International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

A SAMPLE AND FEATURE SELECTION SCHEME FOR GMM-SVM BASED LANGUAGE RECOGNITION

Yan SONG, Li-Rong DAI

Department of EEIS, University of Sci&Tech of China

ABSTRACT

Discriminative training for language recognition has been a key tool for improving system performance. SVM-based algorithms (i.e. GMM-SVM, GLDS-SVM etc.) are important ones for language recognition. The core of these algorithms is to construct the kernel for comparing the similarity of two sequences. It is known that the mismatch between training and test condition will degrade the performance. In this paper, we proposed a novel sample and feature selection scheme under the GMM-SVM framework, which aims at alleviating the duration mismatch problem. The proposed method is evaluated on NIST 03 and 07 language recognition evaluation tasks with improvement over prior techniques.

Index Terms—Language Recognition, GMM-SVM, Feature Selection, Data Selection

1. INTRODUCTION

Gaussian Mixture model (GMM) has become one of the successful methods applied to the language recognition. With the use of Shift-Delta Cepstra Features (SDC), GMM method is of high effectiveness both as individual component and as in fusion with other methods.

Recently, the introduction of discriminative methods into the acoustic system results in significant improvement of performance on language recognition. In [3], the Maximum Mutual Information (MMI) criterion was applied to the update of GMM model for each language. In [4], a SVM method based on the generalized linear discriminant sequence (GLDS) kernel was introduced for both speaker and language recognition, the system is competitive and comparable to other approaches. In [5], a GMM-SVM method was proposed. In GMM-SVM system, the kernel function is derived from the approximation of Kullback-Leibler distance between two distributions. The experiment results on NIST LRE evaluation tasks show advantage of GMM-SVM model for the 30 second duration test [5].

It is known that better recognition performance can be achieved if the training data is matched the test condition. For language recognition, there are several mismatches including speaker, gender, channel and duration of speech segment etc. The channel mismatch can be effectively alleviated by techniques such as Factor Analysis [7] and NAP [8] etc. The gender-dependent or speaker-dependent models can be employed to recognize the test segment in matched condition. In this paper, we focus on the duration mismatch problem under the GMM-SVM framework.

In the NIST LRE evaluation plan [6], there are three segment duration test conditions (including 30, 10 and 3 second) to evaluate system performance of different amounts of speech. It is known that in SVM related language recognition methods, the duration mismatch may cause the degradation of the performance. A straightforward method to tackle this issue is to sub-segment the audio files in training dataset into the segments similar as the test conditions. However, due to the constraints of the memory requirement and computational complexity, the number of training instances is limited, which results in segments with much longer segment duration.

Also, in GMM-SVM the supervector is formed by stacking the adapted mean vectors from the Universal Background Model (UBM). There is only a subset of features relevant to target model in supervector space. The disadvantages of the irrelevant features are two-fold: 1) Increase the memory requirement and computational burden; 2) Degrade the recognition performance.

In this paper, a novel sample and feature selection scheme in GMM-SVM framework is proposed. The training data is hierarchically sub-segmented into segments with different durations. By iteratively refine the SVM model in this hierarchical structure, the model trained on the segments with matched duration condition is obtained. In an iteration of training process, the sample and feature selection scheme was used to improve the efficiency of model training. The experiment results on LRE03 and LRE07 show advantage of our proposed schemes.

The outline of this paper is as follows. In section 2, the overview of our proposed language recognition framework is described. Then the data preprocessing of the training data is presented in Section 3. Next, the GMM-SVM algorithm for language recognition, the feature selection and sample selection scheme are detailed in section 4. In section 5, the experiment results on the NIST03 and NIST 07 language evaluation tasks are shown to prove the advantage of our

proposed scheme, followed by the conclusion and the future work in section 6.

2. FRAMEWORK OVERVIEW

The proposed language recognition framework is illustrated in Figure.1. First, the audio files in training set are hierarchically sub-segmented into segments with different durations, i.e. 3 minute, 30, 10 and 3 second. The SDC features are extracted from these speech segments. In the following recognition process, the distribution of each segment is adapted from the UBM model, and the corresponding supervector is used as the features for SVM training and test.

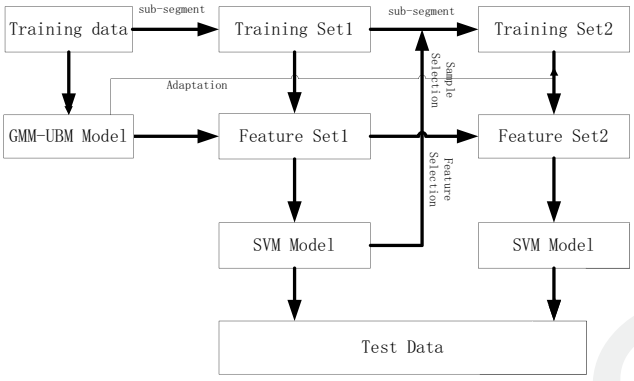


Figure 1: Framework of our Language Recognition system

An iteration of language recognition process consists of three primary steps, including 1) *Model Training*, 2) *Feature Selection* and 3) *Sample Selection*. Given the training set obtained from the 1st-layer speech segments, the SVM classifier is trained for each target language. Then according to the optimized model, the representatives instances are selected, which are used to select the samples from the 2nd-layer speech segments. At the same time, the feature weight is calculated, and the most significant features are selected according to their weights. The process will be iterated several times until the target model, which matches with the test condition, is obtained.

To clarify the idea, we take the NIST 03 and 07 LRE evaluation tasks as the example to present our proposed language recognition process.

3. DATA PRE-PROCESSING

As aforementioned, the mismatch between the training and test data will degrade the performance. In this section, we proposed to form a hierarchical structure of the training data with different durations, as illustrated in Figure 2.

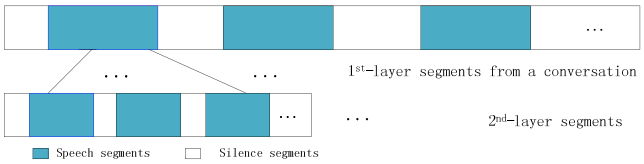


Figure 2: Hierarchical structure of the training data.

The training data we used are provided by NIST. Firstly, an audio file is split into individual conversational sides containing the left and the right channel respectively. An intensity-based speech activity detector (SAD) was applied to both parts. And the audio file is sub-segmented into segments with different durations (i.e. 3 minutes, 30, 10 and 3 second etc.) according to the result of SAD. Short-duration segments are sub-segmented from the long-duration ones. For example, each 3 minute segment can be further divided into the set of segments with 30 second duration.

The acoustic feature sets in our system are 56-D features, including 7-D MFCC coefficient (including coefficient C0) concatenated with SDC feature with N-d-P-k (7-1-3-7), as described in [5]. The RASTA and CMS are used to alleviate channel mismatch. The distribution for each segment is obtained by performing Maximum A Posteriori (MAP) estimation from the UBM with a small relevance factor. Here, only mean vector adaptation is performed during model training, from which the supervector for SVM training is formed.

4. LANGUAGE RECOGNITON BASED UNDER GMM-SVM FRAMEWORK

4.1 GMM-SVM for Language Recognition

An SVM is a two-class classifier constructed from the sum of a kernel function $\mathbf{K}(\mathbf{x},\mathbf{y}) = \mathbf{\Phi}(\mathbf{x}) \cdot \mathbf{\Phi}(\mathbf{y})$, where $\mathbf{\Phi}$ is a nonlinear mapping function. The SVM for classification is as follows

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i t_i K(\mathbf{x}, \mathbf{x}_i) + b = \sum_{i=1}^{n} \alpha_i t_i \mathbf{\Phi}(\mathbf{x}) \cdot \mathbf{\Phi}(\mathbf{x}_i) + b \quad (1)$$

where $\sum_{i=1}^{N} \alpha_i t_i = 0$, $\alpha_i \neq 0$ and $t_i \in \{-1, +1\}$ is the label of the training data. The vectors $\mathbf{\Phi}(\mathbf{x_i})$ are support vectors and obtained from the training set by an optimization process. Since the SVM is a two-class classifier, the language recognition is traditionally treated as a verification problem and the one vs. others strategy is used [9].

In the GMM-SVM framework, the kernel for SVM training is constructed by approximating the KL-distance between the distributions[5]. Suppose that we have a GMM-UBM as follows.

$$g(\mathbf{x}) = \sum_{i=1}^{N} \lambda_i N(\mathbf{x}; \mathbf{m}_i, \mathbf{\Sigma}_i) \quad (2)$$

where λ_i are the mixture weights, $N(\ )$ is a Gaussian distribution, and $\mathbf{m}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance of the Gaussian distribution respectively.

Given two utterances Utt_a and Utt_b, the GMMs g_a and g_b are obtained from the UBM shown in equation (2) using MAP adaptation. The distance between two distributions, g_a and g_b are approximated as

$$d(m^a, m^b) = \sum_{i=1}^{N} \lambda_i (m_i^a - m_i^b)\Sigma_i^{-1}(m_i^a - m_i^b) \qquad (3)$$

And the corresponding kernel function is given as follows,

$$K(Utt_a, Utt_b) = \sum_{i=1}^{N} (\sqrt{\lambda_i}\Sigma_i^{-1/2} m_i^a)(\sqrt{\lambda_i}\Sigma_i^{-1/2} m_i^b)^t = \Phi(Utt_a)\cdot\Phi(Utt_b) \qquad (4)$$

4.2 Feature Selection Scheme

The SVM can be considered as a linear discriminative classifier in the high-dimensional supervector space. Given a SVM classifier, a natural way for selecting the most discriminative features is to use a wrapper method [10]. Suppose that the optimized SVM solution is given in equation (1), which can be reformulated as

$$f(\mathbf{x}) = \mathbf{W}^T\Phi(\mathbf{x}) + b \qquad (5)$$

and

$$\mathbf{W} = \sum_i \alpha_i t_i \Phi(\mathbf{x}_i) \qquad (6)$$

where $\mathbf{W}$ can be considered as the weight of each feature. As mentioned in [10], we can select the most significant features according to the magnitude of the feature weight $|W_i|$.

The experiments results on NIST 03 and NIST 07 LRE tasks show that only a small subset of feature set (about 10%-30%) is needed to obtain the comparable performance. Meanwhile, feature selection can greatly save the training and test time.

4.2 Sample Selection Scheme

Given the training data, the short duration segmentation will created too many training samples, which makes it infeasible to train SVM classifier. We propose a sample selection scheme. In this scheme, the 1st-layer of the hierarchical structure, which contains about 5000 segments with duration about 3 minute, is used as the initial dataset for training SVM classifier. Given the optimized SVM model, there are two schemes to select the samples from the 2nd-layer training data and so on. In our experiments, it is constructed it by sub-segmenting the 1st –layer audio segments.

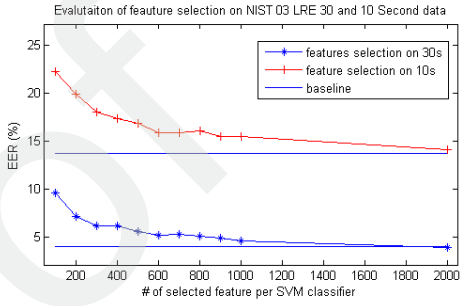
Scheme 1: Given all the 2nd-layer training data, we use the obtained SVM to classification these training data. The samples which are misclassified or close to the separating hyperplane are selected as the training set for next iteration.

Scheme 2: Directly take the support vectors of current SVM as the representative samples, and sub-segment them into the short-duration segments.

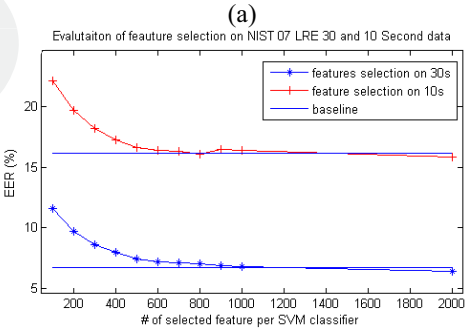
The primary experiments of both schemes give the similar result. However, in scheme 1, the 2nd-layer training samples needs to be processed in advance and the samples are selected according to the classification results, which is time consuming. In experiments, the scheme 2 is taken to select the samples.

5. EXPERIMENTS

In order to evaluate the performance of the proposed language recognition framework, several experiments are conducted on the NIST 03 and NIST 07 LRE 30 and 10 second tasks.



(a)



(b)

Figure 3: Experiment results of different selected features
(a) NIST LRE 03 30, 10 second evaluations
(b) NIST LRE 07 30, 10 second evaluations

The training data set consists of more than 200 hours of telephone speech spanning 14 different languages coming from CallFriend and OHSU. The evaluation data set for NIST 2003 LRE including 3840 utterances for three test conditions. The evaluation data set for NIST 2007 LRE including 7530 utterances spanning the 30, 10 and 3 second conditions. The training data is preprocessed as illustrated in section 3. And to be simplified, only the raw scores are used in performance evaluation.

Experiment 1: Evaluate the performance of feature selection scheme.

In order to evaluate the performance of the feature selection scheme under the GMM-SVM framework, we

conduct the experiments on the training dataset that consists of the segments with duration about 3 minutes. There are about 5000 training samples in total for SVM training.

The experiment results are shown in Figure 3. From Figure 3, we can see that with the increasing number of the selected features for each SVM, the system performance is improved steadily. And with about 2000 features are selected per SVM, which results in about 30000 features, 30% from the total 114688 supervector. The system performance is close to or even better than the baseline.

Experiment 2: Evaluation of the system performance on 30 second training dataset with selected features and selected samples.

In order to further improve the system performance, we applied the samples selection scheme as shown in section 4.2. The informative samples are selected according to the support vectors of each SVM. The experiments on NIST 03 and 07 are illustrated in Table 1.

Year	System	Duration	
		30s	10s
2003	GMM-SVM	4.02	13.75
	Our System	3.33	11.01
2007	GMM-SVM	6.67	16.17
	Our System	5.13	13.7

Table 1: The EER (%) of GMM-SVM and our system on NIST 03 and 07 LRE 30 and 10 second evaluation tasks

There are about 10000 training samples for NIST 03 and 20000 samples for NIST 07 in 2nd-layer of the hierarchical structure. In our experiments, 2000 top-most features are selected per classifier, which results in total 33376 features for NIST 03 and 37405 features for NIST 07. The size of selected training dataset ranges from 3000 to 5000, which results in the similar computational complexity in an iteration of training process. From Table 1, we can see that the system performance is relatively improved about 17%-19% on NIST 03 and 07 30 and 10 second evaluation tasks.

.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel sample and feature selection scheme in the framework of the GMM-SVM. The feature selection aims at finding the most significant features in supervector space and eliminating the irrelevant features. The sample selection scheme exploits the hierarchical structure consisting of the segments with different durations and tries to find a feasible method to tackle the duration mismatch problem under the GMM-SVM framework. The experiment results show the effectiveness of our proposed method.

The future work will be to as follows: 1) find more effective feature selection method by exploring the relationship between the selected features; 2) more efficient sample selection scheme will be studied in the next step. 3) The evaluation of our proposed system on the 3 second duration test data.

7. REFERENCES

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

[2] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in International Conference on Spoken Language Processing, 2002, pp.89–92.

[3] Lukas Burget, Pavel Matejka, and Jan Cernocky, "Discriminative training techniques for acoustic language identification," in Proceedings of ICASSP, 2006, pp. 209–212.

[4] W.M.Campbell, J.P.Campbell, D.A.Reynolds,and A. Solomonoff, "Support Vector Machines for Speaker and Language Recognition," Computer Speech and language, pp.210-219.

[5]Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Acoustic language identification using fast discriminative training," in Proc. Interspeech, 2007.

[6] "The 2007 NIST Language Recognition Evaluation Plan(LRE07)",http://ww.nist.gov/speech/tests/lang/2003/LRE07EvalPlan-v7e.pdf

[7] Patrick Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint Factor Analysis versus Eigenchannels in Speaker Recognition",

[8] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in Proc. ICASSP'05, 2005, pp. I-629–I-632.

[9] Chang, C.-C. and C.-J. Lin (2001). LIBSVM: a library for support vector machines.http://www.csie.ntu.edu.tw/_cjlin/libsvm.

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning,vol. 46, no. 1-3, pp. 389–422, 2002.

[11] Christopher M.Bishop, "Pattern Recognition and Machine Learning", Springer, 2006