The Research of Speaker-Independent Continuous Mandarin Digits Speech-Recognition Based on the Dynamic Search Method of High-Dimension Space Vertex Cover

Wenming Cao [1] ,Xiaoxia Pan [1] , Shoujue Wang [2]

[1]Institute of Intelligent Information System, Information College, Zhejiang University of Technology, Hangzhou 310032, China

csann@zjut.edu.cn

[2]Institute of Semiconductors, Chinese Academy of Science, Beijing 100083, China

Abstract.

In this paper, we present a novel algorithm of speaker-independent continuous Mandarin digits speech-recognition, which is based on the dynamic searching method of high-dimension space vertex cover. It doesn't need endpoint detecting and segmenting. We construct a coverage area for every class of digits firstly, and then we put every numeric string into these coverage-areas, and the numeric string is recognized directly by the dynamic search method. Finally, there are 32 people in experiment, 16 female and 16 male, and 256 digits all together. All these digits are not learned. The correct recognition result is 218, and error recognition result is 26. Correct recognition rate is 85%

1 Introduction

With the development of the technology of speaker-independent small vocabulary isolating speech recognition, its application has been prospected in many fields. But the applied continuous speech recognition system is still rare. Therefore, the research of speech recognition was mainly concentrated on improving the performance of speaker-independent large vocabulary continuous speech recognition since 1990s.

The traditional algorithm of the speaker-independent continuous speech-recognition is mainly depended on the endpoint detection. First, the speech was segmented into small units, syllables or phonemes, and then these units were recognized by the method of Hidden Markov Models (HMMs)[2] or Dynamic Time Warping (DTW). The advantages of these methods are that they can convert the continuous speech into isolated speech and reduce the difficulties of the recognition process. But their method is fragile in the noise, limited in their ability to handle pronunciation variation, and costly for large vocabulary spontaneous speech transcription. That is to say, their ability to represent dynamic behavior is limited [10].

In order to resolve these problems of continuous speech recognition, in this paper, we present a novel algorithm which is based on the Theory of High-Dimension Space Vertex Cover[5],[8]. We classify the corpora firstly and then analyze the point property of the feature space by this means. Because of CASSAN-II neural computer have been widely applied the Theory of High-Dimension Space Geometry to Artificial Neural Networks [3],[7],[6],[4], and soon it succeeds in many fields, such as pattern recognition[9], face recognition[12] and adaptive controller[1],Finally we implement it on the CASSAN-II neural computer[11] invented by Wang Shoujue.

2 The characteristic of continuous speech

(1) Co-articulation exists in the speech units (e.g. phoneme, syllable).

(2) Because there are not clear endpoints between two units, the phone can not be segmented completely or separated from each other.

(3) Both the rate-of-speech[5] and the duration-of-speech are variable, so it is difficult to find an uniform module;

Because of the three characteristics mentioned above, the endpoint detecting and the speech segmentation become difficult. In addition, the environment noise makes the acoustic model establishment even more difficult. Therefore, the decision tree and many phonemic models were used in the traditional Hidden Markov Models method, such as Word-Dependent (WD), Left-Context-Dependent (LCD), Right-Context-Dependent (RCD), context-Dependent (CD, also called TRIPHONE model, triple phone), and so on. This makes the models which appear seldom can't be trained sufficiently, so the recognition rate is lower than those which appeared more.

(4) There are some characters of the nature speech, which is more optional and has a few stochastic phenomena, such as hesitation, pause, filled pauses etc.

3 Feature extraction of the continuous digits speech recognition and the constructing of High-Dimension Space coverage areas

3.1 The collection and establishment of the speech corpora

There are two speech patterns. One is spontaneous speech that is the utterance we speak in our daily life. It is unbending. At least, it has no special prepare in terms of speech pattern. It is always slack, and goes with random events (filled pauses etc.). The other is the reading speech. It is always changeless. Its speech pattern and speech context should be prepared beforehand and accord with the grammar as well.

The continuous speech we adopted in this paper is between the reading and the spontaneous speech. The context of our corpora is the phone numbers. The read pattern is similar to the spontaneous speech. According to the experiment requires, we remove some complex noise but some background noise (e.g. stir of cars in road) is left.

There are two corpora. As for the first corpora, we segment the continuous speech into syllables by hand and then select the better result as "the learning corpora". We must point out that these syllable samples are different from the isolated samples. They have many characteristics of continuous speech. As for the second corpora, we regard all these articulate continuous phone number as "the recognition corpora". Both these corpora are collected in 8000Hz (the Sample Frequency) and 16bits (the Bit Depth).

3.2 Feature extraction method of the learning corpora

There are three steps in this process.

Step 1. Change the wave samples into Mel Frequency Cepstrum Coefficient (MFCC).

First, pre-emphasis processing:

$$x'(n)=x(n)-0.9375x(n-1).$$

Second, hamming windows:

$$x'(n)=[0.54-0.46\cos(2\pi n/255)]x(n)$$

the width of this window is 256, and its offset is 64.

Third, Mel Frequency Cepstrum conversion: the number of Mel filters is 24. We remove the first and the last 7 values, and remain 16 values as feature parameters.

Step 2. Remove the redundant data.

First, suppose that every 16 feature parameters consist of one 16 dimension-vector Ci, i=1,2,3...n, shown as fig.1.

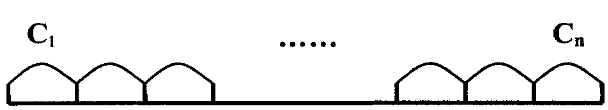

Fig.1. As for a digit sample, here are n 16-dimension-vectors. Every vector has 16 dimensions which are 16 MFCC after step 1.

Second, compute the angle of the two adjacent vectors:

$$\theta_j = ar\cos(\frac{C_j \bullet C_{j+1}}{|C_j| \cdot |C_{j+1}|})$$

When the angle θ_j is below the experiment statistic number 0.13rad, remove Cj or Cj+1 until all $\theta_j \geq 0.13rad$. In this way, we can not only remain more than 8 vectors but also compress data in a certain extent.

Step 3. Normalize a digit sample into a certain length.

First, select one shortest sample from each class of MFCC vectors. These vectors have already been compressed. Choose the optimal 8 vectors in hearing by hand. And regard these 128 values, a new feature vector of High-Dimension Space, as a standard of this class of MFCC vectors.

Second, compare all MFCC vectors of each class with the standard, and choose sequential 8 16-dimension-vectors whose angle with the standard is smallest. And regard this 128-dimension-vector as the feature of this sample. The coverage areas of feature space are constructed by these 128-dimension feature vectors. This process is shown as figure 2.

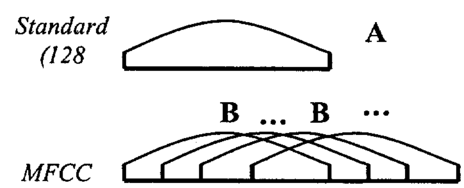

Fig. 2. Choose the most similar 8 vectors (consist of a vector of 128-dimension) from the MFCC vectors.

Suppose,

$$\theta_k = ar\cos(\frac{A \bullet B_k}{|A| \cdot |B_k|})$$

If,

$$\theta_{\min} = \theta_p = \min\{\theta_k, k = 1,2,\cdots n\}$$

Then regard this 128-dimension vector B_p as the most similar one in these vectors $B_k \, (k = 1,2,\cdots n)$.

Therefore, select B_p as this sample's feature vector (128-dimension). This feature vector is one of the samples, which construct the coverage areas of this class of feature space.

3.3 Put the feature vectors into neural networks and construct the coverage areas of every syllable (we see every syllable as one class) in the high-dimension feature space

From 0 to 9 except 1, there are 9 classes of samples. The digit "1" is divided into two classes: "yi" and "yao". Then there are 11 classes of samples totally. Regard each sample class as one set Si（i=1,2,...11）,

$$S_i=\{x_{ij} \mid x_{ij}\in S_i,\ j\in N\}$$

Si is from the learning corpora. And select n samples (every sample xij is one of the point of High Dimension Space) from every sample class as a new sample set S'i. Put the new sample set S'i into neural networks, and adopt the Theory of High-Dimension Space Vertex Cover[4],[8],[11],[12] to construct the coverage areas of the ith class in feature space.

4 The novel algorithm and its implementation of Speaker-Independent Continuous Digit Speech Recognition Based on the Dynamic Search Theory of High-Dimension Space Vertex Cover

4.1 Feature extraction method of recognition corpora

According to the same procedure of the step 1 and step 2 in section 3.2, change wave samples into 16 dimension MFCC vectors and remove the redundant data. And then extract 8 16-dimension vectors as a 128-dimension vector from the beginning of the speech. Extract next 8 16-dimension as the next 128-dimension from the second 16-dimension vector. The rest may be deduced by analogy. Form a series of 128-dimension vectors whose length (the number of 128-dimension vectors) varies with the length of the speech. In this experiment, it varies from 207 to 465.In this procedure, the frame width of Hamming window is 256 sample points, the offset length of every frame is 64 sample points, the number of Mel filter is 24, and the threshold θ=0.13 radian.

4.2 Dynamic search and recognition method of High-Dimension Space Vertex Cover

Regard n 128-dimension feature vectors of continuous speech to be recognized as n points of high dimension space. Here, take the sentence "san wu ba er qi ling ling san" as an example. The length of this sentence is 433, that is n=433. Compute the distance between these n points and every class of the coverage area. The result curve is shown as 11 Figures, we select 2 figures fig.3a-3c. (The x-coordinate is the points' serial number and it is a time-axis. And y-coordinate is the distance between the present point and the present coverage area. The distance will be changed by time). In fig.3a, two distinct minimum points are shown in 79th and 357th points respectively. Since it is in the coverage area of "san", this result indicates that the two positions in time-axis are the pronunciation of "san". Fig.3b indicates the distance between this sentence and the coverage area of "wu". There is only one minimum point, the 109th point. This point stands for "wu" in this sentence. The curves in fig. 3c, 3d and 3e indicate the distance between this sentence and the coverage areas of " ba", "er" and "qi" respectively. The minimum points correspond to the pronunciation of "ba", "er" and "qi" of this sentence respectively.

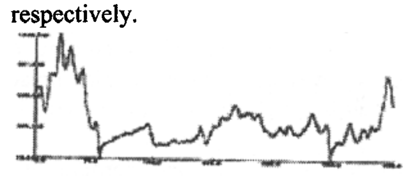

Fig.3a. the distance from "san wu ba er qi ling ling san" to the coverage area of "san"

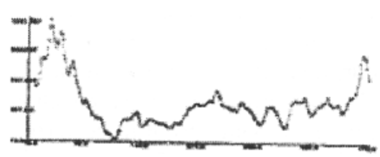

Fig.3b. the distance from "san wu ba er qi ling ling san" to the coverage area of "wu"

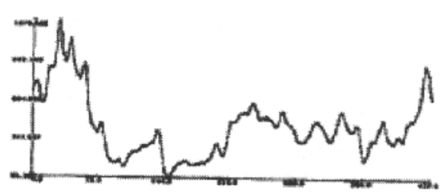

Fig.3c. the distance from "san wu ba er qi ling ling san" to the coverage area of "ba"

Fig.3d. the distance from "san wu ba er qi ling ling san" to the coverage area of "er"

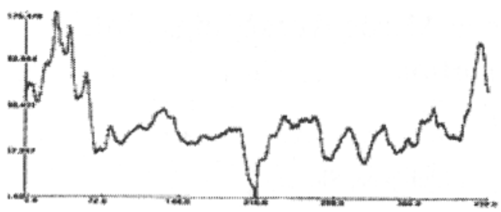

Fig.3e. the distance from "san wu ba er qi ling ling san" to the coverage area of "qi"

All these 11 curves form a mesh surface shown as fig.6.

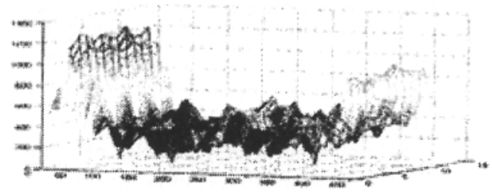

Fig.4. the 3D display of the recognition result

5 Result and discussion

5.1 Statistic result:

There are 32 people in this experiment, 16 female and 16 male, and 256 digits all together. All these digits are not learned. The correct recognition result is 218. and Error result is 26.

5.2 Discussion

We discussed a new method of the Speaker-Independent Continuous Speech-Recognition in this paper. As it dose not need to detect the endpoint and segment speech, it can be used in the recognition of the continuous nature speech. Although the result is not perfect, but we can see the performance of robustness is good. It will be a promising new research direction in the continuous nature speech recognition.

References

[1] Cao, Wenming , Feng, Hao; Zhang, Dongmei; Wang, Shoujue (2002) : An adaptive controller for a class of nonlinear system using direction basis function. Chinese Journal of Electronics, v 11, n 3, July, p 303-306

[2]L. R. Rabiner (1989): A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, vol. 77, 257-286.

[3] Shoujue Wang et al (1998).: The Sequential Learning Ahead Masking (SLAM) model of neural networks for pattern classification, Proceedings of JCIS98, vol. IV, October 1998, RTP, NC, USA, 199–202

[4] Shoujue Wang (2003): A new development on ANN in China - Biomimetic pattern recognition and multi weight vector neurons. Rough sets, fuzzy sets, data mining, and granular computing lecture notes in artificial intelligence, LECTURE NOTES IN ARTIFICIAL INTELLIGENCE, vol. 2639, 35-43

[5]Jing Zheng , Horacio Franco, Andreas Stolcke (2003): Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition. Speech Communication, Vol.41, 273-285

[6] Shoujue Wang(1999): Direction-Basis-Function neural networks, Proceedings of IJCNN'99, 10–16 July, Washington, DC, USA, 1251–2171

[7] Shoujue Wang, Priority ordered neural networks with better similarity to human knowledge representation. Chinese Journal Of Electronics,Vol.8, (1999)1–4.

[8] Shoujue Wang, Wang Bainan(2002): Analysis and Theory of High-Dimension Space Geometry for Artificial Neural Networks, Chinese Journal of Electronics, Vol.30 1-4.

[9] Wenming Cao, Feng Hao and Shoujue Wang (2004): The application of DBF neural networks for object recognition, Information Sciences Vol.160, Issues 1-4 , 22 March (2004)153-160

[10]Martin J. Russell Jeff A. Bilmes (2003): Introduction to the special issue on new computational paradigms for acoustic modeling in speech recognition, Computer Speech and Language, Vol.17, 107 - 112

[11] Wang Shoujue, LiZhao zhou, Chen Xiangdong, WangBainan (2001): Discussion on the Basic Mathematical Models of Neurons in General Purpose Neuro Computer, Acta Electronica Sinica, Vol.29, 577-580

[12] Wang Shoujue Xu Jian Wang Xianbao Qin Hong (2003): Multi-Camera Human-Face Personal Identification System Based on the biomimetic pattern recognition, Acta Electronica Sinica, Vol.31, 1-3