HMM SPEECH RECOGNIZER BASED ON DISCRIMINATIVE METRIC DESIGN

Hideyuki WATANABE *Shigeru KATAGIRI*

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
E-mail: watanabe@itl.atr.co.jp

ABSTRACT

In this paper we apply *Discriminative Metric Design (DMD)*, the general methodology of discriminative class-feature design, to a speech recognizer using a Hidden Markov Model (HMM) classification. This implementation enables one to represent the salient feature of each acoustic unit that is essential for recognition decision, and accordingly enhances robustness against irrelevant pattern variations. We demonstrate its high utility by experiments of speaker-dependent Japanese word recognition using linear feature extractors and mixture Gaussian HMMs. Furthermore, we summarize several other recently-proposed design methods related to our DMD and show that they are special implementations of the DMD concept.

1. INTRODUCTION

A pattern recognizer generally consists of a feature representation of each pattern class and a measurement of membership to each class. Obviously, both modules should be jointly designed with the single objective of minimizing recognition errors for the optimality of the entire recognition process. One general framework to achieve this goal is *Discriminative Metric Design (DMD)* [1]. DMD is used to design an individual feature extractor of each class with the Minimum Classification Error/Generalized Probabilistic Descent method (MCE/GPD) [2] so as to minimize recognition error probability; it emphasizes each salient class feature for accurate recognition, and, accordingly, enhances design robustness against irrelevant pattern variations.

In this paper we specially elaborate upon a DMD formulation for a speech recognizer using HMMs. In this implementation, each state in each HMM has its own feature extractor; the feature of each acoustic unit that is essential for recognition can thus be represented efficiently. We demonstrate its utility by experiments of speaker-dependent Japanese word recognition in the case of linear feature extractors and mixture Gaussian HMMs.

DMD is essentially quite general and can be applied to various system structures, including neural networks. In fact, an HMM recognizer can be easily considered as an extended kernel-based network [3], and therefore the discussions in this paper will be clearly useful for increasing the applicability of neural-network speech recognizers.

In recent literature, several other design methods related to our DMD have been newly investigated [4, 5, 6, 7]. The paper also summarizes these methods from the general DMD viewpoint.

2. BASIC DMD FORMALIZATION

Consider the problem of classifying an input pattern $\boldsymbol{X}$ into one of K classes $\{\boldsymbol{C}_s\}_{s=1}^{K}$ using the following decision rule $\mathcal{C}(\boldsymbol{X})$:

$$\mathcal{C}(\boldsymbol{X}) = \boldsymbol{C}_i \quad \text{if} \quad i = \arg\max_s g_s(\mathcal{T}_s(\boldsymbol{X})), \qquad (1)$$

where $g_s(\boldsymbol{Y}_s)$, named the class-membership measure, indicates the degree to which $\boldsymbol{Y}_s$ belongs to $\boldsymbol{C}_s$, and $\mathcal{T}_s(\boldsymbol{X})$, named the class-feature extractor, represents the extraction of $\boldsymbol{C}_s$'s feature. DMD optimizes each class-specific metric, i.e., both $\mathcal{T}_s$ and g_s, with MCE/GPD so as to minimize the recognition error probability. Figure 1 illustrates the system structure and the training mechanism used in the DMD formalization. Recognizers with DMD therefore can perform robust recognition since each $\mathcal{T}_s$ can represent a salient feature of its corresponding class. In [1], DMD was proven useful through a particular implementation where it optimized the linear transformation $\mathcal{T}_s$ and Euclidean distance g_s in the case of fixed-dimensional $\boldsymbol{X}$.

DMD's most outstanding property is that each class discriminant function $g_s(\mathcal{T}_s(\boldsymbol{X}))$ has its own feature extractor $\mathcal{T}_s(\boldsymbol{X})$. If all class-feature extractors are identical ($\mathcal{T}_1 = \mathcal{T}_2 = \cdots = \mathcal{T}_K = \mathcal{T}$),

DMD becomes equivalent to Discriminative Feature Extraction (DFE) [8], and, furthermore, if $\mathcal{T}$ is fixed during training, DMD becomes equivalent to conventional MCE/GPD classifier design. A clear perspective on discriminative feature design including the above relationships is given in [9].

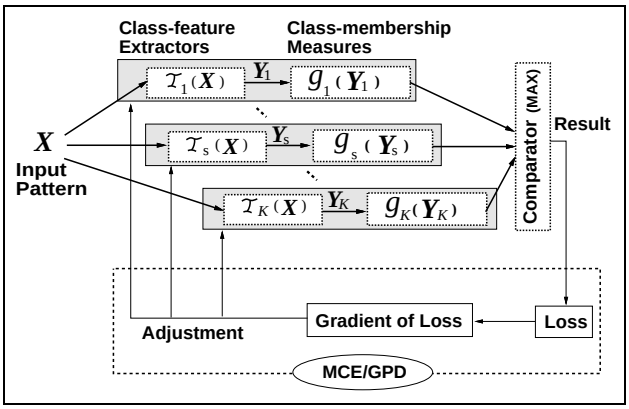

Figure 1. DMD-based pattern recognizer structure and its training mechanism

3. APPLICATION OF DMD TO HMM FRAMEWORK

3.1. Design formulation for HMMs

We apply the DMD concept to the recognition of variable-dimensional (dynamic) patterns, such as speech signals. Such dynamic patterns must be characterized using nonlinear time-warping models such as HMMs. Furthermore, it is desired to model long-durational pattern classes (e.g., words or sentences) by concatenating shorter unit models (e.g., phoneme or syllable models). Here we consider a word recognition task and assume that a word model consists of several phoneme HMMs. For this case, a formulation of DMD training is given as follows.

The discriminant function of $\boldsymbol{C}_s$ (the s-th word class) is a log-likelihood score along the optimal HMM/state path, defined as

$$g_s(\mathcal{T}_s(\boldsymbol{X})) = \sum_{t=1}^{T} \ln\left\{a_{\varphi_{s,t}\theta_{s,t-1}\theta_{s,t}}\right\} + \sum_{t=1}^{T} \ln\left\{b_{\varphi_{s,t}\theta_{s,t}}\left(\mathcal{T}_{\varphi_{s,t}\theta_{s,t}}(\boldsymbol{x}_t)\right)\right\}, \quad (2)$$

where $\boldsymbol{X} = [\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ ...\ \boldsymbol{x}_T\ ] \in \mathcal{R}^{D\times T}$ is an input sequence of D-dimensional vectors,

$$\boldsymbol{\Theta}_s = \{(\varphi_{s,1},\theta_{s,1});(\varphi_{s,2},\theta_{s,2});...;(\varphi_{s,T},\theta_{s,T})\} \quad (3)$$

denotes $\boldsymbol{C}_s$'s optimal HMM/state sequence ($\varphi_{s,t}$: phoneme-HMM index, $\theta_{s,t}$: state index), $a_{h,i,j}$ is the state transition probability of the h-th phoneme HMM, $b_{h,j}(\boldsymbol{y}_{h,j})$ is the output p.d.f. in the j-th state of the h-th phoneme HMM, and $\mathcal{T}_{h,j}(\cdot)$ stands for the feature extractor in the j-th state of the h-th phoneme HMM. In this implementation, $\boldsymbol{C}_s$'s feature extractor $\mathcal{T}_s(\cdot)$ is regarded as the concatenation of each $\mathcal{T}_{h,j}(\cdot)$ along with the optimal path $\boldsymbol{\Theta}_s$:

$$\begin{aligned}\mathcal{T}_s(\boldsymbol{X}) &= \mathcal{T}_s([\ \boldsymbol{x}_1\ ...\ \boldsymbol{x}_T\ ]) \\ &= \left[\mathcal{T}_{\varphi_{s,1}\theta_{s,1}}(\boldsymbol{x}_1)...\mathcal{T}_{\varphi_{s,T}\theta_{s,T}}(\boldsymbol{x}_T)\right]. \quad (4)\end{aligned}$$

The DMD training is done by jointly optimizing both the parameter sets of the models $\{a_{h,i,j},b_{h,j}(\cdot)\}$ ($\boldsymbol{\Lambda}_{\mathcal{G}}$) and the feature extractors $\{\mathcal{T}_{h,j}(\cdot)\}$ ($\boldsymbol{\Lambda}_{\mathcal{T}}$) with MCE/GPD. MCE/GPD generally uses adaptive (sequential) training, whose parameter-updating formulae are given as

$$\boldsymbol{\Lambda}_{\mathcal{G}}^{(n+1)} = \boldsymbol{\Lambda}_{\mathcal{G}}^{(n)} - \varepsilon_n \boldsymbol{U}_{\mathcal{G}} \nabla_{\Lambda_{\mathcal{G}}} \ell(\boldsymbol{X}_n;\boldsymbol{\Lambda}^{(n)}) \quad (5)$$

$$\boldsymbol{\Lambda}_{\mathcal{T}}^{(n+1)} = \boldsymbol{\Lambda}_{\mathcal{T}}^{(n)} - \varepsilon_n \boldsymbol{U}_{\mathcal{T}} \nabla_{\Lambda_{\mathcal{T}}} \ell(\boldsymbol{X}_n;\boldsymbol{\Lambda}^{(n)}) \quad (6)$$

$$\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_{\mathcal{G}}, \boldsymbol{\Lambda}_{\mathcal{T}}),$$

where the superscript $^{(n)}$ stands for the parameter state at the n-th iteration step, $\boldsymbol{X}_n$ is a training pattern picked up randomly at the n-th step, ε_n (> 0) is a learning step size (vanishing with n), $\boldsymbol{U}_{\mathcal{G}}$ and $\boldsymbol{U}_{\mathcal{T}}$ are positive-definite matrices, ∇_λ represents the partial derivative in a parameter $\boldsymbol{\lambda}$, and $\ell(\cdot)$ is the smooth error-counting loss function. Formalizations of the loss function and its derivative are the very same as those of usual MCE/GPD training implementations; e.g., see [10]. Note that, in the same sense as the original MCE/GPD, the training run based on (5) and (6) leads to the optimal (in the sense of minimum recognition error probability) status of $\boldsymbol{\Lambda}$, or in other words, the optimal status of all the metrics, in a probabilistic descent sense. Obviously, this resulting training rule is not restricted to the case where the unit models correspond to phonemes.

If we fix all $\{\mathcal{T}_{h,j}(\cdot)\}$ with identity mapping, the DMD training of HMMs becomes equivalent to conventional Segmental GPD [11]. In Segmental GPD, we have to assign in each state many reference patterns (e.g., multiple Gaussian means) in order to model statistical pattern variations. On the other hand, in DMD, each state's feature extractor $\mathcal{T}_{h,j}(\cdot)$ suppresses such variation factors, and, accordingly, fewer reference patterns are needed; robustness to unknown patterns will increase. For example, in the case of recognizing

fixed-dimensional patterns using distance classifiers, it has been verified that class-feature design contributes more toward increasing robustness than does the assignment of many reference patterns [1].

3.2. Experimental Results

[1] demonstrated the fundamental feasibility of DMD in a task of recognizing static vowel fragment patterns. To study the method in a more realistic environment, we conducted in this paper experiments of speaker-dependent Japanese word recognition using the ATR 5240-word database. Training and testing were performed on the even numbered words and the odd numbered words, respectively.

Each token input to the recognizer was a sequence of 34-dimensional vectors, each consisting of 16 LPC cepstral coefficients and their delta parameters, a power and a delta power. Computation was done using a 20-ms Hamming window with a 5-ms shift.

Acoustic unit models were 26 context-independent left-to-right phoneme HMMs. We specially defined each $b_{h,j}(\cdot)$ as a mixture Gaussian p.d.f., and each $\mathcal{T}_{h,j}(\cdot)$ as a linear transformation matrix $\boldsymbol{W}_{h,j}$. Here the adjustable parameters were mixture weights, mean vectors and covariance matrices in all $\{b_{h,j}(\cdot)\}$ and feature transformation matrices $\{\boldsymbol{W}_{h,j}\}$. All covariance matrices were diagonal, all (non-zero) state transition probabilities $\{a_{h,i,j}\}$ were identical and fixed, and all $\{\boldsymbol{W}_{h,j}\}$ were square matrices (i.e., without dimensionality reduction).

For MCE/GPD training, all parameters of $\{b_{h,j}(\cdot)\}$ were initialized by performing the phoneme-level Segmental K-means training, and all $\{\boldsymbol{W}_{h,j}\}$ were initialized at the identity matrix. In DMD the parameters of both $\{b_{h,j}(\cdot)\}$ and $\{\boldsymbol{W}_{h,j}\}$ were trained by MCE/GPD, whereas in Segmental GPD only the parameters of $\{b_{h,j}(\cdot)\}$ were adjusted.

Table 1 summarizes the recognition rates in terms of phoneme accuracy for one male speaker. In the table, (N, M) denotes the N-state M-mixture HMMs. Both of the DMD-based recognizers with a single mixture and four mixtures achieved the best score for testing sets. Moreover, interestingly, the DMD-based recognizer with a single mixture performed better than did the conventional Segmental GPD-based one with four mixtures, especially over the testing set. This result clearly demonstrates that the DMD effectively designs each salient class feature for accurate recognition, makes the classification models simpler, and consequently enables high-accuracy recognition for unknown future patterns.

Table 1. Recognition rates (phoneme accuracy)

	training	testing
Segmental K-means (2,4)	74.05%	73.21%
Segmental GPD (2,4)	94.10%	90.73%
DMD (2,4)	98.16%	93.95%
DMD (2,1)	97.28%	93.55%

4. RELATED WORKS

In this section we discuss several recent methods related to DMD from the unified viewpoint of general DMD; this convincingly illustrates the DMD's utility in a wide range of applications and will provide us with a clear perspective on feature design.

[4] proposes a design method which finds the optimal feature transformation matrices for Mel-warped short-time spectral input vectors in the case of mixture Gaussian HMMs. It finds model- and state-dependent transformation matrices by MCE/GPD with Discrete Cosine Transformation (DCT), which produces Mel-warped cepstral vectors, as the initial transformation. Since the resulting feature transformation matrices are class-dependent, this design method should be regarded as an implementation of DMD, although in [4] it is placed in DFE. In the experiments, it is demonstrated that the class-specific feature design with MCE/GPD clearly enhances the recognition accuracy rather than designing only the model parameters on the fixed feature transformation (DCT).

[5] also proposes a design framework for finding optimal feature transformation matrices with MCE/GPD, although the input vector sequence differs from that in [4]. Both class-independent and class-dependent training procedures are formulated: the former corresponds to DFE while the latter is an implementation of DMD. In [5], this design method is applied to HMM-based speaker recognition, and the experiments verify the performance improvement of discriminative feature design in both cases of speaker identification and speaker verification. Furthermore, it is interesting to see that the class-dependent feature design (which corresponds to DMD) yields a lower error rate than the class-independent design (which corresponds to DFE).

[6] applies artificial neural networks (ANNs) to nonlinear feature extractors in an HMM-based speech recognition system and adjusts both of the

ANN and HMM parameters by MCE/GPD with a single objective of minimizing the recognition error rate. It also proposes two types of systems: one using a single ANN over all models, which corresponds to DFE, and the other using an individual ANN for each model, which is a DMD implementation. The utility of MCE/GPD-based feature design is demonstrated by experiments in a speaker-independent telephone-based connected digits recognition task. Furthermore, as in [5], the model-specific ANN design, which corresponds to DMD, achieves lower error rates than does the single ANN design, which corresponds to DFE.

[7] investigates a special acoustic model named Frequency-Weighted HMM where the input pattern is a sequence of group-delay spectral vectors [7] and the HMM- and state-dependent frequency-weighting diagonal matrices are optimized by MCE/GPD. This can be considered as a special case of DMD where the class-feature extractors are limited to diagonal matrices. The considerable effect of optimizing weighting matrices is demonstrated in noisy speech recognition experiments. However, in order to explore the salient feature space from the original input pattern space, more general matrices containing rotation factors, rather than diagonal matrices, are desired [1].

5. CONCLUSION

This paper applied Discriminative Metric Design (DMD), the general concept of the discriminative design of class features, to a speech recognizer with connected HMMs. Its high utility was demonstrated by experimental results from a task of speaker-dependent Japanese word recognition using linear feature extractors and mixture Gaussian HMMs. Furthermore, we summarized several other recently-proposed design methods related to our DMD and showed that they are special implementations of the DMD concept.

Since our DMD is general, one can easily extend the special implementations elaborated in the above experiments to more general ones with nonlinear feature extractors [6] or other types of dynamic pattern models in addition to HMM, such as multi-state distance metric [12].

Acknowledgement: The authors would like to express their gratitude to Mr. Hiroaki Tagawa for his valuable contributions to the experimental software development and implementation of the experiments.

REFERENCES

[1] H. Watanabe, T. Yamaguchi and S. Katagiri, "A novel approach to pattern recognition based on discriminative metric design," in *Proc. 1995 IEEE Workshop on Neural Networks for Signal Processing,* pp. 48-57, Sept. 1995.

[2] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing,* vol. 40, No. 12, pp. 3043-3054, Dec. 1992.

[3] S.-Y. Kung, *Digital Neural Networks*, Prentice Hall, 1993.

[4] C. Rathinavelu and L. Deng, "HMM-based speech recognition using state-dependent, linear transformations on Mel-warped DFT features," *Proc. ICASSP 96,* vol. 1, pp. 9-12, 1996.

[5] C.-S. Liu, "A general framework of feature extraction: application to speaker recognition," *Proc. ICASSP 96,* vol. 2, pp. 669-672, 1996.

[6] M. G. Rahim and C.-H. Lee, "Simultaneous ANN feature and HMM recognizer design using string-based minimum classification error (MCE) training," *Proc. ICSLP 96,* vol. 3, pp. 1824-1827, 1996.

[7] M. Ono, O. Uchisasai and H. Matsumoto, "Minimum error classification training of frequency-weighted HMMs," *Proc. Spring Meet. Acoust. Soc. Jpn.*, 3-5-10, pp. 127-128, Mar., 1996 (in Japanese).

[8] A. Biem, S. Katagiri and B.-H. Juang, "Discriminative feature extraction for speech recognition," in *Proc. 1993 IEEE Workshop on Neural Networks for Signal Processing,* pp. 392-401, Sept. 1993.

[9] H. Watanabe, A. Biem and S. Katagiri, "Toward a unified design of pattern recognizers," in *Proc. 1996 IEEE Workshop on Neural Networks for Signal Processing,* pp. 283-292, Sept. 1996.

[10] D. Rainton and S. Sagayama, "Minimum Error Classification Training of HMMs — Implementation Details and Experimental Results," *J. Acoust. Soc. Jpn. (E)*, vol. 13, No. 6, pp. 379-387, Nov., 1992.

[11] W. Chou, B.-H. Juang and C.-H. Lee, "Segmental GPD training of HMM based speech recognizer," *Proc. ICASSP 92,* vol. 1, pp. 473-476, 1992.

[12] E. McDermott and S. Katagiri, "Prototype-based minimum classification error/generalized probabilistic descent training for various speech units," *Computer Speech and Language*, vol. 8, pp. 351-368, 1994.