Spoken Language Identification Using Score Vector Modeling and Support Vector Machine

Ming Li, Hongbin Suo, Xiao Wu, Ping Lu, Yonghong Yan

Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

{ming.li,suohongbin,wuxiao,luping,yanyonghong}@hccl.ioa.ac.cn

10.21437/Interspeech.2007-82

Abstract

The support vector machine (SVM) framework based on generalized linear discriminate sequence (GLDS) kernel has been shown effective and widely used in language identification tasks. In this paper, in order to compensate the distortions due to inter-speaker variability within the same language and solve the practical limitation of computer memory requested by large database training, multiple speaker group based discriminative classifiers are employed to map the cepstral features of speech utterances into discriminative language characterization score vectors (DLCSV). Furthermore, backend SVM classifiers are used to model the probability distribution of each target language in the DLCSV space and the output scores of backend classifiers are calibrated as the final language recognition scores by a pair-wise posterior probability estimation algorithm. The proposed SVM framework is evaluated on 2003 NIST Language Recognition Evaluation databases, achieving an equal error rate of 4.0% in 30-second tasks, which outperformed the state-of-art SVM system by more than 30% relative error reduction.

Index Terms: spoken language identification, support vector machine, score vector modeling

1. Introduction

The goal for language identification (LID) is to determine the language spoken in a given segment of speech. Approaches using phonotactic information, namely PRLM (phoneme recognizer followed by language models) and PPRLM (parallel PRLM), have been shown quite successful [1, 2]. In PPRLM, a set of tokenizers are used to transcribe the input speech into phoneme strings or lattices [3, 4] which are later scored by n-gram language models. Lately, due to the introduction of shifted-delta-cepstral (SDC) acoustic features, promising results using Gaussian Mixture Model (GMM) were reported [5]. The acoustic approach was further improved by using discriminative Maximum Mutual Information (MMI) training for acoustic modeling [6]. It is generally believed that phonotactic feature and spectral feature provide complementary cues to each other, and therefore in both NIST LRE 2003 and 2005, the best systems were combinations of phonotactic and acoustic recognizers whose outputs were fused together to generate the final scores.

Several recent approaches using SVM [7, 8, 9, 10, 11] have attracted much attention as an alternative solution. SVM as a classifier maps input feature vector into high-dimensional space then separate classes with maximum margin hyperplane. Furthermore, its training criteria balance the reduction of errors on training data and the generalizability of the unseen data which makes it generalizes well with small quantities of trainning data.

In phonotactic modeling, the score vector modeling frameworks that map phoneme characteristics into score vectors based on pair-wise strategy and employ a backend SVM classifier to identify each language has demonstrated superior performance over generative language modeling framework [7, 8, 9].

In acoustic modeling, the SVM classifiers with GLDS kernel have shown very competitive performance in the domain of SDC features [10]. The GLDS kernel is based on an explicit expansion in feature space using a monomial basis which gives a very concise way of storing and scoring target models. Nevertheless, because of the high dimension property of each feature vector and the practical limitations of computer memory, the training samples are limited which is not suit for large database training. Recently, with the desire of LID system to be speaker independent, a set of speaker dependent anchor GMM models [11] were trained on SDC features for every speaker in every language, and backend discriminative SVM classifiers are adopted to identify the spoken language based on the anchor GMM outputs. The results show that it is capable to achieve robust speaker independent language identification by compensating for intra-language and inter-speaker variability. However, for every test speech segment, scoring on these entire anchor-GMM language models are computationally expensive and sufficient training data for each person can not be guaranteed in practical application.

In this paper, we study how to efficiently achieve robust speaker independence in LID system with large training database. We propose to use multiple speaker group based GLDS kernel classifiers to construct the discriminative language characterization score vector (DLCSV) and by fusing multiple scores with different weights in DLCSV space to generate a new scoring function with less speaker dependence. Hence, a backend discriminative classifier followed by posterior probability estimation method is adopted.

The organization of the remainder of this paper is as follows. Section 2 describes the GLDS kernel. Section 3 explains our algorithm in detail. Corpora and evaluation methods are given in section 4. Section 5 presents the performance of proposed method and Section 6 follows with a brief summary.

2. Support vector machine with GLDS kernel

An SVM is a two-class classifier constructed from sums of a kernel function $K(\cdot,\cdot)$:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d \qquad (1)$$

where N is the number of support vectors, t_i is the ideal output, α_i is the weight for the support vectors x_i, $\alpha_i > 0$ and

$\sum_{i=1}^{N} \alpha_i t_i = 0$. The ideal outputs are either 1 or -1, depending upon whether the corresponding support vector belongs to class 0 or class 1. By using kernel functions, SVM can be generalized to non-linear classifiers by mapping the input features into a high dimensional feature space.

The original form of the GLDS kernel [10] involves a polynomial expansion $b(x)$, with monomials (between each combination of vector components) up to a given degree p. The GLDS kernel between two sequences of vectors $X = \{x_t\}_{t=1\cdots N_x}$ and $Y = \{y_t\}_{t=1\cdots N_y}$ is denoted as a rescaled dot product between average expansions:

$$K(X,Y) = \frac{1}{N_x}\sum_{i=1}^{N_x} b(x_i)^t \cdot \overline{R}^{-1} \cdot \frac{1}{N_y}\sum_{j=1}^{N_y} b(y_j) \quad (2)$$

$$= \overline{b_x}^t \cdot \overline{R}^{-1} \cdot \overline{b_y} \quad (3)$$

where $\overline{R}$ is the second moment matrix of polynomial expansions and its diagonal approximation is usually used for more efficiency. In addition, the scoring function of GLDS kernel can be simplified with the following compact technique [10].

$$f(\{x_i\}) = (\sum_{i=1}^{N} \alpha_i t_i \overline{R}^{-1}\overline{b_i} + \overline{d})^t \cdot \overline{b_x} = w^t \cdot \overline{b_x} \quad (4)$$

Where $\overline{b_i}^t$ are the support vectors, $\overline{d}$ is denoted as $[d \quad 0 \cdots 0]^t$.

Therefore, by collapsing all the support vectors down into a single model vector w , each target score can be calculated by just a simple inner product which makes this framework suited for those applications with dozens of target languages and critical request of computation complexity very well. However, due to the high dimension property of expansion and the practical limitations of computer memory, the training samples are limited.

3. The proposed LID system

Our focus is on practical considerations that make SVM-SDC technology more effective. In the proposed method described in figure1, multiple speaker group dependent GLDS kernel SVM classifiers are trained to map the expanded cepstral features of speech segments into DLCSVs, which represent both discriminative language information and inter-speaker variability within the same language. Therefore, backend discriminative SVM classifiers are employed to model the probability distribution of each target language in this DLCSV space. Because backend classifiers' output scores are not log-likelihood values, we finally transform the SVM output scores into posterior probabilities as the final language scores.

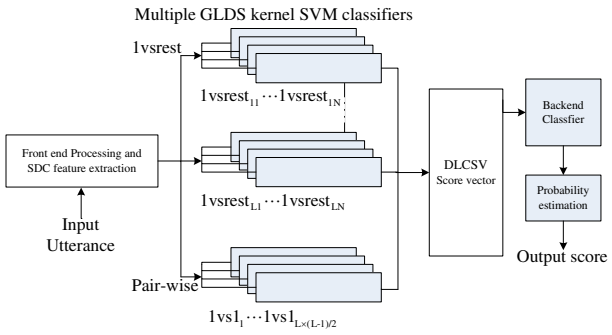

Figure 1: *System overview*

3.1. Features

The features in our system are 7 MFCC coefficients (including coefficient C_0) concatenated with SDC 7-1-3-7 feature, which are in total 56 dimension coefficients per frame. This representation is selected based upon prior excellent results with this choice and the improvement of adding direct coefficients with the C_0 coefficient in this feature vector was studied in [6]. In this paper, SDC feature refers to this 56 dimension feature. After speech activity detection [12], nonspeech frames are eliminated and 56 dimension SDC features are extracted. Then feature warping [13] and cepstral variance normalization are applied on the previously extracted SDC features which results that each feature is normalized to mean 0 and variance 1 on a per-utterance basis.

3.2. Score vector modeling

Score vector modeling approach [7] has been successfully applied to spoken language identification. Each spoken utterance is converted into a feature vector with its attributes representing the statistics of language information, thus a discriminative vector space classifier is built in this score vector space to identify the target language. It is generally agreed upon that the fusion of multiple phonotactic features improves performance. For instance, the PPRLM approach uses parallel recognizers to derive multi phonotactic features.

Recently, SVM classifiers are employed as backend in speaker characterization vectors (SCVs) space to discriminate between the target language's SCVs and the SCVs from the non-target languages. Promising experiment results show that mapping speech segments into SCVs by thousands of speaker-specific anchor-GMM language models can improve the speaker independence of LID systems [11]. However, even using fast score, scoring with thousands of anchor-GMM models is very costly in computation complexity.

In this paper, standard SVM-SDC framework [10] is generalized by employing multiple GLDS kernel classifiers to convert each speech utterance into a score vector, which is the combination of all classifiers' output scores, and represents the extent of matching between each input utterance and all the language specified constraints on the margins. As demonstrated in figure1, both pair-wise and parallel one-versus-the-rest discriminative classifiers are trained based on the GLDS approach. L and N denote the number of target languages and the number of subgroups in each language respectively and the total numbers of GLDS kernel classifiers are $N_{total} = \frac{L\times(L-1)}{2} + L \times N$. CallFriend corpus used for training is extremely large and each SDC feature is explicitly expanded into high-dimension space, thus the training samples are limited for each GLDS classifier. Thereby, we divide each target language data of the CallFriend Corpus into N subgroups, and each of which represent a set of different speakers speaking the same language. By training each group-based one-versus-the-rest or pair-wise classifiers separately, the problem regarding memory limitation is fixed. Moreover, experiments show that significant performance improvement can be obtained by compensating these distortions in the domain of DLCSVs which result from the inter-speaker variability presented by different speaker groups within the same language.

SVMTorch [14] is used to train all these classifiers with GLDS kernel. After discriminative training, multiple classifiers are generated, by combining these multiple classifiers' output scores together, each input speech utterance can be mapped into

a single DLCSV score vector.

A backend radial basis function (RBF) kernel SVM classifier is carried out to discriminate target languages based on the probability distribution in this DLCSV space. The choice of RBF kernel is based on its nonlinear mapping function and small parameters to tune. Furthermore, the linear kernel is a special case of RBF and the sigmoid kernel behaves like RBF for certain parameters [15]. Note that the training data of this backend SVM classifier comes from development data rather than the data used for training GLDS classifiers, and cross validation is employed to select kernel parameters and prevent the overfitting problem.

For testing, after test utterances' DLCSVs are generated, backend SVM classifier estimates the posterior probability of each target language, which is used to calibrate the final outputs.

3.3. Language score calibration

The topic of calibrating confidence scores in the field of multiple-hypothesis language recognition has been studied in [16], and a detail analysis of the information flow and the amount of information delivered to the user through the language recognition system has been performed. We should estimate the posterior probability of each of the M hypotheses and make a maximum-a-posteriori (MAP) decision. In standard SVM-SDC systems [10], log-likelihood ratios (LLR) normalization is applied as a simple backend process and is found to be useful. Suppose $\overline{S} = [S_1 \cdots S_L]^t$ is the vector of L relative log-likelihoods from the L target language for a particular message, and the posterior probabilities for the original hypotheses can be denoted as:

$$P_i = \frac{\pi_i e^{S_i}}{\sum_{j=1}^{L} \pi_j e^{S_j}}, i = 1, 2, \cdots, L \tag{5}$$

where $[\pi_1, \cdots, \pi_L]$ denotes the prior. Considering a flat prior, new log-likelihood normalized score $S_i^{'}$ is denoted as:

$$S_i^{'} = S_i - log(\frac{1}{L-1} \sum_{j \neq i} e^{S_j}) \tag{6}$$

However, the SVM raw scores are not log-likelihood values, thus LDA and diagonal covariance Gaussians are used to calculate the log-likelihoods for each target language [17] and achieve improvement in detection performance [10].

In this paper, we use an alternative approach [18] to estimate the posterior probabilities. Given L classes of data, the goal is to estimate $p_i = p(y = i|x), i = 1, \cdots, L$. In the pair-wise framework, firstly the pair-wise class probabilities are estimated as:

$$r_{ij} = p(y = i | y = i or j, x) \approx \frac{1}{1 + e^{A\hat{f}+B}} \tag{7}$$

where A and B are estimated by minimizing the negative log-likelihood function using known training data and their decision values $\hat{f}$. Then posterior probability p_i can be obtained by optimizing the following problem:

$$min \frac{1}{2} \sum_{i=1}^{L} \sum_{j, j \neq i} (r_{ji} p_i - r_{ij} p_j)^2 \tag{8}$$

$$subject \quad to \sum_{i=1}^{L} p_i = 1, p_i \geq 0 \tag{9}$$

Therefore, the estimated posterior probabilities are applied to performance evaluation. The probability tool of LIBSVM [15] is used in our approach. Experiments show that this pair-wise multi-class probability estimation algorithm is superior over log-likelihood ratios normalization method.

4. Corpora and evaluation

There are 12 target languages in corpora used in this study: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The training data was drawn from the CallFriend corpus (train, development, and test sets) available from the Linguistic Data Consortium (LDC). Each set consists of 20 two-sided conversations from each language, approximately 30 minutes long. Development data was obtained from the 1996 NIST LID development and evaluation sets, and the experiments are done using the NIST LRE 2003 evaluation database. The task of this evaluation was to detect the presence of a hypothesized target language for each test utterance. Test data consisted of speech segments of length 3, 10 and 30 seconds. For each of these durations, 960 true trials and 10560 false trials were generated from the primary evaluation task. Submitted scores are given in the form of Detection Error Tradeoff (DET) curves and equal error rates (ERR).

5. Experiments

In this paper, after front-end processing, 56 dimension SDC features are extracted as in section 3.1 and all monomials up to degree 3 are used in the expansion $b(x)$ which results in an expansion dimension of 32509. The number of target languages L and sub-speaker groups in each language N is 12 and 6, respectively and speakers for these subgroups in each language are selected by K-mean method. Thus, the total number of GLDS SVM classifiers N_{total} is 138. Five types of experiments were conducted to evaluate the performance of each part of proposed methods. Firstly, Score Vector Modeling approach is evaluated by system 1-3. GLCSV-12 denotes that only 12 one-verses-the-rest GLDS classifiers are used to construct the GLCSV space and the duration of each training utterance is 3 minutes, while GLCSV-78 uses both 12 one-verses-the-rest and 66 pair-wise classifiers to map the input speech utterance in GLCSV space. In GLCSV-138 approach, score vector combines multiple classifiers' outputs together, including both 66 one-verse-one and $12 \times 6 = 72$ group based one-verse-the-rest classifiers. The duration of speech segments used for training these 138 classifiers is 30 seconds which is one sixth as long as [10], thus each language is divided into 6 sub-speaker groups to maintain the same training samples for comparison with [10]. Secondly, Log-likelihood normalization and pair-wise posterior probability estimation algorithms are evaluated respectively to calibrate the output language scores. At last, traditional 49 dimension SDC features with the parameters of 7-1-3-7 is replaced by the modified 56 dimension SDC features described in section 2.2 to enhance the capability of language discrimination. Table1 and figure2 demonstrate the equal error rate performance of each of the five systems and comparison with the competitive results of start-of-art SVM systems, GLDS SVM [10] and Anchor GMM [11], is shown in table2.

The experiment results of system 1-3 show that Score Vector Modeling achieves considerable improvement in system performance. There are two basic reasons. Firstly, pair-wise classifiers only need to load training samples from two languages rather than all the twelve target languages, which request less

Table 1: *SVM system results on NIST 2003 30second task.*

SVM-SDC system	**1**	**2**	**3**	**4**	**5**
GLCSV-12	√				
GLCSV-78		√			
GLCSV-138			√	√	√
LLR normalization	√	√	√		
Probability estimation				√	√
SDC (7-1-3-7)	√	√	√	√	
modified SDC (7-1-3-7 + MFCC)					√
EER-NIST03-30s task (%)	7.0	5.9	5.0	4.7	4.0

Table 2: *EER comparison with state-of-art SVM system.*

GLDS SVM	6.1%
Anchor GMM	4.8%
Proposed SVM-SDC system 5	4.0%

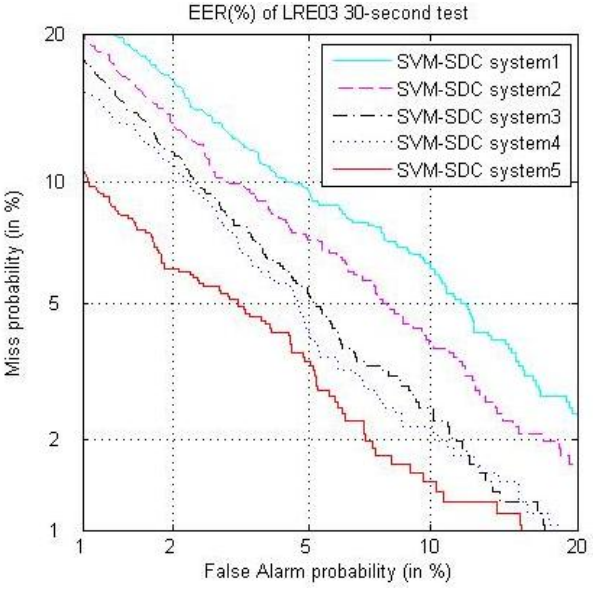

Figure 2: *DET curves on NIST 2003 30second task*

memory and allow training processes to use more samples for each language. Secondly, in GLCSV-138, 12 one-verses-the-rest classifiers are replaced by multiple speaker group based classifiers, which represent both discriminative language information and inter-speaker variability within the same language. By using backend classifiers, this speaker group specified variability can be compensated and make system less speaker dependency.

Further more, SDC feature concatenated with MFCC coefficients achieves significant improvement demonstrated in system 5. The results show that this new SDC feature is also effective in SVM system as well as GMM system[6].

Table1 also shows that the pair-wise posterior probability estimation method adopted in system 4 is comparable to the common employed LLR approach. Because the output scores of backend classifiers are not real log-likelihood values, this alternative language score calibration methods can perform better.

6. Conclusion

In this paper, multiple speaker group based discriminative classifiers are employed to map the speech utterance into DLCSV space efficiently, which represents enhanced language information as well as compensates for intra-language and inter-speaker variability. We applied this new approach to the NIST 2003 language evaluation, and experiment results demonstrate significant improvement by mapping speech segment into this DLCSV score space. Additionally, both modified SDC feature extraction and pair-wise posterior probability estimation methods are proposed to further improve system's performance.

7. References

[1] Zissman, M.A., "Language identification using phoneme recognition and phonotactic language modeling",*Proc. ICASSP*, 1995.

[2] Yan, Y., Barnard, E., "An approach to automatic language identification based on language dependent phone recognition",*Proc. ICASSP*, 1995.

[3] Gauvain, J.L., Messaoudi, A., Schwenk, H., "language recognition using phone lattices",*Proc. ICSLP*, 2004.

[4] Shen, W., Campbell, W., Gleason, T., Reynolds, D., Singer, E., "experiments with lattice-based PPRLM Language identification",*Proc. ODYSSEY*, 2006.

[5] Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, R.A., Deller, J.R., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features",*Proc. ICSLP*, 2002.

[6] Burqet, L., Metejka, P., Cernocky, J., "Discriminative Training Techniques for Acoustic Language Identification",*Proc. ICASSP*, 2006.

[7] Li, H., Ma, B., Lee, C.-H., "A Vector Space Modeling Approach to Spoken Language Identification", *IEEE Trans. Speech and Audio Proc.*, vol.15, pp.271-284, 2007.

[8] White, C., Shafran, I., Gauvain, J.L., "discriminative classifiers for language recognition",*Proc. ICASSP*, 2006.

[9] Zhai, L.F., Siu, M.H., Yang, X., Gish, H., "Discriminatively trained Language Models using Support Vector Machines for Language Identification",*Proc. ODYSSEY*, 2006.

[10] Campbell, W.M., Campbell, J.P.,Reynolds, D.A., Singer, E., "Support vector machines for speaker and language recognition", *Computer Speech & Language*, Vol.20, pp.210-229, 2006.

[11] Noor,E., Aronowitz, H., "efficient language identification using anchor models and support vector machines",*Proc. ODYSSEY*, 2006.

[12] Guo, Y.M., Fu, Q., Yan, Y., "Speech endpoint detection based on sub-band energy and harmonic structure of voice",*Proc. ISCSLP*, 2006.

[13] Allen, F., Ambikairajah, E., Epps, J., "Warped magnitude and phase-based features for language identification",*Proc. ICASSP*, 2006.

[14] Collobert, R., and Bengio, S., "SVMTorch: support vector machines for large-scale regression problems",*Journal of Machine Learning Research*, vol.1, pp.143-160, 2001.

[15] Chang, C.C., Lin, C.J., "LIBSVM : a library for support vector machines",2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[16] Brummer, N., van Leeuwen, D.A., "On Calibration of language recognition scores",*Proc. ODYSSEY*, 2006.

[17] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A., "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification",*Proc. EUROSPEECH*, 2003.

[18] Wu, T.-F., Lin, C.-J., Weng, R.C., "Probability Estimates for Multi-class Classification by Pairwise Coupling",*Journal of Machine Learning Research*, 2004.