



AN ACOUSTIC-PHONETIC EXPERT FOR ANALYSIS AND PROCESSING OF
CONTINUOUS SPEECH IN HINDI

P.Eswar*, S.K.Gupta*, C.Chandra Sekhar*, B.Yegnanarayana* and
K. Nagamma Reddy**

ABSTRACT

We are currently engaged in the development of speech recognition systems for continuous speech in Indian languages. In our opinion the most important block in the system is the phonetic recogniser. We propose to exploit the phonetic nature of Indian languages to design the phonetic recogniser. We consider characters of an Indian language as symbols for the design of speech signal-to-symbol transformation module of our system. In this paper we discuss the limitations of conventional techniques in speech signal-to-symbol transformation and show the importance of knowledge-based signal processing.

INTRODUCTION

The objective of our research is to provide a limited dictation capability to a computer. The system should be independent of speaker, vocabulary and task. We are interested in developing a speech-to-text conversion system for an Indian Language (Hindi). Such an effort, when successful, provides a symbolic representation of continuous speech, which in turn can be developed for several applications such as dictation machine, natural language processing, language translation, etc.

Basically a speech-to-text conversion system consists of two stages. The first stage converts the analog speech into some symbolic form using signal processing and some knowledge of acoustic-phonetics (ref 1). The symbolic form is later converted, in the second stage, into meaningful text using higher level sources of knowledge such as lexical, syntactic, semantic, etc. Approaches followed by earlier systems (ref 2, ref 3) were mostly based on simple signal processing algorithms. The symbols are usually some arbitrary units corresponding to acoustically uniform segments. The number of symbols and the parameter pattern corresponding to each symbol varied from system to system. The systems were dependent to a large extent on higher level knowledge sources to disambiguate the symbol sequence into meaningful text. The disadvantages of these systems are:

1. The complexity of representation of knowledge sources grew with the size of the task and vocabulary.
2. The systems were highly speaker dependent, task dependent, vocabulary dependent and environment dependent.
3. The signal-to-symbol transformation used only parametric knowledge and hence any significant information lost at this stage could not be recovered in the later stages.

* Department of Computer Science and Engineering,
Indian Institute of Technology, MADRAS - 600 036, INDIA.

** Department of Linguistics, Osmania University,
HYDERABAD - 500 007, INDIA.

Consequently these systems had limited practical utility. Current systems (ref 4 to ref 6) seem to lay importance on the signal-to-symbol transformation stage, and thus the emphasis in most of these systems is on the design of a phonetic recognizer. For nonphonetic languages like English, the main problem of choice of symbols still remains. The arbitrariness of the symbol choice creates another problem, that of providing a description of the vocabulary in terms of these symbols. Usually the symbolic description is a laborious process involving several man-months of efforts for any practical system.

We propose to exploit the phonetic nature of Indian languages to design our signal-to-symbol transformation stage. In particular, we have chosen the written characters of one language (Hindi) as symbols. We feel that typically in Indian languages "we write what we speak and we speak what we write". Therefore we expect that most of the meaningful speech sounds of the language can be transformed into unique symbols (characters). This also takes care of the variability of pronunciations, because for a different pronunciation of a given word, there will be a different symbol sequence. The signal-to-symbol transformation stage is expected to capture most of the information available in the input speech.

KNOWLEDGE-BASED APPROACH FOR SIGNAL-TO-SYMBOL TRANSFORMATION

The basic approach in our system is to design the signal-to-symbol transformation stage, where the symbols are the written characters of the language. We have decided to realize this by using an expert system approach for spotting each character. The advantages of this approach are :

1. The speech signal can be processed in a manner dictated by the requirements for spotting the character.
2. We are not tied down to a particular parameter or feature set. All the necessary information for spotting the character can be obtained directly from the speech signal. Hence there is no loss of information at the speech signal-to-symbol transformation stage, as in the previous systems.
3. We are also avoiding creation of complex pronunciation dictionary for converting words into phoneme symbol sequence.
4. The overall complexity of the system does not grow with the size of the vocabulary or task.

The only disadvantage, as we see, is the large number of (typically 5000) characters to be spotted for a given Indian language. But, since each character expert can be implemented independently, we propose to exploit the parallel implementation feature in the final design of our system to overcome this disadvantage.

Each character expert uses knowledge relevant to spot that character. While primarily acoustic-phonetic knowledge is used, the rules may also incorporate either directly or indirectly other knowledge sources such as lexical, syntactic, semantic, etc. The knowledge is incorporated as a set of rules and these rules dictate the parameters and features to be extracted from speech signal.

The rules for each character expert are organised under the following four broad categories.

1. Rules to appropriately locate the possible presence of the character in the input speech.
2. Intrinsic cues to recognise the speech segments of the character through a description in terms of articulatory and acoustic features obtained from an acoustic-phonetic expert.
3. Rules that capture variations in acoustic correlates of speech segments due to the influence of other segments within a character.
4. Rules that describe the variations that the acoustic features of a character undergo in different character contexts.

It is not possible to spot uniquely a character in an utterance because of the ambiguous nature of the speech signal and also the availability of partial knowledge to process the speech signal. Fuzzy logic is used to give confidence measures for the conclusions arrived at each stage of the character expert. The outputs of all character experts are combined by a second level expert which uses some additional language specific constraints to get a unique character sequence.

ILLUSTRATION OF A CHARACTER EXPERT

We describe briefly how the knowledge-base for character experts is organized by considering specifically one Hindi character, namely /dzjo:/(ज्यो).

The following is the description of the character /dzjo:/, as obtained from an acoustic-phonetic expert:

Voiced Palatal Unaspirated affricate(stop+fricative), followed by Voiced High Front semivowel, followed by Voiced Long Half-closed Back Rounded vowel.

The gross features chosen from the description of the character /dzjo:/ to hypothesize its location in the continuous speech are: voiced/nonvoiced, fricative/nonfricative, silence/ nonsilence, burst/nonburst. Rules are written to identify the above features using appropriate acoustic correlates. It is observed from the description of the character that voiced feature should be present throughout the region of the character. Further this voiced region should encompass other features, namely, fricative, burst and silence. Fig.1 illustrates the features determined for a test utterance for hypothesizing the location of /dzjo:/. It can be seen that, for each feature like, voiced/nonvoiced, there may be unidentified regions which can be conveniently used so that the true locations of /dzjo:/ are not missed in the hypothesized regions. The unidentified regions can be reduced by using effective algorithms to detect the features.

The description of /dzjo:/ suggests that rules for intrinsic cues for the affricate /dz/ (ज), semivowel /j/ (य) and the vowel /o:/ (ओ) are to be incorporated in the rule base. Some of the cues are burst frequency spectra for /dz/ and formant structure for semivowel and vowel.

vvvvvvvvv-vvvnnvvvvv----vvvvvnnvvvvvv	(a)	v = voiced n = nonvoiced - = unidentified
nfffnnnf--ffff---ffffnnnnnnff--nnnnnnnn	(b)	f = fricative n = nonfricative - = unidentified
snnnnnnnsnnn---nnnnssssnnns--nnnn--n	(c)	s = silence n = nonsilence - = unidentified
nnbnn-nnnnbbnn-nn-nnnnnbbnnnnbbnn-nnnn	(d)	b = burst n = nonburst - = unidentified

Fig.1 Illustration of identified regions of gross features (a) voicing, (b) frication, (c) silence and (d) burst for an utterance(1.15 sec long)

Manytimes it may not be possible to identify the character /dzjo:/ based on the intrinsic cues alone. There is considerable influence of one speech segment over the other within the character. For example, the formant structure of the semivowel /j/ is influenced by the preceding affricate /dz/ and the succeeding vowel /o:/. Moreover vowels become centralized in continuous speech. We have found that rules pertaining to these variations are more effective in spotting the character /dzjo:/.

The above three sets of rules identify the character /dzjo:/ with a high level of confidence. But sometimes it may be necessary to use rules pertaining to the influence of preceding and succeeding characters in continuous speech. For example, the formant structure of the vowel /o:/ is influenced by the succeeding character, especially if the succeeding character is a stop consonant. Similarly the frication noise of /dz/ varies if the preceding character is a trill. We have collected the rules for a few cases. The rule base of the system must be updated with progressive use of the system for a variety of speech material.

REFERENCES

1. V.W.Zue, Proc. IEEE, Vol.73, No.11, 1602-1634, 1985.
2. D.H.Klatt, JASA, Vol.62, 1345-1365, 1977.
3. W.A.Lea, Trends in Speech Recognition, Prentice-Hall, 1980.
4. R.A.Cole, Proc. Speech Tech '86, Vol.1, No.3, 43-46, 1986.
5. R.De Mori, A.Giordana, P.Laface and L.Saitta, Proc. AAAI Conference, Pittsburg, PA, 107-110, 1982.
6. R.De Mori, P.Laface and Yu Mong, IEEE Trans. on PAMI, Vol.7, No.1, 56-68, 1985.