Unwarping of Continuous Speech Signals

*Cheng-Yuan Liou,*Wen-Pin Tai,**Fang-Ru Hsu, and **Hwai-Tsu Chang

*Dept. of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan ROC.
**Computer and Communication Research Laboratories,
Industrial Technology Research Institute, Hsinchu, Taiwan ROC.

Abstract

Unwarping of varying continuous speech patterns with neural arrays is studied in this work. These curved arrays are designed to solve the nonlinear time-alignment problem. We apply the self-organization algorithm to drive and stretch these flexible arrays in a 2-D plane to fit speech patterns. These arrays plus their stretching parameters constitute a set of time-scaled reference templates. With these templates, the matching between unknown speech patterns and templates can be evaluated accurately. Two perception energies are devised to count the matching credits. One energy counts the time precedent relationships among the patterns. The other energy counts the matched features without considering the precedent relationships. A speech system which can recognize finite Chinese words is developed based on the proposed approach.

1 Introduction

Researches in speech recognition have achieved various applications. Well developed techniques have been integrated in many systems. Unwarping the speech signals to accomplish accurate matching is one of the loci of these researches [1]. In this work, we present a relaxation techniques to improve the matching.

There is much effort expended to the speech recognition with neural networks. The neural phonetic map, developed by Kohonen [2], provides framed speech representations for the phonemes of spoken utterance. The time delay neural network (TDNN) [3] employs a multilayered neural architecture with overlapped input patterns which is capable of tolerating warped speech patterns. Many other methods (for example, [4] and [5]) are also devised to explore the temporal natures of speech patterns.

To represent continuous speech patterns, a linear neuron array [6] has been devised to represent the varying patterns. This array is allowed to collect stable significant features of speech patterns by the self-organization training algorithm [7]. After training, the precedent ordering of speech patterns is also formed on the array. This trained array provides a sequential representation of continuous speech

0-7803-3679-8/96/$5.00 © 1996 IEEE

patterns. This array is used to organize the precedent relationships among the warped patterns. In this work, we use an argumented array to further explore the time intervals among the speech features. This argumented array is allowed to bend and stretch in a 2-D plane during a relaxation process. This will improve the resolution of the recognition in the system.

2 Review of the Recognition System

To ease the contents we briefly review the recognition system in [6].

To solve the nonlinear time-warping problem, 1-D linear arrays are devised in the recognition system. Each Chinese word has its owner array. Each array is a reference template of a word with 100 unwarped features. The matching of unknown speech patterns is determined based on these reference arrays. Fig. 1 shows the recognition system.

From experiments, 100 neurons in an array are enough to represent the features of a Chinese word for a single speaker. For each neuron, there are 15 weighting synapses corresponding to the 15 spectral components of a speech pattern. The synapses of all neurons in a reference array are off-line trained by the self-organization algorithm. The trained array preserves the precedence relationships and preserves stable significant features among the spectral patterns of a Chinese word.

Before the training process, the preparation of spectral patterns is carried out by the signal pre-processing and the feature extraction process. The speech patterns are represented by the melscale coefficients as those in [8]. Each pattern containing 15 normalized melscale coefficients is obtained from speech signals every 9 ms. All patterns for a word are included in a training set for later use. Each array will be trained by a different training set.

To train each reference array, each pattern $\mathbf{x}$ in a training set is randomly selected as the input. Let $\mathbf{w}_i$, $i = 1\ldots100$, denote the synapses of the neurons, where each $\mathbf{w}_i$ is a 15 dimensional vector. After the initialization, there are two basic steps involved in the self-organization training process. For an input pattern $\mathbf{x}$, the best-matching neuron c on the map is determined by the minimum Euclidean distance, i.e.,

$$||\mathbf{x} - \mathbf{w}_c|| = \min_i\{||\mathbf{x} - \mathbf{w}_i||\}, \quad i = 1\ldots100. \quad (1)$$

Then the synapses of the neurons within the neighboring region of the neuron c are adapted to the input pattern $\mathbf{x}$ with different weights by

$$\mathbf{w}'_i = \mathbf{w}_i + \alpha h(c,i)(\mathbf{x} - \mathbf{w}_i), \quad i = 1\ldots100, \quad (2)$$

where α is the learning rate and h is the neighborhood function. Note that these 100 neurons are regularly arranged on a 1-D straight line. The size of function h is defined on this line. All these parameters are determined by experiments. The convergence can be reached by repeating these two training steps.

After the training process, the 1-D reference array will preserve the statistics of speech patterns closely. The precedent relationships among feature patterns are also kept in the array. These properties will be utilized in obtaining the matching credits in the recognition.

When the input patterns of an unknown word are presented during the utterance period, neurons on the correct reference array are excited sequentially from one end of the array to the other end. These excitations along the array will display a monotonously decreasing (or increasing) pattern (see the simulation results) when we cascade the array excitation patterns for the whole utterance period in a 2-D plane. These excitation patterns provide indication of correct reference array.

To evaluate the excitation pattern of neurons and find the best-matched reference template, two perception energies are devised. These two energies statistically measure the similarities of features and features' sequence between the reference array and the speech patterns' sequence.

The first perception energy E_1 estimates the total excitation on each array. The E_1 value is calculated by accumulating the excitations of the active neurons, i.e., $E_1 \equiv (\frac{1}{T}\sum_{\mathbf{x}}\sum_{i\in A(\mathbf{x})}||\mathbf{x}-\mathbf{w}_i||+\sigma)^{-1}$, where T is the duration of speech signals, σ is a small constant, $||\cdot||$ is the Euclidean distance measure, and $A(\mathbf{x})$ is the active neuron set for the input $\mathbf{x}$. Since the significant features are preserved on the trained arrays, the averaged E_1 energy for a correct word will be much larger than that of any other word.

The second perception energy E_2 is designed to determine the right sequence of features. We use a second order polynomial to fit the cascaded array excitations in a 2-D plane. The ranges of curvature and slope of this curve for correct recognition are within predetermined limits. The E_2 can be defined as the inverse of fitting errors, i.e., $E_2 \boxminus (\frac{1}{T}\sum_m e_m \cdot f_m + \sigma)^{-1}$, where e_m and f_m is the excitation and the fitting error of the excitation m. For a correct recognition, the averaged value of E_2 will be larger then that for other words.

Fig. 2 shows the experimental results of the averaged perception energy E_1 in the cases of correct matching and incorrect matching. There are 200 Chinese words in the command set for machine control. We can reasonably define the thresholds of E_1 for correct cases. When the E_1 threshold is satisfied, it implies potential candidate. The similar results for the averaged perception energy E_2 are also shown in Fig. 3.

3 The Argumented Array

To improve E_2, we need to find time intervals between features in the array. Note that each feature represented by a neuron in the array. By including these intervals, the excited array pattern will better approximate a smooth monotonic curve and close to a straight line. We expect these will improve the fitting. The relative time interval between two adjacent neurons along the array can be found by the following relaxation process.

We use two interval synapses which are

added to the 15 trained synapses. The array is allowed to move in a 2-D plane. The positions of the neurons are given in the plane by the two interval synapses. In the process, both interval synapses are adjusted relaxatively according to the precedence relationships among the features. The number of neurons in the array is the same as that in the linear array. In addition to the 15 synapses for speech patterns, each neuron in the array has two more synapses, the interval synapses $\mathbf{p}_i = (p_{i1}, p_{i2})$, $i = 1 \ldots 100$. These 15 synapses are fixed with their trained values and kept unvaried during the relaxation process.

Initially, this array has the same configuration as the trained linear array, except the two interval synapses. The training set contains the same sampled speech patterns. For each randomly selected input pattern from the training set, N neurons with the most activated excitations and N neurons with the least activated excitations are collected. Let L and S denote the neuron sets, respectively. The interval synapses in the two neuron sets, as shown in Fig. 4, are updated by

$$\begin{aligned} \mathbf{p}'_i &= \mathbf{p}_i + \alpha(\pi - \mathbf{p}_i), \quad i \in L, \\ \mathbf{p}'_j &= \mathbf{p}_j - \alpha(\pi - \mathbf{p}_j), \quad j \in S, \end{aligned} \tag{3}$$

where

$$\pi = \frac{1}{N} \sum_{i \in L} \mathbf{p}_i, \tag{4}$$

and α is the training rate which is a decreasing function of iterations, and the number N is also decreased during the relaxation process. These parameters are experimentally determined. We use this process to further unwarp the trained linear array with correct time interval between features.

After relaxation, the distance $||\mathbf{p}_i - \mathbf{p}_{i+1}||$ is the relative time interval between the features represented by neuron i and neuron $i+1$ in the array. We then cascade the same array 100 times in a 2-D plane with intervals proportional to these time intervals. Arrays are cascaded according to the time label t, where t is re-scaled and normalized to a new time label $\tau(t)$,

$$\tau(t) = \sum_{i=1}^{c(t)-1} d(i, i+1) \Big/ \sum_{j=1}^{100} d(j, j+1), \tag{5}$$

where the function $d(i, i+1)$ is defined as the Euclidean distance between the interval synapses $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, i.e., $||\mathbf{p}_i - \mathbf{p}_{i+1}||$. Cascading arrays with these intervals, we always obtain a better fitting of the excited array pattern on the 2-D plane. Fig. 5 and Fig. 6 show the excitation pattern on the linear array and the excitation pattern on the unwarped array, respectively.

4 Simulation Results

We apply the proposed method to accomplish a Chinese continuous speech recognition system. In the system, there are 200 different Chinese words used in voice commands. Each command contains 2 to 7 words. A command may have variables, such as digits. This command set is designed to control the computer system. The training set for each word contains the patterns obtained from a single person with natural utterance of the same word for 20 times. The sampling frequency for speech signals is 8000 Hz. There are about 500 ~ 1000 spectral patterns in each set. Total training cycles for each array are 500 iterations. The real-time system is simulated on a personal computer with DSP boards. We also include the grammars of these commands to improve and speed the recognition. The training process is carried out adaptively. For each new user, it takes 2 hours to reach satisfied recognition. Note that isolated word pronunciation is not necessary during the training process. The utterance of each command may be continuous for both training and recognition.

To illustrate the performance of the proposed model, we carry out 500 experiments for 3 speakers both on the system with linear array and on the system with argumented arrays. Voice commands are randomly selected for tests. The rates of correct recognition for different speakers are averaged. For the two systems, the averaged correct rates are 89.6% and 96.2%, respectively. The simulation results show that the proposed array can be applied to solve the Chinese speech recognition effectively.

5 Acknowledgments

The developments of methodology and technology of this work have been supported and sponsored by National Science Council under contracts numbered *NSC 85-2213-E-002-014*, *NSC 84-2213-E-002-085*, and *NSC 82-0408-E-002-255*. The implementation on VLSI chip of this work is partially supported and sponsored by Industrial Technology Research Institute.

References

1. Waibel, A. and Yegnarayna, B., "Comparative study of nonlinear time warping techniques in isolated word speech recognition systems", *Tech. Rep. Carnegie-Mellon Univ.*, 1981.
2. Kohonen, T., "The neural phonetic typewriter", *IEEE Computer*, vol. 21, 1988, pp. 11-24.
3. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K.J., "Phoneme recognition using time-delay neural networks", *IEEE ASSP*, vol. 37, 1989, pp. 328-339.
4. Watrous, L., Ladendorf, B., and Kuhn, G., "Complete gradient optimization of a recurrent network applied to b,d,g discrimination", *J. Acoust. Soc. Amer.*, vol. 87, 1990, pp. 1301-1309.
5. Tank, D.W. and Hopfield, J.J., "Neural computation by concentrating information in time", *Proc. Nat. Acad. Sci.*, vol. 84, 1987, pp. 1896-1900.
6. Liou, C.Y. and Shiah, C.Y., "Perception of speech signals using self-organization on linear neuron array", *Proc. IJCNN*, Nagoya, Japan, 1993, pp. 251-254.
7. Kohonen, T., *Self-organizing maps*, Springer-Verlag, Berlin, 1995.
8. Davis, B. and Mermelstein, P., "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences", *IEEE ASSP*, vol. 28, 1980, pp. 357-366.

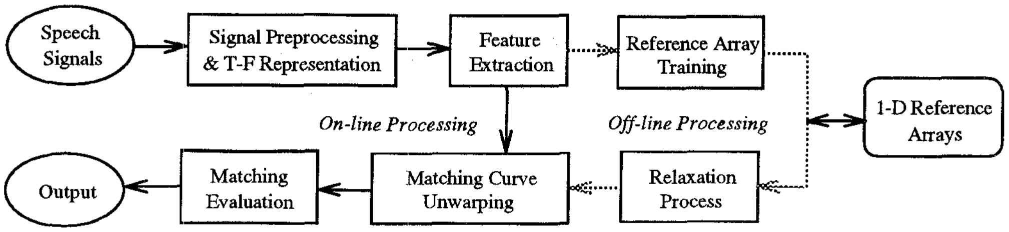
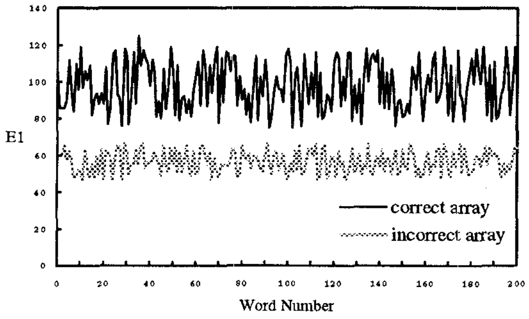


Fig. 1. The recognition system for continuous speech signals.



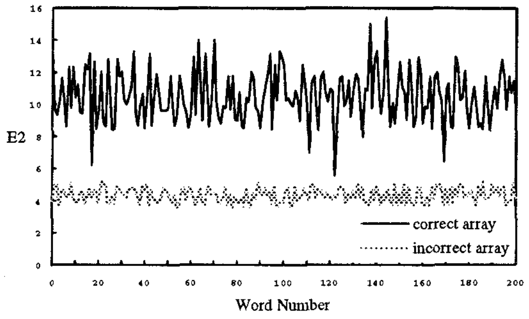


Fig. 2. The averaged E_1 energy.

Fig. 3. The averaged E_2 energy.

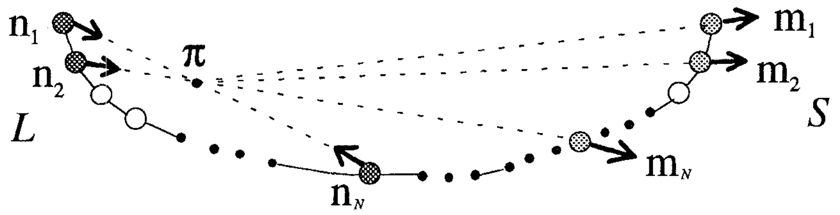
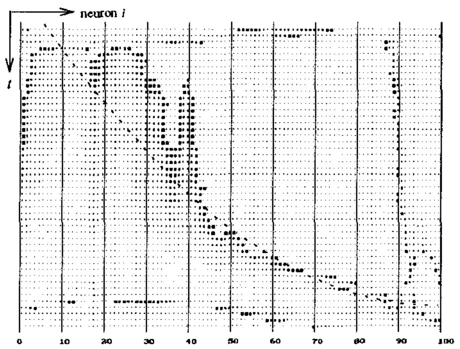


Fig. 4. The adaptation of the interval synapses for the argumented arrays.



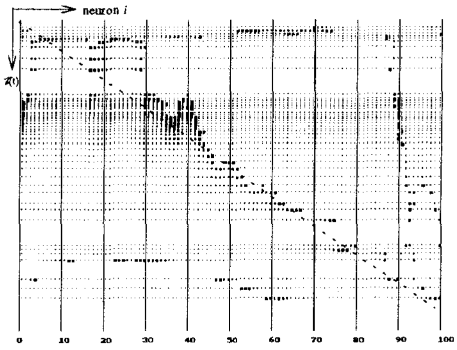


Fig. 5. The excitation patterns on the reference array for recognizing the correct word "ba".

Fig. 6. The excitation patterns on the argumented array for recognizing the correct word "ba".