SPEECH RECOGNITION SYSTEM ON A MICROCOMPUTER

V.G.Lebedev

Lab. of Technical Cybernetics
Novosibirsk State University
Novosibirsk, USSR 630090

S.A.Khamidullin

Dept. of Informatics
Institute of Mathematics
Novosibirsk, USSR 630090

ABSTRACT

Real-time discrete and connected speech recognition system is described. The system includes the following levels:
- pre-processing of speech signal and feature extraction;
- recognition of isolated words;
- recognition and understanding of discrete phrases.

The first level has a hardware implementation and the others have software ones.

A phrase of discrete speech consists of a chain of utterances separated by pauses. Utterances are recognized first, after these and through them phrases are, use being made of semantico-syntactic and pragmatic restrictions.

The word recognition in phrases is carried out by dynamic programming procedure with adaptive corridor. The comparison of the test pattern with all reference patterns is made in parallel. Simultaneously the rejection of misleading candidates over threshold defined in the recognition process is carried out.

Syntactic and semantic districtions are presented in the system as a tree of word connections in phrases. Using this diagram of permissible phrases the system organizes the current word recognition, defines the end of test phrase or goes to the next recognition step and defines in this case the set of permissible phrase continuations.

The system is implemented on microcomputer "Electronica-60M" and also includes pre-processor for the initial speech signal processing and features axtracting. As features the output intensities of six bandpass filters sampled every 16 ms are used.

The 120 words occurring in 140 phrases of 3-7 words each are recognized, with speaker adaptation, in the real time scale with a realiability of 98%.

INTRODUCTION

A system is discussed which is designed to understand phrases of discrete speech. Discrete speech phrases Φ_j , $j=1\div N$, are defined as chains of utterances S_i , $i=1\div K$, separated by short pauses. Here N is the total number of possible phrases, and S_i is a single word form or block of connected words (S_i will further be referred to as words). The minimal duration of pauses between utterances is 100 msec.

The system operates under two main modes: training and recognition. Under the mode of training a set of templates $E=\{E_i\}$ is formed corresponding to the set $S=\{S_i\}$, $i=1\div K$, and under that of recognition to each phrase F_K presented for recognition there corresponds a phrase from a set $\Phi=\{\Phi_j\}$. From the whole set Φ of phrases we will consider only admissible phrases, i. e. phrases meeting the semantico - syntactic and pragmatic restrictions adopted in the system. The syntactic restrictions include among admissible only grammatically correct phrases.

The semantic restrictions admit from the syntactically correct phrases only those making sense.

The pragmatic restrictions make it possible to select from semantically correct phrases only those admissible in particular situations specified by a concrete applied field in which this system of phrase recognition is operating. Note that the process of checking a test phrase for admissibility coincides in the system with that of phrase recognition.

From the now available systems of speech recognition the one under discussion differs in that templates and uses a more effective strategy of taking into account semantico-syntactic restrictions [1]. Unlike the systems of recognizing continuous speech the present system has a restriction consisting in requiring obligatory pauses to separate S_i . On the other hand, introducing this restriction allows the labour input to be essentially reduced as it is no longer necessary to divide phrases into words. An important peculiarity of the system is also the possibility of its quick modification through the means of the operation system (RAFOS or RT-11) maintaining the functioning of programs involved in the recognition system. Most of the present-day systems of recognition on the base of microcomputers function without any operation systems or are created as a special processors and hence are deprived of this possibility.

ARCHITECTURE AND ALGORITHM OF SYSTEM OPERATION

The system is realized on a microcomputer "Electronika-60" involving a pre-processor for the initial speech signal processing and a complex of programs written in the languages MACROASSEMBLER and FORTRAN in the "RAFOS" operation system.

Through a microphone a speech signal is fed to the pre-processor which every 16 msec determines the values of intensities at the outputs six filters covering a band of 400 to 5000 Hz.

The value of the total intensity which is to exceed the assigned threshold determines the start of signal input. The final decision on the start of input is taken if several successive input segments meet this condition. The decision to end the input is taken if several successive segments have a total intensity below the threshold. Otherwise the segments of low intensity correspond to speech pauses.

The threshold value and the required number of segments in the first and the second case are assigned by the user.

Performed parallel with the process of input initial intensity vectors worked out by the pre-processor are the operations:
- replacing the assigned number of input intensity vectors by one averaging,
- uniting the averaged close contiguous vectors of intensity into groups (segments) and the subsequent averaging(segmentation) of the latter.

The secondary averaging is performed for a group of vectors in which the distance between the first vector and all the subsequent ones is less than the so-called segmentation threshold also assigned by the user. After the segmentation the secondary features are constructed:

$$R_\ell = 256\,(P_\ell + G_\ell)/(Q+D)\ ,\quad \ell = 1\div 6\ ,$$

where P_ℓ is the value of averaged intensity in the 1-th band after segmentation, $Q=\sum_\ell P_\ell$, G_ℓ and D are reguleting additives.

The software of the system includes the program of input, training and discrete speech prase recognition, and the programs intended for preliminary vocabulary compilation and for forming a tree of word compatibility in phrases.

The program of input, training and recognition performs the following functions
- distribution of on-line storage among the templates E_i and a test realization, which is defined as the next word from a phrase F_K at the stage of recognition;
- training on the basis of a given vocabulary;
- recognition of discrete speech phrases;
- recording temlates in the file;
- replacing separate templates.

The entire working process is carried on as a dialogue between the user and the computer. To each action required from the user the program supplies prompting requests. This permits a wide range of users to ceadily master the system. In the stage of training the speaker pronounces into the microphone once each word S_i from a given vocabulary. There is a possibility of replacing templates introduced erroneously at the stage of training. The words in phrases are recognized by the method of dynamic programming with an adaptive corridor [2].

Let an utterance S_m consist of ℓ segments. On introduction of the next segment with a number j there occurs a recalculation of all R_i $(i=1,2,\dots,k)$there R_i are the distances from S_m to templates E_i calculated with the help of an algorithm of dynamic programming. Such an arrangement of calculating R_i allows, in the process of recognition S_m , unpromising templates to be cut off. The cutting-off condition has the form: if $R_i/R_{min} > B$ where $R_{min} = \min R_i$ then the template E_i is cut off as unpromising. The cutting-off threshold decreases monotonously with growing j . The cut-off templates take no part in subsequent calculations till the recognition of S_m is over.

Such a behaviour of the cutting-off threshold makes it possible, at the first steps, to exclude from consideration the templates the most different from the test realization by their initial segments. As the number of control realization segments grows account is taken of ever finer distinctions. The recognition comes to an end if at a certain step there remains but one template. Otherwise after the calculation of distances is over a template is chosen with the minimal distance to the test realization. If $R_{min} > R^*$ where R^* is the value of the refusal threshold, the system refuses to recognize the words of S_m .

ALLOWANCE FOR LINGUISTIC RESTRICTIONS

Admissible sequences of words in phrases can be presented in the form of a tree where each branch reflects a continuation of the admissible phrase. Such a presentation is more economical than the conventionally used matrix of word combinability.The tree is built in the form of a two-dimensional array uniting a sequence of units each of which is a set of admissible nodes (word numbers) having a bond with the node of the previous level (the first line of the unit), and reference addresses to the units of the subsequent level (the second line). The reference addresses equal to zero determine the end of the phrase.

The input data describing the phrases are represented as a sequence of lines each of which describes a particular phrase (a group of phrases) or part of it and has the form: $[j]\mathcal{A}_1,\mathcal{A}_2,\dots,\mathcal{A}_i,\dots,\mathcal{A}_k[*]$ where j is the current phrase number, $\mathcal{A}_i'$ - a number or a set of numbers for words that could stand in the i-th place of a phrase, * is the simbol for the phrase continuation in the next line if its description fails to fit into one line.

Example:

1. (10,12), 11(1,2,3,4,5,6,7,8,9), (13,14)
2. (25) 24(23,21)(13,14)
3. (15)(29,92)

The first line of this file describes 36 phrases in which the first place may be occupied by the 10-th or 12-th word, the second by the 11-th, the third by the 1-st, 2-nd,...,or 9-th; the fourth by the 13-th or 14-th word of the fixed vocabulary. The second line describes 4 phrases in which the first place may be occupied by the 25-th word, the second by the 24-th, the third by the 23-d or 21-st, the fourth by the 13-th or 14-th. The program of forming a tree of word compatibility in phrases operates in the dialogue mode and makes it possible to introduce initial data determining the sequence of words in a phrase from the terminal keyboard or from an earlier prepared external file. Taking into account the large variety of identical branches the program eliminates repeated branches which allows the required volume of memory to be reduced 5 to 6 fold.

The array of phrases constructed according to the above example has the form

1	2	3	4	5	6	7	8	9	10	11	12	13	14
28	10	12	25	15	-2	11	-2	1	2	3	4	5	6
4	6	6	21	26	1	8	99	18	18	18	18	18	18
15	16	17	18	19	20	21	22	23	24	25	26	27	28
7	8	9	-2	13	14	-1	24	-1	23	21	-1	29	92
18	18	18	2	0	0	1	23	2	18	18	2	0	0

The initially constructed array describing the tree of word compatibility in phrases would contain 306 words instead of 56 as is the case after the optimization.

EXPERIMENTAL RESULTS

The system was tested on phrases of a problem-oriented vocabulary belonging to the language of an air-traffic-dispatcher. The vocabulary contained 120 words. On the material of 140 phrases made up of 3 to 7 words with speaker adaptation the recognition reliability obtained amounted to 98%. The branching factor varied from 1 to 48 and on the average was equal to 13. The system worked in the real time scale. At present the system is in experimental operation.

REFERENCES

[1] G.Ya.Vysotsky, B.N.Rudny, V.N.Trunin-Donskoy, G.I.Tsemel, "Experience of Speech Control by a Computer", Izvestiya Akademii Nauk SSSR. Tekhnicheskaya kibernetika, Moscow, pp.134-143, no.2, 1970.

[2] H.Sacoe, S.Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. Acous. Speech and Signal Processing, vol. ASSP-26, no.1, pp.43-49, 1978.

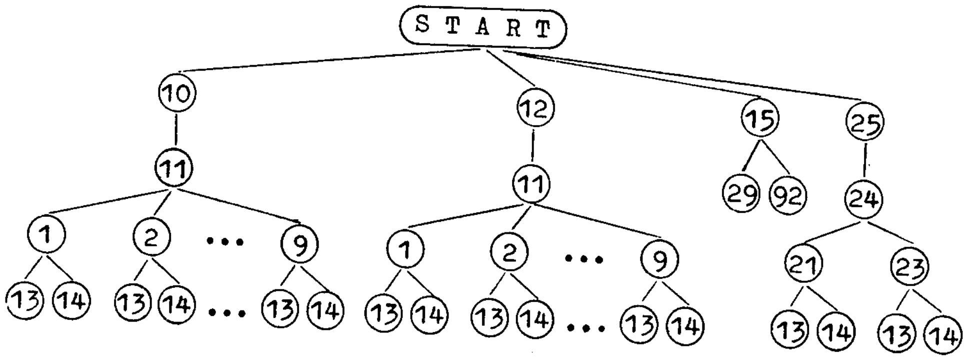

Fig. 1. Example of the tree of permissible phrases.