dPLRM-Based Speaker Identification with Log Power Spectrum

Tomoko Matsui and Kunio Tanabe

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

Abstract

This paper investigates speaker identification with implicit extraction of speaker characteristics relevant to discrimination from the log power spectrum of training speech by employing the inductive power of dual Penalized Logistic Regression Machine (dPLRM). The dPLRM is one of kernel methods like the support vector machine (SVM) and has the inductive power due to the mechanism of the kernel regression. In text-independent speaker identification experiments with training speech uttered by 10 male speakers in three different sessions, we compares the performances of dPLRM, SVM and Gaussian mixture model (GMM)-based methods and show that dPLRM implicitly and effectively extracts speaker characteristics from the log power spectrum. It is also shown that dPLRM outperforms the other methods especially when the amount of training data is small.

1. Introduction

The state of the art speaker recognition method is based on modeling individual speakers with GMMs via a set of reduced data of Mel-frequency cepstrum coefficients (MFCCs) [1]. While the MFCC data have been known to reflect the psycho-physical characteristics [2] and is widely believed to annihilate unwanted fluctuations in speech of individual speakers [3], there is no reason to assume that some useful information for speaker identification might not be lost in the reduced data. Besides, the dimension of MFCC vectors might have been chosen to accommodate the stable computation of estimate of the parameters in the GMM. The Mel-scale might not be needed for speaker identification.

Recently Tanabe proposed dPLRMs based on the penalized logistic regression model with a specific penalty term for bringing about induction-generalization capacity of the machine [4-6]. In our previous work [7,8], we applied dPLRM to text-independent speaker identification and showed that when using MFCCs, the dPLRM-based method was competitive with the GMM and SVM-based methods in small-scale experiments with 10 male speakers. In [9], we first reported speaker identification with implicit extraction of speaker characteristics relevant to discrimination from the log power spectrum by employing dPLRM. It was shown that when using training speech uttered in different sessions, the dPLRM-based method with the log power spectrum was competitive with the GMM-based method with the MFCCs, even though the dPLRM method avoids an elaborate MFCC extraction process such as Mel-scale filtering and dimension reduction.

This paper extends our previous work and examines the effectiveness of the dPLRM-based method with the log power spectrum through comparing the performance with the dPLRM-based method with the MFCCs and the SVM-based method. Both dPLRM and SVM utilize kernel functions and yield a certain duality which leads intrinsically to the kernel regressors. Therefore the two methods are expected to effectively handle nonlinearity in discriminating each speaker and capture important characteristics relevant to discrimination only from training data. However, there are several distinctions between dPLRM and SVM, e.g., dPLRM forms discrimination boundaries depending on the whole set of training data, while SVM forms them in the middle of the margin consisted of support vectors [8]. In this paper, we discuss several cases in which the methods are trained on speech data uttered in one session or different sessions and on coarsely sampled data.

2. dPLRM-based speaker identification

2.1. 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 convenience, we code the class data c_j by *j*-th unit column vector $\mathbf{e}_k \equiv (0,\ldots,1,\ldots0)^t$ 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.

We introduce a mapping from $\mathbf{R}^N$ into $\mathbf{R}^K$,

$$\mathbf{F}(\mathbf{x}) \equiv \mathbf{V}\mathbf{k}(\mathbf{x}) \qquad (2)$$

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

$$\mathbf{k}(\mathbf{x}) \equiv (\mathbf{K}(\mathbf{x}_1,\mathbf{x}),\ldots,\mathbf{K}(\mathbf{x}_N,\mathbf{x}))^t\ , \qquad (3)$$

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

$$\mathbf{p}(\mathbf{x}) \equiv \hat{\mathbf{p}}(\mathbf{F}(\mathbf{x})) \equiv (\hat{p}_1(\mathbf{F}(\mathbf{x})),\ldots,\hat{p}_K(\mathbf{F}(\mathbf{x})))^t, \qquad (4)$$

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

$$L(\mathbf{V}) \equiv -\sum_{j=1}^{N}\log(p_{c_j}(\mathbf{x}_j)) = -\sum_{j=1}^{N}\log(\hat{p}_{c_j}(\mathbf{Vk}(\mathbf{x}_j))) \qquad (5)$$

which is a convex function. 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 ML 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 deal with the problems, the penalty term is introduced and the negative-log-penalized-likelihood

$$PL(\mathbf{V}) \equiv L(\mathbf{V}) + \frac{\delta}{2}\left\| \mathbf{\Gamma}^{\frac{1}{2}}\mathbf{V}\overline{\mathbf{K}}^{\frac{1}{2}} \right\|_F^2 \qquad (6)$$

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

$$\mathbf{\Gamma} = \frac{1}{N}\mathbf{YY}^t \qquad (7)$$

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

$$\overline{\mathbf{K}} = [\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1,\ldots,N}. \qquad (8)$$

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

Due to the introduction of the specific quadratic penalty in (6), the minimizer $\mathbf{V}^*$ of $PL(\mathbf{V})$ is a solution of the neat matrix equation,

$$\nabla PL \equiv (\mathbf{P}(\mathbf{V}) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V})\overline{\mathbf{K}} = \mathbf{O}_{K,N}, \qquad (9)$$

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

$$\mathbf{V}^{i+1} = \mathbf{V}^i - \alpha_i \Delta\mathbf{V}^i, \quad i = 0,\ldots,\infty \qquad (10)$$

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

$$\sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\Delta\mathbf{V}^i(\mathbf{k}(\mathbf{x}_j)(\mathbf{k}(\mathbf{x}_j))^t) + \delta\mathbf{\Gamma}\Delta\mathbf{V}^i\overline{\mathbf{K}} = (\mathbf{P}(\mathbf{V}^i) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V}^i)\overline{\mathbf{K}}. \qquad (11)$$

The detailed algorithm for estimation is shown in [4-6]. Note that we only need to solve an unconstrained optimization of a strictly convex function $PL(\mathbf{V})$ or equivalently, to solve the simple matrix nonlinear equation (9).

2.2. Speaker identification 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 $\mathbf{Y}$. The key matrix $\mathbf{V}^*$ is estimated by dPLRM. Finally the predictor $\mathbf{p}^*(\mathbf{x})$ is obtained.

For testing, 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.

2.3. Effectiveness of the polynomial kernel function

In the previous section, we have introduced the mapping $\mathbf{F}(\mathbf{x})$ a priori for brevity. In fact, dPLRM was introduced by Tanabe as a dual machine of the penalized logistic regression machine PLRM in which $\mathbf{F}(\mathbf{x})$ is represented by

$$\mathbf{F}(\mathbf{x}) = \mathbf{W}\,\boldsymbol{\varphi}(\mathbf{x}) \qquad (12)$$

where $\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}),\ldots,\varphi_m(\mathbf{x}))^t$, each element of which is a certain nonlinear function of $\mathbf{x}$ [4-6]. The PLRM minimizes the penalized likelihood function

$$PL(\mathbf{W}) \equiv L(\mathbf{W}) + \frac{\delta}{2}\left\| \mathbf{\Gamma}^{\frac{1}{2}}\mathbf{W}\mathbf{\Sigma}^{\frac{1}{2}} \right\|_F^2 \quad , \qquad (13)$$

where $\mathbf{\Sigma}$ is a positive definite matrix. It was also shown that dPLRM and PLRM give exactly the same predictor $\mathbf{p}^*(\mathbf{x})$ when $\boldsymbol{\varphi}(\mathbf{x})$, $\mathbf{\Sigma}$ and $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ are appropriately chosen and that the former is computationally far less expensive than the latter. For the speaker identification problem we treat with dPLRM in this paper, we make use of the polynomial kernel function

$$\begin{aligned}\mathbf{K}(x, x') &= (x^t x' + 1)^s \\ &= \sum_{j=0}^{s} sCj\,(x^t x')^j = \sum_{j=0}^{s}\left[sCj\left(\sum_{i=1}^{n}[x]_i[x']_i \right)^j \right]\end{aligned} \qquad (14)$$

which is equivalent to the choice of

$$\begin{aligned}\boldsymbol{\varphi}(\mathbf{x}) = (&1, x_1,\ldots,x_{256}, \\ &x_1^2,\ldots,x_1x_2,\ldots, \\ &x_1^3,\ldots,x_1^2x_2,\ldots,x_1x_2x_3,\ldots, \\ &\ldots\ldots)\end{aligned} \qquad (15)$$

$$\Sigma^{-1} = \mathrm{diag}(1,1,..,1,\\ 1,\ldots,2,\ldots,\\ 1,\ldots,3,\ldots,6,\ldots,\\ \ldots\ldots) \tag{16}$$

in PLRM, where sCj is the number of combinations of s taken j at a time and $[x]_i$ is the i-th degree monomial in the elements of $\mathbf{x} \in \mathbf{R}^n$. If we chose s=5 as is the case with the experiments given in Section 3, the number m of elements of $\varphi(\mathbf{x})$ is so huge as O(10^{10}). Therefore it may be easily seen that the expressive power of the map $\mathbf{F}(\mathbf{x})$ is so high that the map could mimic, if necessary, the operations indicated in the MFCC extraction process. Our experiment suggests that without resorting to human judgment such as the Mel-scale filtering, the dPLRM can automatically construct some kind of nonlinear transformation from the training data although it might not be similar to the MFCC extracting transformation.

3. Experiments

The performances of dPLRM, SVM, GMM-based methods are compared through text-independent speaker identification experiments.

3.1. Data description and experimental conditions

The data has been collected for 10 male speakers who utter several sentences (for four seconds per sentence) and words (for one second per word). Although the texts are common for all speakers, the sentences used for testing are different from those for training. The utterances were recorded at the sampling rate of 16 kHz in six sessions from T0 through T5 over 13 months' period. The interval between T0 and T1 is one month and the other intervals are three months. A 256-dimensional log power spectrum vector and a MFCC vector of 26 components, consisting of 12 Mel-frequency cepstral coefficients plus normalized log energy and their first derivatives, is derived once every 10 ms over a 25.6 ms Hamming-windowed speech segment.

We choose two kinds of training data sets, DS1 and DS2. The set DS1 consists of the data for three sentences, each of which is uttered in Session T0, T1 and T2, respectively, and the set DS2 consists of the data for three sentences uttered in the single session T2. The total duration of the utterances of three sentences is approximately 12 seconds per speaker. For testing purpose, we choose the utterances of the five sentences and the five words from Sessions T3, T4 and T5 and test them individually. For both sentence and word cases, the total case number of the testing is 150 since we have 10 speakers times 5 sentences (or words) and 3 sessions.

The polynomial kernel function (14) is used for the dPLRM. The power is chosen to be s=5 for the log power spectrum data and s=9 for the MFCC data, respectively. In order to execute effective computation with 64-bit precision, the data is so scaled that all the elements of feature vectors lie in the interval [-0.5, 0.5].

In the SVM-based method, the SVM^{light} software is used [10] with the polynomial kernel function (14). The power is chosen in the same way as dPLRM. For each speaker, a one-versus-rest classifier is trained and the speaker who attains the largest positive confidence index averaged over the test speech is regarded as the speaker who uttered the speech for testing.

In the GMM-based method, the mixture model of 16 Gaussian distributions with diagonal covariance was chosen as a speaker model among the competing models with 8, 16 and 24 Gaussian distributions. The parameters were initialized using all training speech for all speakers with the HMM toolkit (HTK) [11], and then estimated with the EM algorithm for each speaker. For testing the GMM method adopts the decision rule to select the speaker who attained the maximum collective log-likelihood.

3.2. Test of DS1-trained methods

Firstly we compare the performances of the methods trained on the set DS1. Table 1 lists the identification rates with the confidence intervals at a confidence level of 90% averaged over the 150 cases.

Table 1. Speaker identification rates (with the confidence intervals at a confidence level of 90%) using the training data of the MFCCs and log power spectrum extracted from three sentences uttered in Session T0/T1/T2 for each sentence.

Testing data	Method	Identification rates (%)	
		MFCC	Log power spectrum
Word speech	dPLRM	90.7 (87.3,94.7)	**92.7 (89.3, 96.0)**
	SVM	91.3 (87.3,94.7)	88.0 (83.3, 92.0)
	GMM	89.3 (85.3, 93.3)	84.0 (79.3, 88.7)
Sent. speech	dPLRM	99.3 (98.7,100)	**100 (99.3, 100)**
	SVM	98.0 (96.0, 99.3)	**100 (99.3, 100)**
	GMM	99.3 (98.7, 100)	99.3 (98.7, 100)

The dPLRM method with the log power spectrum performed the best for word speech. For sentence speech, the SVM method also performed the best. Since the training data contains the information on utterance variations among Sessions T0, T1 and T2, the two kernel methods essentially having high expressive power attain higher success rates with the log power spectrum especially for sentence speech..

3.3. Test of DS2-trained methods

Secondly we test the methods trained on the set DS2. Table 2 lists the identification rates with the same confidence qualification as stated above.

Table 2. Speaker identification rates (with the confidence intervals at a confidence level of 90%) using the training data of the MFCCs and log power spectrum extracted from three sentences uttered in Session T2.

Testing data	Method	Identification rates (%)	
		MFCC	Log power spectrum
Word speech	dPLRM	**88.7 (84.7, 92.7)**	83.3 (78.7, 88.0)
	SVM	86.7 (82.7, 91.3)	73.3 (67.3, 78.7)
	GMM	84.7 (80.0, 89.3)	68.0 (62.0, 74.0)
Sent. speech	dPLRM	**98.7 (97.3, 100)**	97.3 (95.3, 99.3)
	SVM	96.7 (94.0, 98.7)	94.7 (92.0, 97.3)
	GMM	98.0 (96.0, 99.3)	86.7 (82.7, 91.3)

All methods trained on the MFCC data gave higher identification rates than those trained on the log power spectrum data. We note that the performance drops seriously in the order of GMM, SVM and dPLRM when the training data switches from the MFCCs to the log power spectrum. We found some difficulties with the GMM-based method in the estimation process due to the high dimensionality of the 256-dimensional log power spectrum data. Kernel methods of dPLRM and SVM can deal with the high dimensionality more robustly than GMM.

3.4. Test of the methods trained on coarsely sampled data

Table 3 lists the identification rates with the same confidence qualification as stated above. The set DS1 is analyzed with different window shifts. The length of the training data with 20 ms window shift is half of that with 10 ms window shift, and the length of the training data with 30 ms window shift is one-third.

Table 3. Speaker identification rates using GMM trained with the MFCCs and dPLRM/SVM trained with the log power spectrum extracted with different window shifts from three sentences uttered in Session T0/T1/T2 for each sentence.

Training data	Method	Identification rates (%)		
		10 ms shift	20 ms shift	30 ms shift
Word speech	dPLRM	92.7 (89.3,96.0)	91.3 (87.3,94.7)	90.0 (86.0,94.0)
	SVM	88.0 (83.3, 92.0)	84.7 (80.0, 89.3)	83.3 (78.7, 88.0)
	GMM	89.3 (85.3,93.3)	86.0 (81.3,90.7)	85.3 (80.7,90.0)
Sent. speech	dPLRM	100 (99.3,100)	100 (99.3,100)	100 (99.3,100)
	SVM	100 (99.3, 100)	100 (99.3, 100)	99.3 (98.7,100)
	GMM	99.3 (98.7,100)	98.0 (96.0,99.3)	96.7 (94.0,98.7)

It is interesting to note that the dPLRM method trained on such a coarsely sampled data with 30 ms window shift outperforms the GMM method with full 10 ms shift sampled data. Since the dPLRM can handle nonlinearity more effectively with kernel functions and do discriminating learning interdependently, it is expected to work with a smaller amount of training data. Since a discrimination boundary of SVM is formed only by support vectors and the one-versus-rest method is simply used for the multi-class discrimination, it can be considered that SVM is affected by an extremely small amount of training data more than dPLRM. For SVM, the ratios of support vectors to all training vectors are 21.0% in the case of 10 ms shift, 23.6% in the case of 20 ms shift and 25.8% in the case of 30 ms shift and do not heavily vary. The number of support vectors decreases in almost the same ratios (i.e., half and one-third) as the amount of training data.

4. Conclusions

In this paper, dPLRM-based speaker identification without outright pre-processing of speech data was discussed. Comparison was made between the dPLRM, SVM and GMM-based methods in the experiments with training data uttered by 10 male speakers in three sessions. It was shown that the dPLRM method with the log power spectrum was competitive with the GMM-based method with the MFCCs and the SVM-based method. The dPLRM-based method outperforms the other methods especially as the amount of training data becomes smaller.

The evaluation of the method with a larger dataset is left for our future study.

5. Acknowledgements

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research (B) 16300036 and (C) 16500092 and by ISM Project Research 2004.

6. References

[1] http://www.nist.gov/speech/tests/spk/index.htm, NIST Speaker Recognition Evaluations.

[2] S. S. Stevens, “Psychophysics,” John Wiley & Sons, New York, 1975.

[3] H. A. Murthy, F. Beaufays, L. P. Heck and M. Weintraub, “Robust Text-Independent Speaker Identification over Telephone Channels , IEEE Trans. on SAP, vol. 7, no. 5, pp.554-568, 1999.

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

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

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

[7] T. Matsui and K. Tanabe, “Speaker Identification with dual Penalized Logistic Regression Machine,” Proc. Odyssey, pp.363-366, Toledo, 2004.

[8] T. Matsui and K. Tanabe, “Probabilistic Speaker Identification with dual Penalized Logistic Regression Machine,” Proc. ICSLP, pp. III-1797-1800, 2004.

[9] T. Matsui and K. Tanabe, “Speaker Recognition without Feature Extraction Process,” Proc. Workshop on Statistical Modeling Approach for Speech Recognition: Beyond HMM, pp.79-84, Kyoto, 2004.

[10] http://www.cs.cornell.edu/People/tj/svm_light/, the support vector machine software, SVM^{light}.

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