Speaker recognition with penalized logistic regression machines*

Øystein Birkenes (NTNU) and Tomoko Matsui (ISM)

Abstract
We study on speaker recognition using a penalized logistic regression machine (PLRM) [1-3]. Parameters of a multiclass logistic regression model with the log-likelihood values of speaker Gaussian mixture models (GMMs) are discriminatively estimated and the model used for speaker decision. In speaker identification experiments with 74 male speakers in 2001 NIST speaker recognition evaluation corpus, the error rate is reduced from 45.3%, using the conventional GMM-based method, down to 37.6%.

1. Introduction
Speaker recognition technologies are widely used in access control, monitoring, information retrieval, and forensic. GMM-based methods have been typically used for some years ahead especially in text-independent recognition and higher performance is demanded. In the methods, utterance variation is well captured by a mixture of a well-chosen number of Gaussian distributions.

Recently we proposed an isolated-word speech recognition method based on PLRM [4]. In the method, the HMM is combined with multiclass logistic regression resulting in a powerful speech recognizer which provides us with the probability for each word. Here we apply this approach to speaker recognition in order to improve the performance of the GMM-based method.

2. PLRM-based method [4]
A multiclass logistic regression model with parameter $\theta$ of the probability $p_k$ is given by

$$\hat{p}_k = \frac{\exp f_k(x; \theta)}{\sum_{l=1}^K \exp f_l(x; \theta)},$$

where $f_k(x; \theta)$ is a discriminant function for the kth speaker and $K$ denotes the total number of the registered speakers. For the discriminant functions we choose

$$f_k = f_k(x; \theta) = w_k^T \phi(x; \Lambda),$$

where $\phi$ is a nonlinear map from $x$ to an $M$-dimensional real space parameterized by $\Lambda$, and $w_k$ is an $M$-dimensional weight vector. The parameters of the model are jointly denoted by $\theta = (W, \Lambda)$, where $W$ is a $K \times M$-matrix whose rows are $w_k^T$.

The nonlinear map $\phi$ should preserve discriminative information embedded in the speech signals. We use a mapping involving a set of GMMs for the speakers, e.g.,

$$x \mapsto \phi(x; \Lambda) = [l, \phi(x; \lambda_1), \ldots, \phi(x; \lambda_K)]^T$$

where $\phi(x; \lambda_k)$ is the log-likelihood of the GMM with parameter $\lambda_k$ corresponding to the kth speaker, with $\lambda_k$ being the kth column of the matrix $\Lambda$. The inclusion of 1 as the first element of $\phi(x; \Lambda)$ ensures that the discriminant functions in (2) are affine transformations of the GMM log-likelihoods. Figure 1 illustrates our choice of model. The model utilizes the information in all speakers’ GMMs for speaker decision.

Figure 1. Multiclass logistic regression model.

Next we discuss how the parameter $\theta$ of the model can be estimated. Given a set of training data $D = \{(x^{(n)}, y^{(n)})\} (n = 1, \ldots, N)$ with the speaker label $y^{(n)}$, the parameter $\theta$ could be estimated by maximizing the likelihood of the multinomial distribution which is

$$L(\theta; D) = \prod_{n=1}^N \hat{p}_{y^{(n)}}.$$  

---

*「罰金付きロジスティック回帰マシンを用いた話者認識」、ビルケネス・オイストィン (ノルウェーワ工科学), 松井知子 (統数研)
However, due to the limited amount of training data, the maximum likelihood estimate is prone to overfitting, and may lead to poor prediction on unseen speech signals. For that reason, we introduce a penalty \( \Omega(\theta) \) and find an estimate \( \hat{\theta} \) by maximizing the penalized logistic regression likelihood
\[
P_{\hat{\theta}}(\theta; D) = L(\theta; D)\Omega^2(\theta), \tag{5}
\]
where \( \delta > 0 \) is a hyperparameter used to balance the likelihood and the penalty factor.

It was proposed in [1-3] to use a penalty of the form
\[
\Omega(\theta) = \exp\left(-\frac{1}{2}\text{trace}(\Gamma W \Sigma W^T)\right), \tag{6}
\]
where \( \Gamma \) is a \( K \times K \) diagonal matrix whose \( k \)th diagonal element is the fraction of training samples with the \( k \)th speaker, and \( \Sigma \) is an \( M \times M \) positive definite matrix (here \( \Sigma \) is set to the identity matrix).

Although both \( W \) and \( \Lambda \) can be estimated using the coordinate descent method in [4], we first examine to iteratively find the minimum of \( P_{\hat{\theta}}(\theta; D) \) respect to \( W \) [1-3], i.e.
\[
\hat{W} = \min_W P_{\hat{\theta}}(W, \Lambda_0; D). \tag{7}
\]

3. Experiments

The PLRM-based method is evaluated in speaker identification experiments through comparing with the GMM-based method.

3.1. Data description and experimental conditions

The experiments were carried out using 74 male speakers in the evaluation data of the 2001 NIST speaker recognition evaluation corpus (ISBN 1-58563-241-4) which consists of conversational cellular telephone speech. A Mel frequency cepstral coefficient (MFCC) vector of 26 components, consisting of 12 MFCCs plus normalized log energy and their first derivatives, is derived once every 20 ms over a 25.6 ms Hamming-windowed speech segment. The cepstral mean normalization is applied. The total length of the training data is two minutes long per speaker. For each speaker, 10 MFCC files which are from eight to 44 second lengths are created by randomly selecting the MFCC vectors for the training data. The test data comprises 850 MFCC files of varying length not exceeding 60 seconds (roughly 30 seconds on average).

In the GMM-based method, the parameters were initialized using all training speech for all speakers with the HMM toolkit [5], and then estimated with the EM algorithm for each speaker. The number of mixture components per state is 128.

In the PLRM-based method, we used the parameters of the GMMs in the above GMM-based method as \( \Lambda_0 \). Then we optimized \( W \) according to (7) with \( \delta = 0.1 \).

3.2. Result

Table 1 lists the speaker identification rates with the PLRM and GMM-based methods. By using the PLRM-based method, we obtained 17\% of the relative error reduction.

Table 1. Speaker identification rates.

<table>
<thead>
<tr>
<th></th>
<th>PLRM</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62.4%</td>
<td>54.7%</td>
</tr>
</tbody>
</table>

4. Conclusions and future work

A speaker recognition method with PLRM has been discussed and shown to lead higher accuracy on the 2001 NIST speaker recognition evaluation corpus.

Future work includes further investigation on the parameter estimation, e.g., use of the coordinate descent method to estimate both \( W \) and \( \Lambda \). We also plan to evaluate our method in speaker verification.

References