and hypothesized regions for the character /ka:/.

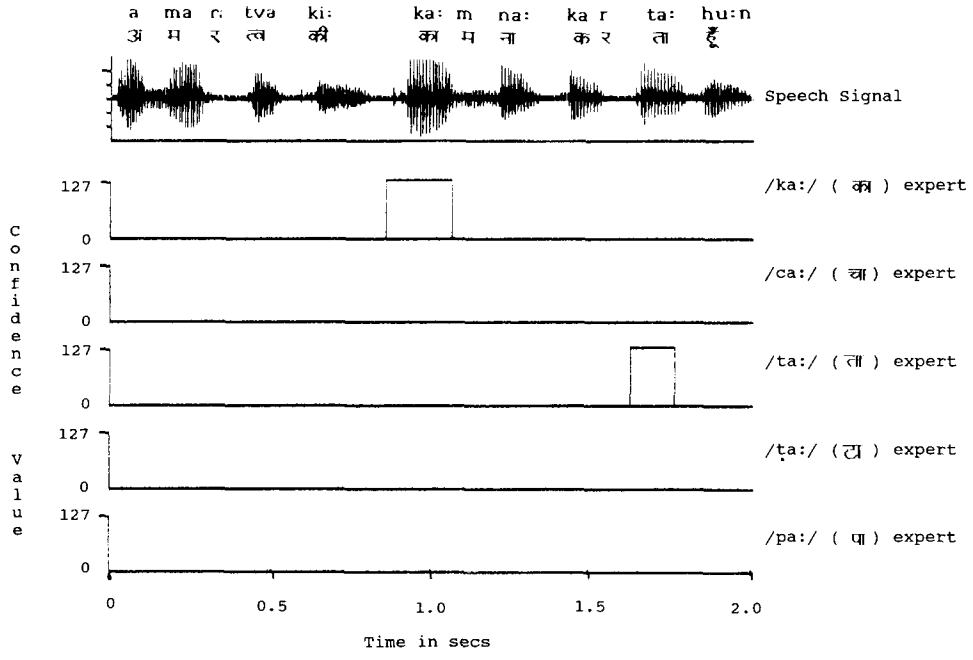

Fig. 5. Confidence value plots of five character spotting experts on an utterance.

cases. This was necessitated when looking for *closure* as a clue for identifying the stop consonants. Separate fuzzy membership functions with appropriate acoustic parameters were used to locate the *voiced* and *voiceless closures*. Fuzzy relations were used for integrating the gross features obtained for hypothesizing the possible presence of the character. Figure 4 shows the confidence value of various gross features in different regions of an utterance. Confidence values are given as integer values in the range 0 to 127. The location rules combine the gross features to arrive at possible locations for the character /ka:/.

The intrinsic and context-dependent rules are then used for hypothesizing the character region. Figure 5 illustrates the confidence value plots for hypothesized regions of five different character spotting expert systems, when they are used on an utterance in Hindi. Since the fuzzy tables for these characters were tuned well by testing them on a large number of utterances, the character experts have spotted the regions correctly with maximum confidence in this case.

Results of the performance of some gross feature experts on a large number (120) of utterances in Hindi are shown in Table 3. The

Table 3. Performance characteristics for gross feature identification

Gross features tested	Gross features identified			
	Vocalic	Silence	Burst	Aspirated
Vocalic	420/420 (127)	—	—	—
Silence	—	120/120 (120)	10/120 (100)	10/120 (100)
Burst	—	10/120 (100)	120/120 (120)	15/120 (100)
Aspirated	—	20/100 (100)	20/120 (100)	120/120 (120)

Notation used in the table: In the figures P/Q, P indicates the number of times a feature is hypothesized and Q indicates the number of times the feature occurred in the data. The figures in parentheses indicate the confidence with which the character is hypothesized. Minimum confidence is indicated for diagonal terms. Maximum confidence is indicated for off-diagonal terms.

entries in the table (P/Q) show that the feature is spotted correctly P times out of Q occurrences in the test data. The confidence value is also given in parentheses. The value given in a diagonal entry is the minimum of the values and the value given in an off-diagonal entry is the

Table 4. Performance characteristics of experts for consonant-vowel (CV) combination (*C* is /k/, /c/, /ṭ/, /t/ and /p/ and *V* is /a:/)

Character experts	Identified characters				
	/ka:/	/ca:/	/ṭa:/	/ta:/	/pa:/
/ka:/	10/10 (120)	—	—	3/7 (100)	—
/ca:/	—	6/6 (115)	—	—	—
/ṭa:/	—	—	6/8 (115)	—	—
/ta:/	2/10 (100)	—	—	17/17 (115)	—
/pa:/	—	—	—	—	6/6 (120)

Notation used is that of Table 3.

maximum of the values. The results indicate that the gross features are spotted correctly most of the time.

Results of the performance of five character experts are shown in Table 4. Since these are confusable characters, there are bound to be some errors in spotting them. Although the results are encouraging, the performance of the character spotting depends critically on the entries in the fuzzy tables. Extensive tuning of these tables is to be done by experimenting the spotting experts on a large number of utterances spoken by different speakers. This tuning is done manually at present. It has to be automated. This means that the system has to be provided with the learning capability to refine the fuzzy tables automatically.

References

[1] P. Eswar, S.K. Gupta, C. Chandra Sekhar, B. Yegnanarayana and K. Nagamma Reddy, An acoustic-phonetic expert for analysis and processing of continuous speech in Hindi, in *Proc. European Conf. on Speech Technology,* Edinburgh, vol. 1 (1987) 369–372.

[2] J.P. Haton, Knowledge based approach in acoustic-phonetic decoding of speech, in: H. Niemann, M. Lang and G. Serger, Eds., *Recent Advances in Speech Understanding and Dialog Systems*, NATO-ASI Series, vol. 46, (1988) 51–69.

[3] D.H. Klatt, Review of the ARPA speech understanding project, *J. Acoust. Soc. Amer.* **62**(6) (1978) 1345–1366.

[4] W.A. Lea, Ed., *Trends in Speech Recognition* (Prentice Hall, Englewood Cliffs, NJ, 1980).

[5] R.De Mori, A. Giordana, P. Laface and L. Saitta, Parallel algorithms for syllable recognition in continuous speech, *IEEE Trans. Pattern Analysis Machine Intelligence.* **7**(1) (1985) 55–68.

[6] D. O'Shaughnessy, *Speech Communication – Human and Machine* (Addison-Wesley, Reading, MA, 1987).

[7] B. Yegnanarayana, C.C. Sekhar, G.V.R. Rao, P. Eswar and M. Prakash, A continuous speech recognition system for Indian languages, in: *Proc. Regional Workshop on Computer Processing of Asian Languages*, Bangkok (1989) 347–356.

[8] L. Zadeh, Ed., *Fuzzy Systems and Their Applications to Cognitive Processes* (Academic Press, New York, 1975).