Discriminative Phonotactics for Dialect Recognition Using Context-Dependent Phone Classifiers

Fadi Biadsy*, Hagen Soltau+, Lidia Mangu+, Jiri Navratil+, Julia Hirschberg*

*Columbia University, NY, USA
+IBM T. J. Watson Research Center, NY, USA

July 1st, 2010
Dialect Recognition

● Similar to language recognition, but use dialects/accents of the same language

● Dialects may differ in any dimension of the linguistic spectrum
  ● Differences are likely to be more subtle across dialects than those across languages
  ● Thus, more challenging problem than language recognition
Motivation: Why Study Dialect Recognition?

- Discover differences between dialects
- To improve Automatic Speech Recognition (ASR)
  - Model adaptation: Pronunciation, Acoustic, Morphological, Language models

- To infer speaker’s regional origin for
  - Forensic speaker profiling
  - Speech to speech translation
  - Annotations for Broadcast News Monitoring
  - Spoken dialogue systems – adapt TTS systems
  - Charismatic speech identification
Multiple cues that may distinguish dialects:

- Phonetic cues:
  - Differences in phonemic inventory
  - Phonemic differences
  - Allophonic differences (context-dependent phones)

Example: /r/
- **Approximant in American English** [ɹ] – modifies preceding vowels
- **Trilled in Scottish English** in [Consonant]–/r/–[Vowel] and in some other contexts

“\text{She will meet him}”

<table>
<thead>
<tr>
<th>MSA:</th>
<th>/s/ /a/</th>
<th>/t/ /u/ /q/</th>
<th>/A/ /b/ /i/ /l/ /u/</th>
<th>/h/ /u/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egy:</td>
<td>/H/ /a/</td>
<td>/t/ /?/</td>
<td>/a/ /b/ /l/ /u/</td>
<td></td>
</tr>
<tr>
<td>Lev:</td>
<td>/r/ /a/ /H/</td>
<td>/t/ /g/</td>
<td>/A/ /b/ /l/ /u/</td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Dialects and Corpora
- CD-Phone Recognizer
- Baselines
- Two Ideas:
  - GMM-UBM with fMLLR
  - Discriminative Phonotactics
- Results
- Conclusions and Future Work
Case Study: Arabic Dialects

(by Arab Atlas)
### Corpora

<table>
<thead>
<tr>
<th>Dialect</th>
<th># Speakers</th>
<th>Test 20% – 30s* test cuts</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulf</td>
<td>976</td>
<td>801</td>
<td>(Appen Pty Ltd, 2006a)</td>
</tr>
<tr>
<td>Iraqi</td>
<td>478</td>
<td>477</td>
<td>(Appen Pty Ltd, 2006b)</td>
</tr>
<tr>
<td>Levantine</td>
<td>985</td>
<td>818</td>
<td>(Appen Pty Ltd, 2007)</td>
</tr>
</tbody>
</table>

- For testing:
  - (25% female – mobile, 25% female – landline, 25% male – mobile, 25 % male – landline)

- Egyptian: Training: CallHome Egyptian, Testing: CallFriend Egyptian

<table>
<thead>
<tr>
<th>Dialect</th>
<th># Training Speakers</th>
<th># 120 speakers 30s* cuts</th>
<th>Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian</td>
<td>280</td>
<td>1912</td>
<td>(Canavan and Zipperlen, 1996) (Canavan et al., 1997)</td>
</tr>
</tbody>
</table>

7 *Exactly 30s
Outline

• Motivation
• Corpora
• CD-Phone Recognizer
• Baselines
• Two Ideas:
  • GMM-UBM with fMLLR
  • Discriminative Phonotactics
• Results
• Conclusions and Future Work
Context-Dependent (CD) Phone Recognizer

- **HMM-triphone-based phone recognizer** using IBM’s Attila system
  - Trained on 50 hours of GALE broadcast news and conversations
- **230 CD-acoustic models** and 20,000 Gaussians

Front-End:
- 13D PLP features per frame
- Each frame is spliced together with four preceding and four succeeding frames followed by LDA $\rightarrow 40D$
- CMVN

Speaker Adaptation:
- fMLLR followed by MLLR

Unigram phone language model trained on MSA
Outline

- Motivation
- Corpora
- CD-Phone Recognizer
- Baselines

- Two Ideas:
  - GMM-UBM with fMLLR
  - Discriminative Phonotactics

- Results

- Conclusions and Future Work
Baselines

- **Standard PRLM**: a trigram phonotactic model per dialect

- **Standard GMM-UBM**:
  - Front-End: Same as the front end of the phone recognizer
  - 2048 Gaussians – ML trained on equal number of frames from each dialect
  - Dialect Models are MAP adapted with 5 iterations -- similar settings of the baseline in (Torres-Carrasquillo et al., 2008)
Results (DET curves of PRLM and GMM-UBM) – 30s Cuts

<table>
<thead>
<tr>
<th>Approach</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRLM</td>
<td>17.7</td>
</tr>
<tr>
<td>GMM-UBM</td>
<td>15.3*</td>
</tr>
</tbody>
</table>

*Comparable to GMM-UBM of (Torres-Carrasquillo et al., 2008) on 3 dialects
Outline

- Motivation
- Corpora
- CD-Phone Recognizer
- Baselines
- Two Ideas:
  - GMM-UBM with fMLLR
  - Discriminative Phonotactics
- Results
- Conclusions and Future Work
Our GMM-UBM Improved with fMLLR

- Motivation: Feature normalization (CMVN and VTLN) improve GMM-UBM for language and dialect recognition
  - (e.g., Wong and Sridharan, 2002; Torres-Carrasquillo et al., 2008)

- Our approach: Feature space Maximum Likelihood Linear Regression (fMLLR) adaptation

- Use a CD-phone recognizer to obtain CD-phone sequence: transform the features “towards” the corresponding acoustic model GMMs (a matrix for each speaker)

- Same as GMM-UBM approach, but use transformed acoustic vectors instead
Results – GMM-UBM-fMLLR – 30s Cuts

<table>
<thead>
<tr>
<th>Approach</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRLM</td>
<td>17.7</td>
</tr>
<tr>
<td>GMM-UBM</td>
<td>15.3</td>
</tr>
<tr>
<td>GMM-UBM-fMLLR</td>
<td>11.0%</td>
</tr>
</tbody>
</table>
Outline

- Motivation
- Corpora
- CD-Phone Recognizer
- Baselines

- Two Ideas:
  - GMM-UBM with fMLLR
  - Discriminative Phonotactics

- Results
- Conclusions and Future Work
Discriminative Phonotactics

- **Hypothesis:** Dialects differ in their allophones (context-dependent phones) and their phonotactics

- **Idea:** Discriminate dialects first at the level of context-dependent (CD) phones and then phonotactics

\[
\begin{align*}
/r/ & \text{ is Approximant in American English } [u] \text{ and trilled in Scottish in } [\text{Consonant}] – /r/ – [\text{Vowel}] \\
\end{align*}
\]

I. Obtain CD-phones
II. Extract acoustic features for each CD-phone
III. Discriminate CD-phones across dialects
IV. Augment the CD-phone sequences and extract phonotactic features
V. Train a discriminative classifier to distinguish dialects
Obtaining CD-Phones

Run our CD-phone recognizer

CD-phone sequence

... [Back vowel]-r-[Central Vowel] [Plosive]-A-[Voiced Consonant] [Central Vowel]-b-[High Vowel] ... * not just /r/ /A/ /b/ ...

Do the above for all training data of all dialects
Each CD phone type has an acoustic model:

e.g., [Back vowel]-r-[Central Vowel]
Obtaining CD-Phones + Frame Alignment

Acoustic frames for second state

<table>
<thead>
<tr>
<th>Acoustic frames:</th>
<th>Front-End</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>CD-Acoustic Models:</th>
<th>CD-Phone Recognizer</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>CD-Phones: (e.g.)</th>
<th>[vowel]-b-[glide]</th>
<th>⋮</th>
<th>[front-vowel]-r-[sonorant]</th>
</tr>
</thead>
</table>
MAP Adaptation of each **CD-Phone Instance**

[Back Vowel]-r-[Central Vowel]

MAP adapt the CD-phone acoustic model GMMs to the corresponding frames \( (r=0.1) \)
MAP adapt the CD-phone acoustic model GMMs to the corresponding frames*

One Super Vector for each CD phone instance:

Stack all the **Gaussian means and phone duration** $V_k = [\mu_1, \mu_2, ..., \mu_N, \text{duration}]$

i.e., summarize the acoustic-phonetic features of each CD-phone in one vector

---

22 *Similar to (Campbell et al., 2006) but at the level of CD-phone*
SVM Classifier for each CD-Phone Type for each Pair of Dialects

[Back Vowel]-r-[Central Vowel]

dialect 1

Super vectors of CD-phone instances of all training speakers in dialect 1

Super vectors of CD phone instances of all training speakers in dialect 2

dialect 2
### Discriminative Phonotactics – CD-Phone Classification

<table>
<thead>
<tr>
<th>Acoustic frames for second state</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic frames:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CD-Acoustic Models:</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-Phones: (e.g.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[vowel]-b-[glide]</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Super Vectors:</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Super Vector 1</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Dialects: (e.g.)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Egy</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>MAP Adapted Acoustic Models:</th>
</tr>
</thead>
</table>

| CD-Phone Recognizer: |

<table>
<thead>
<tr>
<th>MAP Adapt GMMs</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Super Vectors</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Super Vector N</th>
</tr>
</thead>
</table>

| SVM Classifiers:       |

<table>
<thead>
<tr>
<th>Egy</th>
</tr>
</thead>
</table>

### Notes

- **MAP Adapted Acoustic Models**: Super Vectors
- **CD-Phones**: (e.g.)
  - [vowel]-b-[glide]
  - [front-vowel]-r-[sonorant]
- **Dialects**: (e.g.)
  - Egy
CD-Phone Classifier Results

- Split the training data into two halves
- Train 227 (one for each CD-phone type) binary classifiers for each pair of dialects on 1\textsuperscript{st} half and test on 2\textsuperscript{nd}

<table>
<thead>
<tr>
<th>Dialect Pair</th>
<th>Num. of * classifiers</th>
<th>Weighted accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian/Iraqi</td>
<td>195</td>
<td>70.9</td>
</tr>
<tr>
<td>Egyptian/Gulf</td>
<td>196</td>
<td>69.1</td>
</tr>
<tr>
<td>Egyptian/Levantine</td>
<td>199</td>
<td>68.6</td>
</tr>
<tr>
<td>Levantine/Iraqi</td>
<td>172</td>
<td>63.96</td>
</tr>
<tr>
<td>Gulf/Iraqi</td>
<td>166</td>
<td>61.77</td>
</tr>
<tr>
<td>Levantine/Gulf</td>
<td>179</td>
<td>61.53</td>
</tr>
</tbody>
</table>

* performed significantly better than chance (50%)
 Extraction of Linguistic Knowledge

- Use the results of these classifiers to show which phones in what contexts distinguish dialects the most (chance is 50%)

<table>
<thead>
<tr>
<th>CD-Phone ([l-context]−phone−[r-context])</th>
<th>Accuracy</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>[∗]−sh−[∗]</td>
<td>71.1</td>
<td>6302</td>
</tr>
<tr>
<td>[SIL]−a−[∗]</td>
<td>70.3</td>
<td>3935</td>
</tr>
<tr>
<td>[SIL]−?−[Central Vowel]</td>
<td>68.7</td>
<td>1323</td>
</tr>
<tr>
<td>[∗]−j−[∗]</td>
<td>68.5</td>
<td>3722</td>
</tr>
<tr>
<td>[! Central Vowel]−s−[! High Vowel]</td>
<td>68.5</td>
<td>1975</td>
</tr>
<tr>
<td>[!SIL &amp; ! Central Vowel]−E−[!Central Vowel]</td>
<td>67.8</td>
<td>3687</td>
</tr>
<tr>
<td>[Central Vowel]−m−[Central Vowel]</td>
<td>66.7</td>
<td>2639</td>
</tr>
<tr>
<td>[∗]−k−[Central Vowel]</td>
<td>66.4</td>
<td>1433</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>[!SIL &amp; !Central Vowel]−G−[!Central Vowel]</td>
<td>57.5</td>
<td>852</td>
</tr>
<tr>
<td>[!A]−h−[Back Vowel]</td>
<td>57.0</td>
<td>409</td>
</tr>
<tr>
<td>[!Vowel &amp; !SIL]−m−[!Central Vowel &amp; !Back Vowel]</td>
<td>56.2</td>
<td>300</td>
</tr>
</tbody>
</table>

Levantine/Iraqi Dialects
Labeling Phone Sequences with Dialect Hypotheses

Run corresponding SVM classifier to get the dialect of each CD phone

CD-phone recognizer

...  
[Back vowel]-r-[Central Vowel]  
[Plosive]-A-[Voiced Consonant]  
[Central Vowel]-b-[High Vowel]  
...  

...  
[Back vowel]-r-[Central Vowel]  
[Plosive]-A-[Voiced Consonant]  
[Central Vowel]-b-[High Vowel]  
Egyptian  
[Central Vowel]-b-[High Vowel]  
Levantine  
...  
...
Textual Feature Extraction for Discriminative Phonotactics

- Extract the following textual features from each pair of dialects
  - Frequency of annotated CD-Phone bigrams, e.g.,
    
    “[Nasal]–r–[Vowel]_{Iraqi} [Voiced Cons.]–a–[Liquid]_{Gulf}”

  - Frequency of bigrams with only one annotated CD-Phone, e.g.,
    
    “[Nasal]–r–[Vowel] [Voiced Cons.]–a–[Liquid]_{Gulf}”

  - Frequency of annotated unigrams, e.g.,
    
    ![Central Vowel]–E–[Central Vowel]_{Gulf}

  - Frequency of not annotated CD-Phone unigrams and bigrams, e.g.,
    
    “[Nasal]–r–[Vowel] [Voiced Cons.]–a–[Liquid]”

  - Frequency of context independent phone trigrams, e.g.,
    
    “s A l”

- Normalize vector by its norm

- Train a logistic regression with L2 regularizer
Experiments – Training Two Models

- Split training data into two halves
- Train SVM CD-phone classifiers using the first half
- Run these SVM classifiers to annotate the CD phones of the 2\textsuperscript{nd} half
- Train the logistic classifier on the annotated sequences
Discriminative Phonotactics – Dialect Recognition

Acoustic frames for second state

Acoustic frames:

CD-Acoustic Models:

CD-Phones: (e.g.)

MAP Adapted Acoustic Models:

Super Vectors:

Dialects: (e.g.)

Super Vector 1

Super Vector N

[vowel]-b-[glide] Egy

[vowel]-b-[glide] Egy

[vowel]-b-[glide] Egy

[front-vowel]-r-[sonorant]

[front-vowel]-r-[sonorant]

[front-vowel]-r-[sonorant]

Logistic classifier

Egyptian
Baselines

- Standard PRLM: a trigram phonotactic model per dialect

- Standard GMM-UBM:
  - Front-End:
    - 13D PLP features from 9 frames followed by LDA ➔ 40D
    - CMVN
  - 2048 Gaussians – ML trained on equal number of frames from each dialect
  - Dialect Models are MAP adapted with 5 iterations (similar to Torres-Carrasquillo et al., 2008)
Results – Discriminative Phonotactics

<table>
<thead>
<tr>
<th>Approach</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRLM</td>
<td>17.7</td>
</tr>
<tr>
<td>GMM-UBM</td>
<td>15.3</td>
</tr>
<tr>
<td>GMM-UBM-fMLLR</td>
<td>11.0%</td>
</tr>
<tr>
<td>Disc. Phonotactics</td>
<td>6.0%</td>
</tr>
</tbody>
</table>
Results per Dialect

<table>
<thead>
<tr>
<th>Dialect</th>
<th>GMM fMLLR</th>
<th>Disc. Pho.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian</td>
<td>4.4%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Iraqi</td>
<td>11.1%</td>
<td>6.6%</td>
</tr>
<tr>
<td>Levantine</td>
<td>12.8%</td>
<td>6.9%</td>
</tr>
<tr>
<td>Gulf</td>
<td>15.6%</td>
<td>7.8%</td>
</tr>
</tbody>
</table>
Conclusions

- fMLLR to transform the acoustic features significantly improve results for GMM-UBM approach
  - We still need to do more analyses

- The proposed method helps in understanding the linguistic differences between dialects

- Discriminative phonotactics outperforms GMM-UBM-fMLLR in 5% absolute EER.
Future Work

• New SVM Kernel to compute the similarity of all phone supervectors across two utterances ➔ only one SVM classifier for each pair of dialects (IS2010; submitted)

• Test this approach on shorter utterances (3s and 10s)

• Try this approach on dialects/accents of other languages:
  • English accents (American English and Indian English)
  • American English Dialects

• Apply VTLN

• Testing with NAP (need to modify to accommodate for short context Supervectors)
Thank You!

- Acknowledgments:
  - Jason Pelecanos for useful discussions
Case Study: Arabic Dialects – Our Data

- Iraqi Arabic: Baghdadi, Northern, and Southern
- Gulf Arabic: Omani, UAE, and Saudi Arabic
- Levantine Arabic: Jordanian, Lebanese, Palestinian, and Syrian Arabic
- Egyptian Arabic: primarily Cairene Arabic