# Feature Extraction from the Strange Attractor for Speaker Recognition

# 화자인식을 위한 어트랙터로 부터의 음성특징추출

Taesik Kim*
김 태 식*

## ABSTRACT

A new feature extraction technique utilizing strange attractor and artificial neural network for speaker recognition is presented. Since many signals change their characteristics over long periods of time, simple time-domain processing techniques should be capable of providing useful information of signal features. In many cases, normal time series can be viewed as a dynamical system with a low-dimensional attractor that can be reconstructed from the time series using time delay. The reconstruction of strange attractor is described. In the technique, the raw signal will be reproduced into a geometric three dimensional attractor. Classification decision for speaker recognition is based upon the processing of sets of feature vectors that are derived from the attractor. Three different methods for feature extraction will be discussed. The methods include box-counting dimension, natural measure with regular hexahedron and plank-type box. An artificial neural network is designed for training the feature data generated by the method. The recognition rates are about 82%-96% depending on the extraction method.

## 요 약

화자인식을 위한 음성특징을 카오스의 어트랙터와 신경망을 이용해서 추출하는 방법을 제시한다. 기존의 음성신호 표현 방법과 특징 추출법은 음성인식 시스템에서 별 무리가 없이 사용되었으나 2차원 표현에서 오는 한계는 아직까지 극복해야할 과제로 남아있다. 본 연구에서는 최근 각광받고있는 새로운 시그날표현기법인 카오스이론의 스트레인저 어트랙터를 이용하여 음성특징을 추출하는 화자인식시스템에 적용하고자 한다. 입력된 음성신호는 3차원 공간안에서 어트랙터라 불리우는 기하학적인 형태로 표현되는데 이 3차원 어트랙터를 이용하면 기존의 2차원적인 표현으로 부터 얻는 특징보다 더 많은 정보를 추출할 수 있을 것이다. 특징추출 기법은 3가지를 제안하였고 각 기법으로 추출된 특징벡터는 신경회로망을 통해 학습되어 인식률을 실험하였다. 제시한 기법들에 따라 다르나 인식률은 약 82%부터 96%까지 나타났다.

*Dept. of Computer Science, Keimyung University.
접수일자 : 1994년 7월 4일

## Ⅰ. Introduction

In the problem of speaker recognition, one must detect the presence of features from the waveform and classify them to facilitate the determination of whether a particular waveform corresponds to certain speaker's utterance or not [1]. In order to detect features, utterance should be represented into digital waveform as accurately as possible. For representing speech signals, a number of different methods have been proposed ranging from simple sets such as energy and zero crossing rates, to complex representations such as the short-time spectrum, linear-predictive coding, and the homomorphic model [2]. Generally, classification decisions are based upon the statistical distribution of the features of that utterance such as pitch and formant location, the calculation of correlation coefficients, linear predictive coefficients together with any information on the ordering, or time sequence of the features [3]-[6]. The selection of the best parametric representation of speech signal is an important task in the design of a speech recognition system. Normally, these methods are good enough to use in most speech recognition systems. However, some problems there have been time axis distortion and spectral pattern variation. On the other hand, the spectral pattern variation, which is caused by a complex mixture of several effects, is hard to treat. Also, it is possible that traditional signal representation techniques hide some useful information due to its limitation of the presentation technique.

In recent years, traditional methods of time series analysis like power spectra have been augmented by the new method which is called strange attractor. In many cases, normal time series can be viewed as a dynamical system with a low-dimensional attract that can be reconstructed from the time series using time delay. Strange attractors, which are geometric forms that characterize longterm behavior in the state space, have become a popular research topic which has drawn interest not only from computer science, physics and mathematics but also from all other natural science and even the social sicences.

In this paper, a new feature extraction technique using strange attractors for speaker recognition is discussed. The raw speech signal will be reconstructed into strange attractor, then a set of feature vectors, which include some characteristics of a specific speaker's speech, will be generated. Therefore, a set of parameters obtained from attractor can be chosen as the components of a pattern vector of the speech signal, and the pattern vector can be used as inputs for the artificial neural network. Neural networks are quite a general pattern recognition mechanism which, by being fed traning samples of given categories, can learn to achieve a function to discriminate between the categories. Therefore, neural network is suitably applicable to pattern recognition problems where an analytical approach is inapplicable [7].

## Ⅱ. Reconstruction of Strange Attractor

By embedding the time series data into a phase space with time delay coordinates, we can construct orbits. From a measured time series speech data, $\{x_0, x_1, x_2, \dots\}$ where $x_0 = (0)$, $x_1 = x(\tau)$, $x_2 = x(2\tau)$, $x_3 = x(3\tau)$ and $\tau$ is the sampling interval, we can make the following sequence of vectors data with time delay T :

$$
\begin{aligned}
&(x(0),\ x(T),\ x(2T)) \\
&(x(\tau),\ x(\tau+T),\ x(\tau+2T)) \\
&(x(2\tau),\ x(2\tau+T),\ x(2\tau+2T)) \\
&\vdots \\
&(x(k\tau),\ x(k\tau+T),\ x(k\tau+2T)).
\end{aligned}
$$

Plotting these points in three-dimensional space with connecting line segments, we obtain a strange attractor.

A sequence of samples representing a typical speech signal is shown in Figure 1, and its strange attractor is shown in Figure 2.

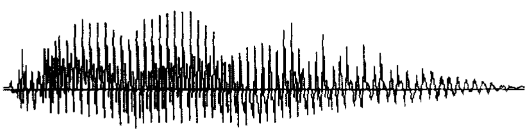

Figure 1. Waveform of the utterance "Ah".

Figure 2. Reconstruction of attractors for the utterance using a time delay of T = 3, viewed from different perspectives.

## Ⅲ. Feature Extraction

There are a large number of potential features one can extract from strange attractors. Many of these tend to be unreliable, and it is difficult to devise detection algorithms for them. Many researchers have attempted detecting various distinctive features with a limited success [8]. In this paper, box counting dimension and natural measure method to detect signal features are discussed.

### 3.1 *Box-Counting Dimension*

The most basic property of an attractor is probably its dimension. A dimension is defined as follows :

Consider a strange attractor which is embedded in a $d$-dimensional space. Let $\{X_i\}_{i=1}^{N}$ be the points of a long time series on the attractor. Cover space with a mesh of $d$ dimensional boxes of size $b^d$. Let $M(b)$ be the number of boxes that contain points of the series $\{X_i\}_{i=1}^{N}$ and let $p_k = N_k/N$ where $N_k$ is the number of points in the $k_{th}$ box. A dimension is then defined by [8]

$$D = \lim_{b \to 0} \lim_{N \to \infty} \log M(b)/\log b.$$

The box-counting dimension proposes a systematic measurement, which applies to any structure in the plane and can be readily adapted for structures in space. To compute box-counting, we put the attractor onto a regular mesh size $s$, and simply count the number of grid boxes which contain part of the attractor. This gives a number, $N(s)$. After counting all the boxes, the geometric region is subdivided into square boxes of smaller linear size $s$. Then we count those boxes which contain a part of the attractor again. When repeating the same procedure using smaller s, we expect to find that the new count $N(s)$. After each iteration we check if the current point is in a box that we have not yet visited, in which case we increase our count by 1. Repeat the whole procedure for a different size $s$ and finally compute as the slope of a $\log(N(s))/\log(1/s)$ diagram to measure its slope $D$, which is the box-counting dimension [8]-[9]. Figure 3 illustrates this procedure for the six speakers using six measurements of each speaker.

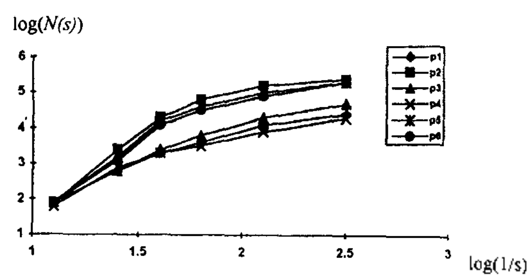


Figure 3. Illustration of Box Count for the Six Speakers.

### 3.2 *Natural Measure*

In the space, boxes can be weighted according to how many times an orbit visits them. Thus, boxes which the orbit passes through very frequently have a stronger impact than boxes which the orbit rarely visits. Let's consider an open subset $B$ of a space $X$ in which an attractor lies. We can count the number of times an orbit $x_0, x_1, x_2, \ldots \in X$ enters the subset $B$, and it is natural to assume that the percentage of all points which are in $B$ stabilizes as we perform more and more iterations. This percentage is called the natural measure $\mu(B)$ for the system. Formally,

$$\mu(B) = \lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \lambda_B(x_k)$$

where

$$\lambda_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise} \end{cases}$$

and $\sum_{k=0}^{n} \lambda_B(x_k)$ is the number of points from the orbit $x_0, \ldots, x_n$ which fall in the set $B$ [8].

The natural measure can be understood as a means of quantifying the mass of a portion of the attractor. In this paper, two kinds of box types are proposed for calculating the natural measure. One type of box forms as a regular hexahedron. When the space is divided into small unit, the box has to be equaled length of x, y, and z-coordinate. The other one is plank-type box which is divided into x-coordinate orientation. Figure 4 illustrates these types.

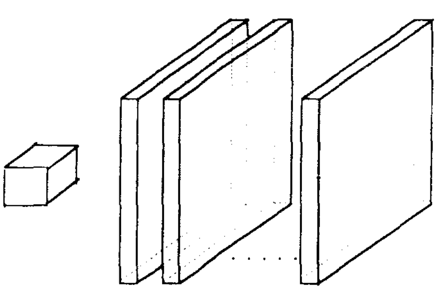

Figure 4. Illustration of Different Types. Small box forms regular hexahedron while the other forms plank.

In our experiment, 198 equal size of small cubes and 51 for the plank-type boxes were generated to get the natural measure. Figure 5 shows the natural measure using the different box types for one of the training data of each speaker.

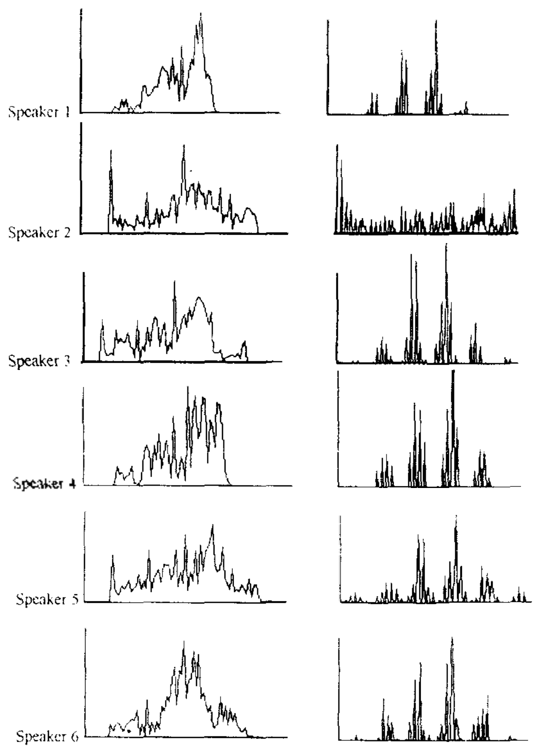


Figure 5. Graphical Representation of the Natural Measure for the Six Speakers. Left column indicates the natural measure using the plank-type box while the other side is using small cube. At each graph, x-axis indicates the box numbers and y-axis indicates the natural measure.

## IV. Experimental Result

### 4.1 Speech database

For performance evaluation, we have used a word "Ah", which was uttered by three female and three male Korean speakers. All utterances were recorded in a quiet room and digitized at a 14kHz sampling rate. The database was then split into a training set and testing set. For each speaker, we used 20 utterances for training and 80 for recognition test.

### 4.2 Neural Network

Neural networks are providing to be useful for difficult tasks such as speech recognition, because they can easily be trained to compute functions from any input space to output space. This section examine the configuration of the neural network that is to perform the proposed method. The input to the network is a feature vector extracted from the attractor. The network architecture is shown in Figure 6.

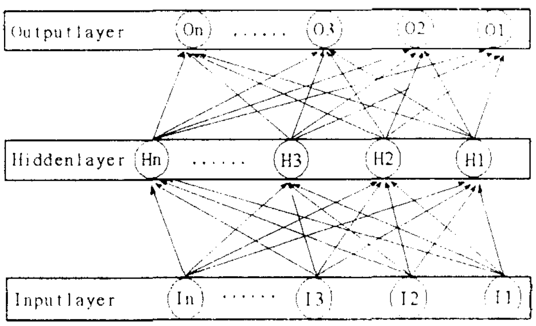


**Figure 6.** Structure of the Neural Network

The network takes the feature vector $[\mu(B)]$. The output vector for hidden layer, denoted by H, is $H = f(W_h[\mu(B)])$, where f( ) is sigmoid function, $W_h$ is weight matrix of hidden layer. The output of the network is given by $W_o H$, where $W_o$ is weight matrix of output layer. The output layer units represent speaker's name. For the box-counting method, the input vector consists of $[\log(N(s))/\log(1/s)]$ for each size. For the natural measure, we used the value of $\mu(B)$ of each block $B$. Training of the network was done by a back propagation algorithm, using an entropy criterion.

### 4.3 Experimental Results

Table 1 shows the results from the recognition experiments as obtained from the testing data. As can be seen, for all six speakers, the natural measure yields considerably higher performance thar box counting dimension. In the natural measure, recognition rates of the plank-type method are much higher than the small cube method.

Experiments also have been carried out to compare the performance of the natural measure method and the *peak transition* method reported in [10]. Evaluation of both methods was carried out using the same speech input data, and the recognition rates of the natural measure is 1% higher than the rates of the peak transition method. Compared with the results, it has been observed that it was possible to improve the recognition performance of any speech recognition system by adapting strange attractor for some other recognition systems.

**Table 1.** Recognition Results for Six Speakers Over test Data Using the Three Methods

| Speaker | Box Counting Dimension | | Natural Measure (Cube) | | Natural Measure (Plank Type Box) | |
|---|---|---|---|---|---|---|
| | number of errors | recognition rate | number of errors | recognition rate | number of errors | recognition rate |
| 1 | 15 | 81.3 | 5 | 93.7 | 4 | 95.0 |
| 2 | 14 | 82.5 | 4 | 95.0 | 3 | 96.3 |
| 3 | 14 | 82.5 | 5 | 93.7 | 2 | 97.5 |
| 4 | 17 | 78.8 | 8 | 90.0 | 3 | 96.3 |
| 5 | 16 | 80.0 | 7 | 91.3 | 4 | 95.0 |
| 6 | 9 | 88.8 | 7 | 91.3 | 1 | 98.8 |
| Total | 85 | 82.3 | 36 | 92.5 | 17 | 96.4 |

## V. Conclusion

A new feature extraction method has been discussed, where strange attractor and neural network are used. A recognition accuracy as high as 96% was obtained. This shows a high potential applicability of the proposed methods. The main advantages of using strange attractor are that they extract features in a simple and elegant way, and that they can model the dynamic properties of speech signal better than traditional linear predictive model. Their main current weakness is that they have poor discrimination even though it has good recognition rates. Future research should concentrate on improving the discriminatory power. Also performance of this model needs to be compared with that many of other techniques for identifying speakers using same input data.

## REFERENCES

1. L. Hong, J. Glass, M. Phillips, and V. Zue, "Phonetic Classification and Recognition Using the Multi-Layer Perceptron," Neural Information Processing System 3, Morgan Kaufmann Publishers, Inc., pp.248-254, 1991.
2. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ : Prentice-Hall, 1978.
3. F. Mark, R. Cole, and K. Roginski, "English Alphabet Recognition with Telephone Speech," Neural Information Processing System 4, Morgan Kaufmann Publishers, Inc., pp.135-142, 1992.
4. H. Patrick and A. Waibel, "Multi-State Time Delay Neural Networks for Continuous Speech Recognition," Morgan Kaufmann Publishers, Inc., pp.199-206, 1992.
5. I. Nathan, "Exploratory Feature Extraction in Speech Signals," Morgan Kaufmann Publishers, Inc., pp.241-247, 1991.
6. R. Linggard, D. Myers, and C. Nightingale, "Neural Network for Vision, Speech and Natural Language," Chapman & Hall, London, 1992.
7. H. Sakoe, R. Isotani and K. Yoshida, "Speaker Independent Word Recognition Using Dynamic Programming Neural Networks," *Proc. IEEE*, 1989.
8. H. Peitgen, H. Jurgens, and D. Saupe, "Chaos and Fractals, New Frontiers of Science," New York, NY : Springer-Verlag New York, Inc., pp.721-727, 1992.
9. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.ASSP-26, no.1, pp.43-49, 1978.
10. T. Kim, "Utterance Transition Approach for Speaker Recognition Using artificial Neural Networks," Bulletin of the Institute for Industrial Science, Keimyung University, vol.16, no.1, pp.193-200, June 1993.

▲Taesik Kim

Taesik Kim received the B.S. degree from the Department of Computer Science, Keimyung University in 1984, the M.S. degree in Computer Science from Moorhead State University in 1987, and the Ph.D degree in Computer Science from North Dakota State University in 1992. He is currently on the Faculty of the Department of Computer Science at Keimyung University. His research interests include expert systems, chaos, and pattern recognition.