TELKOMNIKA, Vol.10, No.7, November 2012, pp. 1604~1609
e-ISSN: 2087-278X
accredited by DGHE (DIKTI), Decree No: 51/Dikti/Kep/2010

Non-Specific Person Continuous Speech Identification in Second Language using BPR

Guangming Xian*[1], Biqing Zeng[1], Qiaoyun Yun[1], Xiongwen Pang[2]
[1]Information Engineering and Technology Department, South China Normal University, Foshan Guangdong 528225, China
[1]School of Computer Science, South China Normal University, Guangzhou 510631, Guangdong, China
*corresponding author, e-mail: xgm20011@126.com

Abstract

Second language speech recognition is an important technical means of man-machine communication system. In this paper, we propose a biomimetic pattern recognition (BPR) algorithm for non specific person continuous speech identification in second language. Feature parameters are extracted directly from single number sample being segmented according to Mel cepstral way. BPR-based connected number recognition experiments show that the proposed method with better identification performance in second language than NN neural network and SVM. From non specific person continuous speech recognition experiments, we show that the BPR algorithm greatly improves word recognition accuracy and has good recognition ability without causing poorer performance in second language.

Keywords*: continuous speech identification; second language; biomimetic pattern recognition; Mel-frequency cepstral coefficient*

1. Introduction

In recent years, voice recognition [1-6] especially continuous speech recognition research has get gratifying progress, positive and practical system development. Researchers are committed to the development of practical system.

The difficulties of continuous speech recognition can be summarized as follow:

(1) Voice unit coarticulation;
(2) As a result of the interlaced fusion of the endpoint, therefore it is difficult to separate;
(3) Because of uncertain speed and uncertain length of word, it is hard to find a unified template to match;
(4) Having similar natural speech characteristics, speaking at random, with a small amount of hesitation, pause and filling the voice phenomenon. The traditional classification method cannot remove these redundant phoneme or syllable which is not in a predetermined state.

In our experiments, the continuous speech is between spontaneous speech and read speech. Its content includes all the linking number. They are read with natural spoken tone and speed. Content are recorded in the relatively quiet laboratory background. The sample frequency is 8000Hz and the bit depth is 16 bit.

According to these continuous speeches with coarticulation which is difficult to separate, we try to find the best continuous speech segmentation by artificial segmentation audition approach and build a continuous speech monosyllable sample library. It must be stressed that single syllable sample database is different from general isolated speech sample library.

A non specific person continuous speech recognition system based on biomimetic pattern recognition (BPR) theory is developed using Mel cepstral way feature extract approach in this paper.

The result of experiments shows that the model applied to continuous speech recognition system performed very well, could successfully recognize small amount of words of continuous speech in second language [7-9] with high recognition rate effectively.

In this thesis, firstly, a brief introduction of the continuous speech recognition [10-13] is given. Secondly, feature extraction method for non specific person continuous speech identification in second language is introduced. Thirdly, the difference between biomimetic pattern recognition [14-15] and traditional pattern recognition are compared. Fourthly,

Received July 21, 2012; Revised September 29, 2012; Accepted October 10, 2012

experimental results can be obtained from speech recognition experiments. Finally, we draw the conclusion of this article.

2 . Feature extraction method for non specific person continuous speech identification in second language

Construction of neural network with the sample feature extraction is done in 3 steps.

Firstly, feature vector can be extracted from single number sample being segmented according to Mel cepstral way. The process is simplified as follows.

(1) The original voice set of study sample is $\{s_i \mid s_i \in s\}$. s_i is the ith class sample collection and x_n is the nth sample point. x_n is preprocessed as follow.

$$x'(n) = x(n) - 0.9375x(n-1) \quad (1)$$

(2) After being processed by Hamming window which its window width is 256 and frame shift is 64, the single number voice frame can be processed as follow.

$$x''(n) = \left[0.54 - 0.46\cos\frac{2\pi n}{255}\right] x'(n) \quad (2)$$

(3) After each frame data being transformed by Mel cepstral way through 24 filter group, 24 Mel-frequency cepstral coefficient (MFCC) can be obtained. Removing first coefficients with obvious energy characteristic and the last 7 coefficients approaching to zero, we can get 16 coefficients left as the feature parameters.

Second, the redundant data is eliminated.

(4) The 16 characteristic parameters construct a vector $C_i(i = 1,2,\cdots,n)$.

(5) Calculate two adjacent angle of16dimensional vectors.

$$\theta_j = \arccos\frac{(C_j, C_{j+1})}{|C_j|\cdot|C_{j+1}|}, j = 1,2,\cdots,n-1 \quad (3)$$

When the angle is less than experimental statistic data 0.13rad, delete vector C_{j+1} or C_j until the adjacent vector angle is greater than or equal to 0.13rad.

Third, data is compressed to a certain length.

(1) We select one of the shortest single digital syllables which are compressed by Mel-frequency cepstral coefficient(MFCC) . The single syllable is intercepted as follow. The best 4 audition consecutive vector (a total of 16× 4 values) are selected using artificial audition method. The high dimension feature vector constructed by 64 numerical values are considered as the reference standard for number MFCC single syllable

(2) comparison between the MFCC single digital syllables in every class all and standard of the class, 64 dimensional vectors are selected from the minimal angle of 4 consecutive 16 dimensional vectors as the MFCC single digital syllables feature vector. The feature space of the covered identification area is constructed.

Let's set

$$\theta_k = \arccos\frac{(A, B_k)}{|A|\cdot|B_k|} \quad (4)$$

if

$$\theta_{min} = \theta_p = \min\{\theta_k, k = 1,2,\cdots,n\} \quad (5)$$

It means that the angle between 64 dimensional vectors B_p and the standard A is minimal.

Therefore, B_p is selected as a feature vector of MFCC single digital syllables to construct a sample in the feature space coverage area. As mentioned above, the process of feature abstraction is descript as shown as fig.1 .

According to the characteristics of the experimental object, we have 11 class samples. We suppose that each set formed by class sample is expressed as $s_i(i = 0,1,\cdots,10)$. As table 1 shown, ten 64 dimensional vectors (a total of 360 sample points) are selected to construct a new network configuration set.

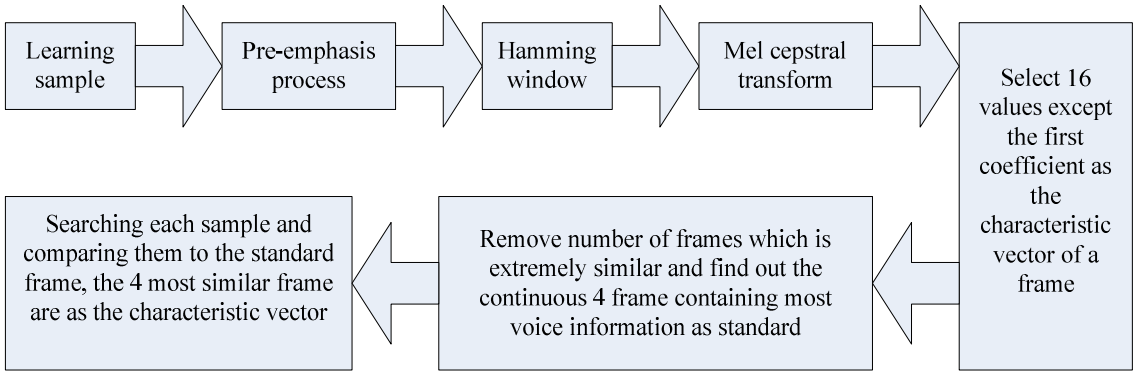

Fig.1 The framework of learning sample characteristics extraction

Table 1. The network structure of total sample library

Class	0	1	2	3	4	5	6	7	8	9
Class No	0	1	2	10	3	4	5	6	8	9
Each person words	10	10	10	10	10	10	10	10	10	10
A total of 36 people	360	360	360	360	360	360	360	360	360	360

(Unit: word)

$$s_i^{'} = \{x_{ij} | X_{ij} \in S_j, j = 1,2,\cdots,360\} \quad (6)$$

For each class network configuration set,

$$s_i^{'}(i = 0,1,\cdots,10) \quad (7)$$

we train the samples and construct a neural network using high dimensional space covering method.

After feature extraction of continuous speech samples to be recognized, 16 dimensional vector strings which have different length can be gotten. We select 4 consecutive 16 dimensional vectors as a window to constitute one of the 128 dimensional feature vectors as a recognition point. We select a 16 dimensional vector as the frame shift, moving their window to form $64\times n$ dimensional feature vector strings which are differ in the length (n is different with the length of the identified speech) .

The feature vectors ($64 \times n$ dimension) extracted from 4 number continuous speeches which are regard as n voice point at a high dimensional space. We compute the minimum coverage distance of each kind from these n voice points in turn.

3. The Difference Between Biomimetic Pattern Recognition and Traditional Pattern Recognition

Traditional pattern recognition is based on the basic mathematical models that all the available information is contained in the training set $(X_1, Y_1), (X_2, Y_2), \cdots (X_l, Y_l)$. That is to say, congener sample points do not have priori knowledge with relation. However, in the actual law of nature is not so. The foundation and the key of biomimetic pattern recognition lies in the introduction of some certain universal laws existing in congener samples and set up recognition principle of non-hypersphere complex geometry covering in a multidimensional space.

In the identification process, two different classification processes exist: one is homologous and another is the logic of knowledge classification process which is not homologous. In this paper, the congener samples are a kind of homologous samples. The following discussion is about continuity rule in homologous congener samples.

For any things in nature to be know (including image, sound, language and status), if there are two homologous congener things which is not exactly the same and the difference of these two things can be gradual or non quantized, the two kind congener things will have at

least a gradual process. And each kind of things in the gradual process belongs to the same class. The continuity rule in the homology sample is called the principle of homologous continuity (PHC).

The mathematical description of PHC are as follows: We suppose that all sample points of class A form a set A. Taking two samples $X, Y \epsilon A, \text{and } X \neq Y \text{ at random}$, for any given $\varepsilon > 0$, there must be a set to meet

$$B = \left\{ X_{1,} X_{2,} X_{3,} \cdots, X_{l,} \middle| \begin{matrix} X_l = X, X_l = Y \\ l \epsilon N \\ \rho(X_m, X_{m+1}) \epsilon \varepsilon \\ 1 \leq m \leq l+1 \end{matrix} \right\} \subset A \quad (8)$$

where $\rho(X_m, X_{m+1})$ is the distance of sample $X_m \text{ and } X_{m+1}$.

In the feature space R^n, the continuity law in congener sample point is beyond the basic assumption between the traditional pattern recognition and learning theory.

The hypothesis is that the available information is contained in the training sample set.

But the continuity law is the law existing in human intuitive understanding scope of the object world. Therefore, PHC is prior knowledge of biomimetic pattern recognition for distributing the sample points. And the perception ability can be improved by PHC. Traditional pattern recognition use the best classification of different samples in the feature space as a target, while bionic pattern recognition use optimal coverage of a kind of samples in feature space distribution as a target.

After the continuity law of congener samples in the feature space being introduced in biomimetic pattern recognition, recognizing a class objects is essentially about analysis and understanding the shape of infinite point set formed by all of this kind of thing in the feature space. The theoretical analysis mathematical tool of pattern recognition is the issue of high dimensional manifolds in point set topology. In order to distinguish with traditional statistical pattern recognition, biomimetic pattern recognition is also known as topological pattern recognition.

In biomimetic pattern recognition, the point set being mapped from any kind of thing in the feature space R^n is considered as a closed set.

In practical engineering application, regardless of what kind of pattern recognition being to solve, sample collection and object identification must cause the random noise.

In the utilization of biomimetic pattern recognition, discrimination covers collection for recognizing class A should use set P_a instead of set A.

$$P_a = \{X | \rho(X, Y) \leq k, Y \in A, X \in R_n\} \quad (9)$$

where k is the selected distance constant. Because set P is n-dimensional, the task of biomimetic recognition is to discriminate whether "image" in the feature space R^n mapping from identified things belongs to set P_a or not.

4. Experimental results

We select 36 people who speak 4 different number continuous speech string (a total of 144 words) in second language to construct training set. 36 people who speak left 6 different number continuous speech string (a total of 216 words) in second language to construct training set A. Another 7 people who speak 10 different number continuous speech string (a total of 70 words) in second language as testing set B.

The performances of 3 kinds of different identification algorithms: Nearest neighbor method, SVM with different kernel function (RBF kernel and lineal kernel) and biomimetic pattern recognition are compared in our study.

In order to measure the identification ability of different algorithms in the same condition, we change the adjustable parameters of different algorithm. Ensuring the correct recognition rate at the almost same level condition, we analysis and compare the error identification rate and false accept rate.

Nearest neighbor method can adjust the parameter of classification threshold. Adjustable parameter of SVM is the penalty factor.

The computational complexity of NN is the number of comparison kernel. The computational complexity of SVM is the number of support vectors. The computational complexity of BPR is the neuron number of φ function.

From the compare results shown in table 2, it is demonstrated that the performance of biomimetic pattern recognition method for non specific person continuous speech identification in second language is better than NN neural network and SVM obviously. BPR has good recognition ability in case of not reducing correct recognition rate.

Table 2. Comparison of different identification methods

Identification algorithm	Computational complexity (SVM or neuron number)	Testing sample A (216 words)			Testing sample B (70 words)	
		Correct Recognition rate (%)	Misclassification rate (%)	False rejection rate (%)	False acceptance rate (%)	Correct acceptance rate (%)
NN	157	174 (80. 6%)	15 (6.9%)	27 (12.5%)	16 (22.9%)	54 (77.1%)
SVM I (linear kernel)	682	176 (81.5%)	4 (1.9)	36 (16.7)	5 (7.1%)	65 (92.9%)
SVM II (RBF kernel)	936	179 (82.9%)	6 (2.7%)	31 (14.4%)	4 (5.7%)	66 (94.3%)
BPR	42	181 (83.8%)	1 (0.5%)	34 (15.7%)	2 (2.9%)	68 (97.1%)

5. Conclusion

Identification approach based on biomimetic pattern recognition has been proposed as solution to non specific person continuous speech recognition technologies in second language. This paper explores an approach for continuous speech recognition which combines the advantages of MFCC and BPR. In this scheme, the recognition is performed from the feature vectors extracted from number continuous speeches which are regard as n voice point at a high dimensional space. In general, our approach outperforms the conventional procedure of NN neural network and SVM obviously in second language. Furthermore, the recognition ability of BPR is high in case of not reducing correct recognition rate.

Acknowledgements

The authors acknowledge the support of the South China Normal University and South China University of Technology. The work was support by the project of research of support vector machine in classification and regression, under project number Guangdong financial education (2008) 342. The work was support by Guangdong province natural science fund (project No. 8151063101000040 and 9451063101002213). The work was also support by humanities and social sciences youth fund project in ministry of education (project No. 10YJC870044).

References

[1]. Hong Kook Kim, Richard V. Cox, and Richard C. Rose. Performance improvement of a bitstream-based bront-end for wireless speech recognition in adverse environments. IEEE transactions on speech and audio processing. 2002, 10(8): 592-604.

[2]. Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gad. significance of the modified group delay feature in speech recognition. IEEE transactions on audio speech and language processing. 2007, 15(1):190-202.

[3]. Wooil Kim, and John H. L. Hansen. Time–frequency correlation-based missing-featurereconstruction for robust speech recognition in band-restricted conditions. IEEE transactions audio, speech and language processing. 2009,17(7):1292-1304.

[4]. Satoshi Nakamura. statistical multimodal integration for audio–visual speech processing. IEEE transactions on neural network. 2002, 13(4):854-866.

[5]. Yu Shao, and Chip-Hong Chang. Bayesian separation with sparsity promotion in perceptual wavelet domain for speech enhancement and hybrid speech recognition. IEEE transaction on systems, man, and cybernetics—part A: system and humans. 2011, 41(2):284-293.
[6]. Claude C. Chibelushi, Farzin Deravi, and John S. D. Mason. A review of speech-based bimodal recognition. IEEE transactions on multimedia. 2002, 4(1):23-37.
[7]. Georgia Andreou· Ioannis Galantomos. Conceptual competence as a component of second language fluency. J Psycholinguist Res . 2009, 38:587–591.
[8]. Paul Miller , Ora Peleg. Doomed to Read in a Second Language: Implications
[9]. for Learning. J Psycholinguist Res . 2010, 39:51–65.
[10]. R. Elliott ,1609 J. R. W. Glauert ,1609 J. R. Kennaway, I. Marshall and1609 E. Safar. Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. Univ Access Inf Soc . 2008, 6:375–391.
[11]. Frank Eisner and James M.Mcqueen. The specificity of perceptual learning in speech processing. Perception & Psychophysics. 2005, 67 (2), 224-238.
[12]. Zica Valsan, Inge gavat and Bogdan sabac. Statistical and hybrid methods for speech recognition in Romanian. International journal for speech of technology. 2002,5, 259–268.
[13]. Dia AbuZeina,Wasfi Al-Khatib, Moustafa Elshafei and Husni Al-Muhtaseb. Cross-word Arabic pronunciation variation modeling for speech Recognition. Int J Speech Technol 2011, 14:227–236.
[14]. Muharram Mansoorizadeh, and Nasrollah Moghaddam Charkari. Multimodal information fusion application to human emotion recognition from face and speech. Multimed Tools Appl 2010, 49:277–297.
[15]. Sarwosri, Herumurti D, Sulistyowati I. The Efficient Classification in Multi Relation Database Using Crossmine. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 2008, 6 (1): 7-14
[16]. Abdul Fadlil. An Automatic Identification System of Human Skin Irritation. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2010, 8(3):255-264.