An approach to speaker identification using DP-matched LVQ neural networks

Munehiro Namba, Hiroyuki Kamata, and Yoshihisa Ishida

*School of Science and Technology, Meiji University,
1-1-1, Higashi-mita, Tama-ku, Kawasaki, 214 Japan*

(Received 23 August 1996)

Automatic speaker identification is the other aspect in speech science, vis-à-vis spoken word recognition. In its traditional realization as a pattern classification task, there are some difficulties, such as the uncertainty of personalities, the dynamic variation of features, and so on. Nevertheless, vector quantization based pattern classifier still gains remarkable interest, because of its simplicity and the recent development in parallel computation technology. There are several reports on the combination technique of LVQ and other method for improvement of the conventional classifier. However, this paper concentrates on a hybrid algorithm of LVQ and DP-matching. The feature normalization capability in both the time and frequency domains of such a method can decrease the incorrect speaker identification which is caused by the variation of feature vectors in a short-term or a long-term. The experiment of 10 speakers identification shows that the proposed method can produce the better reference vectors, and hence it can promote the correct identification.

Keywords: Speaker identification, Speech recognition, LVQ, DTW, Neural Networks

PACS number: 43. 72. Pf

1. INTRODUCTION

The task of speaker recognition is to identify the person who speaks from a speech. In contrast to spoken word recognition, this objective is rather hard to realize, because personalities in speech are not clear in contemporary speech analysis researches.

The dynamic programming based pattern matching (DP-matching) which has been successfully applied to spoken word recognition, can also be utilized in speaker identification. A pioneering work of DP-matching is the paper by Sakoe and Chiba.¹⁾ They have argued the importance of time-normalization, and proposed an optimization algorithm based on the dynamic programming theory.²⁾ In spite of the advantage of DP-matching, however, the selection of the reference patterns has been a serious problem, and hence the number of references tends to increase.

In the recent paper,³⁾ Komori and Katagiri have proposed a novel training algorithm for the references in pattern classification. They have firstly introduced a generalized probabilistic decent algorithm (GPD), then investigated the algorithm in detail with several mathematical proofs. The experiments for the isolated speech tokens recognition have shown that their method can contribute towards increasing the discriminative power of the traditional DP-based speech recognizer. We are interested in a point that the simpler version of their technique leads to a generalized learning vector quantization (LVQ) incorporating a best-path search algorithm based on the dynamic programming, because the feature normalization in both time and frequency domains is necessarily required in speaker identification.

With this motivation, the performance of a hybrid method of DP-matching and LVQ for speaker identification is investigated in this paper. The

training algorithm discussed here is very similar to Komori and Katagiri's technique, however, our algorithm is developed from the artificial neural networks (ANN) topology, and seems to be more simpler than their method.

The next section reviews the traditional pattern recognizer and some related methods. Then, in Section 3, we describe the details of the proposed method. In Section 4, we present some experimental results for speaker identification and effectiveness of our scheme.

2. RECOGNITION METHODS

The overall view of the typical speech/speaker recognition through pattern classification is shown in Fig. 1. It is not reasonable for exploiting speech signals as they are, on comparing the similarity of a certain frame of a spoken word with corresponding prepared references. For the particular objective, speech signals may be redundant as they are, therefore the signals should be transformed into parameters or features through data recollection or reduction process. Their acoustic characteristics usually can be derived from the standard speech analysis and coding methods like linear predictive coefficient, cepstral coefficient, and so on.

The most important part of the recognition is the pattern comparison process. Many parametric and nonparametric models had been applied to it, but the focus of this paper is the vector quantization (VQ) based method. VQ is originally a kind of data compressor that maps a sequence of continuous or discrete vectors into a digital sequence on communication channels, hence this was first introduced as

a speech coding method to minimize required memories.⁴⁾ However, its superiority as a pattern classifier came to be clear soon because a given higher dimensional vector space can be splitted up to several clusters corresponding to the placement of collected vectors in the vector space. In order to make the appropriate codebook vectors, one ordinary averages the distortion of the collected vectors, on the other hand, Kohonen has proposed a novel approach to determine the optimal codebook vectors for the given vector space, now called LVQ. The normal LVQ algorithm defined as LVQ1 does simply arrange the nearest codebook vector according to a given vector iteratively by the following rules.

Let $\mathbf{x}$ be a sample vector picked out from the input data space, and $\{\mathbf{m}_i(t)\}$ be a number of codebook vectors assigned to each class that are properly initialized. The first thing to do is to find the nearest codebook vector by measuring Euclidean distances.

$$c = \arg \left[\min_i (\|\mathbf{x} - \mathbf{m}_i(t)\|) \right] \quad (1)$$

Then each codebook vector is forced to get near or far from the sample vector associating with its class.

a) If $\mathbf{x}$ and $\mathbf{m}_c(t)$ belong to the same class, then

$$\mathbf{m}_c(t+1) = \mathbf{m}_c(t) + \alpha(t)\{\mathbf{x} - \mathbf{m}_c(t)\} \quad (2)$$

b) If $\mathbf{x}$ and $\mathbf{m}_c(t)$ belong to the different class, then

$$\mathbf{m}_c(t+1) = \mathbf{m}_c(t) - \alpha(t)\{\mathbf{x} - \mathbf{m}_c(t)\} \quad (3)$$

c) If $i \neq c$, then

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) \quad (4)$$

where $\alpha(t)$ may be constant or decrease monotonically in time, and $0 < \alpha(t) < 1$.

Adopting LVQ in pattern recognition enables the reference templates to scatter into the given vector space which is formed by collected vectors, so that each one can be a representative of a certain category. From neural networks topology, some more complex learning algorithms have been addressed.⁵⁻⁹⁾

Another strategy to the recognition problem is a dynamic programming based pattern matching method, called DP-matching. Since this technique means a nonlinear time warping in the pattern matching procedure for speech recognition, this is also regarded as a sort of dynamic time warping (DTW) techniques. Because the features of speech

Fig. 1 Speech/Speaker recognition through pattern classification.

Fig. 2 Dynamic time warping in DP-matching.

signals are usually distributed on time-frequency domain in most practical systems, the dimension of the feature vector varies at a case by case. Therefore, it is often unable to get an Euclidean distance between two vectors, each of which has its own dimension. DP gives us a solution for such a problem from a mathematical point of view, “The principle of optimality.”

Assume that two patterns

$$A = \{a_1, a_2, \dots, a_m, \dots, a_M\}$$

$$B = \{b_1, b_2, \dots, b_n, \dots, b_N\}$$

should be compared on the warping function

$$F = \{c_1, c_2, \dots, c_k, \dots, c_K\}$$

as shown in Fig. 2. A new distance can be defined by

$$D_F = \min_F \left[\frac{\sum_{k=1}^K w(k) d(c_k)}{\sum_{k=1}^K w(k)} \right] \quad (5)$$

where $d(c_k)$ is an Euclidean distance between $a_{m|c_k}$ and $b_{n|c_k}$, and $\{w(k)\}$ denotes the penalty value which is positive. F and $\{w(k)\}$ have a variety of sets for selection, however, if the sum of $\{w(k)\}$ is determined so as to be constant (e.g. the city distance), the optimal warping function F defined in Eq. (5) can be obtained by DP.

3. LVQ ON DYNAMIC PATTERN CLASSIFICATION

3.1 A Modified LVQ Neural Network

An LVQ neural network can be viewed as one of the competitive trained networks. It classifies input

vectors on the basis of a set of stored vectors, as we reviewed in the previous section. For speech engineering, the most undesirable point is that LVQ handles only fixed dimensional vectors. Speech signal is a non-stationary stochastic process, so features of speech must be treated as dynamic patterns (i.e., the dimension may vary dynamically), and thus the original LVQ is not suitable for pattern recognition in which patterns are made from speech signals.

The measure at the competition phase of LVQ is the Euclidean distance between vectors. This is the sole reason for why LVQ deals with static patterns. It means that, paradoxically speaking, if something like nonlinear criterion replaces the Euclidean measure in LVQ neural networks, such a modified LVQ should have an ability to discriminate dynamic patterns. In this paper, the distance defined in Eq. (5) is adopted as a novel basis of LVQ, inspired by Komori and Katagiri's GPD. The distance in Eq. (5) is called DP-distance hereafter.

3.2 Training Law

In the similar way to the usual LVQ, the first thing to do is to find the nearest reference vector, however, by measuring DP-distances.

$$c = \arg \left[\min_i \{D(x, m_i)\} \right]$$

where, x is a vector picked out from the input data space and $\{m_i\}$ is a set of reference vectors. The function $D(a, b)$ denotes the DP-distance between a and b . The nearest one, namely the winner vector must be adjusted according to the training rule in Eqs. (2)–(4), but the vectors align on a nonlinear time axis with an optimal sense of DP-matching. An example of matched vectors is illustrated in Fig. 3. In this case, let us suppose that the individual component vector like a_1 or b_1 includes a single

Fig. 3 An example of DP-matched vectors.

LPC cepstral pattern, and each DP-matched correspondence of those component vector becomes as shown in the figure. In order to train the vectors with keeping such relations, irrespective of the vector class, the following two methods are devised.

The first one is a simple method of separate training each component vector based on the above relations. If a component vector in a reference vector corresponds to more than one of the component vectors in a tested vector, all nominees are applied to a component vector in the reference, randomly or sequentially. It means that, in this case, the vector a_1 is trained by both of the vector b_1 and b_2 , randomly or sequentially. Likewise, a_2 and a_3 are trained by b_3 , a_4 by both b_4 and b_5 , and a_5 by b_6 .

The other method is to train the reference vector after adjusting the tested vector to fit for the reference vector. Different from the former method, the reference vector is trained in a lump, but the concept is equivalent. In the same example, the reference vector is trained by following four vectors.

$$\{b_1, b_3, b_3, b_4, b_6\}, \{b_2, b_3, b_3, b_4, b_6\}, \\ \{b_1, b_3, b_3, b_5, b_6\}, \{b_2, b_3, b_3, b_5, b_6\}$$

This seems to be preferable because this training law tries to align the references and inputs entirely, however the number of iterations increases and its computational load cannot be ignored, especially when the dimension of the vector becomes higher. Therefore this paper adopts the former law.

4. EXPERIMENT IN SPEAKER IDENTIFICATION

4.1 Identification System Overview

The speaker identification system inspected in this paper is a neural network consists of three layers, that are the input layer, the competitive layer, and

Fig. 4 Speaker identification system with LVQ.

the linear decision layer. The network is shown in Fig. 4. The competitive layer behaves a modified Kohonen's LVQ layer mentioned in Sect. 3, and all neurons in this layer output their DP-distances to the next layer. The last is the decision layer which is made of linear perceptrons. In general, the decision layer is needless because finding the minimum value in the outputs from the competitive layer is usually enough to estimate the category (*i.e.*, the person). Nevertheless, the consideration of similarities concerning the input vector is known to be effective to characterize, and estimate the true category. The RBF neural networks or LVQ-SLP neural networks include such a concept.^{10,11)}

The training of the network is carried out separately as follows. Similar to the usual LVQ procedure, a number of reference templates should be initialized suitably at first of all. However in this network, it is not necessary for any vector to have the same dimension, at all.

A. Preparation Phase

- 1) Input a sampled feature vector to the input layer.
- 2) Scale up or down the parts of the sample vector to be suitable for DP-distance competition. Actually, each of the component vectors is scaled to a unit length.

B. Competition Phase

- 1) Calculate all the DP-distances between the reference vector and the sample vector.
- 2) Find the most nearest reference vector to determine the winner.

C. Training Phase (the modified LVQ layer)

Update the winner template according to its category (*i.e.*, person). If the winner and the sample vector belong to the same category (*i.e.*, the same person), the winner is brought closer to the sample, otherwise is made more distant from the sample.

- D. If the modified LVQ layer training is not finished, return to A. Otherwise, goes to E.

E. Training Phase (the decision layer)

Update the weights of the decision layer by usual algorithm like the back-propagation method.

4.2 10 Speakers Identification

The proposed network is applied to 10 male speakers identification. There are two steps, learning and testing, in the experiment. In the learning step, each person utters the same Japanese word

“OHAYOU” that means “Good morning” in English, 10 times. Speech signals are sampled by 11.025 [kHz], then segmented by 128-points Hamming window and transformed to a sequence of 32-dimensional mel-scaled LPC cepstral coefficients. Initially, one reference vector is prepared for one person. Although it is known that the classification accuracy is somewhat sensitive to the initial condition in neural networks, in this experiment, the reference of each person is randomly chosen from the collected 10 patterns, for simplicity. After the training, again each person utters the word 10 times that are used in the testing step.

The contour map in Fig. 5(a) illustrates the initial DP-distances between the reference templates and the collected personal feature vectors. The DP-distances are averaged and normalized so that each of the distances to themselves becomes a unit length, for comprehensibility (*i.e.*, $D(A_R, A) = 1$, $D(A_R, B) = D(A_R, B) / D(A_R, A)$, *etc.*). Each alphabet denotes the averaged feature of a person, and the alphabet with subscript R is the reference vector of the person. From this map, we can inspect the adequacy of the references. For instance, besides the dark diagonal belt from top-left to bottom-right, there are some dark regions in this map. Especially, because all the reference vectors are ambiguous to the personality of person G, the undesirable dark belt can be confirmed from the point (A_R, G) to (J_R, G) . In this condition, only 58-percents correct identification rate is achieved.

After the learning step with 60 iterations, the map in Fig. 5(b) can be obtained. The obstacles found

in Fig. 5(a) are diminished, and consequently the diagonal dark belt is emphasized and easily confirmed in the map. With the reference templates here, the identification rate of the given feature vector set becomes 100 percents. In this experiment, we couldn't find the large profit of the decision layer, different from that is presented in the references.^{10,11)} It improves approximately 2–3% through all experiments in this paper. Consequently, we believe that the learning in the modified LVQ layer promotes the correct identification. The effectiveness of the decision layer, however, might be notable in proportion to the number of persons to be identified.

To study and examine whether those reference vectors are really effective to the speaker identification, not only the identification of the trained data set but also of another data set, additional 100 speeches (10 speech per person) are collected within a month, and tested in the next step. The map in Fig. 6(a) shows a terrible result with untrained reference vectors. This is a well-known problem that the references are useless even after a few weeks or months, because of the uncertainty of personalities. The identification rate is 53 percents in this case.

However, in the case of the reference vectors trained by the proposed method in advance, the map in Fig. 6(b) can be obtained. Compared the map with untrained references, this map with trained references is somehow improved. Therefore the correct identification rate becomes 84 percents, such that the presented hybrid technique improves over

Fig. 5 DP-distances in Learning Step: (a) Initial contour map with untrained references; (b) Improved contour map with 60 trained references.

Fig. 6 DP-distances in Testing Step: (a) Contour map with untrained references and newly collected feature vectors; (b) Contour map with the same reference in Fig. 5(b).

Table 1 Identification rate for learning patterns.

(a) With untrained references

	A	B	C	D	E	F	G	H	I	J
A _R	100	0	0	0	0	10	0	0	0	0
B _R	0	40	0	0	10	0	0	0	0	0
C _R	0	0	60	0	0	0	0	0	0	0
D _R	0	0	0	40	0	0	0	0	0	0
E _R	0	0	0	0	30	0	0	0	0	0
F _R	0	0	0	0	0	40	0	0	0	0
G _R	0	60	40	60	60	40	100	20	40	70
H _R	0	0	0	0	0	0	0	80	0	0
I _R	0	0	0	0	0	10	0	0	60	0
J _R	0	0	0	0	0	0	0	0	0	30

[%]

(b) With trained references

	A	B	C	D	E	F	G	H	I	J
A _R	100	0	0	0	0	0	0	0	0	0
B _R	0	100	0	0	0	0	0	0	0	0
C _R	0	0	100	0	0	0	0	0	0	0
D _R	0	0	0	100	0	0	0	0	0	0
E _R	0	0	0	0	100	0	0	0	0	0
F _R	0	0	0	0	0	100	0	0	0	0
G _R	0	0	0	0	0	0	100	0	0	0
H _R	0	0	0	0	0	0	0	100	0	0
I _R	0	0	0	0	0	0	0	0	100	0
J _R	0	0	0	0	0	0	0	0	0	100

[%]

30 percents identification rate.

The distinct identification rates are summarized in Tables 1 and 2. From these tables, it can be confirmed that the proposed method improves the each speaker identification.

Generally, the result obtained here is not enough. As a practical system, over 90 percents correct identification rate is desirable at least. We believe that this unsatisfactory result in the experiment is caused by the inappropriate extraction method of

Table 2 Identification rate for testing patterns.

(a) With untrained references

	A	B	C	D	E	F	G	H	I	J
A _R	100	0	0	0	0	0	0	0	0	0
B _R	0	80	0	0	0	20	0	0	0	0
C _R	0	0	0	0	0	0	0	0	0	0
D _R	0	0	0	0	0	0	0	0	0	0
E _R	0	0	10	0	20	0	0	0	0	0
F _R	0	0	0	0	0	60	0	0	0	0
G _R	0	20	90	100	80	10	100	40	20	70
H _R	0	0	0	0	0	0	0	60	0	0
I _R	0	0	0	0	0	10	0	0	80	0
J _R	0	0	0	0	0	0	0	0	0	30

[%]

(b) With trained references

	A	B	C	D	E	F	G	H	I	J
A _R	100	0	0	0	0	0	0	0	0	0
B _R	0	90	0	0	0	10	0	0	0	0
C _R	0	0	50	0	0	0	0	0	0	0
D _R	0	0	0	80	0	0	0	0	0	0
E _R	0	0	0	0	90	0	0	0	0	0
F _R	0	0	0	0	0	90	0	0	0	0
G _R	0	10	50	20	10	0	100	0	20	40
H _R	0	0	0	0	0	0	0	100	0	0
I _R	0	0	0	0	0	0	0	0	80	0
J _R	0	0	0	0	0	0	0	0	0	60

[%]

feature vector. As it is mentioned in the introduction, what is the individual characteristic in speech, and how to extract it, are still progressive subjects. We will study further on this point, for higher identification rate.

5. CONCLUSION

This paper has discussed the advantages and problems of the dynamic programming based pattern matching and learning vector quantization neural networks in speaker identification, then investigated a combination method in the 10 speakers identification. As the feature normalization in both the time and frequency domains is necessarily required in speaker identification, the presented training technique is capable of reducing the incorrect identification rate. An experimental result has shown that a set of reference templates trained in advance is still effective even after a month, and its simple structure can also contribute to practical systems.

REFERENCES

- 1) H. Sakoe and S. Chiba, "Dynamic programming algorithm for spoken word recognition," IEEE Trans. Acoust. Signal Process. **26**, 43-49 (1978).
- 2) E. V. Denardo, *Dynamic Programming* (Prentice-Hall, Englewood Cliffs, NJ, 1982).
- 3) T. Komori and S. Katagiri, "GPD training of dynamic programming-based speech recognizers," J. Acoust. Soc. Jpn. (E) **13**, 341-349 (1992).
- 4) R. M. Gray, "Vector quantization," IEEE ASSP Mag. April, 4-29 (1984).
- 5) D. O'Shaughnessy, "Speaker recognition," IEEE ASSP Mag. October, 4-17 (1986).
- 6) F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," Proc. ICASSP '85, 387-390 (1985).
- 7) G. R. Doddington, "Speaker recognition—identifying people by their voices," Proc. IEEE **73**, 1651-1664 (1985).
- 8) T. Kohonen, "The neural phonetic typewriter," IEEE Comput., 11-22 (1988).
- 9) T. Kohonen, "Learning vector quantization for pattern recognition," Tech. Rep. TKK-F-A601, Helsinki Univ. Technol. (1986).

- 10) J. He, L. Liu, and G. Palm, "Speaker identification using hybrid LVQ-SLP networks," *Proc. ICNN '95*, Vol. 4, 2052-2055 (1995).
- 11) J. H. L. Kong and G. P. M. D. Martin, "A Review of a hybrid neural network: Kohonen learning vector quantization and counterpropagation," *Proc. ICNN '95*, Vol. 3, 1397-1402 (1995).

Munehiro Namba was born in Yokohama, Japan, on September 14, 1971. He received the B.E. degree in electronics and communication from Meiji University, Kawasaki, Japan, in 1995. He is currently working toward M.E. degree at the Department of Electrical Engineering, Meiji University.

His current research interests are in the area of signal processing, neural networks and speech analysis.

Hiroyuki Kamata was born in Kawasaki, Japan, on December 1st, 1959. He received the B.E., M.E., and Dr. Eng. degrees in electrical engineering from Meiji University, Kawasaki, Japan, in 1982, 1984 and 1987, respectively. In 1987, he joined the School of Science and Technology, Meiji University

as a Research Assistant and became a Lecturer in 1990. He is currently an Associate Professor at the School of Science and Technology, Meiji University. His current research interests are in the field of digital signal processing and its applications. He is a member of Acoustical Society of Japan, ICE and IEEE.

Yoshihisa Ishida was born in Tokyo, Japan, on February 24, 1947. He received the B.E., M.E. and Dr. Eng. degrees in electrical engineering from Meiji University, Kawasaki, Japan, in 1970, 1972, and 1978, respectively. In 1975 he joined the Department of Electrical Engineering, Meiji University, as a Research Assistant and became a Lecturer and an Associate Professor in 1978 and 1981, respectively. He is currently a Professor at the Department of Electronics and Communication, Meiji University. His current research interests are in the area of digital signal processing, speech analysis and recognition, and digital control. He is a member of the Acoustical Society of Japan, the IEEE and the IEICE of Japan.