ACOUSTIC – PHONETIC BASIS OF SPEECH RECOGNITION ALGORITHMS

G. G. RODIONOVA

Computer Centre of the USSR Academy of Sciences,
Vavilova str., Moscow, USSR, 117333

ABSTRACT

The algorithms based on the wider use of accustic-phonetic (APh) information are described. These algorithms include APh-clustering of training set and APh-classification on unknown message. An APh-structure description of speech signal is presented.

1. INTRODUCTION

Automatic speech recognition (ASR) is a key problem of the modern speech technology. In last years a variety of approaches to ASR have been explored and certain progress has been made in this field. This progress is largely due to the use of widely adopted formalistic techniques such as the most popular dinamic-programming (DP) method. DP-technique is based on whole-word template matching making it's performance quite high due to the absence of segmentation error and other advantages. However such problems as large time and storage requirements, dicrimination of similar words, account of coarticulation effects arise. Now it is quite clear that this approach is not promising. An inevitable return to accounting for "human" aspect of speech signal requires the design of acoustic-phonetic information based algorithms. In this paper we briefly describe the algorithms containing a set of procedures used for APh-clustering in training and recognition.

2. PARAMETRICAL DESCRIPTION OF SPEECH SIGNAL

The accuracy of recognition is evidently dependent on reliability of every level of a recognition system, and errors in coding and parametrical description are the most essential.

The speech signal processing hardware has been developed in the Computer center of the USSR Academy of Sciences. This hardware is based on the principle of maximal account of acoustic and phonetic features of a speech wave. Special devices are designed to extract and input into computer a variety of parameters, characterising the 10–20 ms intervals (time-segments) such as:

F_1, F_2 – average first and second formant frequency,

F_0 – pitch frequency,

N_0 – number of zero crossings,

A_0 – total energy etc.

Among these parameters a set of the most informative ones $\{P_i\}$, $i = 1,\dots,4,5,6$ has been extracted. These parameters ensure that the requirements of minimal time and maximal accuracy of recognition are fulfilled.

When an algorithm operates with templates and uses the APh-information, it is necessary that these APh-features provide the distinction between all of the templates in the conditions of real-time processing. We have developed computer programs to obtain some lexical parameters which reflect the phonetic structure of a message. So the parameters show the presence (or absence) of certain phoneme-like subword units (ph-segments) and their order in a word. As a result of the procedures the secondary characteristics from a set of primary ones have been obtained; e.g.:

$R_1 = 1$, if a word S_i contains the noise consonant (NC) segment of a duration τ^i, which does not exceed a threshold value, $\tau_{th.}$: $\tau^i < \tau_{th}$, $N_0^i \leq N_{0th}$, $A_0^i < A_{th}$;

$R_4 = 1$, if the stressed vowel is of the "a" type, i.e. $\tau^i > \tau_{th}$, F_1^i and F_2^i are lying in their standart domain of mean values, $A_0^i > A_{th}$, etc.

Vector $W = \{R_i\}$, $i = 1,\dots,8,16$ reflects a certain information about message phonetic structure, for example, the word "сани" is characterised by vector:

$$W = \{1,1,0,1,0,0,1,0\}.$$

This means:

$R_1 = 1$: there is a noise consonant (NC) during the given word realization (WR);

$R_2 = 1$: this NC is in the beginning of the word;

$R_3 = 0$: there are no second NC in the word;

$R_4 = 1$: the stressed vowel is of the "a" type;

$R_5 = R_6 = 0$: the stressed vowel is not of "y" or "и" type;

$R_7 = 1$: the NC has the energy maximum in the high frequency region;

$R_8 = 0$: R_8 is not computed when $R_3 = 0$; if $R_3 = 1$, then the value of R_8 depends on N_0 of the second NC.

Such a description is rather rough, but it is of a reliable nature. When the number of secondary features is equal to 16 or 24, the vector W is more informative, but in this case the description has some evident desadvantages. Thus, the original speech signal is presented by means of full description (FD), being two kinds of parameter, having different levels of extracting and different powers of adequacy.

Namely FD consists of:

(a) primary description - a temporal matrix $\|P_T\|$, where T is the message duration in 10 ms time-segments and

(b) secondary description - a vector W of binary lexical features (ph-segments).

This representation is more accurate then the one obtained with whole-word templates, where phonetical variations can be expressed only by adding other templates. It also shows better the distinctions between similar words. Such a description provides a more natural way of dealing with acoustic-phonetic information and, on the other hand reduces considerably the required amount of training material.

3. APh -CLUSTERING OF A TRAINING SET

The recognition system software consists of two parts: a teaching one, which provides a training set and a recognizing part, destined to carry out the search operations. The training set is formed by means of pronouncing every position of given vocabulary $\{S\}$ (a word or a word combination with their attributes indicated: word code, speaker name, etc.). Input and full representation (primary and secondary) for every utterance is made. These descriptions are stored in two computer memory domains. Then the training set of the length M: $\{S_i\}$, i = 1,..,M, (which is at the beginning structureless) is clustered on phonetical features base, i,e, is divided into J structural clusters C_j, j = 1,...,J. The clustering process is performed with the aid of vector W components, lying in the nodes of a binary logical tree (BLT). These components are previously arranged and BLT is constructed after the user manner. It should be noted that the total number of terminal clusters J is not equal to 2^k, where k is the length of vector W. It is so since every branch of BLT does not contain all of the theoretically possible nodes due to the special nature of W. Thus we can estimate the mean number of templates, N_j, in the cluster C_j:

$$N_j \approx M/J.$$

In case of phonetical nonbalanced vocabulary this estimation may turn out to be rather approximate, but this fact is not of a great importance. The point of the method is that a cluster, C_j, contains a template which is relevant to unknown utterance with the same phonetical label j. APh-clustering is carried out automatically with the help of especially developed procedures of speech recognition system.

4. RECOGNITION ALGORITHM BASED ON APh-CLUSTERING

Spoken message pertaining to a given vocabulary $\{S\}$ is recognized by looking for a relevant pattern through a subset of templates that maximizes a measure of similarity with the input signal. The main feature of algorithms under consideration is that the search for a maximal similar candidate is made within the templates that form one cluster without any resortion to the remaining templates. The sample to check the recognition algorithm was defined by a given vocabulary $\{S\}$. First, the imput utterence S^i was transformed into primary parameter description, the matrix $\|P_T\|^i$. For the same message a secondary parameter sequence - vector W^i was calculated (in real-time) and an APh-classification was made by the values of vector W^i's components. So the S^i got a structure label, i.e. it was marked by the number of "its" cluster, j. APh-classification procedure was performed by using the same learning binary logical tree as in the learning stage. The fact that both the template and the searched for descriptions belong to the same terminal APh-cluster C_j makes it possible to restrict the search for relevant candidate $\tilde{E}$ to objects of C_j only: $\tilde{E} \in C_j$.

The choise of search strategy is of great importance to the outcome (error rates and recognition time), but the described algorithms are independent of this strategy. In our case the relevant candidate $\tilde{E}_j$ was found by comparing the parametrical matrix $\|P_r\|^i$ with the matrices of templates composing cluster C_j.

If APh-clustering technique is well developed, the algorithms under consideration not only shorten the recognition time on the average by a factor of J times, but also improve the accuracy. The latter takes place because the smaller the number of processed templates the lower the error in classification.

CONCLUSIONS

Adequate description of a speech object is always of great importance. But in the problems on speech recognition dealing with an object that is highly variable in time and in the parameter space, the question of optimal formalization of this object is decisive. The algorithms described may be of interest for those who develop speech recognition systems and who realize that the role of acoustic-phonetic information should be strengthened.