A CMLLR Supervector Kernel for SVM Language Recognition

Shan Zhong, Jia Liu
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing, China
zhongshan00@mails.thu.edu.cn, liuj@tsinghua.edu.cn

ABSTRACT
This paper explores the use of constrained maximum likelihood linear regression (CMLLR) transforms as features for language recognition. Modeling is carried out through support vector machine (SVM). This work proposes a novel CMLLR supervector kernel. Results on the NIST LRE09 task show that feature-domain CMLLR transforms contain more language dependent information than model-domain MLLRs, and the proposed CMLLR supervector kernel outperforms some other ones. We also compare our CMLLR-SVM system with some state-of-the-art systems, and combine them for a further improvement.

Index Terms—language recognition; CMLLR; SVM

1. INTRODUCTION
Recent studies show that different levels of speech features contain specific language information including acoustic parameters, prosody, phonotactic information, and lexical knowledge [1]. The most widely used features are phonotactic and spectral features, which are generally believed to provide complementary language cues to each other [2]. Phonotactic features capture the lexical constraint of admissible phonetic combination in a language. One typical implementation is parallel phone recognizer followed by language modeling (PPRLM) which employs several tokenizers to transcribe the input speech into phonemes and performs scores on phoneme strings or lattices using n-gram language model [3]. The spectral features of speech are collected as independent vectors. The collection of vectors can be extracted as shifted delta cepstral (SDC) features or Gaussian supervector (GSV) features, and then modeled by a Gaussian mixture model (GMM) or a support vector machine (SVM) to form GMM-SDC [4] and GSV-SVM systems [5]. Because different recognition methods contain different hierarchical language cues, these are often considered to work together to get the best performance [1].

In our previous work, we introduce maximum likelihood linear regression (MLLR) as features for language recognition task [6]. The transforms have been successful used as features for speaker recognition, achieving performance competitive to state-of-the-art cepstral systems [7-8]. Our study showed that the MLLR transforms also contain language dependent information, and can be used for language recognition purpose. In this paper, we further research on using constrained MLLR (CMLLR) transforms for language recognition and construct a suitable framework, which uses CMLLR to adapt the means and covariances of language dependent GMM UBMs to a given utterance and then uses the entries of the transform as features. Because the choice of kernels is the core of an SVM system, we also propose a novel CMLLR supervector (CMLLRSV) kernel.

The paper is organized as follow. In Section 2 we give a brief overview of CMLLR and MLLR supervector (MLLRSV) kernel. Section 3 introduces the framework of CMLLR-SVM system and the new CMLLR supervector (CMLLRSV) kernel. Experimental results are shown in Section 4 followed by the conclusion given in Section 5.

2. BACKGROUND

A. CMLLR
MLLR is the most commonly used approach for adaptation with limited adaptation data. The fundamental idea of MLLR is to estimate affine transforms to represent the transformation relation between a speaker independent modeling space and that of a specific speaker with ML estimation,

\[
\hat{\mu} = A\hat{\mu} + \hat{b} = W\xi
\]

where \(\hat{\mu}\) and \(\hat{\mu}\) represent the mean vectors before and after adaptation, \(A\) and \(b\) stand for the transform parameters using ML criterion. \(\xi\) is the extended mean vector \([\mu^T, b^T]\) and \(W\) represents the extended transform matrix \([\xi^T, A]\) estimated using the expectation maximization (EM) algorithm.

In theory MLLR can be used to implement adaptation through normalization by applying the regression matrix to the input feature vectors instead of the mixture component means. In [9] a constrained variant of MLLR (CMLLR) was proposed. In this formulation, it is assumed that mean and covariance parameters are governed by one transforms as follows:

\[
\hat{\mu} = A^T\mu + \hat{b}^T
\]

\[
\hat{\Sigma} = A^T\Sigma A
\]

where \(A^T = A^{-1}\) and \(b^T = -Ab\). The CMLLR parameters are estimated using a procedure similar to that used for mean-only MLLR parameter estimation, but are applied in the feature domain as \(\hat{o}(t) = A\hat{o}(t) + \hat{b} = W\xi\). where \(\xi = [o_1 o_2 \ldots o_T]^T\) is the extended observation vector at time \(t\). Although this derivation follows similar arguments to the standard MLLR transform, the nature of the transformation generated is significantly different. Instead of generating a transform which relocates the models in the acoustic space (transformation of the mean
vectors), the observation vectors are transformed in an attempt to locate the appropriate vectors in the correct region of the acoustic space.

B. MLLR supervector (MLLRSV) kernel

At the most basic level, SVM is a two-class hyperplane based classifier operating in a high-dimensional space related nonlinearly to the original input space. Given an observation $x \in X$ and a kernel function $K$, an SVM function $f(x)$ is

$$f(x) = \sum_{i=1}^{N} \alpha_i \rho_i K(x, x_i) + d = \sum_{i=1}^{N} \alpha_i \rho_i \phi(x) \cdot \phi(x_i) + d.$$  \hfill (3)

$K(x, y)$ are an inner product expressible as $\phi(x) \cdot \phi(y)$ where $\phi : x \rightarrow y \in Y$ for some expansion space $Y$. We compare the output of the SVM in (3) to a threshold in order to produce a decision. The $x_i, \rho_i,$ and $\alpha_i$ are obtained through a training process. The $x_i$ are called support vectors and the $\rho_i$ are the target class values: +1 for in-class and -1 for out-of-class.

For language recognition, the goal is to determine the language of an utterance from a set of known languages. For two utterances, $utt_1$ and $utt_2$, we adapt the GMM UBM, via MLLR adaptation of the means, to obtain two new GMMs, resulting in GMM supervectors, $\mu_1$ and $\mu_2$.

The mixture function of GMM is $\psi(\omega_n, \mu_n, \Sigma_n)$, where $n = 1 \sim N$ and covariance is diagonal. So the GMM supervector (GSV) kernel can be summarized as [10],

$$K_{GSV}(utt_1,utt_2) = \sum_{n=1}^{N} \left( \omega_n \Sigma_n^{-1/2} \mu_{n,utt_1} \right)^T \left( \omega_n \Sigma_n^{-1/2} \mu_{n,utt_2} \right).$$  \hfill (4)

Replacing the adapted means with the affine transform of the UBM means, the MLLRSV kernel can be rewritten as [11],

$$K_{MLLRSV}(utt_1,utt_2) = \sum_{n=1}^{N} \left( \omega_n \Sigma_n^{-1/2} \mu_{n,utt_1} \right)^T \left( \omega_n \Sigma_n^{-1/2} \mu_{n,utt_2} \right)$$

$$= \sum_{n=1}^{N} \left( \Delta_n^{-1/2} \left( \mu_n + b \right) \right)^T \left( \Delta_n^{-1/2} \left( C \mu_n + d \right) \right)$$

$$= \sum_{n=1}^{N} \left( \Delta_n^{-1/2} A \mu_n \right)^T \left( \Delta_n^{-1/2} C \mu_n \right) + \sum_{n=1}^{N} \left( \Delta_n^{-1/2} A \mu_n \right)^T \left( \Delta_n^{-1/2} d \right)$$

$$+ \tau \cdot Q \cdot \tau^T$$

where $\tau$ is the MLLRSV, $\Delta_n = \omega_n \Sigma_n^{-1}$ and $Q$ is a block diagonal matrix consisting of M blocks $B_k$ of size $(M + I) \times (M + I)$. $B_k = \begin{pmatrix} r_k & r_k \end{pmatrix}$, where $r_k = \sum_{n=1}^{N} \Delta_{nk} \mu_n \mu_n^T$, $r_k = \sum_{n=1}^{N} \Delta_{nk} \mu_n$ and $\delta_k = \sum_{n=1}^{N} \Delta_{nk}$.

3. CMLLR-SVM

A. CMLLR-SVM construction

Generally speaking, there are a lot of target speakers with little training data per speaker for speaker recognition task. However, language recognition is characterized by the completely opposite: little target languages with abundant training data. According to this characteristic, we train language-dependent GMMs separately.

One of the main advantages of CMLLR is that the transforms can be integrated used as a compensation technique and a new feature for language recognition. So we propose to use CMLLR in two steps. As a first step, CMLLR is used as a feature domain compensation technique that attempt to compensate for speaker and channel variability by moving acoustic features closer to existing GMM models with the use of linear transforms. As shown in Fig. 1, CMLLR transforms are applied to training utterances and a new language-dependent GMM is constructed using the transformed features. This new model is used as the initial model for EM training of the GMMs. With this process, it is possible to iterate the CMLLR transform estimation using the language-dependent model from prior iteration as the adaptation model similar to iterative SAT used in ASR training. In this paper we use a single iterative CMLLR estimation.

As a second step, CMLLRs are estimated by using GMMs and eventually rearranged as vectors to be modeled by SVM classifier for training and testing after normalization. The multi-class CMLLR supervector not only contains the specific information of its own language type, but also contains the difference between other languages. Because such supervectors include more discriminative information between different languages, their classification performance is better than the supervectors based on a single UBM. The CMLLR-SVM framework in Fig. 2 is summarized as follow:

1) To $L$ languages, train $GMM_i'$ for each language with $N$ Gaussian mixtures using maximum mutual information (MMI) criterion.
2) To every speech segment, calculate CMLLR transform on corresponding $GMM_i'$ with EM iterative algorithm.
3) CMLLR transform $W_i'$ are estimated on all GMMs, and then combined to form supervectors $\hat{W}$ after normalization. Because different feature normalization methods influence the performance of SVM classification, we use the rank normalization method following [12] which is reported to be very effective in the linear transform based SVM framework.
4) \( \tilde{W} \) are then used as features for training and testing in an SVM classifier.

The framework above is easily extended to multi-class case by training several GMMs for each language. Each GMM generates a corresponding adaptation transform. All the ordinal transforms are then combined to form the supervector feature for every input utterance.

The main advantage of (9) over (6) is that the number of multiplications it requires only depends on the small size of the GMM feature vector and not on the large number of GMM mixtures. And \( \tilde{\xi}_{jkl} \) only depends on the UBM, so it can be calculated in advance.

4. EXPERIMENTS

A. Speech corpus

All data used for the experiments are recorded over telephone lines and radio broadcasts. Test data came from National Institute of Standards and Technology (NIST) language recognition evaluation (LRE) 2009. There are 23 target languages: Amharic, Bosnian, Cantonese, Creole (Haitian), Croatian, Dari, English (American) English (Indian) Farsi, French, Georgian, Hausa, Hindi, Korean, Mandarin, Pashto Portuguese, Russian, Spanish Turkish, Ukrainian, Urdu and Vietnamese, and also some outset test segments. Each language consists of test segments in 3 length groups: 30, 10, and 3 seconds. The training and develop data consist of Voice of America (VOA) radio broadcasts and conversational telephone speech (CTS). The CTS data come from CallFriend, CallHome, OHSU, LRE07Train, OGL, and the VOA data mainly come from VOA3. The CTS data only contain 13 target languages, so the training data are not balanced. The development data consist of lid03e1, lid05d1, lid05e1, lid07e1, VOA data with NIST annotations and some VOA3 data without NIST annotations.

B. Results

Equal error rate (EER) was used to summarize all the results below. The first experiment was designed to evaluate the performance of CMLLRSV kernel, comparing with the linear kernel and polynomial kernel. The results in table 1 showed that CMLLRSV kernel was much better than the linear and polynomial kernels. The fact indicated that the choice of SVM feature expansion and an associated choice of kernel were very important.

**Table 1. EER (%) for CMLLRSV, Linear, and Polynomial Kernels**

<table>
<thead>
<tr>
<th>Kernel</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>5.35</td>
<td>7.85</td>
<td>18.29</td>
</tr>
<tr>
<td>Polynomial</td>
<td>5.58</td>
<td>8.07</td>
<td>19.11</td>
</tr>
<tr>
<td>CMLLRSV</td>
<td>3.23</td>
<td>7.64</td>
<td>16.62</td>
</tr>
</tbody>
</table>

We then compared the performance of MLLR and CMLLRSV features using corresponding associated kernels. It is interesting that the results in Table II imply completely opposite conclusion to the use of CMLLR and...
MLLR features for speaker recognition. Reference [7] showed that MLLR features were much more effective than CMLLRs in speaker recognition. But in language recognition task, it seemed that feature-domain CMLLR transforms contained more language dependent information than model-domain MLLRs.

The last experiment was designed to compare our (C)MLLR-SVM systems with GMM-SDC, PPR-BT and GSV-SVM in Table II. It showed the fusion results for each individual systems and all-combination system. CMLLR-SVM was competitive with the other individual systems (a, b, c and d) in terms of EER. Both GMM-SDC and GSV-SVM worked better than CMLLR-SVM at 30s test duration though. The advantage of CMLLR-SVM was demonstrated in short-time test conditions. CMLLR-SVM system leaved EER at the same level with other two individual systems at 10s and 3s test duration. Including MLLR-SVM and CMLLR-SVM in final fusion system brought further improvement over the baseline in EER. The gain indicated that (C)MLLR-SVM are complementary with the other two systems.

### Table 1. EER (%) for the Individual Baseline and All-Combination Systems

<table>
<thead>
<tr>
<th>System</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-SDC (a)</td>
<td>5.2</td>
<td>8.61</td>
<td>18.9</td>
</tr>
<tr>
<td>GSV-SVM (b)</td>
<td>4.68</td>
<td>9.00</td>
<td>19.99</td>
</tr>
<tr>
<td>PPR-BT (c)</td>
<td>5.23</td>
<td>10.40</td>
<td>24.73</td>
</tr>
<tr>
<td>PPR-VSM (d)</td>
<td>3.84</td>
<td>8.67</td>
<td>27.50</td>
</tr>
<tr>
<td>MLLR-SVM (e)</td>
<td>8.63</td>
<td>9.14</td>
<td>20.25</td>
</tr>
<tr>
<td>CMLLR-SVM (f)</td>
<td>4.72</td>
<td>7.58</td>
<td>17.04</td>
</tr>
</tbody>
</table>

| a+b+c+d | 2.95 | 5.28 | 16.44 |
| a+b+c+d+e+f | 2.54 | 4.75 | 13.89 |

Figure 3. DET curve for the individual systems and the combination.

### 5. Conclusion

This paper introduced a new feature extraction method for language recognition based on CMLLR transforms which were used for adaptation originally, and also proposed a novel CMLLR-SV kernel. Results on the NIST LRE09 corpus showed that feature-domain CMLLR was more effective than model-domain MLLR for language recognition purpose. The proposed CMLLR-SV kernel was much better than linear and polynomial kernels. We compared (C)MLLR-SVM systems with state-of-the-art systems, and combined them for a further improvement. The system comparison and fusion experiment indicated that our (C)MLLR-SVM systems were comparable and complementary with other systems.

### 6. Acknowledgment

This work is supported by National Natural Science Foundation of P. R. China and Microsoft Research Asia (60776800), National High Technology Research and Development Program of China (863 Program) (2006AA010101,2007AA042223, 2008AA022Z414).

### 7. References


