A KIND OF CONTINUOUS DIGIT SPEECH RECOGNITION METHOD

Wenming Cao
1Institute of Intelligent Information System, Information College, Zhejiang University oj Technology, Hangzhou 310032, China

Abstract: In the light of descriptive geometry and notions in set theory, this paper redefines the basic elements in space such as curve and surface and so on, presents some fundamental notions with respect to the point cover based on the High-Dimension Space(HDS) point covering theory, finally takes points from mapping part of speech signals to HDS, so as to analyze distribution information of these speech points in HDS, and various geometric covering objects for speech points and their relationship. Besides, this paper also proposes a new algorithm for speaker independent continuous digit speech recognition based on the HDS point dynamic searching theory without endpoints detection and segmentation. First from the different digit syllables in real continuous digit speech, we establish the covering area in feature space for continuous speech. During recognition, we make use of the point covering dynamic searching theory in HDS to do recognition, and then get the satisfying recognized results. At last, compared to HMM-based method, from the development trend of the comparing results, as sample amount increasing, the difference of recognition rate between two methods will decrease slowly, while sample amount approaching to be very large, two recognition rates all close to 100% little by little. As seen from the results, the recognition rate of HDS point covering method is higher than that of in HMM-based method, because, the point covering describes the morphological distribution for speech in HDS, whereas HMM-based method is only a probability distribution, whose accuracy is certainly inferior to point covering.

Keywords: High-Dimension Space, High-Dimension Space Covering Theory, Continuous Speech of Speaker-Independent

1. INTRODUCTION

In the recent years, Wang Soujue[1-4] proposed a point covering theory in high dimensional space (HDS), which uses the HDS descriptive geometry as tools, maps multidimensional digital signals to be points in HDS, and discuss relationship among multidimensional digital signals via analysis of point distribution and geometric object characteristics in HDS. Since this method is simple and feasible, and is able to accelerate computation by parallel operation, it has get remarkable achievements in many application fields [5-14].

In the light of descriptive geometry and notions in set theory, this paper redefines the basic elements in space such as curve and surface and so on, presents some fundamental notions with respect to the point cover based on the HDS point covering theory, finally, as an example, considers points generated by mapping part of speech signals to HDS to analyze distribution information of these speech points in HDS and various geometric covering objects for speech points and their relationship. Besides, this paper also proposes a new algorithm for speaker independent continuous digit speech recognition on the basis of the HDS point dynamic searching theory without endpoints detection and segmentation. First from the variant digit syllables in real continuous digit speech, we establish the covering area in feature space for continuous speech. During recognition, we make use of the point covering dynamic searching theory in HDS to do recognition, and then get the satisfying recognition results. At last, compared to HMM-based method, from the development trend of the comparing results, as sample amount increases, the difference of recognition rate between two methods will decrease slowly, while sample amount approaching to be large enough, two rates all close to 100% little by little. As seen from the results, the recognition rate of HDS point covering method is higher than that of in HMM-based method, because, the point covering describes the morphological distribution of speech in HDS, whereas HMM-based method is only a probability distribution, whose accuracy is certainly inferior to point covering.

2. SPEECH MORPHOLOGICAL ANALYSIS IN HDS

2.1 Cover

For geometry objects a and β, $\alpha \subset \beta$ in n-D space, β is said to be a cover of a or a is covered by β, β is the covering graph, a is the covered graph.

If for subspaces a and β, congruent transformation $P:\alpha\rightarrow\gamma$ exists, let $\gamma\subset\beta$, we say β can cover a or a can be covered by β. If after covering, $\gamma=\beta$, say **congruent cover**.

If a is the set of discrete points in n-D space, the feasibility using points to do congruent cover is almost zero. In order to process it quickly, we adopt fixed graph to cover it, say one cover which has the minimum volume of cover graph is the minimum cover of this graph, if using triangle to do covering, say minimum triangle cover, if sphere, say minimum sphere cover. If points in the set of covering points have been distributed to the border line of covering graph, that is, the covering graph cannot be shrank, we call it as vertex cover. This type of cover for the point set is said to be the point cover, vertex cover is one of point covers, which is the minimum point cover in the same class graph with different size.

Since signal is usually some discrete points in the computer signal processing, we focus on the point cover in HDS, therefore, if no special declaration, the cover below is the point cover.

2.2 Speech distribution in HDS

◆ A continuous single digit segment in feature space is distributed inside a flat and slender hyper-rectangular, tone and rhyme in different pronunciation on two ends, the transient region in the middle:

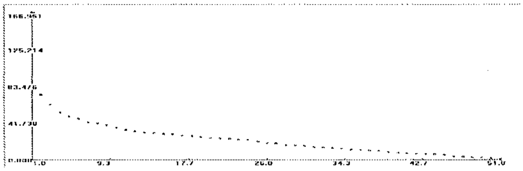

Fig. 1 PCA for a continuous digit “qi” in space, where horizontal axis is 51 mutually orthogonal axes from 52 points via PCA, vertical ordinate is axes length.

Distribution of the same speech segment from different speakers is much sparser in feature space, which might be covered by covering product of the surface jointed by multi-plane and hypershpere.

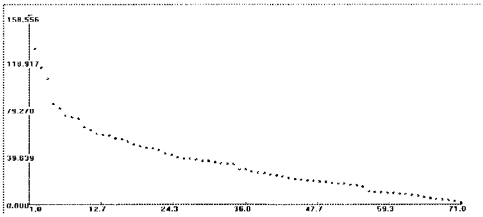

Fig. 2. PCA in feature space for the same digit speech “qi” from 24 speakers, each one choose 3 points totally 72 points. Horizontal axis is the position of 71 principal axes, vertical ordinate is the length of every principal axis

- ◆ What kind of cover should we use for speaker-independent speech recognition? The answer is to adopt the local covering principle, use the above-mentioned curve or surface to do the one by one cover for points in space, and then establish the covering area for one such speech by the cover product of surface and hypersphere. Thus the computation complicacy is reduced and the distribution of speech points might be depicted much realer. In the case of covering same amount of points completely, we apply a covering method of a smaller-volume manifold product (product of surface and hypersphere), so that probability of points in other classes which are dropped to this covering area would be reduced. The different continuous words are also continuously distributed in feature space. the graded procedure exists from one to another feature area, this sort of transformation is just the interface of the different speech, which is unobvious, just like the interface of continuous speech in time domain (the co-articulation effect of continuous speech), the dividing line but the pronounce area is not clear, we therefore easily find the speech covering area in the transformation domain, that is the foundation of the dynamic searching based on point covering.

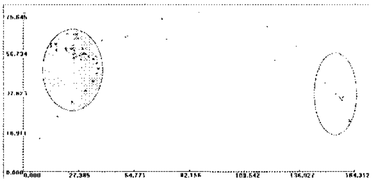

Fig 3. Distribution in the feature space of two continuous words “yi qi”. The former solid points are the speech points of “yi”, the latter hollow points are of “qi”, the horizontal axis is the link line of two points with the farthest

distance, vertical ordinate is the vertical distance from point to horizontal axis.

2.3 Covering method for different classes of speech

As the above statement, the covering method for speech in the different classes is summarized as follows:

construct the speech sample points in the covering area, select one representative point in continuous single digit speech (In Fig. 8b, one point in the transient area from tone to rhyme, choose one middle point with the clear pronunciation and strong periodicity in syllable without tone), this point is distinguishable in the time domain, i.e., its pronunciation can be recognized clearly by ears.

Apply the local covering principle, let covering basic element be the plane, the covering set is just a curve, hence the computation load is reduced and also the covering area is more approximate to the speech point distribution.

Assume dimensionality of the space in which speech points are via space mapping be N, use the covering product of surface and (N-2)-D hypersphere for the last covering area, the threshold of covering product is the radius of hypershpere.

Use neuron computer CASSANN-II to compute, the number of covering basic elements is co-determined by the sample number in the covering area within the built current network and the training algorithm of neural network.

Select the threshold of covering product between the average distance from points in one class to the center of its own covering area and the minimum distance from points in other classes to the central axis plane.

3. EXPERIMENTAL RESULT AND DISCUSSION

3.1 Statistical result and discussion in the experiment

The continuous speech sample used by built network is from 28 people. Select randomly 10 sample points in each digit class for each person, 280 sample points in each class, which builds up the digital neural network covering area in this class, experimental result is given in Tab. 1.

Table 1 Recognition result of randomly selected sample network

	Set A	Set B	Sum
Recognizable digit amount	225	32	256
digit amount of accuracy recognition	191	20	211
digit amount of misrecognition	18	8	26
Accuracy recognition rate	85.27%	62.5%	82.42%

The continuous speech sample used by built network is from 24 people, everyone picks up 10 sample points, but which is only by simple screening, these sample points by direction tracing to the source, the original speech segment is the digit read which can be recognized by ears. Applying 240 sample points to created network for the digit of each class, Tab.5-6 is the experimental result.

Tab.2 Recognition result of built sample network by screening

	Set A	Set B	Sum
Recognizable digit amount	200	64	264
Digit amount of accuracy recognition	177	52	229
Digit amount of misrecognition	14	6	20
Accuracy recognition rate	88.5%	81.25%	86.74%

3.2 Comparing result and discussion with HMM model

Under the case of fewer training samples, the speech recognition approach based on the HDS point covering theory is much better than HMM-based approach:

Firstly, in theory, the traditional HMM model is based on statistical model, whose accuracy completely depends on accuracy of probability statistics, and accuracy of probability statistics depends on massive data. Therefore, once the number of one type model is limited, algorithm accuracy based on statistical model is sure to be suspected. Method using neural network model building based on HDS point covering theory, by the efficient information melting ability of neural network and morphological analysis method and with a small amount of samples, can find the covering area in HDS for one type sample, thus the high recognition rate is obtained.

Secondly, during experiments, train the same training sample (according to the number of training sample in each class, there are five groups 48,96,144,192,240 to attend comparison) via HDS point covering and

HMM model method, then recognize the same recognized sample by these two methods respectively. Results are showed in table 3-9

Tab 3 Optimal state amount used by HMM model building and Gauss density function amount

Training sample amount	48	96	144	192	240
State amount	4	5	5	5	5
GDF	4	3	6	4	6

Tab.4 Comparison of recognition rate for speech of every class when 48 training samples in each class

digit	point covering method	HMM
ling	0.863931	0.7927
yi	0.923221	0.8408
er	0.852252	0.8054
san	0.978462	0.9154
si	0.969474	0.8895
wu	0.952712	0.9249
liu	0.849257	0.7495
qi	0.940202	0.8289
ba	0.936508	0.8825
jiu	0.917793	0.8423
yao	0.940828	0.7707

Tab. 5 Comparison of recognition rate for speech of every class when 96 training samples in each class

Digit	Point covering	HMM
ling	0.887689	0.8099
yi	0.973783	0.8783
er	0.861261	0.8288
san	0.989231	0.9369
si	0.957895	0.9453
wu	0.986092	0.9527
liu	0.921444	0.7771
qi	0.977921	0.8381
ba	0.968254	0.9079
jiu	0.967342	0.8795
yao	0.967456	0.9379

Tab.6. Comparison of recognition rate for speech of every class when 144 training samples in each class

Digit	Point covering	HMM
ling	0.946004	0.8963
yi	0.973783	0.9382
er	0.888288	0.8937
san	0.987692	0.9538
si	0.973684	0.9547
wu	0.968011	0.9555
liu	0.921444	0.8641
qi	0.966881	0.8684
ba	0.946032	0.927
jiu	0.977477	0.9324
yao	0.961538	0.9497

Tab.7. Comparison of recognition rate for speech of every class when 192 training samples in each class

Digit	Point covering	HMM
ling	0.952484	0.8531
yi	0.973783	0.9644
er	0.918919	0.8775
san	0.986154	0.9646
si	0.972632	0.9579
wu	0.987483	0.9805
liu	0.942675	0.8705
qi	0.973321	0.8657
ba	0.984127	0.946
jiu	0.974099	0.9279
yao	0.985207	0.9527

Tab. 8. Comparison of recognition rate for speech of every class when 240 training samples in each class

Digit	Point covering	HMM
ling	0.961123	0.9006
Yi	0.986891	0.9494
Er	0.933333	0.9045
san	0.989231	0.9738
Si	0.981053	0.9695

Wu	0.973574	0.9777
Liu	0.974522	0.896
Qi	0.965961	0.8896
Ba	0.993651	0.946
Jiu	0.974099	0.9313
yao	0.985207	0.9482

Tab.9. Comparison of overall recognition rate under the variant training sample amount in each class

Training sample amount in each class	Point covering	HMM
48	0.927203	0.8441
96	0.955939	0.8859
144	0.959907	0.9221
192	0.970033	0.9245
240	0.974001	0.9357

4. CONCLUSION

Research for HDS geometric theory and its application in the continuous digit speech recognition, making an investigation on a new sort of approach, experiments show it is promising for solving stability of Speaker-independent continuous speech recognition system. In those cases, e.g., microphone close to speakers and under circumstance with parts of road background noise etc., the word recognition rate still reach 86.74% (which is close to the result based on HMM method), however, for the speech recognition from untrained speakers, its recognition rate is still 81.25 % (impossible for HMM method to have better stability like this). The method present here is no need to go through endpoints detection and segmentation, the recognized speech can be directly projected to the built covering area, and by a dynamic searching algorithm along time axis to find recognition results. This algorithm is simple and feasible, with fast computation speed, and can solve the contradiction of the processed signals in HDS having the same dimensionality necessarily and signals of same class having the irregular length on the time, i.e., solve difficulty about the continuous speech signal having the different length under the signal sequence at variant speech speed. This method differs from the conventional pattern matching and dynamic time warping, and the optimal route searching, so that it is a new research direction based on HDS covering theory.

REFERENCES

1. Wang ShouJue, Bionic(Topological) Pattern Recognition——A New Model of Pattern Recognition Theory and Its Applications, ACTA ELECTRONICA SINICA, 2002, 30(10): 1417-1420
2. Wang ShouJue, Wang Bianan, Analysis and Theory of High-Dimension Space Geometry for Artificial Neural Networks, ACTA ELECTRONICA SINICA, 2002,30(1): 1-4
3. Wang ShouJue, A new development on ANN in China – Biomimetic pattern recognition and multiweight vector neurons, LECTURE NOTES IN ARTIFICIAL INTELLIGENCE 2003,2639: 35-43
4. Wang Shoujue, Xu Jian, Wang Xianbao, Qin Hong, Multi-camera Human-face Personal Identification System Based on the Bionic Pattern Recognition, ACTA ELECTRONICA SINICA,2003,31(1): 1-3
5. Cao WM, Hao F, Wang SJ, The application of DBF neural networks for object recognition, INFORMATION SCIENCES, MAR 22 2004, 160 (1-4): 153-160
6. Shi Jingpu, Chen Ji, Chen Xiaodong, Chen Chuan, Wang Shoujue, The Speaker Verification System Based on Neurocomputer and Its Application, ACTA ELECTRONICA SINICA, 1999,27(10):27-29
7. Wang Shoujue, Li Zhaozhou, Wang Bonan, Deng Haojiang, Feed-forward Neural Network Modeling for Noise Rejection, JOURNAL OF CIRCUITS AND SYSTEMS, 2000, 5(04):21-26
8. Deng Haojiang, Wang Shoujue, Xing Cangju, Li Qian, Research of Text-Independent Speaker Recognition Using Clustering Statistic, JOURNAL OF CIRCUITS AND SYSTEMS, 2001,6(03):77-8
9. Xing Cangju, Qu Yanfeng, Wang Shoujue, Face Detection on Gray-Scale Static Image with Complex Background, JOURNAL OF COMPUTER-AIDED DESIGN & COMOPUTER GRAPHICS,2002, 14(05):401-403
10. Qu Yangfeng, Li Weijun, Xu Jian, Wang Shoujue, Fast Multi-Pose Face Dection in A Complex Background, OURNAL OF COMPUTER-AIDED DESIGN & COMOPUTER GRAPHICS, 2004,16(01):45-50
11. Deng Haojiang, Wang Shoujue, Du Limin, Text-Independent Speaker Verification on Using Priority Ordered Radial Basis Function Networks, JOURNAL OF ELECTRONICS AND INFORMATION TECHNOLOGY, 2003, 25(09):1153-1159
12. Wenming Cao Feng hao and Shoujue Wang, An adaptive controller for a class of nonlinear system using direction basis function. ACTA Journal Of Electronics 2002, 23: 43-48.
13. Xu, Jian, Li, Weijun, Qu, Yanfeng, Qin, Hong, Wang, Shoujue, Architecture Research and Hardware Implementation on Simplified Neural Computing System for Face Identification, Proceedings of the International Joint Conference on Neural Networks, 2003, 2:948-952