Development of the Approach to Recognition of Speech With a Support on Visual Images

Potapova R.K., Sobakin A.N.

Department of Applied and Experimental Linguistics, Moscow State Linguistic University, 125445, Moscow, Ostozenca, 38, Russia
e-mail: potapova@linguanet.ru; mglu@online.ru
Fax: (095) 246-28-07;
Tel.: (095) 201-56-97

ABSTRACT

This paper describes the concept of receipt of useful phoneme-acoustic information on the bases of nature of visual images (and its parts) configuration by means of associative connections [1] specially for similarity of local-time character.

1. INTRODUCTION

The tasks of speech recognition with difficult conditions of transmission and by means of not precise methods of representation of speech signal need the research of this kind.

Except for a hierarchical way of the account of feature units the parallel analysis of the quantitative characteristics of offered attributes is possible. **The method** realizing parallel way of classification of words, is based on use of a matrix of affinity received for the relation of belonging to classes of words. The given descriptions of two words concern to one class, if taxonomic distance between the descriptions in space is least.

2. METHOD AND EXPERIMENT

With the purpose of demonstration of serviceability of an offered technique we shall consider procedure of classification of the descriptions of ten words (for figures) above mentioned. As the initial data we use values of related features for them [2].

In connection with that a number of features used for the description of the images of words, is given by several values minimal and maximal, for several figures of the image of a word etc.), with the purpose of simplification of the further calculations we shall take advantage only of their average values.

According to the offered attributes we shall receive a matrix of distances (tab. 1).

For realization of procedure of classification we use the feature of a belonging for threshold value equal 10.

Having applied the chosen threshold value concerning all elements of a matrix of distances (see tab. 2), we received a logic matrix of the given relation, by replacement of elements exceeding 10: "zeros", and all others – "ones" (tab. 2).

Having analysed a logic matrix, we see, that the given relation of the belonging are derivated by some variants of classes of equivalence (groups) of words. In these classes are included words, to which correspond individual elements of a matrix of the relation.

As a result we received:
{1.4,8}, {2,7,9,0}, {3,5}, {6} or
{1.4,8,0}, {2.7,9,}, {3}, {5.6}.

The multialternativeness of classification is a consequence of intransity of logic matrix of the given relation. It is easy to be convinced of it, having applied to it criterion of check of property of transitivity. It is necessary to apply to elimination of multialternativeness a method of transitive short circuit consisting in multiplication of a logic matrix on themselves so long as its elements will cease to change.

Division of initial set of the images of words into groups more large, than the separate word, is received by virtue of that circumstance, that we used in example only four features from seven, used at direct classification.

Received preliminary results of use offered in the paper attributes of the approached description of speech images of separate words confirm an opportunity of their division with the help of this system attributes.

At the same time it is necessary to note preliminary character of such conclusion owing to limited volume of sample of images used for experiment both by amount of realizations, and on number of the speakers.

For a substantiation of more universal character offered or not how many modified system of features will be carried out more extensive experiment: the experimental and number of the subjects increased.

Besides the preliminary conclusion about an opportunity of development of the automated technique of classification of the images of words and its realization can be made on the basis of a parallel way of classification.

Processing of parametrical representation of words received with use of band-spectrograph can be considered in quality of one of tasks of statistical processing of the graphic information having two levels of "brightness" in each "point" of the image.

Each realization of a word of one speaker in such graphic representation has its own frame, the beginning and the ending of this one is defined accordingly by beginning and ending of pronouncing of the given speech formation.

The beginning of a word in all experiments was determined on occurrence of the first not "zero" in one of frequency channels from F1 up to F2 (in frequency interval it corresponds with frequencies from 100 Hz up to 4000 Hz). Not "zero" values of each of the specified ranges correspond to excess of accumulation for 20 ms of energy of a speech signal of some threshold value.

Thus, size of threshold value, its choice in relation to level of a signal are very important for steady and reliable finding of the left border of the frame. From the further description of a way of statistical processing of frame it becomes clear, as far as this parameter is important for all procedure of statistical recognition in whole.

The right border of frame of the image of a word was determined similar on last value distinct from zero, of a level of a signal, in same frequency ranges also depend on the same threshold value.

Each realization of a word frame was put in conformity the contrast (graphic) image, on the basis of which statistically were created model sequences.

Creation of model sequences: for creation of model sequences each frame with image of a word was exposed to preliminary transformation which consist in the following.

The image was transformed into a sequence of value of brightness (zero and one) by means of transformation of everyone vertical frequency section of a speech signal in a horizontal vector. Such transformation was carried out by "turn" of a vertical vector on 90 degrees clockwise, that transformed a vertical vector in horizontal. Its left values corresponded with the bottom frequency ranges, and right – high frequency ranges. Each subsequent vector incorporated with previous on the right and, thus, long horizontal vector – was formed line (sequence from "zero" and "ones").

After the described preliminary transformation of each frame of the image we have a set of sequences for set of pronouncings of ten words by ten speakers. This set of sequences serves a statistical material for formation of model sequences.

Formation of a model sequence for each word (class) was carried out by calculation for each position of a sequence of probability of occurrence in it of unit on ten speaker realizations. In practice not probability of occurrence of unit, and size appropriate to numerator of this probability was used. In other words, was counted up amount of units in each position of sequences for the various speakers.

Sequence, received on the basis of statistics, was accepted for reference for the given word (class).

Reference sequences for everyone thus were created words (class). In the given experiment number of classes of recognition coincided with amount of words of recognition and was equaled to ten. As much was created of reference sequences.

At a stage of recognition the showed(presented) sample of a word as a sequence of "zero" and "ones" is compared to the standards and the measure of similarity " of a represented sample for each standard is calculated ". The algorithm of calculation " measures of similarity " creates in set of binary sequences metric space being base for each algorithm of races cognition.

In considered space of binary sequences the metrics as the sum of conditional probabilities of occurrence of "one" or "zero" is offered in each position of a sequence.

This sum of conditional probabilities in view of the made remarks is counted up as follows. If in the presented standard in some position there is a "one", in the sum the number of "ones" from reference is added from sequences of a checked class in the same position. If in the presented standard in a considered position there is a "zero", in the sum the difference is added. Last action corresponds to addition in the sum of probability of occurrence of "zero" (additional probability) in the given position.

The complete summation is made on length of the presented word. In the volume a case, when the word comes to an end before the standard (model), procedure of summation comes to an end. If the standard is shorter than a word, it is supplemented in "zero", that corresponds to "zero" probability of occurrence of "ones" in these positions.

The received sum is normalized on the length of presented word, and this sum is accepted for a measure of similarity of the presented word.

The greater value of a measure of similarity corresponds to greater "similarity" of the presented word to the given class. The maximal value of measures of similarity on all classes will correspond hypothetically to belonging of showed word to this class.

3. CONCLUSION

The described algorithm of recognition was examined on a material to ten words realized by ten speakers.

Experimental results of comparison of ten words pronounced by the speaker C., with reference sequences are given in the table 3. It is necessary to note,

that the showed speech samples of the speaker C. not participated in reception of the samples for comparison.

As it is visible from the given table, the greatest value of a measure of similarity are located on the main diagonal of a matrix. It means in the whole serviceability of such approach in recognition of the images of words.

At the same time, there are separate difficult cases of comparison for offered measures of similarity. In particular, the measure of similarity of a word "eight" with various sample sequences has no the brightly expressed maximal value, that can be explained as large variability of separate speaker realizations of the word which has served with a basis for creation of the sample of the appropriate class ("eight").

Variability of speaker realizations in this class is explained, on the one hand, by various degree of a reduction of past-stressed vowel, on the another – by various degree of coarticulation, penetrating all word as a whole.

The possible ways of increase of reliability of algorithm of recognition consist in expansion of statistical base of creation of the samples and improvement of techniques of comparison of speaker realizations with the samples.

REFERENCES

1. R. K. Potapova. "Ob odnom podkhode k klassifikatsiji reche-zritelnych obrazov na baze assoziativnykh svjazej. Mater. VI Vseros. Konf. "Nejrokompjutery i ikh primenenije". M., 2000.
2. R.K. Potapova. "Rech: kommunikatsija, informatsija, kibernetika." Radio i Svjax. M., 1997.

APPENDIX

Table 1. A matrix of distances

Word	1.	2.	3.	4.	5.	6.	7.	8.	9.	0
1.	0	11,25	24,29	1,02	21,4	16,05	14,09	7,28	12,17	6,20
2.		0	21,26	12,03	16,28	10,68	3,61	13,91	9,30	8,21
3.			0	23,85	6,64	11,05	21,01	18,5	13,18	18,00
4.				0	21,77	16,08	14,86	7,24	12,50	6,65
5.					0	6,35	15,34	17,85	9,73	15,24
6.						0	10,45	12,88	4,95	9,46
7.							0	16,43	9,63	10,35
8.								0	10,60	6,44
9.									0	6,06
0										0

Table 2. A logic matrix of distances

Word	I	2	3	4	5	6	7	8	9	0
I	1	0	0	I	0	0	0	1	0	I
2		1	0	0	0	0	I	0	I	I
3			1	0	I	0	0	0	0	0
4				I	0	0	0	I	0	I
5					I	I	0	0	I	0
6						I	0	0	I	I
7							I	I	0	I
8									0	I
9									I	I
0										I

Table 3. Experimental results of comparison of ten words, pronounced by the speaker C., with reference sequences.

Word	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
"0"	6,11	4	4,69	3,7	5,37	3,96	4,18	3,35	5,02	4,99
"1"	4,43	6,54	3,25	2,63	5,65	3,12	2,94	2,05	4,31	4,1
"2"	5,28	3,07	6,77	4,55	3,48	5,73	4,84	3,79	4,11	5,08
"3"	2,84	3,3	3,81	7,84	4,08	5,35	5,47	5,93	4,53	5,65
"4"	4,41	5,28	3,69	4,5	7,51	3,77	4,44	3,45	4,74	5,74
"5"	3,98	3,47	4,37	5,34	3,32	7,15	4,32	4,26	3,10	5,01
"6"	4,51	4,57	5,46	6,7	6,4	5,77	7,54	5,52	5,8	6,21
"7"	2,73	2,89	3,41	6,63	3,86	4,65	3,91	7,06	3,48	5,02
"8"	4,98	4,86	4,04	4,76	5,62	4,25	4,47	3,52	6,06	6,05
"9"	4,31	3,31	4,48	5,59	5,96	4,89	5,93	4,73	5,42	7,43