Dialect Recognition using Adapted Phonetic Models

Wade Shen, Nancy Chen and Douglas Reynolds

MIT/Lincoln Laboratory
244 Wood St.
Lexington, MA 02420
{swade, nancy.chen, dar}@ll.mit.edu

Abstract

In this paper, we introduce a dialect recognition method that makes use of phonetic models adapted per dialect without phonetically labeled data. We show that this method can be implemented efficiently within an existing PRLM system. We compare the performance of this system with other state-of-the-art dialect recognition methods (both acoustic and token-based) on the NIST LRE 2007 English and Mandarin dialect recognition tasks. Our experimental results indicate that this system can perform better than baseline GMM and adapted PRLM systems, and also results in consistent gains of 15-23% when combined with other systems.

Index Terms: Dialect Recognition, Language Recognition, Phonetic Models, GMMs

1. Introduction

The problem of language recognition from speech lends itself to a variety of modeling approaches at different levels of the linguistic hierarchy [1]. The best systems, as evaluated by NIST [2] in recent years, have made use of multiple techniques that exploit both acoustic and phonotactic information.

For the automatic language recognition problem, all of these techniques rely on language-specific differences in the underlying speech for discrimination between languages. Both acoustic and phonotactic characteristics of speech may differ across languages, warranting the combination of different modeling techniques that make use of these characteristics. This intuition has been proven in practice as many research sites, in prior work, have shown that it is beneficial to employ a fusion of multiple acoustic and phonotactic systems at the score level [3]. A number of techniques have also been suggested that incorporate both types of information within a single system. In [4], the authors describe a phonotactic extension to a standard acoustic model. In [5] the authors extend the standard phonotactic approach (P-PRLM) using acoustic posterior probabilities.

In this paper we describe a method of deriving an acoustic system from a single PRLM phonetic model. We apply this system individually and in combination with an adapted PRLM model to the NIST LRE 2007 dialect recognition task. We show that these models can achieve similar performance to PRLM from which they are derived and, in combination, the two systems can reduce the dialect recognition error rate by 19.07% (relative). These results are similar for both the Mandarin dialect and English dialect tasks.

The remainder of this paper is organized as follows. In section 2 we describe the derivation of language-specific acoustic models. Section 3, details an efficient scoring procedure for these models and in section 4 we apply these models to the NIST LRE 2007 dialect recognition tasks for both Mandarin and English.

2. Adapted Phonetic Models for Dialect Recognition

The dialect recognition task, like language recognition, is a detection task. As is customary, we make decisions via a likelihood ratio. Given a sequence of test observations \( O \) and a target dialect model \( \lambda_d \) and a decision threshold \( \theta \), we hypothesize that \( O \) was produced by \( \lambda_d \) if:

\[
\log \left( \frac{L(O|\lambda_d)}{\sum_{\lambda \neq \lambda_d} L(O|\lambda)} \right) > \theta
\]

(1)

Where \( \frac{1}{H} \) is used to normalize duration differences across utterances as frame independence is typically assumed \( L(O|\lambda_d) = \prod_{t=1}^{T} p(o_t|\lambda_d) \).

In phonotactic systems, \( p(o_t|\lambda_d) \) is modeled using n-gram language models over phonetic sequences. For acoustic systems, this function is often modeled as mixtures of Gaussians over shifted delta cepstra observations [6] using a likelihood computation of this form:

\[
p(o_t|\lambda_d) = \sum_{i} c_i \mathcal{N}(o_t; \mu_i, \Sigma_i)
\]

(2)

In typical HMM-based phonetic recognizers used for language recognition, a similar GMM is used as the state-dependent observation model \( p(o_t|s_i) \) (where \( s_i \) is an HMM state). Generally, for such recognizers, the observation model of a given language is not trained specifically to a dialect.

In this paper we extend existing phonetic HMMs used by a single PRLM tokenizer to create dialect-specific acoustic models. We do so by exploiting the observation models of each phone state within this HMM to create dialect-specific phonetic models. This is done without dialect-specific phonetic supervision. The resulting models are then used to compute per-frame likelihoods for making dialect recognition decisions.

The tokenizers used for language recognition are typically standard HMM model defined by:

\[
p(o_t|s_i, \lambda) = \sum_{j} c_{ij} \mathcal{N}(o_t; \mu_{ij}, \Sigma_{ij})
\]

(3)

\*This work was sponsored by the Department of Defense under Air Force Contract F18625-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Nancy Chen is also supported under NIH/NIDCD grant nos. DC02978 and T32DC00038.

We choose a model $\lambda$ that is assumed to be dialect-neutral (designated as the  \texttt{root} HMM), then adapt each model probability to each target dialect using dialect-marked data without phonetic labeling. We apply an unsupervised form of MAP adaptation \cite{mula2013unsupervised} of gaussian mean parameters to build these dialect-specific models. To arrive at transcripts for MAP adaptation, we use the \texttt{root} tokenizer to decode dialect-specific data. For a given dialect $d$, the resulting dialect-specific HMM models are defined as follows:

$$
\mu_{ij}^d = \frac{r}{\tau + \gamma_{\text{ij}}(Q)} \mu_{ij} + \frac{\gamma_{\text{ij}}(Q)}{\gamma_{\text{ij}}(Q) + r} \bar{\mu}_{ij}^d
$$

$$
\gamma_{ij}(Q) = \sum_{t=1}^{\left|Q\right|} \gamma_{ij}(t)
$$

$$
\bar{\mu}_{ij}^d = \frac{\sum_{t=1}^{\left|Q\right|} \gamma_{ij}(t) q_i}{\sum_{t=1}^{\left|Q\right|} \gamma_{ij}(t)}
$$

Where $Q$ is the set of dialect-specific observations, $r$ is the MAP relevance factor and $\left\{ \mu_{ij}^d, \bar{\mu}_{ij}^d \right\}$ are the dialect-specific mean, covariance and weights for each gaussian $j$ within state $i$. $\bar{\mu}_{ij}^d$ is simply the ML estimate of the dialect-specific mean given example vectors $Q$.

The resulting models can be used for two purpose:

1. **Adapted Phonetic Model Scoring** – Direct scoring of the acoustic likelihood $\gamma_{\theta}(O|\theta) = L(O|\lambda_d)$, where $\lambda_d$ is a dialect-specific acoustic model trained via MAP adaptation.

2. **Parallel PRLM Using Adapted Tokenizers** – Adapted tokenizers are then used to generate phonetic sequences to be modeled via language models.\textsuperscript{2}

### 2.1 Training Procedure Details

We perform unsupervised MAP adaptation to arrive at dialect specific models. In order to minimize the variability associated with non-dialect factors, we apply channel compensation during both the decoding and dialect model estimation process. This is done through the use of CMLLR transforms as described in \cite{cho2012towards}.

The training procedure works as follows given a dialect-specific training set $Q$:

1. For each utterance $q$ in $Q$, perform an unadapted decoding using the \texttt{root} model.
2. Estimate channel/speaker normalizing transforms $T_q$ for utterance $q$.
3. Perform adapted decoding of $q$. The resulting 1-best hypothesis is used as a reference transcription for MAP adaptation.
4. Using $Q$, transforms $T_q$ and the prior model set (i.e. \texttt{root}), MAP adapt dialect specific models.
5. Iterate MAP adaptation replacing prior models with MAP-adapted models from previous iterations.

The training procedure incorporates SAF-style channel/speaker normalization. Experimentally, we found that this process is critical to achieving state of the art performance. For all experiments described below the relevance factor ($r$) was fixed to 16. We found significant performance improvements on our development data when iterating the MAP estimation process. In subsequent experiments, three iterations of adaptation were used.

### 3. Efficient Scoring

During scoring, our goal is to evaluate the log likelihood ratio presented in equation 1. In order to facilitate fast scoring during test time, we do not compute the needed likelihoods directly. Instead, we compute an efficient approximation over a single phone lattice. This is done by computing an approximation to the following likelihood of the phonetic model given a target dialect model:

$$
L(O|\lambda_d) = \prod_{t=1}^{T} p(o_t|\lambda_d)
$$

$$
= \sum_{\lambda_d} \alpha_T(s_t|\lambda_d)
$$

$$
\approx \sum_{\forall i \in \lambda_c} \alpha_T(s_t|\lambda_d)
$$

where $\alpha_T(s_t|\lambda_d)$ is the forward probability leading to state $s_t$ of the observation sequence $O$ given a dialect model $\lambda_d$. Equation 9 limits the computation of $\alpha_T$ to states observed in the phone lattice $L$.

We can make use of phone lattices generated by the \texttt{root} tokenizer used for PRLM-based dialect recognition and rescore them to generate dialect-specific lattices using dialect-adapted acoustic models constructed during training. This is more efficient than redecoding each target. These dialect-specific lattices can then be used to compute $\alpha_T(s_t|\lambda_d)$.

The process we used for all experiments reported in this paper is described below.

- **Lattice Generation**: This is done using the \texttt{root} tokenizer to decode input audio segments without the use of any language models. During this step, speaker/channel adaptation is performed using CMLLR. Heavy pruning is employed to keep the lattice small so as to minimize processing requirements during rescoring.

- **Rescoring**: Within each lattice, root tokenizer acoustic model scores are replaced by their dialect-specific counterparts.

- **Likelihood Computation**: The rescoring lattice is then used to compute a full forward probability using the SRI lattice-tool \cite{Li2008} to yield an approximation $L(O|\lambda_d)$.

We found that applying CMLLR during decoding with the root tokenizer to be essential for good performance. Note that the lattice generation step is needed for PRLM-LID based on CMLLR adaptation. As a result this can be shared between a PRLM system and the proposed adapted phonetic model system. No additional decoding is needed beyond this step to score the MAP-adapted phonetic models or the P-PRLM system based on these models. Table 1 shows a timing analysis of the needed dialect-specific lattice rescoring steps. It shows that this process is significantly more efficient than full decoding. As a result the added computational cost of the proposed systems is minimal when PRLM is used.

<table>
<thead>
<tr>
<th>Scoring Method</th>
<th>Time</th>
<th>% of RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Decoding</td>
<td>24.2</td>
<td>80.7%</td>
</tr>
<tr>
<td>Lattice Rescoring</td>
<td>3.0</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

Table 1: Full decoding vs. lattice rescoring times on 30s utterances
4. Experiments

The models described above were tested as part of the NIST LRE 2007 evaluation campaign. For all of this effort, we ran a number of experiments comparing and combining the adapted phonetic model system with other acoustic and phonotactic systems.

4.1. Data

We evaluated the performance of our system on the 30 second, NIST LRE 2007 Mandarin and English dialect recognition tasks. In both cases, the task is a dialect detection and we present results in terms of equal error rate. For Mandarin, the test set consisted of 80 and 78 true trials of Mainland and Taiwanese dialects, respectively. For English, the LRE07 test set included 80 trials of American English and 160 trials of Indian English.

We trained systems for each language using data from the CALLHOME corpus, the LRE05 test set, and data from OGI's foreign accented English and LDC's MIXER and FISHER corpora [10][11][12]. In total, 104 and 20.14 hours of data were used to adapt and train PRLM and adapted phonetic models for English and Mandarin respectively. A small set of 1,400 trials from MIXER and LRE05 data was reserved as a development set for training of a backend classifier.

In each task, the performance of the adapted phonetic model system is compared with a baseline GMM configured to match our LRE 2005 submission and a PRLM-based phonotactic system using the root model as described above. The baseline GMM system was trained with 2,048 components, MAP-adapted to each target dialect from SDC features [7-3-7-3] [4] with feature normalization and RASTA filtering. Each GMM was trained using the training sets described above. The baseline PRLM system uses CMLLR adaptation, and 3-gram LMs estimated from lattice counts.

4.2. English Dialect Recognition

We trained an English root phonetic models using data drawn from Switchboard-II phase 4 (Cellular). In total, 23 hours of word transcribed data was used to train 47 phone models. Table 2 shows the configuration of these models.

<table>
<thead>
<tr>
<th>Front End</th>
<th>PLP-13 + 1st and 2nd Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Models</td>
<td>47 phone models, 3-states, 31g per state</td>
</tr>
<tr>
<td>Adaptation</td>
<td>SAT training + Test-time CMLLR</td>
</tr>
<tr>
<td>Training Data</td>
<td>23 hours (word transcription)</td>
</tr>
</tbody>
</table>

Table 2: SWB-CELL training configuration

This recognizer was used to adapt Indian and American English acoustic models using the aforementioned English training set. Table 3 shows the performance of the adapted acoustic model system in comparison with GMM and adapted PRLM baselines.

The performance of the adapted phonetic model system outperforms the the GMM baseline by 24% relative once a backend is applied. The adapted phonetic model also outperforms the PRLM by 10.25%.

4.3. Mandarin Dialect Recognition

We trained a Mandarin root phonetic models using data from the CALLHOME Mandarin. In total, 18.85 hours of word transcribed data was used to train 38, 3-state phone models. Table 4 shows the configuration of these models. As with the English adapted phonetic models, this root model was used to create dialect-specific phonetic models via MAP adaptation.

<table>
<thead>
<tr>
<th>Front End</th>
<th>PLP-13 + 1st and 2nd Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Models</td>
<td>38 phone models, 3-states, 31g per state</td>
</tr>
<tr>
<td>Adaptation</td>
<td>SAT training + Test-time CMLLR</td>
</tr>
<tr>
<td>Training Data</td>
<td>18.85 hours (word transcription)</td>
</tr>
</tbody>
</table>

Table 4: CALLHOM-Mandarin training configuration

Table 5 shows the performance of GMM and PRLM baselines compared to the proposed adapted phonetic model. As with the English task, the adapted PRLM performance is similar to that of the two baseline systems. That said, we did not observe performance improvements from this system on its own after applying a backend. This may be due to the limited training data for both model adaptation and backend training.

<table>
<thead>
<tr>
<th>System</th>
<th>EER (no BE)</th>
<th>EER (w/BE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>22.67%</td>
<td>19.56%</td>
</tr>
<tr>
<td>Adapted PRLM</td>
<td>19.42%</td>
<td>22.05%</td>
</tr>
<tr>
<td>Adapted Phonetic Model</td>
<td>24.07%</td>
<td>19.65%</td>
</tr>
</tbody>
</table>

Table 5: LRE07 Mandarin dialect task (30s) with and without backend

4.4. Fusion Experiments

Prior studies have shown that the fusion of acoustic and phonotactic systems can be beneficial for language recognition problems [3]. In the section, we compare the performance of a baseline acoustic system (GMM-based) with the proposed adapted phonetic models in combination with a standard PRLM-based phonotactic model. Figures 1 and 2 show post-backend results for all systems using including fusions of multiple subsets for English and Mandarin respectively. System F1 combines systems based on a single root tokenizer (i.e. the adapted PRLM and adapted phonetic models). System F2 combines the adapted PRLM and GMM baseline systems. Finally, F3 combines all three. Table 6 summarizes these results for both languages at the EER operating point.

<table>
<thead>
<tr>
<th>System</th>
<th>Eng</th>
<th>Man</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>14.38%</td>
<td>19.56%</td>
</tr>
<tr>
<td>Adapted PRLM</td>
<td>12.19%</td>
<td>22.05%</td>
</tr>
<tr>
<td>Adapted Phonetic Model</td>
<td>10.94%</td>
<td>19.65%</td>
</tr>
<tr>
<td>F1 = Adapted PRLM + Phonetic Model</td>
<td>10.31%</td>
<td>17.94%</td>
</tr>
<tr>
<td>F2 = Adapted PRLM + GMM</td>
<td>11.25%</td>
<td>19.54%</td>
</tr>
<tr>
<td>F3 = ALL</td>
<td>9.06%</td>
<td>16.41%</td>
</tr>
</tbody>
</table>

Table 6: Fusion performance (@ EER) of systems for English and Mandarin tasks (with backend)
In both languages, there is significant improvement when adding the phonetic model. \( F_1 \) and \( F_2 \) show two different combinations of phonotactic and acoustic systems. The use of adapted phonetic models improves the PRLM baseline (combination \( F_1 \)) by 15.42\% (English) and 22.72\% (Mandarin). This is better than adding GMM models to a PRLM baseline (combination \( F_2 \)) by 8.3\% and 12.8\% respectively for English and Mandarin.

The fusion of all systems reduces the EER of the best single system by 17.18\% (English) and 16.10\% (Mandarin).

![DET curves for the English dialect task](image)

**Figure 1: DET curves for the English dialect task**

![DET curves for the Mandarin dialect task](image)

**Figure 2: DET curves for the Mandarin dialect task**

suggests, this may further improve the performance of our systems. We intend to pursue this in future research.

### 5. Discussion

In summary, the adapted phonetic model system is capable of good performance for the dialect recognition problem without phonetically word transcribed data. We show that it can be run with little additional overhead when a PRLM-based model is used. Furthermore, this model can be combined with PRLM to improve performance. In fusion experiments, we found that the combination of this system with PRLM outperformed combinations of PRLM with GMM-based models and the combination of all three systems could further improve performance.

Note that the number of gaussian parameters for the adapted phonetic system is 4,371 (English) and 3,534 (Mandarin), nearly twice that of the baseline GMM. This may be part of the performance improvement, though it is unlikely to account for all of it. In prior development experiments, there has generally been little additional gain from adding gaussian components beyond 2,048. It is possible that performance gains from the adapted phonetic model are due to the phonetic allocation gaussians and the CMLLR channel compensation done during PRLM adaptation. GMMS too have been shown to benefit from similar types of channel compensation. A comparison of these methods would be interesting for follow-on research.

As our phonetic models were derived from ASR-based models, we did not experiment with SDC-based features. As [6]

### 6. References