Automatic Dialect/Accent Recognition

Fadi Biadsy

April 12th, 2010
Outline

- Problem
- Motivation
- Corpora
- Framework for Language Recognition
- Experiments in Dialect Recognition
  - Phonotactic Modeling
  - Prosodic Modeling
  - Acoustic Modeling
  - Discriminative Phonotactics
Problem: Dialect Recognition

- Given a speech segment of a predetermined language

  \[ \text{Dialect} = \{D_1, D_2, \ldots, D_N\} \]

- Great deal of work on language recognition

- **Dialect and Accent recognition** have more recently begun to receive attention

- Dialect recognition more difficult 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
  - Speech to speech translation
  - Annotations for Broadcast News Monitoring
  - Spoken dialogue systems – adapt TTS systems
  - Charismatic speech
- Call centers – crucial in emergency situations
Motivation: Cues that May Distinguish Dialects/Accents

- Phonetic cues:
  - Differences in phonemic inventory
  - Phonemic differences
  - Allophonic differences (context-dependent phones)
- Phonotactics: Rules/Distribution that govern phonemes and their sequences in a dialect

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

"She will meet him"

<table>
<thead>
<tr>
<th>MSA:</th>
<th>/s/</th>
<th>/a/</th>
<th>/t/</th>
<th>/u/</th>
<th>/q/</th>
<th>/A/</th>
<th>/b/</th>
<th>/i/</th>
<th>/l/</th>
<th>/u/</th>
<th>/h/</th>
<th>/u/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egy:</td>
<td>/H/</td>
<td>/a/</td>
<td>/t/</td>
<td>/?/</td>
<td>/a/</td>
<td>/b/</td>
<td>/l/</td>
<td>/u/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lev:</td>
<td>/r/</td>
<td>/a/</td>
<td>/H/</td>
<td>/t/</td>
<td>/g/</td>
<td>/A/</td>
<td>/b/</td>
<td>/l/</td>
<td>/u/</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Motivation: Cues that May Distinguish Dialects/Accents

- Prosodic differences
  - Intonational patterns
  - Timing and rhythm

- Spectral distribution (Acoustic frame-based features)

- Morphological, lexical, and syntactic differences

Subjects rely on intonational cues to distinguish two German dialects (Hamburg urban dialects vs. Northern Standard German) (Peters et al., 2002)
Outline

- Problem
- Motivation
- **Corpora**
- Framework for Language Recognition
- Experiments in Dialect Recognition
  - Phonotactic Modeling
  - Prosodic Modeling
  - Acoustic Modeling
  - Discriminative Phonotactics
- Contributions
- Future Work
- Research Plan
Case Study: Arabic Dialects

- 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
Corpora – Four Dialects – DATA I

- Recordings of spontaneous telephone conversation produced by native speakers of the four dialects available from LDC

<table>
<thead>
<tr>
<th>Dialect</th>
<th># Speakers</th>
<th>Total Duration</th>
<th>Test Speakers</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulf</td>
<td>965</td>
<td>41h</td>
<td>150</td>
<td>Gulf Arabic conversational telephone Speech database (Appen Pty Ltd, 2006a)</td>
</tr>
<tr>
<td>Iraqi</td>
<td>475</td>
<td>26h</td>
<td>150</td>
<td>Iraqi Arabic conversational telephone Speech database (Appen Pty Ltd, 2006b)</td>
</tr>
<tr>
<td>Egyptian</td>
<td>398</td>
<td>76h</td>
<td>150</td>
<td>CallHome Egyptian and its Supplement (Canavan et al., 1997) CallFriend Egyptian (Canavan and Zipperlen, 1996)</td>
</tr>
<tr>
<td>Levantine</td>
<td>1258</td>
<td>79h</td>
<td>150</td>
<td>Arabic CTS Levantine Fisher Training Data Set 1-3 (Maamouri, 2006)</td>
</tr>
</tbody>
</table>
Outline

- Problem
- Motivation
- Corpora
- **Framework for Language Recognition**
- Experiments in Dialect Recognition
  - Phonotactic Modeling
  - Prosodic Modeling
  - Acoustic Modeling
  - Discriminative Phonotactics
- Contributions
- Future Work
- Research Plan
Probabilistic Framework for Language ID

- Task:

\[
\arg\max_i P(L_i|\bar{a}, \bar{f})
\]

\(\bar{a}\): Frame-based spectral features
\(\bar{f}\): Frame-based prosodic features

- Hazen and Zue’s (1993) contribution:

\[
\arg\max_i P(L_i|C, S, \bar{a}, \bar{f})
\]

\(C\): Most likely underlying linguistic unit sequence hypothesis
\(S\): Corresponding segmentation

\[\iff \arg\max_i P(L_i) P(C|L_i) P(S, \bar{f}|C, L_i) P(\bar{a}|C, S, \bar{f}, L_i)\]

Prior  Phonotactic  Prosodic model  Acoustic model
Outline

- Problem
- Motivation
- Corpora
- Framework for Language Recognition

- Experiments in Dialect Recognition
  - **Phonotactic Modeling**
  - Prosodic Modeling
  - Acoustic Modeling
  - Discriminative Phonotactics

- Contributions
- Future Work
- Research Plan
**Hypothesis:** Dialects differ in their phonotactic distribution

\[
\arg\max_i P(D_i) \ P(C|D_i) \ \frac{P(S, f|C, D_i)}{P(C, S, f, D_i)}
\]

- Early work: Phone Recognition followed by Language Modeling (PRLM) (Zissman, 1996)

- Training: For each dialect \(D_i\):
  
  - Run a phone recognizer
  
  - Train an n-gram model: \(\lambda_i\)

- Examples:
  
  - dh uw z hh ih n d uw ey...
  - f uw v ow z l iy g s m k dh...
  - h iy jh sh p eh ae ey p sh...
Phonotactic Approach – Identification

Test utterance:

Run the phone recognizer

\[
\arg\max_i P(C = c_1, \ldots, c_T; \lambda_i)
\]

uw hh ih n d uw w ay ey uh jh y eh k oh v hh ...
Applying Parallel PRLM (Zissman, 1996)

- Use multiple \((k)\) phone recognizers trained on multiple languages to train \(k\) n-gram phonotactic models for each language of interest

- Experiments on our data: 9 phone recognizers, trigram models
Our Parallel PRLM Results – 10-Fold Cross Validation

![Graph showing test utterance duration in seconds vs. percentage.](image-url)

- **Accuracy**
- **Gulf F-Measure**
- **Iraqi F-Measure**
- **Levantine F-Measure**
- **Egyptian F-Measure**

Test utterance duration in seconds: 5, 15, 30, 45, 60, 120

%: 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100
Outline

- Problem
- Motivation
- Corpora
- Framework for Language Recognition

**Experiments in Dialect Recognition**
- Phonotactic Modeling
- **Prosodic Modeling**
- Acoustic Modeling
- Discriminative Phonotactics

- Contributions
- Future Work
- Research Plan
Prosodic Differences Across Dialects

- **Hypothesis**: Dialects differ in their prosodic structure
  - What are these differences?

- **Global Features**
  - Pitch: Range and Register, Peak Alignment, STDV
  - Intensity
  - Rhythmic features: $\Delta C$, $\Delta V$, $\%V$ (using pseudo syllables)
  - Speaking Rate
  - Vowel duration statistics

- Compare dialects using descriptive statistics
New Approach: Prosodic Modeling

\[
\arg\max_i P(D_i) \cdot P(C \mid D_i) \cdot P(S, \tilde{f} \mid C, D_i) \cdot P(\tilde{u} \mid C, S, \tilde{f}, D_i)
\]

- Pseudo-syllabification

- Sequential local features at the level of pseudo-syllables:

- Learn a sequential model for each prosodic sequence type using an ergodic continuous HMM for each dialect
New Approach for Prosodic Modeling

- Dialect Recognition System:

  ![Diagram of Dialect Recognition System]

  - Prosodic Feature Extractor
  - F0 Mean HMM Egyptian
  - F0 Mean HMM Levantine
  - Rhythm HMM Iraqi
  - Normalized Log likelihoods
  - Logistic Backend Classifier
  - Hypothesized Dialect

Global Features
Prosodic Modeling – Results (2m test utterances)

• 10-fold cross-validation (on Data I)

Accuracy (%)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>25</td>
</tr>
<tr>
<td>Global Features***</td>
<td>54.8</td>
</tr>
<tr>
<td>+ Vowel Dur. ***</td>
<td>60</td>
</tr>
<tr>
<td>+ Local Features***</td>
<td>72</td>
</tr>
<tr>
<td>Parallel PRLM ***</td>
<td>83.5</td>
</tr>
<tr>
<td>+ Prosodic Model *</td>
<td>86.33</td>
</tr>
</tbody>
</table>

* $p<0.05$; *** $p<0.001$
Outline

- Problem
- Motivation
- Corpora
- Framework for Language Recognition
- **Experiments in Dialect Recognition**
  - Phonotactic Modeling
  - Prosodic Modeling
  - **Acoustic Modeling**
  - Discriminative Phonotactics
- Contributions
- Future Work
- Research Plan
**Baseline: Acoustic Modeling**

- **Hypothesis:** Dialect differ in their spectral distribution

\[
\arg \max_i P(D_i) P(C|D_i) P(S, \tilde{f}|C, D_i) P(\tilde{a}|C, S, \tilde{f}, D_i)
\]

- Gaussian Mixture Model – Universal Background Model (GMM-UBM) widely used approach for language and speaker recognition (Reynolds et al., 2000)

- \(a_i\): 40D PLP features

- I. Train GMM-UBM using EM
- II. Maximum A-Posteriori (MAP) Adaptation to create a GMM for each dialect
- III. During recognition

\[
\arg \max_i P(\tilde{a}; \lambda_i)
\]
### Corpora – Four Dialects – DATA II

<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)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Canavan et al., 1997)</td>
</tr>
</tbody>
</table>
NIST LREC Evaluation Framework

- Detection instead of identification: given a trial and a target dialect
  - Hypothesis: Is the utterance from the target dialect?
    - Accept/reject + likelihood

- DET curves: false alarm probability against miss probability
  - Results are reported across pairs of dialects
  - All dialects are then pooled together to produce one DET curve
  - Trials 30s, 10s, and 3s long

- Equal Error Rate (EER)
Results (DET curves of PRLM and GMM-UBM) – 30s Cuts (Data II)

<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>

![Graph showing DET curves for PRLM and GMM-UBM.]
Our GMM-UBM Improved with fMLLR

- **Motivation:** VTLN and channel compensation improve GMM-UBM for speaker and language recognition.

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

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

- **Intuition:** Consequently produce more compact models.

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

<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

- Problem
- Motivation
- Corpora
- Framework for Language Recognition

**Experiments in Dialect Recognition**
- Phonotactic Modeling
- Prosodic Modeling
- Acoustic Modeling
- *Discriminative Phonotactics*

- Contributions
- Future Work
- Research Plan
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.

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

/r/ is Approximant in American English [ɹ] and trilled in Scottish in [Consonant] – /r/ – [Vowel]
Obtaining CD-Phones

Run Attila context-dependent phone recognizer \textit{(trained on MSA)}

Context-dependent (CD) phone sequence

... 

[Back vowel]-r-[Central Vowel] 
[Plosive]-A-[Voiced Consonant] 
[VCentral]-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>
<tbody>
<tr>
<td>CD-Acoustic Models:</td>
<td>CD-Phone Recognizer</td>
</tr>
<tr>
<td>CD-Phones: (e.g.)</td>
<td>[vowel]-b-[glide]</td>
</tr>
</tbody>
</table>
MAP Adaptation of each **CD-Phone Instance**

2. MAP adapt the universal background model GMMs to the corresponding frames

*Back Vowel*]-r-[*Central Vowel*]
MAP Adaptation of each **CD-Phone Instance**

2. MAP adapt the universal background 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., a sequence of features with unfixed size to fixed-size vector*
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 speakers in dialect 1

Super vectors of CD phone instances of all speakers in dialect 2
Acoustic frames for second state

**Acoustic frames:**

**CD-Acoustic Models:**

**CD-Phones: (e.g.)** [vowel]-b-[glide] [front-vowel]-r-[sonorant]

**MAP Adapted Acoustic Models:**

**Super Vectors:**

**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>
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>[<em>]–sh–[</em>]</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>[<em>]–j–[</em>]</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>[Nasal]–A–[Anterior]</td>
<td>68.1</td>
<td>5459</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>![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

CD-phone recognizer

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

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

...  
[Back vowel]-r-[Central Vowel] Egyptian  
[Plosive]-A-[Voiced Consonant] 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.,
    \[
    \text{“[Nasal]–r–[Vowel]}_{Iraqi} \quad \text{[Voiced Cons.]–a–[Liquid]}_{Gulf}
    \]
  - Frequency of bigrams with only one annotated CD-Phone, e.g.,
    \[
    \text{“[Nasal]–r–[Vowel]} \quad \text{[Voiced Cons.]–a–[Liquid]}_{Gulf}
    \]
  - Frequency of annotated unigrams, e.g.,
    \[
    \text{[!Central Vowel]–E–[Central Vowel]}_{Gulf}
    \]
  - Frequency of not annotated CD-Phone unigrams and bigrams, e.g.,
    \[
    \text{“[Nasal]–r–[Vowel]} \quad \text{[Voiced Cons.]–a–[Liquid]}
    \]
  - Frequency of context independent phone trigrams, e.g.,
    \[
    \text{“s A l”}
    \]
- 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 2nd half
- Train the logistic classifier on the annotated sequences
### Discriminative Phonotactics – Dialect Recognition

**Acoustic frames for second state**

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

#### CD-Acoustic Models:

<table>
<thead>
<tr>
<th>CD-Phones: (e.g.)</th>
<th>CD-Phone Recognizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>[vowel]-b-[glide]</td>
<td></td>
</tr>
<tr>
<td>[front-vowel]-r-[sonorant]</td>
<td></td>
</tr>
</tbody>
</table>

#### MAP Adapted Acoustic Models:

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

#### Super Vectors:

<table>
<thead>
<tr>
<th>Super Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super Vector 1</td>
</tr>
</tbody>
</table>

#### Dialects: (e.g.)

<table>
<thead>
<tr>
<th>SVM Classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>[vowel]-b-[glide]</td>
</tr>
<tr>
<td>[front-vowel]-r-[sonorant]</td>
</tr>
</tbody>
</table>

**Logistic classifier**

**Egyptian**
### 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>Discriminative Phonotactics</td>
<td>6.0%</td>
</tr>
</tbody>
</table>

The diagram shows the performance of different approaches in terms of EER, with Discriminative Phonotactics achieving the lowest EER of 6.0%.
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>
Comparison to the State-of-the-Art

- **State of the art system:** (Torres-Carrasquillo et al., 2008)
  - Two English accents: EER: 10.6%
  - Three Arabic dialects: EER: 7%
  - Four Chinese dialects: EER: 7%

- **NIST Language Recognition 2005:** (Mathjka et al., 2006) – fusing multiple approaches:
  - 7 Languages + 2 accents: EER: 3.1%
## Research Plan

<table>
<thead>
<tr>
<th>Month</th>
<th>Tasks</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 2010</td>
<td>Further analyses of the discriminative phonotactic approach</td>
<td>Defense proposal</td>
</tr>
<tr>
<td>Apr 2010</td>
<td>Compare all approaches for 11 Arabic sub-dialects</td>
<td>Bi-phone system</td>
</tr>
<tr>
<td>May 2010</td>
<td>Build the new Bi-phone system using HTK</td>
<td></td>
</tr>
<tr>
<td>Jun 2010</td>
<td>Test different techniques for biphone acoustic models on Arabic dialects</td>
<td>Experiment with different languages: Chinese, Spanish, American vs. Indian English, and American English Dialects</td>
</tr>
<tr>
<td>July 2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aug 2010</td>
<td>Work with IBM to Improve Arabic ASR using the best approach for dialect ID</td>
<td></td>
</tr>
<tr>
<td>Sep 2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct 2010</td>
<td></td>
<td>Write Dissertation</td>
</tr>
<tr>
<td>Nov 2010</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Acknowledgments

Thank You!
Prosodic Differences Across Dialects

- **F0 differences**
  - Levantine and Iraqi speakers have higher pitch range and more expanded pitch register than Egyptian and Gulf speakers
  - Iraqi and Gulf intonation show more variation than Egyptian and Levantine
  - Pitch peaks within pseudo-syllables in Egyptian and Iraqi are shifted significantly later than those in Gulf and Levantine

- **Durational and Rhythmic differences**
  - Gulf and Iraqi dialects tend to have more complex syllabic structure
  - Egyptian tend to have more vocalic intervals with more variation than other dialects, which may account for vowel reduction and quantity contrasts
For each CD phone sequence:

1. Get the frame alignment with the acoustic model’s states