Isolated Word Recognition using Modified Dynamic Averaging Method

변형된 Dynamic Averaging 방법을 이용한 단독어인식

Eui-Bung Jeoung,* Young-Hyuk Ko,** Jong-Arc Lee*

정 의 붕,* 고 영 혁,** 이 종 악*

ABSTRACT

This paper is a study on isolated word recognition by independent speaker, we propose DTW speech recognition system by modified dynamic linear averaging method as reference pattern.

57 city names are selected as recognition vocabulary and 12th LPC cepstrum coefficients are used as the feature parameter.

In this paper, besides recognition experiment using modified dynamic linear averaging method as reference pattern, we perform recognition experiments using causal method, dynamic averaging method, linear averaging method and clustering method with the same data in the same conditions for comparison with it.

Through the experiment result, it is proved that recognition rate by DTW using modified dynamic averaging method is the best as 97.6 percent.

요 약

본 논문은 단일화자에 대한 단독어 음성 인식에 대한 연구이다. 우리는 표준패턴으로서 변형된 dynamic linear averaging 방법을 이용한 DTW 음성 인식 시스템을 제안한다. 57개의 도시명이 인식 대상 어휘로 선정되었고 12차 LPC cepstrum 계수는 특징 파라미터로 사용되었다. 이 논문은 표준패턴으로 변형된 dynamic linear averanging 방법을 이용하여 인식 실험을 하는 것 이외에도, 같은 데이터 같은 조건에서 causal 방법과 dynamic averaging방법, linear averaging방법, clustering방법을 이용하여 실험하였다. 실험결과로 변형된 dynamic linear averaging 방법을 이용한 DTW 음성인식이 97.6%로 가장 좋은 인식률을 보였다.

I. Introduction

Speech recognition is performed by analyzing speech waveform, extracting phonetic features and

*Dept. of Electronics Engineering, Kon Kuk University
**Dept. of Electronics Eng., Dong Shin University

translating themin to linguistic symbols.

In this paper, we propose new method for making a templates of DTW recognition system. DTW is tempalte matching method which stores acoustic features as templates and compares with test pattern and reference pattern. The reference pattern means reference data which is compared with test data in speech recognition system. [1]

The method to make a reference pattern from tokens is causal method, clustering method, linear averaging method, dynamic averaging method, and modified dynamic averaging method. In these methods, recognition rate by DTW using modified dynamic averaging method is the best as 97.6 percent.[4]

It is reasonable for small or medium scale of vocabulary in the speech recognition system to use recognition units as word unit than subword units like phoneme or syllable. Therefore, this paper carry out the recognition of isolated word of word unit from 57 city names.

For ending point detection, ZCR and Energy parameters is used and 12-order LPC cepstrum coefficients was used for feature vectors.

Ⅱ. Reference pattern generation

1. Causal method[2]

Causal method is a reference pattern of word from several tokens settled beginning and ending point, that means taking an optional token and express 12th LPC cepstrum coefficients.(It is called the reference word template with it)

It has merit that an implementation is easy and also it takes short time but on the other hand it has fault that the reference pattern has noise during pronouncing. Thereforen, take one or more tokens not to make that an unreliable token becomes a reference pattern, and make template of a reference word, sometimes it can be the template of a reference word.

At this time, you'd better to have a interval to reduce an external noise during recording.

2. Averaging method.

Let's suppose that use pronouncing M times per lone word as trainingdata in order to create a reference pattern of a word. The averaging method is a method that takes 12th LPC cepstrum coefficient by the average value can be got after finding out beginning and ending point and getting tokens of M numbers by averaging 12th LPC cepstrum coefficients.

This method can reduce making an unreliable reference pattern because it can get the average value even if one(or two) of M numbers of token is differnt from others. However, characteristic of M numbers of token is very different from others, at this moment characteristic of an actual word can be had a distortion.

For such kinds of reason, this method is useful for speech recognition system of speaker dependent that training data has be different even through pronounce same word for M times to make a reference pattern. In this case, it is needed normaliz ation of time axis before taking an average value. According to method of normalization, averaging method is divided into dynamic average and linear average.

(1) Linear average[3]

Linear average is the method to get averaging value that make each token linear time normaliz ation after seeking token of each word from tra-ining data.

The algorithm is as follows.

i) Seek frame length, $N_i(i=1,\cdots, M)$ between

number of token, M and ith token and then get the averaging length of the number of token, M.

$$N=\frac{1}{M}\sum_{i=1}^{M}N_i$$

ii) Obtain normalized feature vector of ith token from i=0 to M.

$$R_i(m)=\alpha\ R_i(k)+R_i(k+1)\ (m=1,\cdots,M)$$

$$K=[1+\frac{N_i-1}{N-1}\ (m-1)]$$

$$\alpha=K-\frac{N_i-1}{N-1}\ (m-1)$$

iii) Finally, seek the averaging feature vector of linear time normalization and make predict coefficient or reflect coefficient used by this as reference pattern.

$$R^*(m)=\frac{1}{M}\sum_{m=1}^{M}R_i(m),\ m=1,\cdots,M$$

Actually, dynamic average method needs many times to peroform so generally we use linear average method.

(2) Dynamic Average

Dynamic average is the method that finds out an optimum path using dynamic programming first and then traces an averaging value of tokens according to it.

The algorithm is as follows.

i) Get the averaging length of M after looking for token numbers, M and the frame length of token of ith, N_i.

$$N=\frac{1}{M}\sum_{i=1}^{M}N_i$$

ii) $|N_F-N|=\min|N_i-N|\ i=1,\cdots,M$

Get K'th token satisfied with above, and set n=1.

iii) When use the dynamic programming, put kth token which is from (ii) on reference axis and putnth token on test axis, and then get an optimum path w(m) $(m=1,\cdots,N_k)$.

Next, get the averaging value according to the averaging value.

$$R^*(m)=\frac{1}{2}[R_k(m)+R(w(m))],\ (m=1,\cdots,N_k)$$

iv) If nth token is the last in total number, M, make a reference pattern after seeking predict coefficient with $R^*(m)$ $(m=1,\cdots,N_k)$. If not, $R^*(m)$ becomes new $R_k(m)$ $(m=1,\cdots,N_k)$ and goes (iii) through n=n+1.

3. Clustering method[5]

It is very usual method and the order is as below. First, get beginning point and ending point from training data and then make tokens cluster by clustering method. Next, get averaging token from each cluster. There are many methods to find reference token from clustering method.

The representative algorithm is as follows.

i) Seek token M from training data and frame length of ith token $N_i(i=1,\cdots,M)$, then get number of reference word template, K, the maximum iteration NCI and distortion threshold I_{max}. Set k=1, TRY=1.

ii) The averaging length of total number of

token, M

$$N=\frac{1}{M}\sum_{i=1}^{M}N_i$$

iii) Seek j which is satisfied with $|N_j-N|=\min|N_i-N|$, i=1,…,M make jth token as reference token R(m) (m=1,…,N_j)

iv) Set up are refernce token $R_{m+1}(m)$ (m=1,…,N_j) by linear average method.

v) If n is larger than the maximum iteration number, NCT, it goes to (vi), If not, get distortion value from R(m) and $R_i(m)$, (m=1,…,N_j) and set n=n+1. If distortion value is larger than distortion threshold T_{max}, it goes to (iv) and if it is smaller then goes to (vi).

vi) Check TRY, if it is 1, then goes to (vii), if not, goes to (viii).

vii) Get reference token R(m) and distortion value from total number of token and when it is larger then T_{max} cancel that token and set M=N-1. Set TRY=2 and goes to (ii).

viii) Get predict coefficient or reflect coefficient then make reference word template.

ix) Check it is same or not with reference word template what R wants, if same, finish whole courses and if not, get new number of token M which was cancelled tokens in (vii) and N_i(i=1,…M), then goes to (ii).

The above can take one or more reference word template, per one word, so it has a good merit to use not noly speech recognition system of speaker dependent but also speech recognition system of speaker independent.

4. Modified dynamic Average method.

Modified dynamic average method is the method that finds out an optimum path using dynamic programming first as dynamic average method and then traces an averaging value of tokens according to it. But it unlikes method that traces an averaging value of tokens.

The algorithm is as follows.

i) Get the averaging length of M after looking for token numbers, M and the frame length of token of ith, N_i.

$$N=\frac{1}{M}\sum_{i=1}^{M}N_i$$

ii) $|N_k-N|=\min|N_i-N|$ i=1,…,M

Get k'th token satisified with above, and set n=1.

iii) When use the dynamic programming, put kth token which is from (ii) on reference axis and put nth token on test axis, and then get an optimum path w(m) (m=1,…,N_k).

Next, get the averaging value according to the optimum path.

$$R'(m)=\frac{1}{M}\sum_{i=1}^{M}[R_k(m)+R_i(w(m))],\ (m=1,\cdots,N_k)$$

Ⅲ. Experiment result

In speech recognition of the independent speaker according to modified dynamic average method as reference pattern, we chose out 37 city name as the recognition vocabulary.

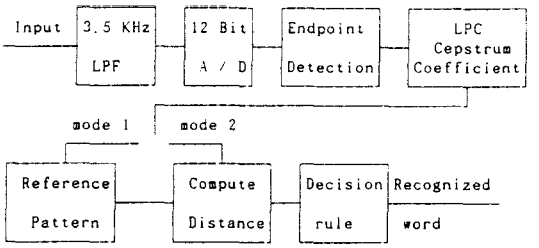
Modified dynamic average method is made by taking one among words spoken three times

three men respectively and we recognize with the remained data which is not used by training data of each speaker.

And also, in case of the other methods, we used the same data for comparison.

1. construction for recognition system

Fig. 1 represents the speech recognition system according to proposed modified dynamic average method, and all data fixed sampling frequency as 8KHz, LPF as 3.5KHz, feature parameter as 12th LPC cepstrum coefficient.



mode 1: tranining set
mode 2: test set

Fig. 1 Block diagram of recognition system

2. Recognition result.

In this experiment, reference pattern by DTW is selected one template in each word. Recognition rate by generating reference pattern is showed in table 1.[9]

Table 1. Recognition rate according by each method.

(Unit: %)

Ref. Pattern	Speaker A	Speaker B	Speaker C	Rec. Rate
Causal method	84.0	81.0	84.0	83.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Clustering method	96.0	96.0	96.5	96.6
Modified dynamic average method	97.0	98.0	98.0	97.6

Ⅳ. Conclusion

In this paper, we performed speech recognition of independent speaker by DTW, with 57 city names, and compared recognition experiment by design method of each reference pattern.

As an result of comparison experimentation, recognition accuracy by causal method is about 83.4 percent, linear average method is about 96.6 percent, dynamic average method is about 96.8 percent, clustering method is about 96.6 percent and modified dynamic average method is about 97.6 percent.

Therefore, it is proved that it is the best to use modified dynamice average method proposed in this paper as reference pattern.

Reference

1. H.sakoe and S.chiba "Dynamic programming optimization for spoken word recognition", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp.43-49, Feb. 1978.
2. L.R. Rabiner and J.G. Wilpon, "A simplified, robust traning procedure for speaker trained, isolated word recognition system", J.Acoust. SOC, Amer., vol.68, pp. 1271-1276, NOV. 1980.
3. L.R. Rabiner, "On creating reference templates for speaker independent recognition of isolated word", IEEE Trans. Acoust., Speech, Signal processing, vol. ASSP-26, pp.34-42, Feb. 1978.
4. J.T. Tou, R.C. Gonzalez, Pattern recognition Principles, Addison Wesley Publishing Company, Inc. 1971.
5. S.E. Levison, L.R. Rabiner, A.E. Rosenberg and J.E. Wilpon, "Interactive Clustering Techniques for Selecting Speaker Independent Reference Techniques for Isolated Word Recognition", IEEE Trans. on ASSP [illegible]
6. C.S. Myers, R.R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-29, pp. 284-297, Apr. 1981.

▲Eui-Bung Jeoung

Eui-Bung Jeoung was born in Seoul, Korea.

1984.2 : Departments of Electronics Engineering, Won Kwang University(B.S.)

1986.8 : Departments of Computer Engineering, Kwang Woon University(M.S.)

1987~ : Ph. D course from Departments of Electronics Engineering, Kun Kuk University.

▲Young-Hyuk Ko

1977.3 : Departments of Electronics Engineering Kun Kuk University(B.S.)

1981.3 : Departments of Electronics Engineering Kun Kuk University(M.S.)

1990.2 : Departments of Electronics Engineering Kun Kuk University(Ph.D.)

1990.2~ : He has been a professor in the Departments of Information Communication Dong shin University

▲Jong-Arc Lee

He received the B.S. degree in electrical engineering from the Han-Yang University in 1966, and the M.S., Ph.D. degree in electrical engineering from the Yon-Sei UNIV, Seoul, in 1970 and 1974, respectly. Since 1978, he has been a professor in the Kun-Kuk University.