公 開

002

TR-IT-0077

A UNIFIED APPROACH TO PATTERN RECOGNITION

片桐 滋
Katagiri Shigeru

1994. 11

Pattern recognition is a complex process consisting of several subprocesses, such as feature extraction and classification. Naturally, design efforts should be devoted to the entire process of recognition in a manner consistent with task goal achievement, i.e., achievement of high recognition accuracy. Recent approaches based on Artificial Neural Networks(ANNs) have contributed to increasing recognition accuracy. However, they are still insufficient due to the lack of a global scope for the design of an overall recognition system. This paper reviews the present situation of the ANN-based approach from the methodological viewpoint of the Generalized Probabilistic Descent method and clarifies a number of technical issues to be further investigated.

This is an internal-use version of the paper for the 1994 International Symposium on Artificial Neural Networks, Dec 15--17, 1994.

A UNIFIED APPROACH TO PATTERN RECOGNITION

Shigeru KATAGIRI
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-02 Japan
phone: +81 7749 (5) 1052
fax: +81 7749 (5) 1008
email: katagiri@hip.atr.co.jp

Abstract

Pattern recognition is a complex process consisting of several subprocesses, such as feature extraction and classification. Naturally, design efforts should be devoted to the entire process of recognition in a manner consistent with task goal achievement, i.e., achievement of high recognition accuracy. Recent approaches based on Artificial Neural Networks (ANNs) have contributed to increasing recognition accuracy. However, they are still insufficient due to the lack of a global scope for the design of an overall recognition system. This paper reviews the present situation of the ANN-based approach from the methodological viewpoint of the Generalized Probabilistic Descent method and clarifies a number of technical issues to be further investigated.

1 Introduction

The Artificial Neural Network (ANN) is being established as an important basic system structure for pattern recognition. However, its superiority has not been fully demonstrated, as seen in its unfavorable competition with the Hidden Markov Model (HMM) in the speech recognition field. Has ANN already reached its performance limit? We think not, and we believe that the frustrating situation of ANN originates in the fact that previous ANN-based recongition attempts were too closely tied to the procedural frameworks of ANN. It is now necessary to shed this superficial understanding.

From the above viewpoint, we aim in this paper to offer a critical review to the current ANN approach to pattern recognition. Our discussion will be mainly based on a new design methodology, which is called the Generalized Probabilistic Descent method (GPD) and is recently becoming common in the speech pattern recognition area. We consequently argue that the most important future investigation issue is to find ANN structures which allow one to appropriately model the nature of given patterns as well as the overall recognition decision process.

For clarity of discussion, this paper focuses on speech pattern recognition tasks. The paper is organized as follows. Section 2 provides a basis of our discussion, i.e., a fundamental framework of the Discriminant Function Approach (DFA) to speech pattern recognition. Section 3 describes the GPD concept by introducing its exemplar implementation. Section 4 first surveys recent ANN approaches to speech pattern recognition and then considers issues to be investigated in the future. The paper is summarized in Section 5.

2 Discriminant Function Approach to Speech Pattern Recognition

2.1 Bayes decision theory

Speech (pattern) *recognition* is a process for mapping a *dynamic* (variable length) instantiation, which belongs to one of the prescribed M speech classes C_j $(j = 1, \cdots, M)$, to a class index. We specially consider a so-called statistical design approach in which one statistically trains the set of adjustable recognizer parameters Ψ (in other words, designs recognizers) over a preset design sample set, aiming to achieve the *optimal* decision status (the best in recognition accuracy for all of the possible future instantiations).

The most fundamental framework in this statistical approach is the Bayes decision theory, and its essence is summarized in the following Bayes decision rule.

Bayes decision rule: To minimize the overall risk $\mathcal{L}$, compute all of the (M) possible conditional risks $L(C(u_1^{T_o}) \mid u_1^{T_o})$, and select class C_j for which the corresponding conditional risk is minimum. Here, $u_1^{T_o}$ is a dynamic speech instantiation (waveform) of length T_o, and $C(\cdot)$ represents recognition operation; the conditional risk is given as

$$L(C(u_1^{T_o}) \mid u_1^{T_o}) = \sum_k \ell_k(C(u_1^{T_o}))1(u_1^{T_o} \in C_k)p(C_k \mid u_1^{T_o}) \tag{1}$$

by using a loss $\ell_k(C(u_1^{T_o}))$ which occurs when recognizing $u_1^{T_o}$, and then the overall risk becomes

$$\mathcal{L} = \int L(C(u_1^{T_o}) \mid u_1^{T_o})p(u_1^{T_o})du_1^{T_o} \tag{2}$$

(it is assumed that the integral in (2) exists); $1(\mathcal{A})$ is the indicator function as

$$1(\mathcal{A}) = \begin{cases} 1 & (\text{if } \mathcal{A} \text{ is true}) \\ 0 & (\text{otherwise}) \end{cases} \tag{3}$$

The statistical validity of using this rule is clear, and we thus define our ultimate design goal as finding the state of trainable recognizer parameters that can approximate the rule as accurately as possible.

2.2 Practical approaches

Fundamentally, speech recognizers can be as opaque as the hearing process of humans. If there is a system structure having a sufficiently large representation capability, and if a sufficiently large number of design samples are available, such an opaque recognizer would successfully realize a direct estimation of an *a posteriori* probability such as $p(C_k \mid u_1^{T_o})$. However, these assumptions are rarely true in a realistic situation where all of the available resources are limited. Developing a practical approach which emulates the full execution of the Bayes decision rule by using only limited resources has long been a pattern recognition research subject.

2.2.1 Transparent system structure

In most cases, realistic design attempts assume that systems are transparent. Accordingly, the problem of designing an overall large-scale recognizer has been replaced by the problem of designing practical size subprocesses. Figure 1 illustrates a typical speech recognizer. The recognizer consists of 1) the feature extractor (*feature extraction* subprocess) which converts an instantiation to some feature pattern, and 2) the classifier (*classification* subprocess) which maps this feature pattern to a class index. The classifier is here composed of a language model and an acoustic model. We define that Φ, Λ, and Υ denote the sets of adjustable parameters for the feature extractor, the acoustic model, and the language model, respectively ($\Psi = \Phi \cup \Lambda \cup \Upsilon$).

The feature extraction subprocess converts $u_1^{T_o}$ to a dynamic sequence consisting of the *static* (fixed-dimensional) F-dimensional acoustic feature vectors, $x_1^T = (\chi_1, \chi_2, \cdots, \chi_\tau, \cdots, \chi_T)$, where T is the length of this dynamic feature pattern and χ_τ is the τ-th static feature vector in the sequence. Scientific expertise about speech production and hearing is utilized to realize this subprocess. In particular, cepstrum and bank-of-filters have been widely used in recent works.

The classification subprocess assigns a class index to the converted feature pattern. This assignment is usually performed by using the rule as

$$C(x_1^T) = C_i, \quad \text{iff } i = \arg\max_j p(x_1^T \mid C_j)P(C_j), \tag{4}$$

which is a special case of the Bayes decision rule (the two-value $(1-0)$ error count loss is used as $\ell_k(C(x_1^T))$). The accurate execution of this rule will consequently lead to the minimum state of average

recognition error probability (the minimum recognition error situation). Note in (4) that, based on the Bayes rule of probability, a direct (but almost unachievable in reality) estimation of the *a posteriori* probabilities is replaced by the estimation of the conditional probabilities (density functions) and *a priori* probabilities that can easily be executed by using the well-formalized Maximum Likelihood (ML) method. Actually, in most speech recognizers, the conditional probabilities are estimated by using ML-based HMM modeling and the *a priori* probabilities are statistically computed by using probabilistic models such as the trigram.

2.2.2 Classifier design

Design interest has recently been directed to the classification subprocess, which is closely related to the recognition result. Studies on the feature extraction for recognition seem to be somewhat inactive.

The purpose of designing a classifier using (4) is to approximate the decision based on (4) as accurately as possible. Investigated approaches to this design are mainly divided into the Bayes approach and the DFA.

The Bayes approach aims to estimate the conditional probabilities and the *a priori* probabilities by using the ML principle (essentially the same as the Minimum Distortion principle); i.e., design in this approach attempts to estimate $p(x_1^T \mid C_j)$ as accurately as possible by using $p_\Lambda(x_1^T \mid C_j)$, which is a function of Λ, and to estimate $P(C_j)$ as accurately as possible by using $P_\Upsilon(C_j)$, which is a function of Υ. However, this Bayes approach is less satisfactory due to its generally low classification accuracy. This dissatisfaction comes from the following root cause: the functional form of class distribution (the conditional probability density) that primarily determines the estimation quality is intrinsically unknown and the maximization of the conditional probability (likelihood) that attempts to model the overall sample distribution of an individual class is not necessarily direct with regard to the minimization of classification errors (the accurate estimation of the class boundary).

Meanwhile, an alternative approach, the classical DFA, has been attracting much interest. This re-advent of interest in the classical approach was actually caused by ANN-based attempts of pattern classification. Most of the recent ANN classifiers can be categorized as a method of this approach. In the approach, a discriminant function $g_j(x_1^T;\Lambda)$ that measures the class membership of x_1^T (the degree to which x_1^T belongs) is introduced. The discriminant function does not need to be a probabilistic function; it can be various kinds of measures, such as distance and similarity. Moreover, especially concerning the acoustic model, the rule

$$C(x_1^T) = C_i, \quad \text{iff } i = \arg\max_j g_j(x_1^T;\Lambda) \tag{5}$$

is used in place of (4). The approach concerning the language model has been less studied, in reality. Usually, training is performed for classifier parameters (acoustic model parameters), aiming at reducing losses, each associated with a classification result.

2.3 Discriminant function approach

The DFA that evaluates the classification results in the design stage seems to be more direct with regard to the minimization of classification errors than the Bayes approach. Truly, the DFA aims to execute the overall Bayes decision rule accurately instead of estimating the individual probability functions in (4). Focusing on this directness to the decision, one may be more interested in the DFA.

There are four fundamental issues of the DFA: 1) the functional form of the discriminant function (classifier structure or measure), 2) the design (training) objective, 3) the optimization method, and 4) the consistency with unknown samples (training robustness or generalization). Many studies have been reported about these issues. Among them, it can be concluded that mainly based on the practicality and effectiveness of implementation, the most promising way is to design the state of Λ with simple and practical gradient optimization methods so as to directly reduce some loss properly reflecting a classification result.

Readers may refer to a number of excellent text books, such as [6], [7] and [23], to study the discussions in Section 2 in more detail.

3 The Generalized Probabilistic Descent Method

3.1 Formalization concept

The development of GPD was motivated by the considerations cited in Section 2.3. GPD was developed as a comprehensive solution to the problems in the DFA. A fundamental concept of the GPD formalization is to directly embed the overall process of recognizing a dynamic speech instantiation $u_1^{T_0}$ in a *smooth* (at least first differentiable in adjustable recognition parameters) functional form suitable for the use of a practical optimization method, especially a gradient-based optimization method, instead of attempting to accurately classify the static acoustic feature vectors of speech or the dynamic sequence of these vectors. It aims to thoroughly formalize the entire process in Figure 1, i.e., the feature extraction subprocess converting $u_1^{T_0}$ to x_1^T as well as the classification subprocess classifying x_1^T, in a manner consistent with the design goal of minimizing the final recognition errors.

GPD is an extended version of the classical, adaptive discriminant function design method called the Probabilistic Descent Method (PDM). A key point in its formalization is to overcome the unsmoothness problem of the original PDM formalization. To achieve this, GPD uses the L_p norm form and a sigmoidal function which has been widely used in ANN applications. Another key point is to provide a way to consistently optimize the overall recognition process consisting of the localized subprocesses. To achieve this, GPD also uses the chain rule of differential calculus. For example, in Figure 1, the acoustic model and the language model, which are independent of each other, are updated based on the partial differential in the corresponding adjustable parameters, respectively. The adjustable parameters of the feature extraction subprocess that is located under the acoustic model are updated based on the chain rule.

Since GPD originates in PDM, it adaptively updates the recognizer parameters. However, the formalization concept of GPD can apply to batch-type updating, such as the steepest descent method, without any loss of its mathematical rigor.

As HMM is widely used, the acoustic and language models are often represented by a probabilistic model form. Therefore, the updating rule must meet a probabilistic constraint, such as the condition that the sum of the state transition probabilities should be equal to one (1) at each state. GPD successfully provides the updating rule for such constrained situations [13].

3.2 Implementation example for recognition using a distance network

The GPD design concept shown above can be implemented for most reasonable recognizer structures. For clarity of presentation, we describe the GPD implementation by using a distance network recognizer which includes the Prototype-Based Minimum Error Classifier [18] [19]. This recognizer is fundamentally equivalent to a distance computation-based ANN, i.e., a distance network (or radial basis function network) incorporating the state transition-based dynamics. The input is a dynamic speech wave form $u_1^{T_0}$. Our design purpose is to accurately recognize such dynamic patterns appearing in the future.

Figure 2 illustrates the architecture of the recognizer which was used. The feature extraction subprocess is a function that converts a fixed duration speech segment to an F-dimensional acoustic feature vector. Among various possibilities of feature representations, we use the cepstrum computation-based conversion $f(\Phi)$, shifting a preset length time window over the dynamic input wave pattern and computing the cepstrum coefficients at each window position [3], where Φ is a set of parameters adjustable with GPD. The acoustic model in the classification subprocess is an S-state left-to-right phoneme model, in which each state has a set of F-dimensional reference vectors; $\Lambda = \{\{\mathbf{r}_{bs}^j\}_{b=1}^{B_s^j}, \{a_{s_\iota,s_\kappa}^j\}\}$, where s is the state index, B_s^j denotes the number of C_j's reference vectors at state s, and a_{s_ι,s_κ}^j denotes the transition probability of moving from s_ι to s_κ. The language model is a set of weight coefficients $\{W(\theta;\Upsilon)\}$, each determining the syntactic or semantic certainty of a concatenation of the phoneme models (equivalent to the state sequence θ). The figure suggests that this recognizer is similar, in strucuture, to an HMM recognizer widely used in speech recognition.

The main concept of GPD is to achieve a formalization consistent with the goal of a given task. Our formalization for recognizing dynamic speech patterns starts with the decision rule

$$C(u_1^{T_0}) = C_i, \quad \text{iff } i = \arg\max_j g_j(u_1^{T_0};\Psi) \tag{6}$$

which is similar to (5).

The feature extraction subprocess produces an F-dimensional acoustic feature vector χ_τ at the window position τ, shifting the window. This χ_τ is compared with each reference vector, $\mathbf{r}_{bs}^j$, and provides a kind of likelihood (possibility)

$$p_o(C_j, \chi_\tau, s) = \left\{ \sum_{b=1}^{B_s} e^{-d(\chi_\tau, \mathbf{r}_{bs}^j)\zeta} \right\}^{1/\zeta}, \tag{7}$$

where $d(\chi_\tau, \mathbf{r}_{bs}^j)$ represents a distance between the two vectors such as the Euclidian distance, and ζ is a positive constant. Clearly, $p_o(C_j, \chi_\tau, s)$ indicates the degree (possibility) to which the acoustic feature vector χ_τ belongs to state s of class C_j. The degree to which the input $u_1^{T_0}$ belongs to class C_j, corresponding to a phoneme symbol sequence such as word and sentence, is consequently represented by the discriminant function

$$g_j(u_1^{T_0}; \Psi) = \left[\sum_\theta \{P_\theta(P_o(C_j, T_0, S))\}^\xi \right]^{1/\xi}, \tag{8}$$

where $P_o(C_j, T_0, S)$ is a matrix of which the (τ, s) element is $p_o(C_j, \chi_\tau, s)$, $P_\theta(P_o(C_j, T_0, S))$ is an aggregate likelihood that is accumulated, conditioned by $\{W(\theta; \Upsilon)\}$, along the DTW search path θ defined over the matrix, and ξ is a positive constant.

Here, let ζ in (7) and ξ in (8) go to ∞, respectively. Clearly, (7) approximates the operation selecting the closest reference pattern and (8) approximates the operation searching for the best (maximum likelihood) path. It turns out that the L_p norm form successfully achieves a smooth function approximation of the unsmooth search operations in the decision process.

The next step of the formalization is to represent the decision rule by a smooth adjustable functional form. To achieve this, GPD defines a misclassification measure. Let us assume that a sample $u_1^{T_0} \in C_k$ is given for training. Among many possibilities, we consider

$$d_k(u_1^{T_0}; \Psi) = -g_k(u_1^{T_0}; \Psi) + \left[\frac{1}{M-1} \left\{ \sum_{j, j \neq k} g_j(u_1^{T_0}; \Psi) \right\}^\mu \right]^{1/\mu}, \tag{9}$$

where μ is a positive constant. Importantly, $d_k() > 0$ emulates misclassification and $d_k() < 0$ emulates correct classification. Similar to ζ and ξ, controlling μ provides various derivatives of the rule. In particular, as μ becomes closer to infinity, (9) comes to better resemble the classification rule (6).

Similar to the conventional methods of the DFA, training of the adjustable parameters is performed by minimizing the losses. The loss is defined as a smooth, monotonically increasing function of the misclassification measure as

$$\ell_k(u_1^{T_0}, \Psi) = l_k(d_k(u_1^{T_0}; \Psi)) = \frac{1}{1 + e^{-(\alpha d_k(u_1^{T_0}; \Psi) + \beta)}}, \quad (\alpha > 0), \tag{10}$$

where α and β are constants. Note that this sigmoidal loss is a smooth version of the most important, in the sense of recognition (classification), error count loss.

3.3 Probabilistic descent theorem

Ideally, the state of Ψ should be searched for by using the expected loss (overall risk) consisting of the above defined smooth losses; $L(\Psi) = \sum_k P(C_k) \int \ell_k(u_1^{T_0}; \Psi) p(u_1^{T_0} | C_k) du_1^{T_0}$. GPD uses the adaptive adjustment based on the following probabilistic descent theorem (originally developed for classification in [1])

Probabilistic Descent Theorem: Assume that a sample $u_1^{T_0}(t) \in C_k$ is given at the time index t of the design stage. If the adjustment amount of the recognizer parameters $\delta\Psi(u_1^{T_0}(t), C_k, \Psi(t))$ is set as

$$\delta\Psi(u_1^{T_0}(t), C_k, \Psi(t)) = -\epsilon \mathbf{U} \nabla \ell_k(u_1^{T_0}(t); \Psi(t)), \tag{11}$$

then

$$E[\delta L(\Psi)] \leq 0 \tag{12}$$

holds true, where $\mathbf{U}$ is a positive-definite matrix, ϵ is a small positive real number, and $\Psi(t)$ represents the state of the parameters Ψ at the time index t.

Here, in particular, one should note that the adjustment of the feature extraction subprocess parameters Υ is performed based on the losses back-propagated through the computations at the higher-layer acoustic model. The theorem is further described in [1] and [12].

Consequently, the full computation based on this theorem will lead to at least the state of Ψ that corresponds to the local minimum of the overall loss, in other words, the local minimum recognition error situation.

3.4 Design optimality

3.4.1 Practical finite length adaptive training

It is obviously impractical to strictly observe the infinitely-repeated, probabilistic descent adjustment. Therefore, in realistic situations, the learning coefficient $\epsilon(t)$ is often set to a finite-length, monotonically decreasing function, and consequently the GPD training performs only a finite repetition of adjustments. Also, this finite-length minimization search is applied to the empirical average loss $L_0(\Psi) = \frac{1}{N}\sum_k \sum_t \ell_k(u_1^{T\circ}(t); \Psi) 1(u_1^{T\circ}(t) \in C_k)$ instead of the expected loss $L(\Psi)$.

Given the reality that only finite samples and a finite repetition of the training adjustments are possible, the GPD optimality intrinsically requiring infinite adjustment repetition seems to be useless. However, the contribution of GPD is obvious. The widely used Bayes approach can never circumvent errors in estimating class distributions, and, in particular, these errors usually spread over the entire sample space. Reducing these spread errors is obviously inadequate: It is inconsistent with the recognition error reduction. In contrast with this, even a finite run of the GPD training reduces the smooth recognition error counts in a manner directly corresponding to the decision rule given to the task (e.g., the recognition rule (6)). Concerning this feature, GPD is clearly distinct from the conventional DFA methods as well as the Bayes approach methods.

3.4.2 Increase of robustness to unknown samples

The GPD formalization smooths the empirical average loss that is conventionally uncontinuous, and consequently achieves a design result having high robustness to unknown samples.

To discuss this point, we consider an illustrative task. The task is to classify static one-dimensional samples as one of two classes ($\circ$ and $\times$). Design samples are distributed on the horizontal axis of Figure 3. Shown in the figure are several curves of the empirical average loss consisting of (10), each computed over these samples with a different value of α. These loss curves were computed, assuming that a classifier having one reference pattern for each class was used. These curves alter continuously almost everywhere, though the error count must originally change only at the design sample positions. It is thus obvious that the smooth GPD design virtually performs learning (adjustment) at sample space places other than the design sample positions, or in other words, increases the robustness to unknown samples.

3.4.3 Global optimum search

Figure 3 also illustrates another interesting use of the smooth formalization. In the figure, the smaller α becomes (the gentler the sigmoid function of (10)), the smoother the loss curve; the larger α becomes (the sharper the sigmoid function of (10)), the more minute the loss curve. It seems that the smoothness successfully fills undesirable local minima of the loss curve and can thus be useful for accelerating the adjustment convergence and circumventing its locality. Therefore, GPD proposes a global optimum search method that lets one use the smoother setting in the beginning of the training stage and decrease the smoothness as the training proceeds. The training is first performed in a coarse but global-search mode, and gradually changes to a fine-tuning search. This is similar to the idea presented in [10].

4 Present Situations and Future Issues of the ANN-Based Approach

The descriptions in Section 3 clearly show that GPD provides reasonable and practical solutions to the three issues of the DFA, i.e., the design objective, the optimization method, and the consistency with unknown samples. So, how can one evaluate the recent ANN approach from this new viewpoint? We will explore the technical issues which require further investigation through such a GPD-based review.

4.1 Survey of the present situation

We first try the taxonomy-type investigation of pattern recognition techniques using ANN. Among the various possibilities, we arrived at the following manifolds:

1. Discriminative (discriminant function approach) ANN having a time delay structure [17][25][26]
2. A structural hybrid system of ANN and Dynamic Time Warping (DTW) achieved by dynamic programming or the Viterbi algorithm [4][15][20][24]
3. Discriminative HMM based on the ANN design concept [2][5][11][21]

Actually, these applications all contributed, to some degree, toward increasing recognition accuracy. However, the viewpoint of GPD obviously asserts that those contributions are insufficient.

First, as is obvious from the above manifolds, the usage of ANN was rather limited. The examples using DTW additionally, categorized in the second item above, clearly point out this problem. In those cases, ANN was first designed so that one could increase the classification accuracy of short segments, each being a part of the dynamic input feature pattern. DTW was then used to integrate the output from this ANN over the entire input. Obviously, the localized classification capability of such ANN does not directly correspond to the final recognition performance of the entire input pattern. Also, DTW is based on the minimum distortion principle, and this principle is not consistent with the minimization of recognition errors.

Moreover, the viewpoint of GPD clearly shows that these ANN applications suffer from a fundamental mathematical difficulty; That is, the heuristic algorithms, such as the error correction training and the adaptive learing version of the error back-propagation, are not guaranteed to be optimal in their convergence results, and also neither the squared error loss minimization nor the mutual information maximization is equivalent to the classification error minimization.

4.2 Future issues

Considering the above mentioned, present situation of ANN applications and the generality of GPD, one can note that a solution for improving these existing ANN approaches to speech pattern recognition is to adaptively optimize an overall recognition process that is formalized in a smooth functional form that includes ANN computation procedures, based on the probabilistic descent theorem. Actually, a recent example in [3] demonstrates a fundamental use of this solution. Therefore we think that, in the future, research should shift to the first issue of the DFA, i.e., the discriminant function form selection.

Previously, there have been many comparative studies posing "ANN vs HMM", "ANN vs DTW", or "HMM vs DTW". However, we think that these discussion frameworks are often dangerous enough to cause us to lose sight of the essence. The issues of the system structure (discriminant function selection) and the design objective have been carelessly jumbled up. One should remember that the truly important issue is the selection of the discriminant function form.

In this paper, we have used the distance network. However, considering that the commonly-used, Euclidean distance is a measure based on the Gaussian likelihood function, one may note that the distance network can be replaced by a likelihood network having the same state transition structure [14]. This likelihood network will basically use a kernel function, such as the Gaussian kernel, and then come to be closer, in structure, to a radial-basis function network with the state transition structure or a continuous HMM. The multi-layer perceptron is different in structure from that of these kernel function networks; but it is not so different in its practical representation capability. Through the above considerations, it

turns out that the structural difference between ANN and the conventional systems such as HMM is no longer significant.

In the beginning of the ANN research, there were special expectations for achieving highly-performing pattern recognizers by using opaque but powerful ANNs. However, to analyze theoretical potential is one thing [8][9], and to design practical sytems by using available finite resources is another. Taking this point into account together with the above cited close relation between ANN and HMM, we can consider it reckless to make up a discriminant function for such an opaque structure. Therefore, even in the case of using the ANN structures, one should probably use a transparent discriminant function that is based on the careful analysis of the nature of the patterns to be recognized. One should also have a global design scope covering the entire recognition process. In particular, for speech pattern recognition, it will be necessary to develop an ANN structure that can appropriately model the acoustical phenomenon of speech. It will also be necessary to develop an ANN framework for language modeling. As shown in [3] and [16], new investigations motivated by these points have actually been started. Thus, ANN research may proceed to a new advanced stage via the recurrence of orthodox analytic investigations of the pattern nature.

5 Summary

In this paper we have critically reviewed the recent ANN approach to pattern recognition from a novel, unified pattern recognizer design method called the Generalized Probabilistic Descent method. The paper has concluded that the most important future issue is to find an ANN structure which is sufficiently suited to model the overall recognition process as well as the nature of patterns.

References

[1] S. Amari; IEEE, Trans. Electronic Computers, Vol. EC-16, No. 3, pp. 299-307 (1967 6).

[2] L. Bahl, P. Brown, P. de Souza, and R. Mercer; IEEE, Proc. ICASSP88, Vol. 1, pp. 493-496 (1988 4).

[3] A. Biem, and S. Katagiri; IEEE, Proc. ICASSP93, Vol. 2, pp. 275-278 (1993 4).

[4] H. Bourlard, and C. Wellekens; IEEE, Trans. PAMI, Vol. 12, No. 12, pp. 1167-1178 (1990).

[5] J. Bridle; Speech Communication, Vol. 9, pp. 83-92 (1990).

[6] R. Duda, and P. Hart; "Pattern Classification and Scene Analysis", John Wiley and Sons (1973).

[7] K. Fukunaga; "Introduction to Statistical Pattern Recognition", Academic Press (1972).

[8] K. Funahashi; Neural Networks, Vol. 2, No. 3, pp. 183-191 (1989).

[9] E. Hartman, J. Keeler, and J. Kowalski; Neural Computation, Vol. 2, pp. 210-215 (1990).

[10] J. Hopfield, and D. Tank; Biological Cybernetics, Vol. 52, pp. 141-152 (1985).

[11] H. Iwamida, S. Katagiri, E. McDermott, and Y. Tohkura; ASJ, J. Acoust. Soc. Jpn. (E), Vol. 11, No. 5, pp. 277-286 (1990 9).

[12] B.-H. Juang, and S. Katagiri; IEEE, Trans. SP., Vol. 40, No. 12, pp. 3043-3054 (1992 12).

[13] S. Katagiri, C.-H. Lee, and B.-H. Juang; ASJ, Proc. Conf., pp. 141-142 (1990 9).

[14] S. Katagiri, C.-H. Lee, and B.-H. Juang; in "Neural Networks for Signal Processing", IEEE, pp. 11-20 (1991 9).

[15] S. Katagiri, and C.-H. Lee; IEEE, Trans. SAP, Vol. 1, No. 4, pp. 421-430 (1993 10).

[16] S. Katagiri, B.-H. Juang, and A. Biem; in "Artificial Neural Networks for Speech and Vision (ed. R. Mammone)", Chapman and Hall, pp. 278-293 (1994).

[17] E. McDermott, and S. Katagiri; IEEE, Trans. SP, Vol. 39, No. 6, pp. 1398-1411 (1991 6).

[18] E. McDermott, and S. Katagiri; IEEE, Proc. ICASSP92, Vol. 1, pp. 417-420 (1992 3).

[19] E. McDermott, and S. Katagiri; J. Applied Intelligence, Vol. 4, pp. 245-256 (1994).

[20] M. Miyatake, H. Sawai, Y. Minami,and K. Shikano; IEEE, Proc. ICASSP90, Vol. 1, pp. 449-452 (1990 4).

[21] L. Niles, and H. Silverman; IEEE, Proc. ICASSP90, Vol. 1, pp. 417-420 (1990 4).

[22] N. Nilsson; "The Mathematical Foundations of Learning Machines", Morgan Kaufmann Publishers (1990).

[23] L. Rabiner, and B.-H. Juang; "Fundamentals of Speech Recognition", Prentice Hall (1993).

[24] H. Sakoe, R. Isotani, K. Yoshida, K. Iso, and T. Watanabe; IEEE, Proc. ICASSP89, Vol. 1, pp. 29-32 (1989 5).

[25] K. Unnikrishnan, J. Hopfield, and D. Tank; IEEE, Trans. SP, Vol. 39, No. 3, pp. 698-713 (1991 3.).

[26] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang; IEEE, Proc. ICASSP88, Vol. 1, pp. 107-110 (1988 4).

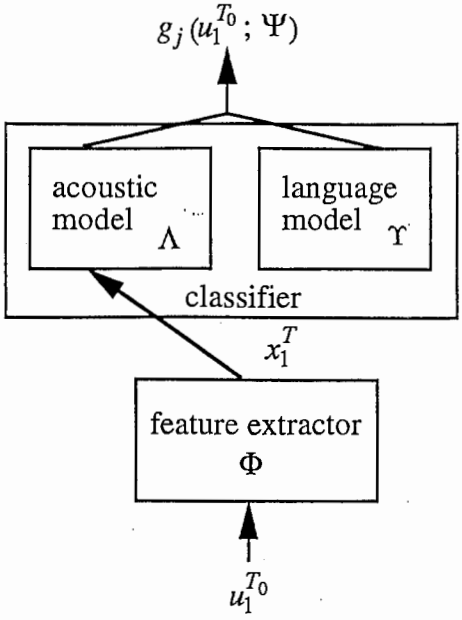

Figure 1 Typical structure of a speech recognizer.

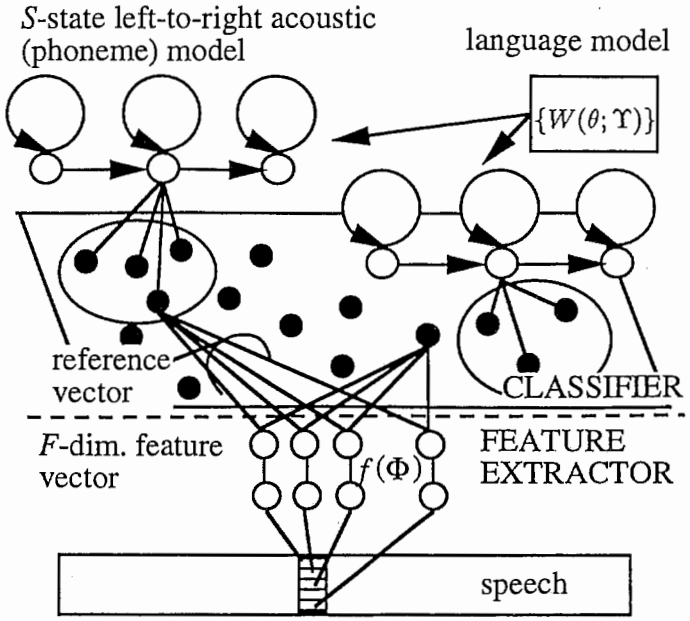

Figure 2 Speech recognizer structure based on the distance network.

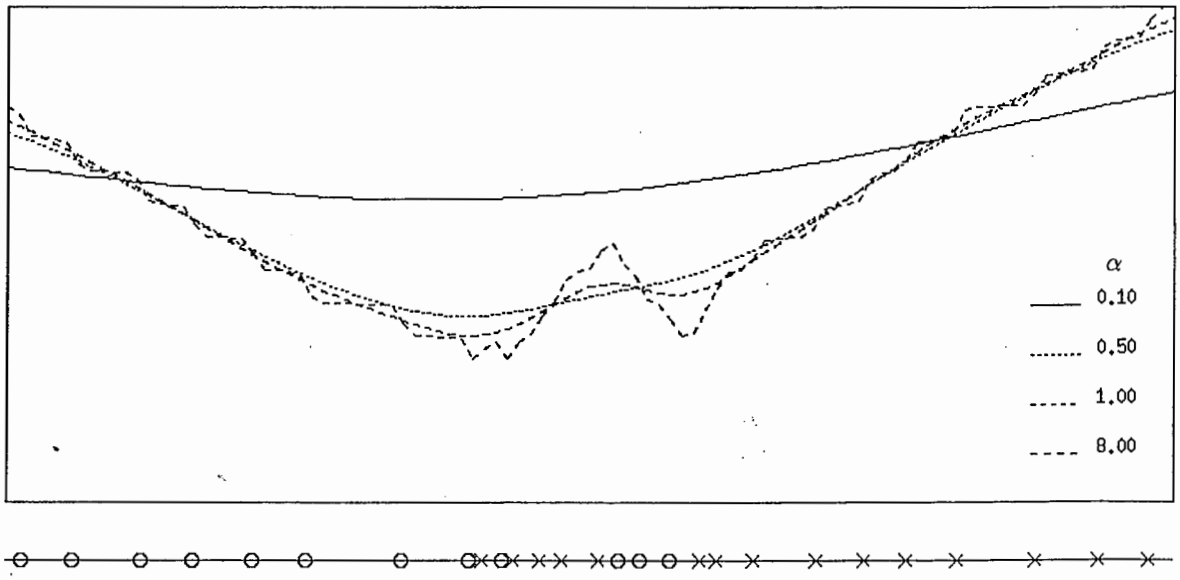

Figure 3 The relation in smoothness between the smooth classification error count loss and its corresponding empirical average loss.