ELSEVIER

Pattern Recognition Letters 23 (2002) 1735–1746

Pattern Recognition Letters

www.elsevier.com/locate/patrec

On the use of nearest feature line for speaker identification

Ke Chen [a,*], Ting-Yao Wu [b], Hong-Jiang Zhang [c]

[a] *School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*
[b] *National Laboratory of Machine Perception, The Center for Information Science, Peking University, Beijing 100871, China*
[c] *Microsoft Research Asia, 5/F, Sigma Center, No. 49, Zhichun Road, Hai Dian District, Beijing 100080, China*

Received 4 July 2001; received in revised form 28 November 2001

Abstract

As a new pattern classification method, nearest feature line (NFL) provides an effective way to tackle the sort of pattern recognition problems where only limited data are available for training. In this paper, we explore the use of NFL for speaker identification in terms of limited data and examine how the NFL performs in such a vexing problem of various mismatches between training and test. In order to speed up NFL in decision-making, we propose an alternative method for similarity measure. We have applied the improved NFL to speaker identification of different operating modes. Its text-dependent performance is better than the dynamic time warping (DTW) on the Ti46 corpus, while its computational load is much lower than that of DTW. Moreover, we propose an utterance partitioning strategy used in the NFL for better performance. For the text-independent mode, we employ the NFL to be a new similarity measure in vector quantization (VQ), which causes the VQ to perform better on the KING corpus. Some computational issues on the NFL are also discussed in this paper.

Keywords: Nearest feature line; Speaker identification; Dynamic time warping; Vector quantization; Nearest neighboring measure

1. Introduction

Nearest feature line (NFL) is a new pattern classification method, first proposed by Li (1998). In particular, it performs better on the condition that only limited data are available for training. The basic idea underlying the NFL approach is to utilize all the possible lines consisting of any pair of feature vectors (prototypes) in a given training set to encode the feature space in terms of the ensemble characteristics and the geometric relationship. As a simple yet effective algorithm, the NFL has shown good performance in face recognition (Li, 1998), audio classification and retrieval (Li, 2000), and image classification (Li et al., 2000). The NFL takes advantage of both the ensemble and the geometric features of samples for pattern classification. In contrast to a nearest neighbor (NN) classifier, the NFL makes better use of the ensemble information for decision-making.

Speaker identification is a task to determine an unknown voice token as belonging to one of registered speakers. There are usually two operating

* Corresponding author. Tel.: +44-121-414-4769; fax: +44-121-414-4281.
E-mail address: k.chen@cs.bham.ac.uk (K. Chen).

0167-8655/02/$ - see front matter
PII: S0167-8655(02)00147-2

modes; i.e., text-dependent and text-independent modes. By text-dependent, the same text is used in both training and testing. For the text-independent mode, any text is allowed during speaker identification. There are numerous sources of variability leading to poor speaker identification performance. The mismatch factors include session variability, health, educational level and intelligence, speech effort level, speaking rate as well as experience (Doddington et al., 2000). Therefore, speaker identification is extensively recognized as a challenging pattern classification problem.

Dynamic time warping (DTW) is a typical method to align two sequences of different lengths, and thus, applicable to both speech and speaker recognition for temporal template matching (Sakoe and Chiba, 1978; Furui, 1981). For speaker identification, the DTW distinguishes between two different speakers by means of speaker characteristics conveyed in verbal information of the given text. Thus, the frame-by-frame alignment makes the DTW applicable to only a text-dependent task. Unfortunately, such a frame-by-frame alignment leads to an expensive computational cost. In contrast, vector quantization (VQ) (Linde et al., 1980) provides a way to take advantage of speaker features regardless of verbal information, and therefore, becomes a useful method for text-independent speaker identification (Soong et al., 1985). In VQ, an NN criterion is employed for decision-making by measuring the similarity between a testing pattern and every codeword achieved during training. Thus, the geometric relationship among codewords is not sufficiently taken into consideration in the decision-making process.

Theoretically, speaker recognition belongs to the category of non-verbal speech classification. Although both instantaneous and transitional information turns out to be useful for speaker recognition (Soong and Rosenberg, 1988), our previous work showed that the use of such transitional (inter-frame) information might not be involved with a strict temporal alignment (Chen et al., 1996). Moreover, some instantaneous information carried by certain frames within an utterance can play a more important role than that of others in text-dependent speaker identification (Chen et al., 1996). In this paper, we explore the use of NFL for speaker identification without a strict temporal alignment. In this paper, we attempt to investigate the performance of the NFL in speaker identification and examine how the NFL approach performs in such a vexing problem of various mismatches between training and test. We propose an utterance partitioning strategy for the NFL to yield better text-dependent performance. On the other hand, we employ the NFL as a new similarity measure in VQ for the text-independent mode. In reality, a VQ method may need numerous codewords to model the original problem. Since an exhaustive search in the feature-line space is necessary for decision-making, the NFL measure could introduce higher computational cost to decision-making. In order to alleviate the problem, we propose a new distance calculation algorithm to reduce the computational cost. Some computational issues related to the NFL will also be discussed in this paper.

In simulations, we use two benchmark speech databases of Linguistic Data Consortium, Ti46 and KING, designed especially for speaker recognition. The former is used for the text-dependent mode, and the latter is used for the text-independent mode. As a consequence, text-dependent simulation results indicate that the NFL performs better than the DTW and its computational cost is much lower than that of DTW during decision-making. On the other hand, text-independent simulation results suggest that the VQ with a help of the NFL as a similarity measure leads to the better performance in contrast to the standard VQ itself for the same problem.

This paper is organized as follows: Section 2 briefly reviews the NFL and presents a new distance calculation algorithm. Section 3 describes the methodology of our experiments, and Section 4 reports simulation results. Some issues related to the NFL are discussed in Section 5 and conclusions are drawn in the last section.

2. Nearest feature line

In this section, we first briefly review NFL, and then, present a new distance calculation algorithm for the NFL to find the feature line of the nearest

distance from an unknown pattern to feature lines composed of the prototypes registered in a pattern classification system.

2.1. Brief review

NFL assumes that there are at least two prototypes in every class. The line passing through two feature points in the same class can extrapolate or interpolate to form a feature line in the feature space, as illustrated in Fig. 1.

We consider two feature points $\boldsymbol{f}_i^s$ and $\boldsymbol{f}_j^s$ in the feature space belonging to class s. Let $\boldsymbol{f}_i^s = (f_{i,0}^s, f_{i,1}^s, \ldots, f_{i,m}^s, \ldots, f_{i,M}^s)$, $0 \leqslant m \leqslant M-1$, where M is the dimension of this feature point. The distance d between the feature line $\overline{\boldsymbol{f}_i^s \boldsymbol{f}_j^s}$ passing through $\boldsymbol{f}_i^s$ and $\boldsymbol{f}_j^s$ and a query feature point $\boldsymbol{f}_x$ is calculated as

$$d(\boldsymbol{f}_x, \overline{\boldsymbol{f}_i^s \boldsymbol{f}_j^s}) = \|\boldsymbol{f}_x - \boldsymbol{p}_{i,j}^s\|, \tag{1}$$

where $\boldsymbol{p}_{i,j}^s$ is the projection point of $\boldsymbol{f}_x$ (c.f. Fig. 1) and $\|\cdot\|$ is the Euclidean norm. Thus, $\boldsymbol{p}_{i,j}^s$ can be achieved by

$$\boldsymbol{p}_{i,j}^s = \mu \boldsymbol{f}_i^s + (1-\mu)\boldsymbol{f}_j^s = \boldsymbol{f}_i^s + \mu(\boldsymbol{f}_j^s - \boldsymbol{f}_i^s), \tag{2}$$

where

$$\mu = \frac{(\boldsymbol{f}_x - \boldsymbol{f}_i^s)^{\mathrm{T}}(\boldsymbol{f}_j^s - \boldsymbol{f}_i^s)}{(\boldsymbol{f}_j^s - \boldsymbol{f}_i^s)^{\mathrm{T}}(\boldsymbol{f}_j^s - \boldsymbol{f}_i^s)}. \tag{3}$$

We can find the feature line of minimal distance by traversing all i and j ($i \neq j$), and thus, name it NFL. As a consequence, the testing pattern is recognized as belonging to the class represented by the prototypes constituting the NFL.

2.2. A new distance calculation algorithm

We denote $\boldsymbol{f}_i^s$, $0 \leqslant i \leqslant N_s - 1$, as training feature points belonging to class s in the feature space, and N_s is the number of points in class s. $\boldsymbol{f}_x$ is a query feature point, $\boldsymbol{p}_{i,j}^s$, $0 \leqslant i \leqslant N_s - 1$, $i < j \leqslant N_s - 1$, is the projection of $\boldsymbol{f}_x$ onto the feature line passing through $\boldsymbol{f}_i^s$ and $\boldsymbol{f}_j^s$. $\theta_{i,j}^s$ is the angle formed at the intersection of feature lines $\overline{\boldsymbol{f}_x \boldsymbol{f}_i^s}$ and $\overline{\boldsymbol{f}_x \boldsymbol{p}_{i,j}^s}$. Then we can get a cluster of lines all passing through $\boldsymbol{f}_i^s$ belonging to class s, $0 \leqslant i < N_s - 1$, as illustrated in Fig. 2.

Then, the distance between $\boldsymbol{f}_x$ and $\overline{\boldsymbol{f}_i^s \boldsymbol{f}_j^s}$ can be calculated as

$$d(\boldsymbol{f}_x, \overline{\boldsymbol{f}_i^s \boldsymbol{f}_j^s}) = \|\boldsymbol{f}_x - \boldsymbol{f}_i^s\| \sin\theta_{i,j}^s, \tag{4}$$

where $\|\boldsymbol{f}_x - \boldsymbol{f}_i^s\|$ is the norm which is unchangeable in the same cluster. When $\sin\theta_{i,j}^s$ is a minimum, that is,

$$|\cos\theta_{i,j}^s| = \frac{|(\boldsymbol{f}_x - \boldsymbol{f}_i^s)^{\mathrm{T}}(\boldsymbol{f}_j^s - \boldsymbol{f}_i^s)|}{\|\boldsymbol{f}_x - \boldsymbol{f}_i^s\| \|\boldsymbol{f}_j^s - \boldsymbol{f}_i^s\|}$$

is maximum, $d(\boldsymbol{f}_x, \overline{\boldsymbol{f}_i^s \boldsymbol{f}_j^s})$ is the shortest in this cluster. Since $\|\boldsymbol{f}_j^s - \boldsymbol{f}_i^s\|$ can be achieved off-line during training, thus we only need to compute $|(\boldsymbol{f}_x - \boldsymbol{f}_i^s)^{\mathrm{T}}(\boldsymbol{f}_j^s - \boldsymbol{f}_i^s)|$ in testing phase. Let

$$k_i^s = \arg\max_{i<j\leqslant N_s-1} \frac{|(\boldsymbol{f}_x - \boldsymbol{f}_i^s)^{\mathrm{T}}(\boldsymbol{f}_j^s - \boldsymbol{f}_i^s)|}{|\boldsymbol{f}_j^s - \boldsymbol{f}_i^s|},$$

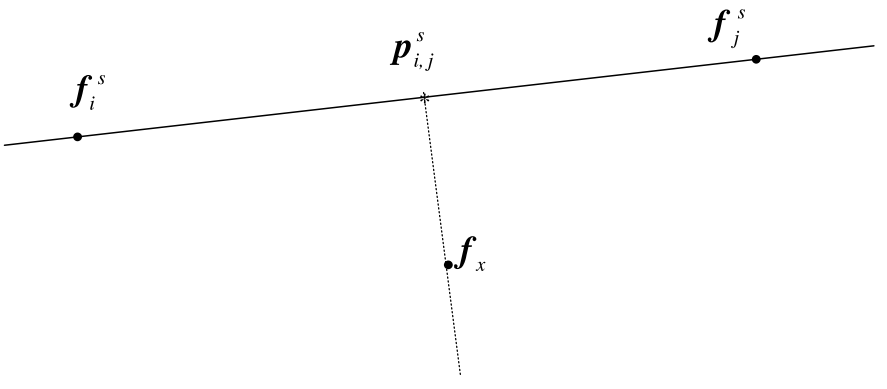

Fig. 1. Feature line $\overline{\boldsymbol{f}_i^s \boldsymbol{f}_j^s}$ and the query feature point $\boldsymbol{f}_x$.

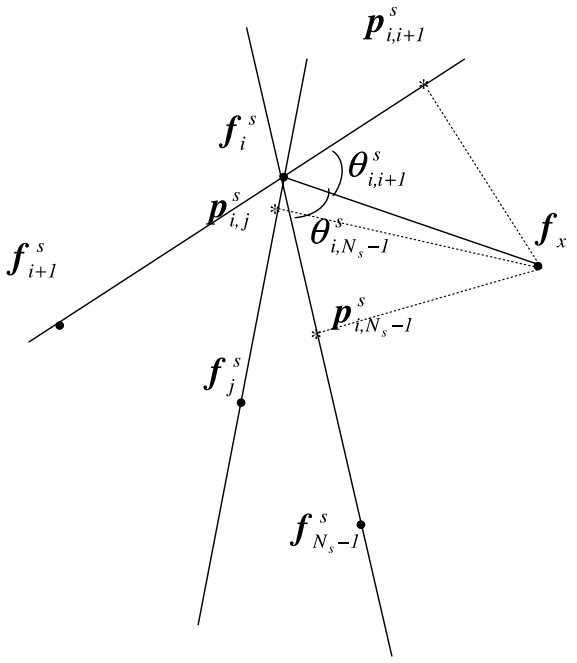

Fig. 2. A new distance calculation algorithm for NFL.

thus, the nearest distance in cluster i belonging to class s can be calculated as

$$d(\boldsymbol{f}_x,\overline{\boldsymbol{f}^s_i\boldsymbol{f}^s_{k^s_i}}) = \|\boldsymbol{f}_x-\boldsymbol{f}^s_i\|\sqrt{1-\cos^2\theta^s_{i,k^s_i}} = \sqrt{\|\boldsymbol{f}_x-\boldsymbol{f}^s_{k^s_i}\|^2-\left(\frac{|(\boldsymbol{f}_x-\boldsymbol{f}^s_{k^s_i})^{\mathrm{T}}(\boldsymbol{f}^s_i-\boldsymbol{f}^s_{k^s_i})|}{\|\boldsymbol{f}^s_i-\boldsymbol{f}^s_{k^s_i}\|}\right)^2}. \tag{5}$$

Then, the nearest distance d^s_{NFL} in terms of class s is

$$d^s_{\mathrm{NFL}} = \arg\min_{0\leqslant i<N_s-1} d(\boldsymbol{f}_x,\overline{\boldsymbol{f}^s_i\boldsymbol{f}^s_{k^s_i}}). \tag{6}$$

The decision rule throughout all classes is defined as follows:

$$s^* = \operatorname*{argmin}_{1\leqslant s<S}(d^s_{\mathrm{NFL}}), \tag{7}$$

where S is the number of registered speakers and s^* is the resulting speaker identity in terms of the current testing pattern.

We reduce the computational cost through introduction of a new distance calculation algorithm presented above to the NFL. Now we analyze why our algorithm yields a faster search. Suppose there are N_s training prototypes in class s and each of them is an M-dimensional feature vector. Then there will be $N_s(N_s-1)/2$ feature lines. As a result, the cost of the original NFL is $(3M+1)\times N_s(N_s-1)/2$ multiplication operations; i.e., it needs

$$\frac{(3M+1)}{2}N_s^2+\mathrm{O}(N_s)$$

multiplication operations, while our distance calculation algorithm takes only $(M+1)(N_s-i)$ multiplication operations in cluster i. For all i, $0\leqslant i<N_s-1$, our algorithm can find the NFL in this class with

$$\sum_{i=0}^{N_s-2}(M+1)(N_s-i)+(M+2)(N_s-1) = \frac{M+1}{2}N_s^2+\mathrm{O}(N_s)$$

multiplication operations. Thus, our computational cost is only 1/3 of the original one. Here, we emphasize that the time analysis for our distance calculation algorithm is only applicable to the search for the NFL during decision-making for an unknown pattern.

It is worth stating that our new distance calculation algorithm merely speeds up the search during decision-making but never changes the performance of the original NFL. Since the nearest distance d^s_{NFL} is fixed given a set of prototypes registered in a system, the decision-making in NFL is to find this nearest distance such that a tested pattern can be labeled by the corresponding prototypes' owner identity. Therefore, recognition results are identical regardless of distance calculation algorithms. Thus, we shall not differentiate the improved NFL from the original one hereinafter while we report recognition rates of the NFL.

3. Methodology

In this section, we present our experimental methodology. First, we give a brief description on two benchmark databases, Ti46 and KING. Then the acoustic analysis for different operating modes is presented. It is followed by the NFL classification including an utterance partitioning strategy

for the text consisting of multiple phonemes in text-dependent speaker identification.

3.1. Database

Ti46 database is a benchmark corpus for text-dependent speaker recognition. It contains isolated speech words from 16 speakers, eight males and eight females, and moreover, is divided into two sets: Ti20 and Ti_alpha. Digits (from "0" to "9") and ten simple English words (such as "go", "help") uttered by these speakers are collected in Ti20, and 26 alphabet letters are collected in Ti_alpha. In the Ti46 database, all utterances have been divided by its designers into two sets; i.e., training and test. For each word/alphabet, the training set contains at least five utterances uttered by each speaker at different times and there are ten utterances for each speaker in the testing set.

The KING corpus is used in text-independent speaker identification. There are only 49 speakers who own the complete data in 10 recording sessions, labeled by $S01, S02, \ldots, S10$, and all speakers are male. Each session was recorded in both a wide-band and a narrow-band channel. We employ only the wide-band set in the simulations reported in this paper.

3.2. Acoustic analysis

Prior to feature extraction, the speech data are pre-emphasized with the weight 0.95 and are blocked into frames. Each frame has 256 samples (around 23 ms) with 11.5 ms frame shift. The feature used is the statistical parameters of 19-order Mel-scaled ceptrum coefficients (MFCCs). The 19-order adaptive component weighted ceptrum coefficients (Assaleh and Mammone, 1994) are superposed on MFCCs. Adaptive component weighted ceptrum is a robust feature to discriminate the speakers through emphasizing the formants of speaker. Then the means and standard deviations of all the feature vectors corresponding to a piece of speech are estimated to form a 38-dimensional feature vector. In other words, the set of feature vectors corresponding to the piece of speech are further processed to form an integration feature vector through the use of their statistics in our simulations.

3.3. NFL classification

For an utterance for text-dependent mode or an acoustic segment for text-independent mode, the corresponding integration feature vector based on its statistics forms a prototype in the feature space. As reviewed in Section 2, the NFL approach demands at least two prototypes belonging to a specific class. Thus, we construct feature lines by exhaustively combining any pair of prototypes belonging to a speaker, which creates an NFL speaker model. For text-independent mode, in particular, a codeword would be viewed as a prototype if the VQ procedure operates on feature vectors prior to the NFL classification. Thus, the ultimate task of decision-making in the NFL is to find the NFL.

For text-dependent mode, our empirical studies indicate that the above NFL classification may not produce good performance for a text consisting of multiple phonemes. The possible explanation to the phenomenon is that the use of one prototype based on statistics may not represent the utterance of more than one phoneme well. Thus, we propose an utterance partitioning strategy for such an utterance; that is, we tend to partition an utterance into a number of acoustic segments consistent with the number of phonemes if it contains more than one phoneme. Moreover, each acoustic segment forms a prototype after acoustic analysis and, thus, such an utterance would be represented by a number of prototypes and feature lines are constructed on the basis of those prototypes. As well known, it is a non-trivial task to extract phonemes from speech, which is difficult and time-consuming (Huang et al., 2000; Rabiner and Juang, 1993). To avoid introducing a higher computational load, we do not use a precise phoneme extraction procedure. Instead we simply partition each utterance of multiple phonemes into several clips of equal length to form prototypes. To some extent, we expect that such an utterance partitioning strategy, which can be regarded as an approximate procedure of phoneme extraction, could be helpful to

facilitate the NFL classification for such utterances.

4. Simulation results

In this section, we report comparative results on the Ti46 and KING corpuses. The text-dependent results are first presented. It follows by text-independent results. For evaluating performance, we use two testing methods; that is, an unknown voice token is identified by either the one best testing procedure, where the identity is inferred by the NFL based on the top candidate, or the three best testing procedure, where the identity is determined in terms of top three candidates produced by the NFL. All the simulations are performed by means of ANSI C programs on a personal computer (Microsoft Window 98 platform and 800 MHz Pentium III).

4.1. Text-dependent results

In order to investigate the performance of NFL, we use three, four, and five utterances of a text, respectively, to construct the feature-line space. Those utterances of a single word/alphabet for training are randomly selected from the training sets in Ti46 to form three, four, and five prototypes. Multiple trials are performed in our simulations for reliability; that is, the same experiment is repeated three times and the averaging result is reported here. In the testing phase, all of ten utterances in the testing sets for each word/alphabet are used. As a result, the overall averaging recognition rates of one-best-test and three-best-test on Ti_alpha and Ti20 subsets are illustrated in Fig. 3(a) and (b), respectively. For comparison, we also apply DTW to the same problem and show its performance in the same figure. From Fig. 3(a) and (b), we observe that the performance of NFL

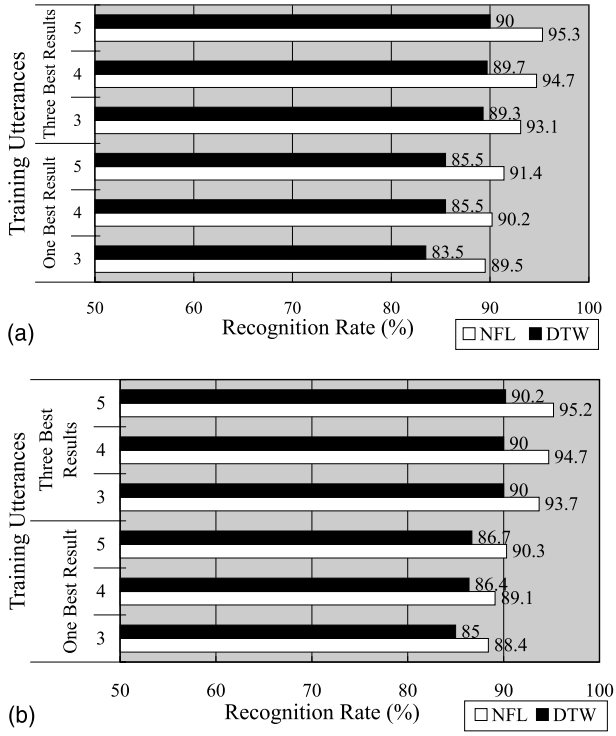

Fig. 3. The text-dependent recognition rates (%) of NFL and DTW in terms of different training utterances on the Ti46 database. (a) Results of the Ti_alpha set, (b) results of the Ti20 set.

Table 1
The CPU time (s) taken by the NFL with our improved algorithm (INFL), the original NFL, and the DTW, respectively, during decision-making on the Ti46 database in terms of different prototypes

No. of prototypes	Method		
	INFL	NFL	DTW
Three utterances	10	35	9929
Four utterances	12	45	12 639
Five utterances	16	53	20 237

is better than that of DTW in text-dependent speaker identification although the temporal information is not considered in the NFL approach. On the other hand, the computational cost of NFL is much less than that of DTW as shown in Table 1, where the total decision-making time taken by the NFL with our distance calculation algorithm and the DTW is reported on the Ti46 database. In terms of CPU time, in general, tens of seconds are only taken by the NFL to make decisions for all utterances in the testing set, while the DTW has to take around a few hours for the same task. For comparison, the time taken by the original NFL is also listed in Table 1. The computational cost of an NFL classifier regardless of the distance calculation algorithm becomes slightly higher as the number of prototypes increases, but it is still significantly lower than that of DTW. From Table 1, it is also evident that the time taken by our algorithm is roughly three times shorter than that of the original NFL, which is highly consistent with our formal analysis presented in Section 2.

For more details, we now report simulation results for each word/alphabet in the Ti46 database. In our simulations, the proposed partitioning strategy is applied prior to the NFL decision-making, and both one-best and three-best tests have been conducted. Figs. 4 and 5 illustrate recognition rates of all word/alphabet on the Ti_alpha and Ti20 sets, respectively, in terms of different segments and testing methods. In order to investigate effects of our partitioning strategy, we partition an utterance up to three segments of equal length without use of any prior knowledge on phonetics of alphabet/word.

From Fig. 4, it is observed that the recognition rates of some alphabets of more than one phoneme, e.g. "f", "m", and "s", are considerably lower than that of just one phoneme, e.g. "a", "e", and "i" without the use of our partitioning strategy (results corresponding to one segment in Fig. 4). When our two-segment partitioning strategy is applied, the recognition rates on those alphabets of two phonemes, e.g. "f", "m", and "s", have been considerably improved, while the recognition rates on those alphabets of only one phoneme, e.g. "a", "e", and "i", slightly decrease. For alphabet "f", in particular, more than 20% gain in the one-best test and a gain near 20% in the three-best test are achieved by our two-segment partitioning strategy, as illustrated in Fig. 4(a) and (b). Similarly, the improvement for words of two phonemes on the Ti20 set has been achieved by the two-segment partitioning strategy, as illustrated Fig. 5. In this circumstance, the recognition rates of these two-phoneme letters are considerably raised, which demonstrates the effectiveness of our utterance partitioning strategy. From Figs. 4 and 5, however, further simulations by partitioning an utterance into three segments of equal length indicate that over-segmentation causes the performance of NFL to be dramatically degraded for those alphabet/word of less than three phonemes, which suggests the necessity in the proper use of such an utterance partitioning strategy. Issues on our utterance partitioning strategy will be further discussed later on.

4.2. Text-independent results

As mentioned above, VQ provides a way to take advantage of speaker features regardless of verbal information. In VQ, decision-making is usually performed with the NN criterion (Linde et al., 1980). If we treat every codeword of VQ one prototype in the feature space, then NFL can be achieved through comparing distances between the query feature point and feature lines passing through any pair of codewords. Apparently, the NFL criterion leads to an alternative way for decision-making in VQ.

In text-independent simulations, we first use the NFL as a substitute of NN in VQ. We conduct two simulations with different training sets to investigate the performance of VQ based on the

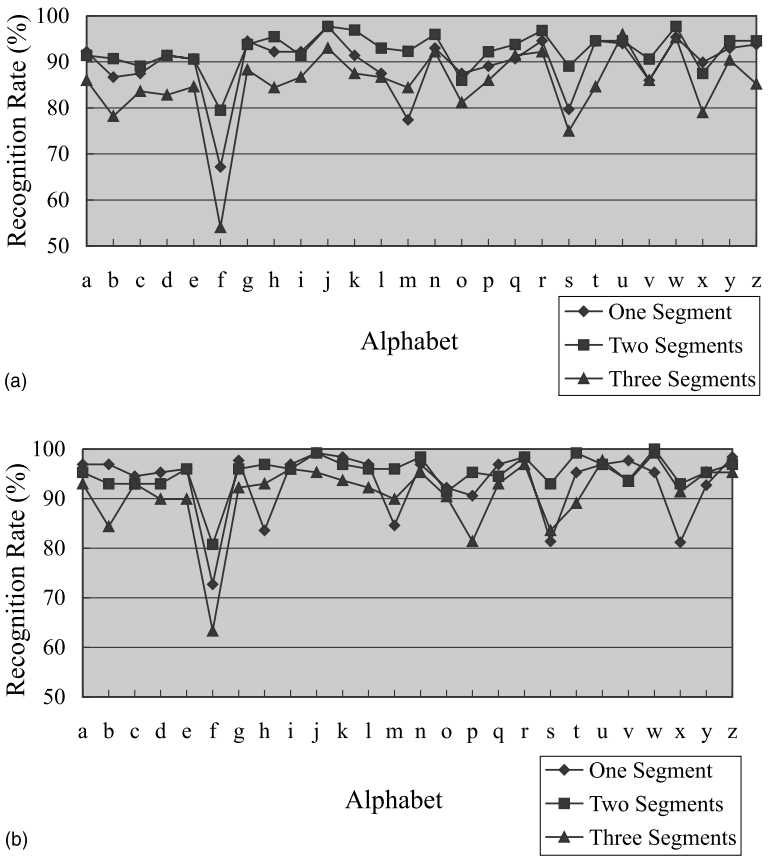

Fig. 4. The text-dependent recognition rates (%) on each single alphabet using NFL with three training utterances on the Ti_alpha set in terms of different segments, (a) results of one-best test, (b) results of three-best test.

NFL criterion. All training sets are chosen from the KING wide-band set. One is to use only S01 session for training and the remaining nine sessions for test. The other is to use three sessions (S01, S02, and S03) for training and the remaining seven sessions for test. The sophisticated VQ algorithm (Linde et al., 1980) is employed to obtain the codewords for each speaker. We denote VQ + NFL as VQ with the NFL criterion and VQ + NN as VQ with NN criterion in these two simulations. Fig. 6(a) and (b) show the comparative results in terms of different capacities of codewords.

The simulation results in Fig. 6 show that VQ + NFL is better than VQ + NN evidently. In particular, it is observed from Fig. 6(a) that it performs considerably better when fewer training samples, often resulting in a small capacity of codewords, are available. Here, we emphasize that as a new measure the NFL makes a VQ perform better against voice aging, as evident in Fig. 6(a) where the data recorded in a session are merely used. However, the computational cost of VQ + NFL is more expensive than that of VQ + NN.

The final simulation is to use NFL individually as a classifier to investigate its performance in the text-independent case. We also use three wide-band sessions (S01, S02, and S03) for training. Since there are about 30 s speech data in one session, we divide the speech into K segments. Each segment can be viewed as independent data to form a feature point or a prototype. Therefore, the

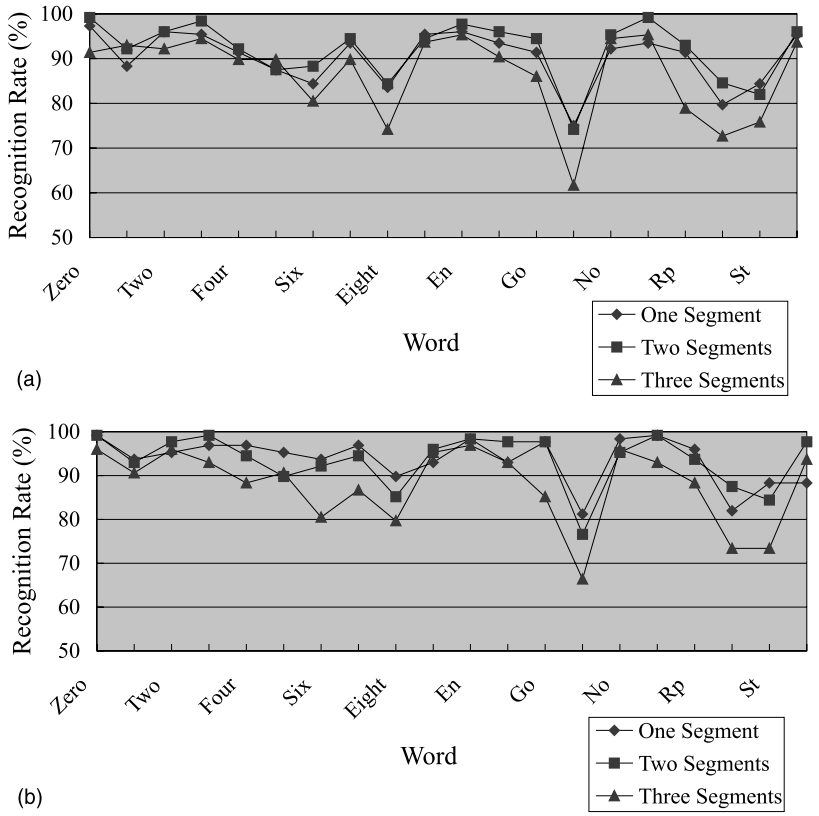

Fig. 5. The text-dependent recognition rates (%) on each single word using NFL with three training utterances on the Ti20 set in terms of different segments. (a) Results of one-best test, (b) results of three-best test.

total number of prototypes used for training is $3 \times K$. In the testing phase, all audio streams are partitioned into acoustic segments of the same length for training. Then every segment in test produces one recognition result. We use the majority voting result as the final result of this whole stream. Fig. 7 shows the accuracy of text-independent speaker identification by using the NFL method directly. In contrast, the performance of the NFL approach is unsatisfactory while the identification accuracy of VQ + NN reaches 83.5% in the same training set when the capacity of codewords is 16 (cf. Fig. 6(b)). Note that for comparison we only show the performance of the NFL on the condition that the number of prototypes resembles that of codewords in VQ. In other words, Fig. 7 depicts the recognition rates of the NFL as the number of prototypes is from 15 to 33 corresponding to $K = 5, \ldots, 11$, which covers the capacities of VQ, 16 and 32, used in our simulations. Further simulations by using less or more prototypes have been done as well though the results are not presented in Fig. 7. As a consequence, the performance of NFL in such circumstances is worse in general. Like codewords in VQ, too few number of segments is unlikely to model a speaker's characteristics since speaker's information represented by a statistical measure may be blurred, while too many number of segments results in a loss of speaker information since such a segment is likely too short to carry speaker's characteristics. In either of two circumstances, speaker's characteristics may not be modeled well due to an improper selection of prototypes.

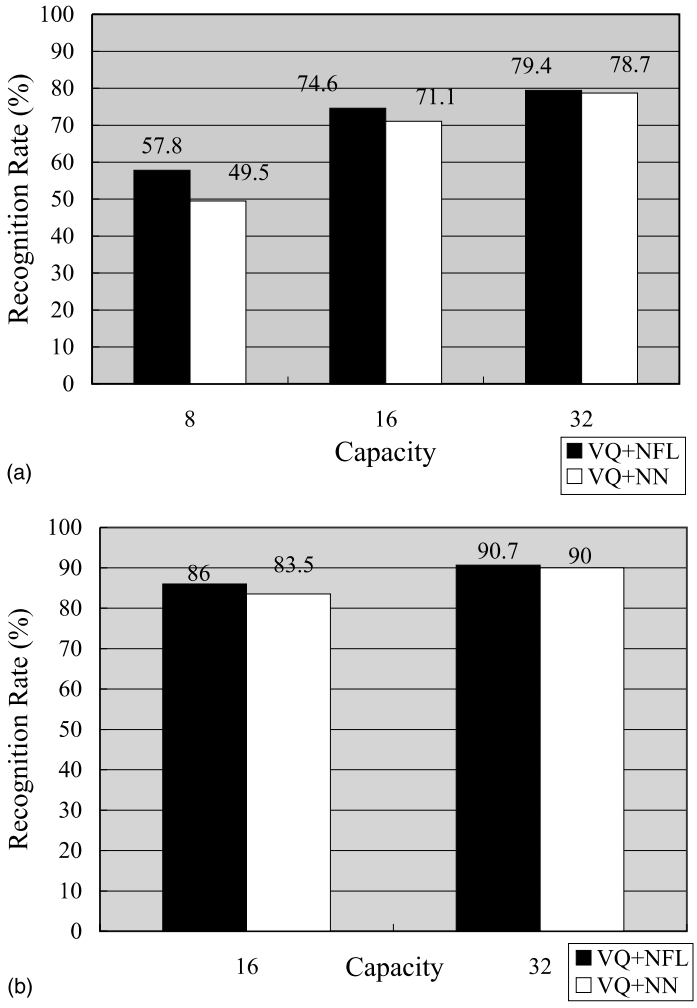

Fig. 6. The text-independent recognition rates (%) of VQ + NFL and VQ + NN on the KING database (wide-band set) in terms of different capacities of codewords. (a) Results by the use of only one session for training, (b) results by the use of three sessions for training.

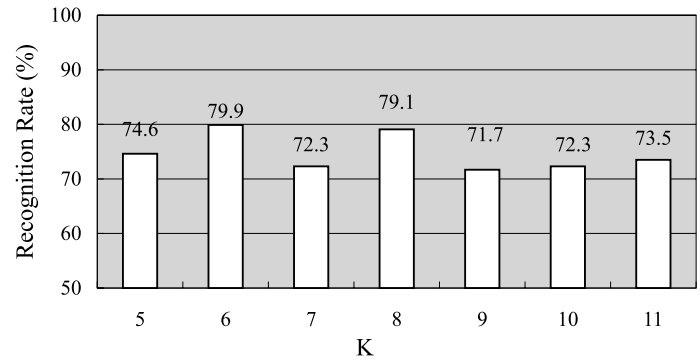

Fig. 7. The text-independent recognition rates (%) on the KING database (wide-band set) through the use of NFL individually.

5. Discussion

In this section, we discuss some issues on the NFL performance for different operating modes and its potentials.

Our studies have shown that the performance of NFL is consistently better than that of DTW regardless of testing methods. As well known, DTW takes a temporal alignment to match two sequences in an accurate way. Thus, it is dedicated to distinguish between two different sequences well, which very much facilitates isolated word speech recognition. Unfortunately, its salient feature causes DTW to be sensitive to intra-speaker variability, which violates an ingrained spirit, tolerance of intra-speaker variability, in speaker recognition. Thus, the performance of DTW is unsatisfactory in text-dependent speaker identification especially on the condition of severe mismatch. In contrast, there does not exist any temporal alignment in the NFL and a prototype encoding statistical information is robust against mismatch. As argued by Chen et al. (1996), a non-temporal alignment approach might be more effective in text-dependent speaker recognition since speaker's information may not uniformly distribute in each piece of an utterance. Here, the NFL provides another evidence to support this argument.

Our simulation results show that the NFL approach yields good performance in the text-dependent case without a strict temporal alignment but does not in the text-independent case. A possible reason is that those prototypes carry not only the speaker's characteristics but also the verbal information. In text-dependent case, both the characteristics of a speaker and verbal information are used simultaneously and, therefore, the performance of the NFL is satisfactory. On the contrary, in the text-independent case, we aim to emphasize the individual characteristics of speakers along with neglecting the verbal information. Unfortunately, the mismatch in verbal information may cause the NFL not to capture the speakers' characteristics well in this circumstance.

The motivation of our utterance partitioning strategy is to capture speaker's characteristics in terms of intrinsic acoustic units. As argued by Nolan (1983), the phonetic information plays a critical role in speaker recognition. Obviously, our partitioning strategy tends to facilitate encoding speaker's characteristics on the basis of a single phoneme. It is well known that the exact extraction of phonemes from voice is a challenging problem and such an algorithm is often time-consuming. From a computational viewpoint, our partitioning strategy is viewed as a preliminary simplification for extracting an acoustic unit, where an utterance of alphabet/word is simply partitioned into several segments of equal length. Apparently, this strategy does not guarantee that a resulting segment always corresponds to an acoustic unit. Thus, it becomes clear on why the three-segment partition strategy leads to worse performance for most of alphabets and words in the Ti46 database. As a result, our utterance partitioning strategy is simply an attempt towards better representing speakers' characteristics for a kind of classifiers without temporal alignment like the NFL, which points out a possible way to improve the performance of NFL in text-dependent speaker identification. No doubt, this topic is worth studying in the future.

Feature extraction is always a central issue in pattern recognition. In our method, a prototype is formed based on statistics of feature vectors corresponding to an utterance in the text-dependent case or a segment of the utterance resulting from a partitioning strategy. The equal-length partitioning strategy in our simulations is so simple that some segments may contain little speakers' information but tend to convey some verbal information, in particular, in the text-independent case. In the current method, unfortunately, all the segments are treated equal in producing prototypes. The use of NFL in this way may not make full use of speakers' information carried in voice. When severe mismatch conditions are involved, the NFL will not perform well as indicated in our text-independent results. To compensate mismatch in the verbal aspect, we introduce a VQ method to text-independent speaker identification, which results in the improved performance of NFL. Although how to effectively extract and represent speakers' characteristics for the NFL is still an open problem, it can be expected that the NFL could reach better performance in speaker identification if prototypes are properly created.

6. Conclusion

In this paper, we have applied the NFL classifiers to speaker identification where unlike other problems there are many sources of variability. To speed up the NFL, we propose an improved distance calculation algorithm. Both theoretical analysis and empirical evaluation show that the NFL along with our algorithm takes shorter time during decision-making (around three times shorter than that of the original NFL) without change of recognition performance. Our simulation results on benchmark databases show that the NFL performs well for the text-dependent mode, but fails to yield satisfactory results for the text-independent mode. For text-dependent mode, moreover, the proper use of our utterance partitioning strategy yields significant improvement for those two-phoneme alphabets and words. In contrast to DTW, the NFL approach, no matter whether our distance calculation algorithm is applied, leads to better performance and, in particular, takes much shorter time during decision-making. Our further simulation results indicate that with the help of the NFL, the performance of VQ can be improved for text-independent mode though an additional computational cost is introduced to decision-making. On the basis of our studies, we suggest that the NFL should be used in the text-dependent circumstance when only limited training data are available. As a future topic, how to extract proper speakers' features regardless of verbal information is worth to be studied, which would critically determine whether the NFL approach is applicable to text-independent speaker recognition.

Acknowledgements

An extended abstract of this paper was presented at the International Conference on Neural Information Processing, Shanghai, 2001. Authors are grateful to Stan Li for a discussion on his NFL concept and an anonymous reviewer for his/her comments that improve the presentation of this paper. This work was in part supported by an MSR grant on non-verbal speech analysis and an NSFC grant (60075017) to KC.

References

Assaleh, K.T., Mammone, R.J., 1994. New LP-derived features for speaker identification. IEEE Transactions on Speech and Audio Processing 2, 630–638.

Chen, K., Xie, D.H., Chi, H.S., 1996. A modified HME architecture for text-dependent speaker identification. IEEE Transactions on Neural Networks 7, 1309–1313.

Doddington, G.R., Przibocki, M.A., Martin, A.F., Reynolds, D.A., 2000. The NIST speaker recognition evaluation—Overviews, methodology, systems, results, perspective. Speech Communication 31, 225–254.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics Speech, and Signal, Processing 29, 254–272.

Huang, X.D., Acero, A., Hon, H., Meredith, 2000. Spoken Language Processing. Prentice Hall, New York.

Li, S.Z., 1998. Face recognition based on nearest linear combinations. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Santa Barbara, CA, pp. 839–844.

Li, S.Z., 2000. Content-based classification and retrieval of audio using the nearest feature line method. IEEE Transactions on Speech and Audio Processing 8, 619–625.

Li, S.Z., Chan, K.L., Wang, C., 2000. Performance evaluation of the nearest feature line method in image classification and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1335–1339.

Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantizer design. IEEE Transactions on Communication 28, 84–95.

Nolan, F., 1983. The Phonetic Bases of Speaker Recognition. Cambridge University Press, Cambridge.

Rabiner, L., Juang, B.H., 1993. Fundamental of Speech Recognition. Prentice Hall, New York.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 27, 43–49.

Soong, F.K., Rosenberg, A.E., 1988. On the use of instantaneous and transitional spectral information in speaker identification. IEEE Transactions on Acoustics, Speech, and Signal Processing 36, 871–879.

Soong, F.K., Rosenberg, A.E., Rabiner, L.R., Zhuang, B.H., 1985. A vector quantization approach to speaker identification. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Tampa, FL, pp. 387–390.