IMPROVED GMM-BASED LANGUAGE RECOGNITION USING CONSTRAINED MLLR TRANSFORMS

Wade Shen and Douglas Reynolds

MIT/Lincoln Laboratory
Information Systems and Technology Group
244 Wood Street
Lexington, MA 02420, USA
{swade,dar}@ll.mit.edu

ABSTRACT

In this paper we describe the application of a feature-space transform based on constrained maximum likelihood linear regression for unsupervised compensation of channel and speaker variability to the language recognition problem. We show that use of such transforms can improve baseline GMM-based language recognition performance on the 2005 NIST Language Recognition Evaluation (LRE05) task by 38%. Furthermore, gains from CMLLR are additive with other modeling enhancements such as vocal tract length normalization (VTLN). Further improvement is obtained using discriminative training, and it is shown that a system using only CMLLR adaption produces state-of-the-art accuracy with decreased test-time computational cost than systems using VTLN.

Index Terms— Language Recognition, LID, GMM, Adaptation, Maximum Likelihood Linear Regression, MMI

1. INTRODUCTION

The use of spectral features for language recognition has proven to be successful using a number of different modeling paradigms [1][2][3]. To a greater or lesser degree, all of these modeling paradigms are sensitive to acoustic variability that arise from speaker, gender, session or channel differences. A number of supervised methods have been proposed in the speaker and language recognition literature for compensation of channel and session variation [4][5][6]. Unsupervised methods such as vocal tract length normalization (VTLN) and maximum likelihood linear regression (MLLR) for speaker, gender and channel compensation of Gaussian models have also been successfully applied for speech recognition [7][8].

In [9], we found that the use of MLLR and VTLN techniques could also be applied to improve the performance of phonotactic LID systems by regularizing the token sequences/lattices generated by these systems. In this paper, we describe the application of a feature-space implementation of constrained MLLR (CMLLR) to GMM-based spectral language recognition. We report on experiments combining these methods with other advanced modeling methods including discriminative training (based on the Maximum Mutual Information (MMI) criterion) and VTLN (successfully applied in [11]). Our results suggest that the use of unsupervised CMLLR can significantly improve LID performance and that the gains from CMLLR are complementary to VTLN. Applying discriminative training appears to reduce much of this additive gain, but a CMLLR-only system with MMI training produces state-of-the-art accuracy on the 30 second LRE05 test.

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. GMM-BASED LANGUAGE RECOGNITION

In a GMM-based language recognition system, each language to be recognized is modeled by an M-th order GMM with parameters $\lambda_l = \{w_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}, i = 1, \ldots, M$. The model parameters for language l are estimated using spectral based features, $O = \{o_1, \ldots, o_T\}$, extracted from a collection of speech utterances spoken in language l. During recognition, the language model likelihoods for a test utterance, $p(O|\lambda_l) = \prod_t p(o_t|\lambda_l)$, are used to form a likelihood ratio score from which a decision can be made to accept or reject the hypothesis that the utterance was spoken in a particular language k:

$$LR(k|O) = \left(\frac{p(O|\lambda_k)}{\sum_{l \neq k} p(O|\lambda_l)} \right)^{\frac{1}{T}} \underset{\text{reject}}{\overset{\text{accept}}{\gtrless}} \theta \qquad (1)$$

Often, a backend fuser, trained with development data, is used to calibrate and combine the likelihood scores from the language models [3]. To maintain focus on compensation of the GMM features and models, we will not be using a backend fuser in this paper.

The GMM-based language recognition system operates by capturing the underlying sound classes as reflected in the spectral feature distributions for each language. Thus these systems are susceptible to feature variability due to non-language factors, such as speaker and channel. Utilizing compensation techniques from the areas of speech and speaker recognition, GMM LID systems can produce robust and accurate performance. A number of enhancements make this possible:

1. Use of LID-specific features (e.g. Shifted Delta Cepstra [2]).

2. Application of speaker and gender normalization techniques [11][12][13]

3. Application of discriminative training methods [10][11]

In this paper, we propose the use of CMLLR transforms that attempt to compensate for speaker and channel variability by moving acoustic features closer to existing GMM models with the use of linear transforms. This differs from other methods like Channel Factors compensation that have been successfully applied to speaker recognition (see [12]) that attempt to project out feature subspaces associated with channel and session factors.

3. MAXIMUM LIKELIHOOD LINEAR REGRESSION

As commonly used in automatic speech recognition (ASR) for speaker/channel adaptation, Maximum Likelihood Linear Regression (MLLR) applies a linear transform to model parameters estimated on a per utterance/speaker basis so as to maximize the likelihood of the transformed model given the utterance [14]. In its simplest form, MLLR can be applied to the Gaussian mean parameters of the GMM model. A linear transform $\mathbf{W}$ is applied so as to shift and rotate each Gaussian component of the model, with covariance parameters left unaltered. Different transforms can be applied to individual Gaussians, or classes of Gaussians [15].

The MLLR transform applied to the Gaussian mean vector $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} = \mathbf{A_r}\boldsymbol{\mu} + \mathbf{b_r} = \mathbf{W_r}\boldsymbol{\xi} \tag{2}$$

where $\boldsymbol{\xi} = [1\ \mu_1\ \mu_2\ \ldots\ \mu_n]^T$, n is the dimensionality of the observation features, and $\mathbf{W_r} = [\mathbf{b_r}\ \mathbf{A_r}]$ is the transform for Gaussians of class r. $\mathbf{W_r}$ is found using the EM algorithm [15].

In [7] 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:

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \mathbf{A'_r}\boldsymbol{\mu} + \mathbf{b'_r} \\ \hat{\boldsymbol{\Sigma}} &= \mathbf{A'_r}\boldsymbol{\Sigma}\mathbf{A'^T_r} \end{aligned} \tag{3}$$

where $\mathbf{A'_r} = \mathbf{A_r^{-1}}$ and $\mathbf{b'_r} = -\mathbf{A_r}\mathbf{b_r}$. The CMLLR parameters are estimated using a procedure similar to that used for mean-only MLLR parameter estimation [15], efficiently applied in the feature domain as $\mathbf{o}(t) = \mathbf{A_r}\hat{\mathbf{o}}(t) + \mathbf{b_r} = \mathbf{W_r}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta} = [1\ o_1\ o_2\ \ldots\ o_n]^T$.

For GMM LID, we propose to use CMLLR as a feature domain compensation technique that can be applied to both the training and testing data. CMLLR requires a "target" model (the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in Equation 3) to which the features are adapted. While it is possible to compute language-model-specific CMLLR transformed features, this would require two recognition passes for each model and could produce language specific score biases and scales. Instead, we use a language-independent GMM, trained using a pooling of portions of training data from all the language models in the system, as the CMLLR target, thus requiring only one CMLLR transform for each utterance. During training, CMLLR transforms are applied to all utterances and a new language-independent GMM is constructed using the transformed features. This new model is used as the initial model for EM training of the language-dependent GMMs. With this process, it is possible to iterate the CMLLR transform estimation using the language-independent model from each prior iteration as the adaptation model[1]. During recognition, a CMLLR transform is applied to the test features that are then used to compute the model likelihoods.

4. EXPERIMENTS

4.1. Data

The results reported below were obtained on the NIST LRE05 evaluation set and protocol[2]. The primary LRE05 data covers 7 languages[3] drawn from a previously unexposed corpus collected by OHSU. Performance, in terms of the language-weighted equal error rate (EER), is reported on the 30 second test set of 2,413 utterances.

Models were trained using a cross-corpus training set with data drawn primarily from LDC's CALLFRIEND corpus (1996 train partition). Additional data from the OGI-22 multilingual telephone and Foreign Accented English corpora was also used for training. In total, the training set included data from 81 hours of speech from seven languages with 7-19 hours per language.

All audio files from training and evaluation were preprocessed to remove silence using a GMM-based speech activity detector.

4.2. Recognizer System Configurations

4.2.1. Shifted Delta Cepstra Feature Extraction

For these experiments we used Shifted Delta Cepstra (SDC) features extracted every 10ms in a 7-1-3-7 configuration [2] with 7 static cepstra (including c0) appended [11]. The base MFCC cepstra were extracted over the 300-3100 Hz band, passed through a RASTA filter and warped to a $N(0,1)$ Gaussian over a 3 second moving window.

4.2.2. Vocal Tract Length Normalization

We use VTLN implemented as a maximum-likelihood grid search over warping factors α. In our implementation α is used to modify the mel-scaling used to compute filter bank centers as follow:

$$f_{mel} = 2595 * log_{10}(1.0 + \frac{f}{700\alpha}) \tag{4}$$

Twenty discrete (and equally spaced) αs ranging from 0.75 to 1.25 are scored for each train/test utterance and the maximum likelihood α is chosen.

[1] This is similar to iterative SAT used in ASR training. In this paper we use single iteration CMLLR estimation.

[2] See http://www.nist.gov/speech/tests/lang/2005/ for LRE05 details

[3] English, Hindi, Japanese, Korean, Mandarin, Spanish, and Tamil

4.2.3. Model Training Methods

We tested models trained under both ML and MMI criteria. For ML training, models of mixture order 512 were trained per language with 10 EM iterations starting from a common, language-independent initial model. For MMI training, models were trained from existing ML models using 15 iterations of the extended Baum-Welch algorithm with a learning rate of $E = 1$ and a posterior exponent of $K = 6$ [11]. MMI statistics were computed over speech segments > 2 seconds.

4.3. Results

In Table 1 we show the results for various combinations of VTLN and CMLLR compensations applied to train and test data for ML trained models. Full DET plots for the systems with EERs in boldface in the table are shown in Figure 1. A few observations can be made from these results:

1. Application of VTLN and/or CMLLR significantly improves baseline performance.
2. CMLLR alone produces better results than VTLN alone.
3. Further gains are obtained from joint application of VTLN and CMLLR.
4. For both VTLN and CMLLR, performance gains over baseline can be obtained with application in either train or test data alone.
5. Most gain comes form application of VTLN and/or CMLLR to train data.

We note that gains from training alone are particularly important for applications that require maximum processing speed during recognition. The performance of such systems can be improved with train time only compensation. Further, the use of the less computationally intensive CMLLR appears to provide better performance gains compared to VTLN.

As with SAT training in ASR, the application of CMLLR seems to regularize data during training, allowing for better modeling by GMMs. This may be due to corpus-specific channel conditions that exist in our cross-corpus training set.

Next we applied MMI training to three of the ML models: no compensation, VTLN, CMLLR and VTLN+CMLLR. Results for recognition with these models for combinations of VTLN and CMLLR applied to test data is shown in Table 2. The full DET curves for the three systems with boldface EERs are shown in Figure 2.

For no compensation, VTLN applied in train and test, and CMLLR applied in train and test, the MMI reduces the EER by >50% relative when compared to the corresponding ML models. The combination of VTLN and CMLLR in train and test has a smaller relative reduction of 34% and the additive gain from both VTLN and CMLLR seen with ML models is erased. After MMI training, performance with no compensation applied to test data is similar for all train time compensations. The application of any test time compensation with the MMI models are also similar, though CMLLR results are slightly better (but not significantly).

Table 1. Results for 30 second test duration with ML Trained GMMs (512 mixtures)

Train		*Test*		EER (%)
VTLN	CMLLR	VTLN	CMLLR	
–	–	–	–	**14.0**
√	–	√	–	**10.2**
√	–	–	–	11.3
–	–	√	–	11.4
–	√	–	√	**8.6**
–	√	–	–	10.8
–	–	–	√	11.3
√	√	√	√	**6.8**
√	√	–	–	8.6
–	–	√	√	11.5
√	√	√	–	7.5
√	√	–	√	7.2

Table 2. Results for 30 second test duration with MMI Trained GMMs (512 mixtures, 15 iterations)

Train		*Test*		EER (%)
VTLN	CMLLR	VTLN	CMLLR	
–	–	–	–	**6.9**
√	–	–	–	4.9
–	√	–	–	**5.0**
√	√	–	–	5.0
√	–	√	–	4.6
–	√	–	√	**4.2**
√	√	√	√	4.5

5. CONCLUSION

We have shown that feature-domain CMLLR transforms provide an effective speaker and channel compensation for GMM-based language recognition systems. With MMI training and CMLLR compensation, we demonstrated a 4.2% EER on the LRE05 30 second test. We show that that CMLLR-only compensation is more effective than VTLN-only compensation and is computationally more efficient. Surprisingly, compensations (both VTLN and CMLLR) provides performance gains when applied to just train or test data exclusively, with more gain coming from train data compensation. This can be important for applications in which recognition is computational constrained. Future work will examine using CMLLR compensated features for other LID and SID classifiers such as SVM-GLDS and SVM-GMM as well as combinations with NAP and LFA speaker/channel compensations.

6. REFERENCES

[1] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. on SAP*, 4(1), Jan. 1996.

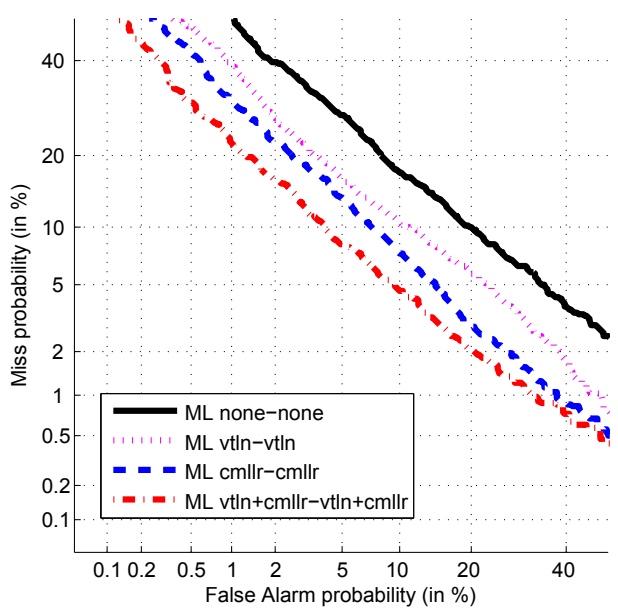

Fig. 1. *DET curves for ML trained GMM systems with VTLN and CMLLR applied in various combinations (LRE05 30 second Primary)*

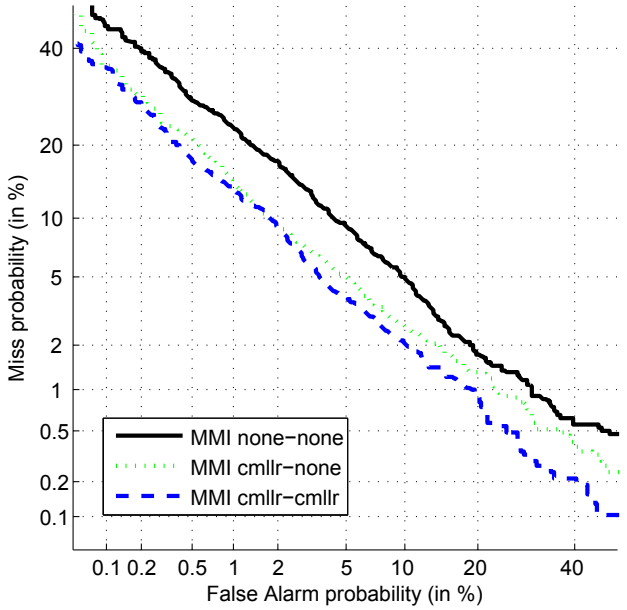

Fig. 2. *DET curves for MMI trained GMM systems with no compensation and CMLLR applied in train only and train and test (LRE05 30 second Primary)*

[2] P. A. Torres-Carrasquillo, et al, "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features," *In Proc. of ICSLP*, September 2002.

[3] E. Singer et al., "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification," *In Proc. Eurospeech*, 2003.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," *In Proc. of ICASSP*, 2005.

[5] A. Solomonoff, C. Quillen, I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," In Proc. of ICASSP, 2005/

[6] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class Covariance Normalization for SVM-based Speaker Recognition," Proc. of ICSLP, Pittsburgh, PA, 2006.

[7] V. V. Digalakis, D. Rtischev, L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on SAP*, 1995.

[8] S. Wegmann, D. McAllester, J. Orloff, B. Peskin, Speaker Normalization On Conversational Telephone Speech, *In Proc. of ICASSP*, 1996.

[9] W. Shen and D. Reynolds, "Improving Phonotactic Language Recognition with Acoustic Adaptation," *In Proc of Interspeech*, 2007.

[10] Q. Dan and W. Bingxi, "Discriminative training of GMM for language identification," *ISCA and IEEE SSPR*, 2003.

[11] P. Matějka Pavel, L. Burget, P. Schwarz, J. Černock'y, "Brno University of Technology System for NIST 2005 Language Recognition Evaluation", *In Proc. of Odyssey 2006*, San Juan, PR, 2006, p. 57-64

[12] C. Vair et al., "Channel Factors Compensation in Model and Feature Domain for Speaker Recognition," *Odyssey Speaker and Language Recognition Workshop*, 2006.

[13] F. Castaldo et al., "Language Identification using Acoustic Models and Speaker Compensated Cepstral-Time Matrices," *In Proc. of ICASSP,* 2007.

[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, 1995.

[15] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech and Language*, 1998.

[16] T. Anastaskos et al., "A Compact Model for Speaker Adaptive Training," *in Proc. of ICSLP*, 1996.