ELectronics EXpress

IEICE Electronics Express, Vol.7, No.6, 390–396

A bayesian logistic regression approach to spoken language identification

Haipeng Wang[a)], **Xiang Xiao, Xiang Zhang, Jianping Zhang, and Yonghong Yan**

Thinkit Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, China

a) *wanghaipeng@hccl.ioa.ac.cn*

Abstract: This paper presents a novel token-based approach for spoken language identification (LID) using bayesian logistic regression model, which takes into account prior distribution for parameters of logistic regression models in order to avoid overfitting. Speech utterances are first decoded into token sequences, and then we design a hierarchical system which utilizes bayesian logistic regression model to perform LID task on these token sequences. Experiments conducted on the NIST LRE 2007 database show that the proposed approach provides quite competitive performance compared to other state-of-the-art token-based approaches.

Keywords: language identification, token-based approach, bayesian logistic regression, hierarchical system design

Classification: Science and engineering for electronics

References

[1] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, 1996.

[2] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition Using Phone Latices," *8th Int. Conf. Spoken Language Process.*, Jeju Island, Korea, pp. 25–28, Oct. 2004.

[3] M. Li, H. Suo, X. Wu, P. Lu, and Y. Yan, "Spoken Language Identification Using Score Vector Modeling and Support Vector Machine," *Interspeech 2007*, Antwerp, Belgium, pp. 350–353, Aug. 2007.

[4] W. Campbell, F. Richardson, and D. Reynolds, "Language recognition with word lattices and support vector machines," *Proc. ICASSP, 2007*, Hawai'i, USA, April 2007.

[5] H. Li, B. Ma, and C. H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp.271–284, 2007.

[6] A. Genkin, D. D. Lewis, and D. Madigan, "Large-Scale Bayesian Logistic Regression for Text Categorization," *TECHNOMETRICS*, vol. 49, no. 3, pp. 291–304, 2007.

[7] T. Zhang and F. J. Oles, "Text categorization based on regularized linear classification methods," *Information Retrieval*, vol. 14, no. 1, pp. 5–31,

© IEICE 2010

DOI: 10.1587/elex.7.390
Received September 19, 2009
Accepted February 25, 2010
Published March 25, 2010

2001.
[8] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," *Machine Learning-International Workshop Then Conf.*, pp. 412–420, 1997.
[9] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification," *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006*, pp. 1–6, 2006.

1 Introduction

Automatic spoken language identification(LID) is the task of determining the language identity of a particular speech segment [1, 2, 3]. Token-based approaches have proved quite effective for LID tasks. These approaches first decode speech utterances into sequences of tokens such as phones, and then perform LID task based on these token sequences. A traditional token-based technique is phone recognition followed by language modeling (PRLM) [1], which employs n-gram language models to derive phonotactic scores. PRLM approach can be improved consistently by using lattices generated by phonetic tokenizers [2]. Recently another token-based approach has become prevailing to LID tasks. This approach, sharing similarity with representing text documents as term vectors, converts token sequences into feature vectors which represent the co-occurrence statistics of token sequences over a fixed bag-of-n-grams [4]. Then support vector machines (SVMs) are used to perform classification on these feature vectors. We refer this approach as Ngram-SVM method [4, 5].

In this paper, we introduce bayesian logistic regression (BLR) models for LID tasks. BLR models combine prior knowledge with logistic regression model in a bayesian framework [6]. Before using BLR models, we obtain similar sparse feature vectors as Ngram-SVM method. We consider the sparseness of feature vectors as the prior knowledge of the BLR models, and we design our system in a hierarchical structure, which consists of a number of BLR classifiers followed by a back-end score merger. Compared to the traditional PRLM approach, our proposed method (Ngram-BLR) owns the advantage of discriminative feature representation. Compared to Ngram-SVM approach, our proposed method is not so sensitive to those extreme data in the training set because of the prior knowledge incorporated in BLR models. Meanwhile, it is also expected that the proposed classification method will complement those state-of-the-art approaches.

The rest of this paper is organized as follows. In Section 2, we review the concept of logistic regression and introduce the bayesian approach in particular. In Section 3, we present the proposed system design of our Ngram-BLR approach. In Section 4, we present the experiments conducted on the 2007 NIST LRE database. Finally, we conclude our work and discuss the future in Section 5.

© IEICE 2010
DOI: 10.1587/elex.7.390
Received September 19, 2009
Accepted February 25, 2010
Published March 25, 2010

2 Bayesian logistic regression model

2.1 Logistic regression

A general logistic regression model is a linear model which tries to model the conditional probability of a binary outcome

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)} \tag{1}$$

where x_i denotes the feature vector for example i and y_i is the class label (or a binary response) of example i. For binary classification problems, $p(y_i = +1|\mathbf{x}_i)$ means the probability that the ith example belongs to the *positive* events group, while $p(y_i = -1|\mathbf{x}_i)$ means the probability that the ith example belongs to the opposite *negative* events group. Then the accumulated likelihood in the training data is

$$\prod_{i=1}^{n} \frac{1}{1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)} \tag{2}$$

Therefore the maximum likelihood estimation of the parameter $\mathbf{w}$ is equivalent to maximize the log-likelihood function as following

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left\{ -\sum_{i=1}^{n} \ln(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)) \right\} \tag{3}$$

where n denotes the number of training examples.

2.2 Bayesian logistic regression with Gaussian prior

In the bayesian approach to logistic regression, the model parameters are determined not only by training data, but also the prior distribution of the regression coefficients. A typical MAP (maximum a posteriori) criteria is to modify the objective function as the sum of the log likelihood of the training data and the log of the prior distribution of the model parameters [6]:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left\{ -\sum_{i=1}^{n} \ln(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)) + \ln p(\mathbf{w}) \right\} \tag{4}$$

where $p(\mathbf{w})$ denotes the prior probability of model parameter vector $\mathbf{w}$.

Gaussian prior is the most widely used prior distribution to logistic regression. In the application to LID tasks, we assume a Gaussian prior with mean $\mathbf{0}$ and variance $\sigma^{\mathbf{2}}$. The mean of $\mathbf{0}$ is due to the sparseness of the term vectors.

$$p(w_j|\sigma_j) = \mathbf{N}(0, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-w_j^2}{2\sigma_j^2}\right), \qquad j = 1, ..., d \tag{5}$$

where w_j is the jth component of parameter vector $\mathbf{w}$ and d is the dimension of vector $\mathbf{w}$. Therefore, we deduce the MAP optimization criteria in this case as:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left\{ -\sum_{i=1}^{n} \ln(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)) - \sum_{j=1}^{d} \left(\frac{1}{2}\ln 2\pi + \ln\sqrt{\sigma_j^2} + \frac{w_j^2}{2\sigma_j^2} \right) \right\} \tag{6}$$

© IEICE 2010
DOI: 10.1587/elex.7.390
Received September 19, 2009
Accepted February 25, 2010
Published March 25, 2010

To solve MAP optimization estimation of parameter $\mathbf{w}$, we propose to use an CLG algorithm in [7] for its efficiency and easy implementation. In each iteration of CLG algorithm, we only need to find the optimization value for a single variable while holding all other variables constant.

3 System design using BLR models

3.1 Feature representation

Before applying BLR models to LID tasks, we represent speech signals by discriminative feature vectors similar with the feature representation method in [4]. Each speech utterance is first converted into a token lattice using a token recognizer. Although 1-best token sequence can be produced from the lattice, we directly produce expected counts of token n-grams included in the pre-fixed bag-of-n-grams [5]. The bag-of-n-grams, which describes a discriminative feature space, is selected based on the mutual information [8] of token n-grams of different languages.

Expected counts are calculated over all possible hypotheses by a forward-backward algorithm. Then the probability of a token n-gram t_j in a lattice L is:

$$p(t_j|L) = \frac{count(t_j|L)}{\sum_{i=1}^{d} count(t_i|L)} \qquad j = 1, ..., d \tag{7}$$

where d is the number of n-grams in the bag-of-n-gram, which equals the dimension of the feature vector. We use a so-called *TFLLR* method [4] to weight all these probabilities. This weighting method normalizes the probability of each n-gram by the square root of its probability in the whole training set:

$$\hat{p}(t_j|L) = \frac{p(t_j|L)}{\sqrt{p(t_j|ALL)}} \qquad j = 1, ..., d \tag{8}$$

where $\hat{p}(t_j|L)$ denotes the weighted probability of the jth token n-gram in the bag-of-n-grams, and $p(t_j|ALL)$ represents its probability in the whole training set. Consequently, by appending all these weighted n-gram probabilities in a single vector, we obtain the sparse feature vector used for BLR models.

3.2 Hierarchical system structure

After feature vectors are extracted from the speech signals, the LID task becomes a vector-based classification problem. Since BLR model is inherently binary classifier, it is straight to implement an one-versus-rest [4] system design for LID task. However, the training of BLR models suffers heavy memory requirement when a myriad of training samples are utilized. Another popular system design is one-versus-one approach [5]. Let L denotes the number of languages, this approach builds $\frac{L*(L-1)}{2}$ pair-wise classifiers followed by a score classifier. In our implementation, we generalize the one-versus-one idea to design a hierarchical system structure.

An overview of the system design is demonstrated in Fig. 1. As can be seen, we build multiple BLR model classifiers, including both pair-wise BLR

© IEICE 2010
DOI: 10.1587/elex.7.390
Received September 19, 2009
Accepted February 25, 2010
Published March 25, 2010

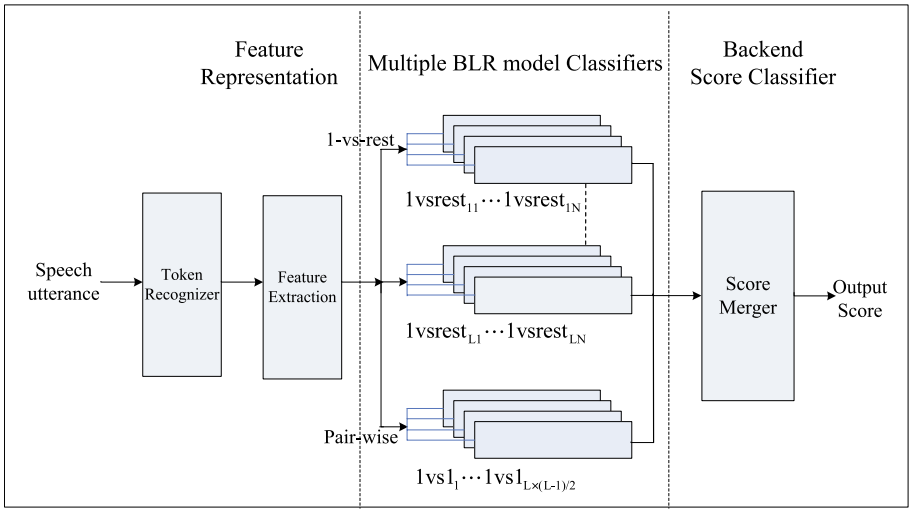


Fig. 1. system design overview

classifiers and one-versus-rest BLR classifiers. We regard all these classifiers as a group of score generators, and each *small* BLR classifier is a component score generator. For each language, we split the rest language training data into N subgroups, considering the great amount of training data. Therefore, there are $L*N$ one-versus-rest classifiers, and in total, there are $\frac{L*(L-1)}{2}+L*N$ component BLR classifiers built for the system. Following the group of score generator, a score merger is built and its function is to combine score outputs into a score vector, and perform final classification decision based on the score vector. In our implementation, LDA+Gaussian backend [9] is used as the backend score merger.

4 Experiments

4.1 Experimental setup

The token recognizer used in our experiment is the Hungarian tokenizer released by BUT. This tokenizer is used along HTK toolkit to produce token lattices across the training data, development data, and evaluation data. The evaluation data set in our experiments is the NIST 2007 Language Recognition Evaluation corpus consisting of 7530 utterances. We evaluated our system for the 30 and 10 seconds tasks under the NIST 2007 closed condition which covers 14 languages (Arabic, Bengali, English, Farsi, French, German, Hindi, Japanese, Korea, Mandarin, Spanish, Tamil, Thai, and Vietnamese). The training data is drawn from the whole CallFriend corpus, NIST 1996 LRE corpus, part of NIST 2003 LRE corpus and part of NIST 2005 LRE corpus. The development data set includes part of NIST 2003 LRE corpus and part of NIST 2005 LRE corpus.

A speech activity detector is applied to discard noise speech segments for all the speech utterances. We segment each utterance in the training data and development data to 30 seconds segments and 10 seconds segments according to the 30 and 10 seconds tasks. For each language, we select the 1500 n-grams ($n \leq 4$) into the bag-of-n-gram based on mutual information.

© IEICE 2010
DOI: 10.1587/elex.7.390
Received September 19, 2009
Accepted February 25, 2010
Published March 25, 2010

Equal error rates (EER) is used as the evaluation metric.

4.2 Variance setting methods

As described in section 2.2, we assume the mean of the Gaussian prior distribution to be 0, and in this part we compare two methods for the prior settings of the variance σ^2: a 5-fold Cross-validation (CV) method and a Norm-based method. For cross-validation method, we test values for Gaussian variance σ^2 ranging from 0.001 to 100 by multiples of 10. In each run, the value that maximize posterior predictive log-likelihood for the training set is selected. The norm-based method [6] is simply:

$$\sigma_j^2 = \frac{dn}{\sum_{i=1}^{n} \|\mathbf{x}_i\|_2^2} \qquad j = 1, ..., d \tag{9}$$

where d is the dimension of feature vector, n is the size of the training data set, and $\|\mathbf{x}_i\|_2$ represents the Euclidean norm of the feature vector of training sample $\mathbf{x}_i$. Note that both of these two methods set a common prior variance distribution for each feature value.

Table I shows the comparison of these two methods. As can be seen, these two methods exhibit close performances. Cross-validation performs slightly better than norm-based method. However, it is important to note that cross-validation requires much heavier computation cost for the BLR model training, so we determine to use norm-based method to set the prior value of Gaussian variance in the following experiments.

Table I. EER (%) for CV and norm-based method

	10 sec	30 sec
cross-validation	8.28	2.57
norm-based method	8.49	2.68

4.3 Comparison with PRLM and Ngram-SVM

We compare the performance of our proposed system (Ngram-BLR) with the start-of-the-art lattice-based PRLM approach and Ngram-SVM approach. The same Hungarian tokenizer is used for both PRLM and Ngram-SVM approaches. For PRLM approach, 4-gram language models are trained with Witten-Bell discounting method similar with [2]. For Ngram-SVM approach, we apply the same feature vectors and similar hierarchical system structure as Ngram-BLR in order to make them comparable and the SVM*light* package is used.

Table II summarizes the results of the comparison. Compared to the traditional PRLM approach, our proposed Ngram-BLR method achieves a relative EER reduction of 17.3% on 10 seconds task and 26.6% EER reduction on 30 seconds task. Compared to Ngram-SVM approach, our method obtains very close performance on 30 seconds task while 7.9% relative EER reduction is gained on the 10 seconds task. This is because feature vectors on 10 seconds

© IEICE 2010
DOI: 10.1587/elex.7.390
Received September 19, 2009
Accepted February 25, 2010
Published March 25, 2010

Table II. EER (%) for different approaches

	10 sec	30 sec
PRLM	10.26	3.65
Ngram-SVM	9.22	2.47
Ngram-BLR	8.49	2.68
PRLM + Ngram-BLR	8.32	2.45
Ngram-SVM + Ngram-BLR	8.18	2.26

task are more sparse than 30 seconds task, and thereby the prior knowledge described in section 2.2 performs more effective for 10 seconds task. We also make linear fusion of our proposed method with these two state-of-the-art approaches. As can be seen from Table II, combination between our proposed method and either PRLM or Ngram-SVM leads consistent further reduction of EER. This demonstrates the complementary performance of our method.

5 Conclusion

In this paper, we have presented a novel bayesian logistic regression approach to LID task, based on a hierarchical system design. We apply this approach to the NIST 2007 LRE database, and our approach exhibits quite competitive performance and proves to be complementary with other approaches. Future work includes: 1) further research of the prior distribution of the BLR models and 2) generalizing this system design to other applications.

Acknowledgments

This work is partially supported by MOST (973 program, 2004CB318106), the National Natural Science Foundation of China (10574140, 60535030), the National High Technology Research and Development Program of China (863 program, 2006AA01010, 2006AA01Z195).

© IEICE 2010
DOI: 10.1587/elex.7.390
Received September 19, 2009
Accepted February 25, 2010
Published March 25, 2010