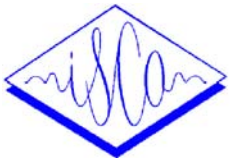


ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Speaker Identification with Dual Penalized Logistic Regression Machine

Tomoko Matsui and Kunio Tanabe

The Institute of Statistical Mathematics
Tokyo, Japan
{tmatsui, tanabe}@ism.ac.jp

Abstract

This paper proposes a novel speaker identification method based on the dual Penalized Logistic Regression Machine (dPLRM) for general multi-class discrimination. The machine employs kernel functions which implicitly map an acoustic feature space to a higher dimensional space. Each speaker is discriminatively identified in this space implicitly. The penalized logistic regression model used in dPLRM provides a reliable estimate of probability of each identification decision. Text-independent speech data recorded by 10 male speakers in four sessions over nine months was used to evaluate the new approach. The proposed method effectively reduced the error rate of the conventional GMM-based approach.

1. Introduction

Speaker recognition techniques are widely applied not only to secure access controls of information service systems but also to such problems as the speaker detection problem in speech dialogue and speaker indexing problem with large audio archives. The demand has been increasing for techniques with higher-accuracy.

The conventional method for text-independent speaker recognition is based on the Gaussian mixture model (GMM) [1]. In this approach, utterance variation is well captured by a mixture of a well-chosen number of Gaussian distributions. However, the estimation of the mixture distribution tend to be unreliable especially when the number of training data is relatively small. In such a case additional fine-tuning process of the estimated mixture is needed for gaining higher discriminative power [2].

The support vector machines (SVMs) [3] have been successfully applied to various pattern recognition problems: handwritten digit recognition, face detection, text categorization and speaker recognition [4-8]. The SVMs are basically designed for two-class discrimination, and some special techniques are necessary for multi-class discrimination.

The dual Penalized Logistic Regression Machine (dPLRM), which was employed in this paper for the speaker identification problem, was proposed by Tanabe [9,10] based on penalized logistic regression model with a specific penalty term for bringing about induction-generalization capacity of the machine. Maximization of the dual penalized logistic regression likelihood leads intrincically to the kernel regressors as is the case with SVMs which employ a quadratic programming model. One of the most notable advantage of the machine over the other methods is that it can give probabilistic predictions.

In this paper, we show a new approach to the speaker identification problem and demonstrate the power of dPLRM. In the following section, we briefly sketch the machine dPLRM. In section 3, we introduce our speaker identification procedure. In section 4, our proposed method is evaluated in text-independent speaker identification experiments. In section 5, we discuss the choices of kernel functions and assess the estimate of the probability on each identification decision.

2. Dual penalized logistic regression machine

Let $\mathbf{x}_j$ is a column vector of size n and c_j takes a value in the finite set $\{1,2,\ldots,K\}$ of classes. The learning machine dPLRM feeds a finite number of training data $\{(\mathbf{x}_j, c_j)\}_{j=1,\ldots,N}$, and then produces a conditional multinomial distribution $\mathrm{M}(\mathbf{p}^*(\mathbf{x}))$ of c given $\mathbf{x} \in \mathbf{R}^n$, where $\mathbf{p}^*(\mathbf{x})$ is a predictive probability vector whose k-th element $p_k{}^*(\mathbf{x})$ indicates the probability of c taking the value k.

For mathematical convenience, we code the class data c_j by j-th unit column vector $\mathbf{e}_k \equiv (0,\ldots,1,\ldots 0)^t$ of size K and define an $K \times N$ constant matrix $\mathbf{Y}$ by

$$\mathbf{Y} \equiv [\mathbf{y}_1;\cdots;\mathbf{y}_N] \equiv [\mathbf{e}_{c_1};\cdots;\mathbf{e}_{c_N}] \qquad (1)$$

whose j-th column vector $\mathbf{y}_j \equiv \mathbf{e}_{c_j}$ indicates the class to which the data $\mathbf{x}_j$ is attached. While SVM determines a single valued dichotomous discriminative function

$$\mathbf{f}(\mathbf{x}) \equiv \mathbf{v}\mathbf{k}(\mathbf{x}) \qquad (2)$$

where $\mathbf{v}$ is a row vector of size N, we introduce a multi-valued function

$$\mathbf{f}(\mathbf{x}) \equiv \mathbf{V}\mathbf{k}(\mathbf{x}) \qquad (3)$$

mapping $\mathbf{R}^N$ into $\mathbf{R}^K$, where $\mathbf{V}$ is an $K \times N$ parameter matrix which is to be estimated by using the training data set $\{(\mathbf{x}_j, c_j)\}_{j=1,\ldots,N}$, $\mathbf{k}(\mathbf{x})$ is a map from $\mathbf{R}^n$ into $\mathbf{R}^N$ defined by

$$\mathbf{k}(\mathbf{x}) \equiv (\mathbf{K}(\mathbf{x}_1, \mathbf{x}), \ldots, \mathbf{K}(\mathbf{x}_N, \mathbf{x}))^t \ , \tag{4}$$

and $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ is a certain positive definite kernel function. Then we define a model for multinomial probabilistic predictor $\mathbf{p}(\mathbf{x})$ by

$$\mathbf{p}(\mathbf{x}) \equiv \hat{\mathbf{p}}(\mathbf{f}(\mathbf{x})) \equiv (\hat{p}_1(\mathbf{f}(\mathbf{x})), \ldots, \hat{p}_K(\mathbf{f}(\mathbf{x})))^t , \tag{5}$$

where $\hat{p}_k(\mathbf{f}(x)) \equiv \dfrac{\exp(\mathbf{f}_k(\mathbf{x}))}{\sum_{i=1}^{K} \exp(\mathbf{f}_i(\mathbf{x}))}$ is the logistic transform.

Under this model assumption, the negative-log-likelihood function $L(\mathbf{V})$ for $\mathbf{p}(\mathbf{x})$ is given by

$$L(\mathbf{V}) \equiv -\sum_{j=1}^{N} \log(p_{c_j}(\mathbf{x}_j)) = -\sum_{j=1}^{N} \log(\hat{p}_{c_j}(\mathbf{V}\mathbf{k}(\mathbf{x}_j))) \tag{6}$$

which is a convex function (see [9,10]). This objective function $L(\mathbf{V})$ is of discriminative nature, and that if the kernel function is appropriately chosen, the map $\mathbf{f}(\mathbf{x})$ can represent a wide variety of functions so that the resulting predictive probability $\mathbf{p}(\mathbf{x})$ can be expected to be close to the reality. A predictive vector $\mathbf{p}^*(\mathbf{x})$ could be obtained by putting $\mathbf{p}^*(\mathbf{x}) = \hat{\mathbf{p}}(\mathbf{V}^{**}\mathbf{k}(\mathbf{x}))$ where $\mathbf{V}^{**}$ is the the maximum likelihood estimate which minimize the function $L(\mathbf{V})$ with respect to $\mathbf{V}$.

However, over-learning problems could occur with $\mathbf{V}^{**}$ with the limited number of training data. In order to obtain an induction power of dPLRM, the penalty term is introduced and the negative-log-penalized-likelihood

$$PL(\mathbf{V}) \equiv L(\mathbf{V}) + \frac{\delta}{2} \left\| \mathbf{\Gamma}^{\frac{1}{2}} \mathbf{V} \overline{\mathbf{K}}^{\frac{1}{2}} \right\|_F^2 \tag{7}$$

is minimized to estimate $\mathbf{V}$ where $\|\cdot\|_F$ is the Frobenius norm. The matrix $\mathbf{\Gamma}$ is an $K \times K$ positive definite matrix. A frequent choice of $\mathbf{\Gamma}$ is given by

$$\mathbf{\Gamma} = \frac{1}{N} \mathbf{Y}\mathbf{Y}^t \tag{8}$$

which equilibrates a possible imbalance of classes in the training data. The matrix $\overline{\mathbf{K}}$ is the $N \times N$ constant matrix, given by

$$\overline{\mathbf{K}} = [\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1,\ldots,N} . \tag{9}$$

The δ is a regularization parameter and can be determined by the empirical Bayes method.

Note that by maximizing the convex $PL(\mathbf{V})$, our method can handle multi-class problem without suffering from the difficulties encountered by a pairwise classification method employed in SVM, while SVM maximizes the margin between two classes of data, hence cannot treat multi-class situations at once.

Due to the introduction of the specific quadratic penalty in (7), the minimizer $\mathbf{V}^*$ of $PL(\mathbf{V})$ is a solution of the neat matrix equation,

$$\nabla PL \equiv (\mathbf{P}(\mathbf{V}) - \mathbf{Y} + \delta \mathbf{\Gamma}\mathbf{V})\overline{\mathbf{K}} = \mathbf{O}_{K,N} , \tag{10}$$

where $\mathbf{P}(\mathbf{V})$ is an $K \times N$ matrix whose j-th column vector is the probability vector $\mathbf{p}(\mathbf{x}_j) \equiv \hat{\mathbf{p}}(\mathbf{V}\mathbf{k}(\mathbf{x}_j))$. The matrix $\mathbf{Y}$ is given in (1).

The minimizer $\mathbf{V}^*$, which gives the probabilistic predictor $\mathbf{p}^*(\mathbf{x}) \equiv \hat{\mathbf{p}}(\mathbf{V}^*\mathbf{k}(\mathbf{x}))$, is iteratively computed by the following algorithm.

Algorithm: Starting with an arbitrary $K \times N$ matrix $\mathbf{V}^0$, we generate a sequence $\{\mathbf{V}^i\}$ of matrices by

$$\mathbf{V}^{i+1} = \mathbf{V}^i - \alpha_i \Delta \mathbf{V}^i , \quad i = 0, \ldots, \infty \tag{11}$$

where $\Delta \mathbf{V}^i$ is the solution of the linear matrix equation,

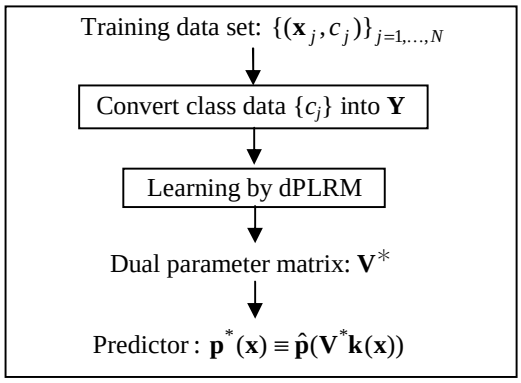


Figure 1. Training procedure.

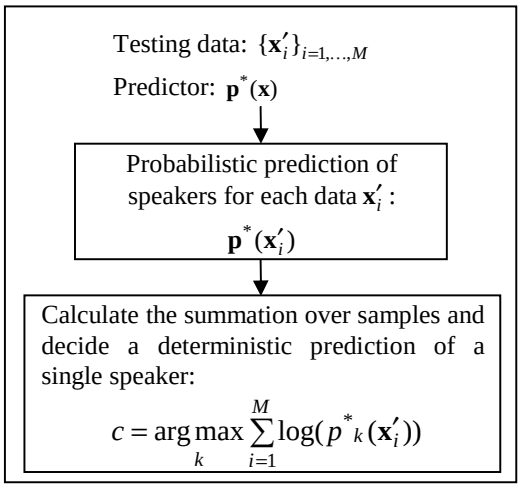


Figure 2. Testing procedure.

$$\sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)]-\mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\Delta\mathbf{V}^i(\mathbf{k}(\mathbf{x}_j)(\mathbf{k}(\mathbf{x}_j))^t) \quad (12)$$
$$+\delta\Gamma\Delta\mathbf{V}^i\overline{\mathbf{K}}=(\mathbf{P}(\mathbf{V}^i)-\mathbf{Y}+\delta\Gamma\mathbf{V}^i)\overline{\mathbf{K}}.$$

The detailed algorithm for estimation is shown in [9-11].

Note that we only need to solve an unconstrained optimization of a strictly convex function *PL*(**V**) or equivalently, to solve the simple matrix nonlinear equation (10). SVM, however, has to solve a series of inequality constrained maximization problems in multi-class situations.

3. dPLRM-based Speaker identification

Fig. 1 shows the training procedure. The training data set $\{(\mathbf{x}_j, c_j)\}_{j=1,\ldots,N}$ which covers all the speakers' data is collected. The class data $\{c_j\}$ is converted into matrix **Y**. The key matrix $\mathbf{V}^*$ is estimated by dPLRM. Finally the predictor $\mathbf{p}^*(\mathbf{x})$ is obtained.

Fig. 2 shows the testing procedure. The predictive probability $\mathbf{p}^*(\mathbf{x}'_i)$ is calculated for each data $\mathbf{x}'_i$. Then we sum up the log-probability for each class over samples and choose the class which attains its maximum as the speaker who utters the testing data.

4. Experiments and results

The proposed method was evaluated in text-independent speaker identification. The performance was compared with that of the conventional GMM-based method.

4.1. Data and system description

The database consists of 10 male speakers. Each speaker utters several sentences and words as listed in Table 1. Each sentence is approximately four seconds in duration and each word one second. The texts are common for all speakers. The same set of sentence and word speech was repeatedly recorded in four sessions (T0 to T3) over nine months and sampled at 16 kHz. A feature vector of 26 components, consisting of 12 mel-frequency cepstral coefficients plus normalized log energy and their first derivatives, was derived once every 10 ms over a 25.6 ms Hamming-windowed speech segment.

For the dPLRM training, three sentences from session T0 (12 second speech in total) were used for estimating $\mathbf{V}^*$. The following polynomial function was used as the kernel function.

$$\mathbf{K}(x,x')=(x^t x'+1)^s \quad (13)$$

The power *s* was nine. The dPLRM parameters α and δ were experimentally set to 1.0 in (11) and 7.7e-5 in (7), respectively.

In the testing, five sentences and five words from three sessions T1 to T3 were individually tested. The case number is 150 for each sentence and word data. The sentences for testing were different from those for training and were the same for all testing sessions.

In the GMM-based method, diagonal covariance models were used as speaker models. The parameters were initialized using all training speech for all speakers with the HMM toolkit (HTK) [12], and then estimated with the EM algorithm using the three sentences for each speaker. For testing, the speaker who attained the maximum collective log-likelihood was regarded as the speaker who uttered the testing data.

Table 1. Training and testing sentences and words (the Hepburn system of romaji for Japanese scripts)

	Contents
Training: sentences	1. seno takasaha hyakunanajusseNchi hodode mega ookiku yaya futotteiru 2. oogoeo dashisugite kasuregoeni natte shimau 3. tashizaN hikizaNha dekinakutemo eha kakeru
Testing: sentences	1. tobujiyuuo eru kotoha jiNruino yume datta 2. hajimete ruuburubijutsukaNe haittanoha juuyoneNmaeno kotoda 3. jibuNno jitsuryokuha jibuNga ichibaN yoku shitteiru hazuda 4. koremade shouneNyakyuu mamasaN bareenado chiikisupootsuo sasae shimiNni micchakushite kitanoha musuuno boraNtiadatta 5. giNzakeno tamagoo yunyuushite fukasase kaichuude sodateru youshokumo hajimatteiru
Testing: words	1. mouichido 2. torikaeshi 3. teisei 4. horyuu 5. shoukai

4.2. Results

Tables 2 lists the numbers of sentences and words identified correctly for each session and the accuracy rates and the confidence intervals (%) averaged over sessions T1 to T3 when compared with the GMM-based method. For the GMM-based method, 16-Gaussian-mixture models were used for sentence speech and 24-Gaussian-mixture models for word speech, since those models showed the best performance in the preliminary experiments using 8, 16, 24 and 32-Gaussian-mixture models as listed in Table 3. Our method outperformed the GMM-based method especially for word speech. These results indicate that the dPLRM can precisely capture the speaker characteristics with a small amount of data and effectively discriminate each speaker distribution.

Figure 3 shows the speaker identification rates averaged over sessions T1 to T3 for each speaker. Especially for word speech, the dispersion in the rates for our method was smaller than that for the GMM-based method. It can be considered that our method tends to stably identify any speakers.

Table 2. The numbers of (A) sentences and (B) words identified correctly for each session and the accuracy rates ± the confidence intervals (%) averaged over sessions T1 to T3.

(A) sentence speech

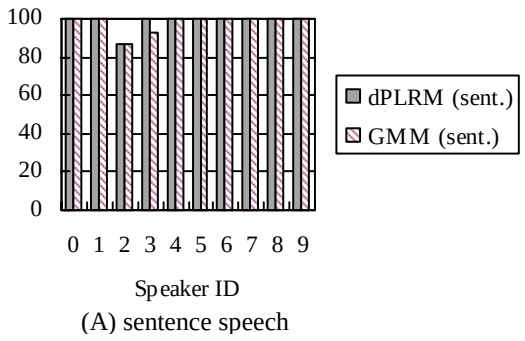
Method	T1	T2	T3	Average
dPLRM	50	48	50	98.7 ± 0.9
GMM (16)	49	48	50	98.0 ± 1.2

(B) word speech

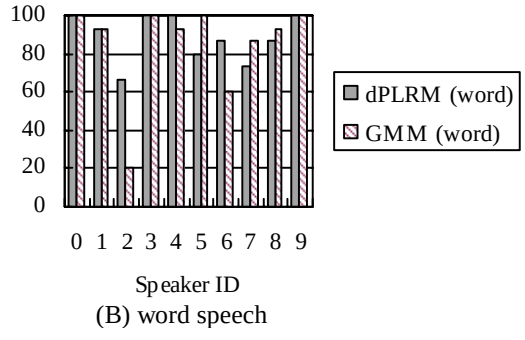
Method	T1	T2	T3	Average
dPLRM	46	42	45	88.7 ± 2.9
GMM (24)	43	42	42	84.7 ± 3.5

Table 3. Accuracy rates (%) of speaker identification averaged over sessions T1 to T3 using the GMM-based method with 8, 16, 24 and 32-Gaussian mixtures.

Testing	8	16	24	32
sentence speech	96.7	**98.0**	95.3	95.3
word speech	76.7	84.0	**84.7**	81.3



(A) sentence speech



(B) word speech

Figure 3. Accuracy rates (%) of speaker identification averaged over sessions T1 to T3 using (A) sentence and (B) word speech for each speaker.

5. Discussion

We investigate our dPLRM-based method from the viewpoints of the kernel functions. The polynomial kernel functions (13) with several powers and the Gaussian kernel function,

$$\mathbf{K}(x, x') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{h}\right) \tag{14}$$

were compared in performance in the case where the estimation step (11) was terminated at the 10-th iteration.

Table 4 lists the accuracy rates of speaker identification averaged over sessions T1 to T3 for the sentence and word speech data. The polynomial kernel function with the power of nine performed the best, and much better than the Gaussian kernel function. For the polynomial kernel function, the performance with the power of 11 was worse than that with the power of nine because of the lack of precision with 32-bit computers which we used in the experiments.

Table 4. Accuracy rates (%) of speaker identification with several kernel functions.

Kernel function		Sentence	Word
Polynomial	$s = 5$	97.3	84.7
	$s = 7$	97.3	86.0
	$s = 9$	98.7	88.7
Gaussian $h = 0.7238$		92.0	82.0

6. Conclusions

This paper proposed a new speaker identification method based on dPLRM. In dPLRM, the polynomial kernel function with the power of nine was shown to be effective for speaker discrimination. In the experiments with 12-second speech for training and speech with several sessions for testing, our dPLRM-based method was shown to perform better than the conventional GMM-based method.

Our future work includes the evaluation of dPLRM with a larger dataset, the use for speaker verification and end-point detection, and the comparison in performance with conventional discriminative methods and SVM.

7. REFERENCES

[1] D. A. Reynolds, "Speaker Identification and Verification Using Mixture Speaker Models," Speech Communication, 17, pp. 91-108, 1995.

[2] G. R. Doddington, M. A. Przybocki, A. F. Martin and D. A. Reynolds, "The NIST Speaker Recognition Evaluation – Overview, Methodology, Systems, Results,

Perspective," Speech Communication, 31, pp. 225-254, 2000.

[3] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[4] M. Schmidt and H. Gish, "Speaker Identification via Support Vector Classifiers," Proc. ICASSP, Atlanta, 1996.

[5] M. Schmidt, "Identifying Speakers With Support Vector Networks," Proc. Interface, Sydney, 1996.

[6] S. Fine, J. Navratil and R. A. Gopinath, "A Hybrid GMM/SVM Approach to Speaker Identification," Proc. ICASSP, Salt Lake City, 2001.

[7] W. M. Campbell, "A Sequence Kernel and its Application to Speaker Recognition," NIPS-14, pp. 1157-1163, 2001.

[8] V. Wan and S. Renals, "SVMSVM: Support Vector Machine Speaker Verification Methodology," Proc. ICASSP, Hong-Kong 2003.

[9] K. Tanabe, "Penalized Logistic Regression Machines: New methods for statistical prediction 1," ISM Cooperative Research Report 143, pp. 163-194, 2001.

[10] K. Tanabe, "Penalized Logistic Regression Machines: New methods for statistical prediction 2," Proc. IBIS, Tokyo, pp. 71-76, 2001.

[11] K. Tanabe, "Penalized Logistic Regression Machines and Related Linear Numerical Algebra," *KOKYUROKU 1320*, Institute for Mathematical Sciences, Kyoto University, pp. 239-249, 2003.

[12] http://htk.eng.cam.ac.uk, the hidden Markov model toolkit (HTK).