COMPUTER RECOGNITION OF ISOLATED WORDS IN FIXED LENGTH FEATURE SPACE

ALGIMANTAS RUDŽIONIS

Speech Research Laboratory
Kaunas Polytechnical Institute
Kaunas, Lithuania, USSR 233028

ABSTRACT

After the bounaaries of separate words were defined, each word was divided linearly into K_S static and K_D dynamic segments. With a fixed number of spectral components L, arbitrary reference words have equal volumes. Several vocabularies of digits and of 100 geographic names were read, and recognition accuracy was estimated in relation to K_S, K_D, L and to the number of repetitions R. It is shown, that with a satisfactory training, even small computers can recognize about 100 words with a 1 to 2% error rate without any devices of increasing their operation speed.

I. INTRODUCTION

A good number of studies on the main parameters of speech recognition systems, and to the main factors determining their success have appeared recently /1-4/.We find such studies as /3, 4/ revealing the difficulties encountered in attemps to evaluate the advantages of different parameter descriptions of signals, as well as of different methods of comparison.
Some of the recent works consider comparison of words divided into fixed number of segments /5,6/. This approach ensures considerably higher recognition speeds,as compared to the non-linear time mode.Perhaps because of the insufficient local similarities of words, the advantages of dynamic transformations of time function are not always exploited.
The present study was aimed at an evaluation of speaker dependent recognition. when separate words are described by fixed numbers of static and dynamic segments of the speech signal, and when the reference of any word is of the same dimension.This mode of description opens new potential ways of word comparison, including polydimensional rating. We consider here the most simple, but not the less important parameters: number of static K_S and dynamic K_D segments, number of spectral components L and effect of training as number of repetitions R.
Our results on several vocabularies of Russian words suggest a continous decrease of the error rate with increasing K_S, K_D, L, R. In particular, which $K_S \sim 8$, $K_D \sim 4$, $L \sim 8$ and satisfactory training, on-line recognition of random vocabularies of about 100 words is possible without any increase in the processing speed, at the error rate of 1 to 2 %.

2. GENERAL REMARKS

The suggested approach has the undoubtful advantage of speed. The effects of small values of K_S, K_D, L, R on its reliability is to be estimated yet. Intuition sugthat higher values of these parameters should extract more information from the speech signal. But we are also interested in the least admissible values and expect significant effects of the vocabulary content. Let us now consider some special aspects of the problem.

2.1. Extration of Features from Speech Signals

Normal logarithmic spectral from the filter bank were chosen as primary features. A word was first described by sequence of spectral vectors $S_K(\ell)$ every T seconds, where K=1,2,...,K_W number of spectral vector; K_W - duration of a word expressed by the number of vectors, and ℓ =1,2,...,L number of a spectral component.For a word divided into K_S static segments, its division points φ_s are

$$\varphi(s) = INT[1+(s-1)(K_w-1)/K_s] \qquad (1)$$

INT stands for a whole part, S=1,2,...K_S+ +1. Static segments $F_S(\ell)$ are formed by averaging spectral vector in intervals K= $= \varphi_s , \dots , \varphi_{s+1}$

$$F_s(\ell) = \frac{1}{\varphi_{s+1}-\varphi_s+1} \sum_{k=\varphi_s}^{\varphi_{s+1}} S_k(\ell) \qquad (2)$$

The filter analyser used contained L=24 filters, discributed on a semi-logarithmic scale. Dimensional variation of features (L) over the frequency scale was performed either by averaging subsequency spectral frequency components, or by choosing a certain part of component say, in the range of the telephone channel, which is further denoted by L_T.

2.2. Introduction of Dynamic Segments

A certain part of useful information in recognizing speech signals may be gained from changes of the signal spectrum. Its most simple estimation is through the spectrum dynamics

$$\tilde{S}_k(\ell) = \sum_{j=1}^{J} |S_{k+j}(\ell) - S_k(\ell)| \qquad (3)$$

where K=1,...,K_W-J. For the K_D dynamic segments $F_D(\ell)$, in (2) $S_K(\ell)$ is replaced by $\tilde{S}_K(\ell)$. One of the other possible estimations of spectrum dynamics is weighting by linear functions. But its result was not significantly better, so that (3) was used thanks to its simplicity. The absolute value of (3) has an effect of limitation, but we used it successfully to denote preservation of feature values inside one byte, without any additional operation. Vectors $S_K(\ell)$ were followed every 10 ms with assumed J=3. Size of a reference word was in our case equal to L(K_S K_D) bytes and did not depend on the duration K_W of a word. Recognition was carried out by Euclidian distance minimum.

2.3. The Level of Training

The often applied one-time reading of a vocabulary as a means of training for isolated words recognition systems suffers a high degree of randomization of reference words. As a main parameter in our study we chose the estimate of the training set, that is the number of repetitions R for a reference word. The level of training is significantly dependent on the distance measure and on the dimension of the reference word. From practical considerations we consider here only small values of R. On the other hand, a proper level of training can reveal the information efficiency of separate features, and the application limits of the suggested approach.

2.4. Vocabularies Studied

The resultant recognition accuracy is significantly influenced by the vocabulary length and especially by its content. To combine practical interest with fundamental solutions, we selected the following vocabularies: w_1- names of 100 towns in the Soviet Union, included as a demonstration of the recognition accuracy on vocabularies of medium length; w_2 - Russian digits from zero to thousand (38 words); w_3 - Russian digits from zero to nine (10 words); w_4 - Russian digits from eleven to twenty (10 words). Vocabularies w_3 and w_4 constitute parts of vocabulary w_2. Vocabulary w_3 consists of most frequent in computer recognition words, and

vocabulary w_4 is a most difficult vocabulary, because all its words have a common stressed second part.

3. EXPERIMENTAL RESULTS

Most of the results refer to a single speaker. Hardware consisted of a dynamic microphone, a spectral analyser and a microcomputer. Each word was repeated several times, but direct recognition was estimated repeatedly on the second reading, until a statistically significant relation was found. Where possible, several parallel reference vocabularies were formulated. The number of control outputs for each test point was from 500 to 8000.

The main studied parameters were: K_S, K_D - numbers of static and dynamic segments; L - number of spectral components (in the telephone channel range L_T); R - number of training repetitions.per word; N - vocabulary length; B - number of bits for a vector component of a reference or of an input; N_C - number of control inputs; T - period of following the spectral vectors from filter analyser ms.

Number of static segments. Fig.1 shows recognition accuracy for different K_S. A significant effect of the vocabulary structure is evident, Fig.1b, combined with the effect of the level of training Fig.1a. Vocabulary w_4 is by far more difficult, than vocabulary w_3, because in w_4 information is concentrated in the first unstressed parts. Vocabulary w_2, which is nearly four times longer, is nearly as difficult as vocabulary w_4 for R=I, but its recognition accuracy improves largely with the number of repetitions, Fig.Ia. In general, recognition accuracy is unsatisfactory at K_S 8.

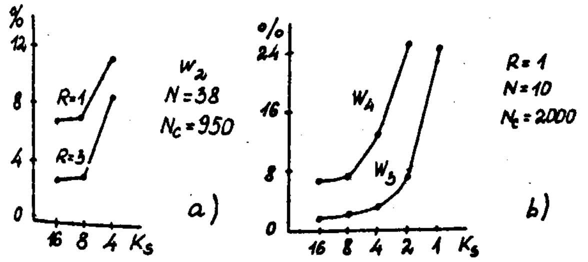

Fig.1. Recognition errors as a function of the number of static segments. (L_T=8, T=10)

Dimension of spectral representation. Typical results in Fig.2 support the expected increase of the recognition accuracy at larger L.

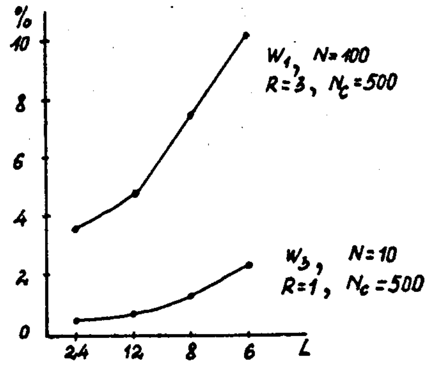

Fig.2. Recognition errors under the effect of spectral resolution (K_S=4, T=20).

Level of system training. The effect of R as also estimated on vocabularies w_3 and w_4 which described by K_S=4 static segments. The error rate in vocabulary w_4 was never lower than p≈7% because of its incomplete information, even at levels of training as high as R=I00. Vocabulary w_3 was nicely recognized at R=3 (Fig.3.

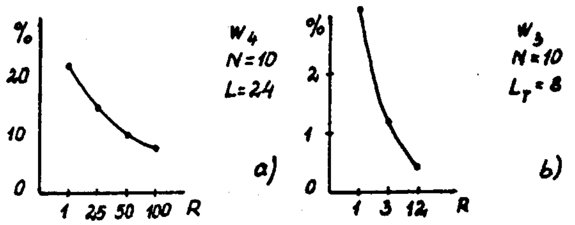

Fig.3. Recognition errors of w_3 and w_4 at different levels of training (K_S=4, T=20).

Effect of spectrum dynamics. Introduction of dynamic segments was shown in 2.2. Let the words be presented K_S=6 static and K_D=4 dynamic segments. A proper level of training was ensured by numerous repetitions (R=20). It follows from table I, that dynamic segments carry 2 to 4 times less information than static ones, but combined application of both gives 3 times error rates, than static segments only. Note also, that with a satisfactory training, the accuracy of recognition is nearly independent of the vocabulary length.

Table 1. Recognition errors of three vocabularies from their static dynamic and mixed segments, % (L_T=8, R=20, $N_C \approx$ 1000)

K_S	K_D	w_1 N=100	w_2 N=38	w_4 N=10
6	-	6,5	5,3	5,6
-	4	15,2	19,5	11,6
6	4	1,8	1,6	2,2

<u>Presentation accuracy of reference words.</u> The volume of a reference word depends on the number of bits, which are given for each component of the vectors sequences. On the other hand one might expect that the more reliable reference words are also less sensitive to the accuracy of their presentation. Fig.4 shows the recognition errors of 50 words (vocabulary w_2 plus 12 control words) and of vocabulary w_4 of ten words, when the reference words and the input words were represented by a different number of bits (B). The lower curve, Fig. 4a represents reference words of increased reliability thanks to higher R. Here a decrease of B down to 3 bit has no significant influence on the result. Vocabulary w_4 is highly sensitive to a decrease B of even at R=100.

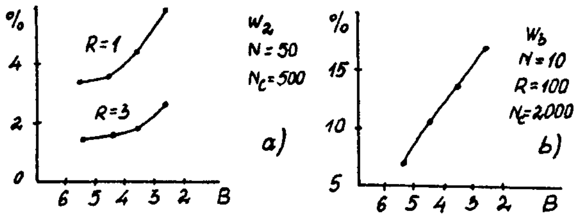

Fig.4. Recognition error rate of the bit number in the reference sequence (L=24, K_S=4).

4.CONCLUSIONS

1. Recognition of words in finite spaces of features with small reference sequencies (8 to 16 segments) and proper levels of training can be suggested for vocabularies of up to 100 words and medium processing speed.
2. Introduction of dynamic segments considerable (3-fold) descrease of the erros rate. A continous decrease of the error rate was observed with increasing K_S, K_D, L and R.

/1/ N.R.Dixon, H.F.Silverman, "What are the significant variables in dynamic programming for discrete utterance recognition?", ICASSP 81, pp.728-731.

/2/ A.Weibel and Yegnanarayana, "Comparative study of Nonlinear Time Warping Techniques in Isolated Word Speech Recognition Systems", IEEE Trans.ASSP-31, pp.1582-1586, No.6, Dez.,1983.

/3/B.A.Dautrich, L.R.Rabiner, T.B.Martin, "On the Use of Filter Bank Features for Isolated Word Recognition",ICASSP 83, pp.1061-1064.

/4/ N.Noncerino, F.Soong, L.Rabiner, D. Klatt,"Comparative Study of Several Distortion Measures for Speech Recognition", ICASSP 85, pp.25-28.

/5/ D.Burton, "Applying Matrix Quantization to Isolated Word Recognition", ICASSP 85, pp.29-32.

/6/ H.Iizuka, "Speaker Independent Telephone Speech Recognition", ICASSP 85, pp.842-845.