Automatic Dialect/Accent Recognition

Fadi Biadsy

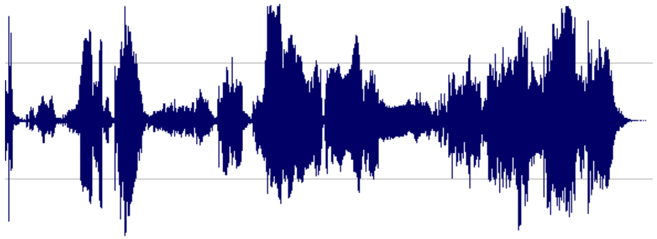
April 12th, 2010

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



Dialect = {D1, D2,…,DN}

- 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?

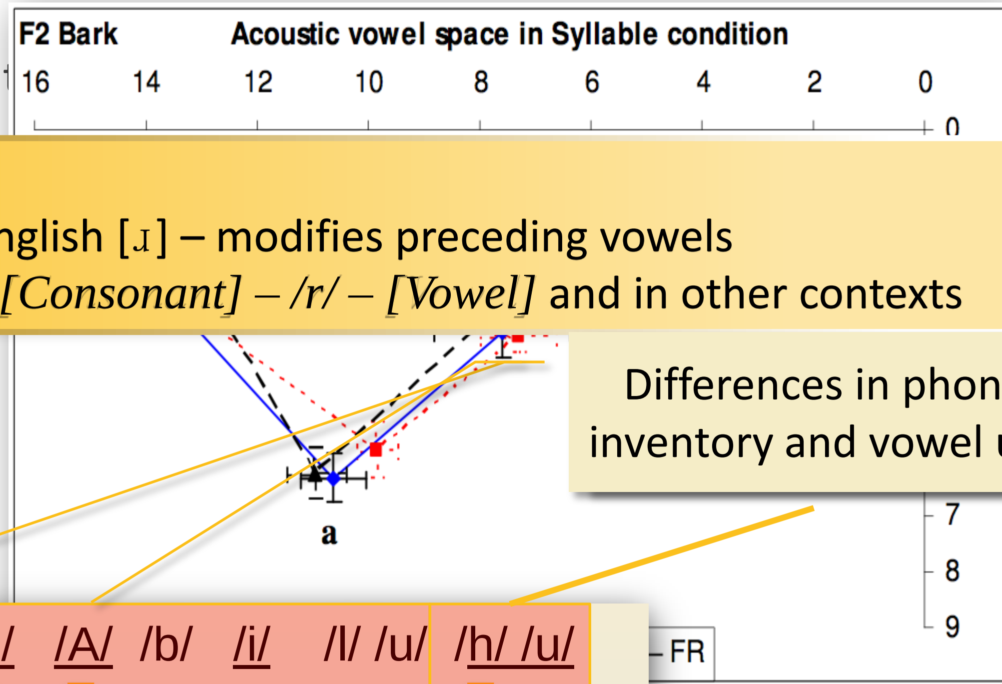
- Discover differences between dialects
- To improve Automatic Speech Recognition (ASR)
  - Model adaptation: Pronunciation, Acoustic, Morphological, Language models

- 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 (cont
- Ph
  a d



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

Differences in phonetic inventory and vowel usage

"She will meet him"

MSA:	/s/ /a/	/t/ /u/ /q/	/A/	/b/	/i/	/l/ /u/	/h/ /u/
Egy:	/H/ /a/	/t/ /?/	/a/	/b/		/l/	/u/
Lev:	/r/ /a/ /H/	/t/ /g/	/A/	/b/		/l/	/u/

2005)

Motivation: Cues that May Distinguish Dialects/Accents

- Prosodic differences
  - Intonational patterns
  - Timing and rhythm
- Spe
- 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

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

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:

$$\arg\max_i P(L_i|\vec{a}, \vec{f})$$

$\vec{a}$: Frame-based spectral features
$\vec{f}$: Frame-based prosodic features

- Hazen and Zue's (1993) contribution:

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

$$\arg\max_i P(L_i|C, S, \vec{a}, \vec{f})$$

$$\Leftrightarrow \arg\max_i \underbrace{P(L_i)}_{\text{Prior}} \underbrace{P(C|L_i)}_{\text{Phonotactic}} \underbrace{P(S, \vec{f}|C, L_i)}_{\text{Prosodic model}} \underbrace{P(\vec{a}|C, S, \vec{f}, L_i)}_{\text{Acoustic model}}$$

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

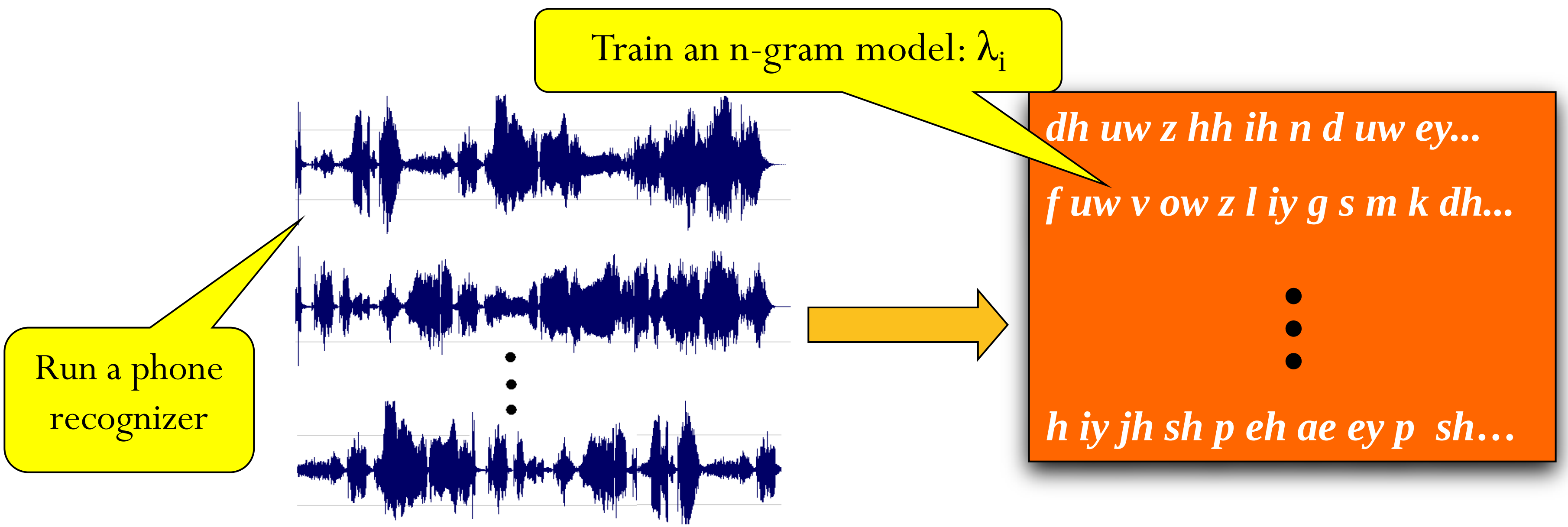
Phonotactic Approach

- **Hypothesis:** Dialects differ in their phonotactic distribution

$$\operatorname*{argmax}_i \cancel{P(D_i)}\; P(C|D_i)\; \cancel{P(S,\vec{f}|C,D_i)}\; \cancel{P(\vec{a}|C,S,\vec{f},D_i)}$$

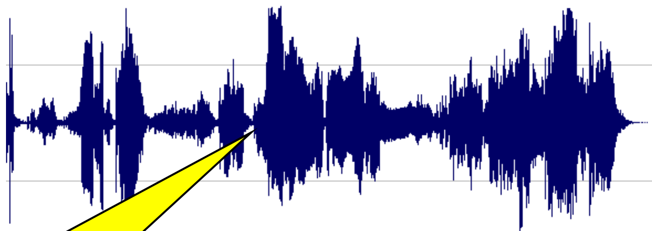
- Early work: Phone Recognition followed by Language Modeling (PRLM) **(Zissman, 1996)**

- Training: For each dialect D_i :



Phonotactic Approach – Identification

Test utterance:



Run the phone recognizer

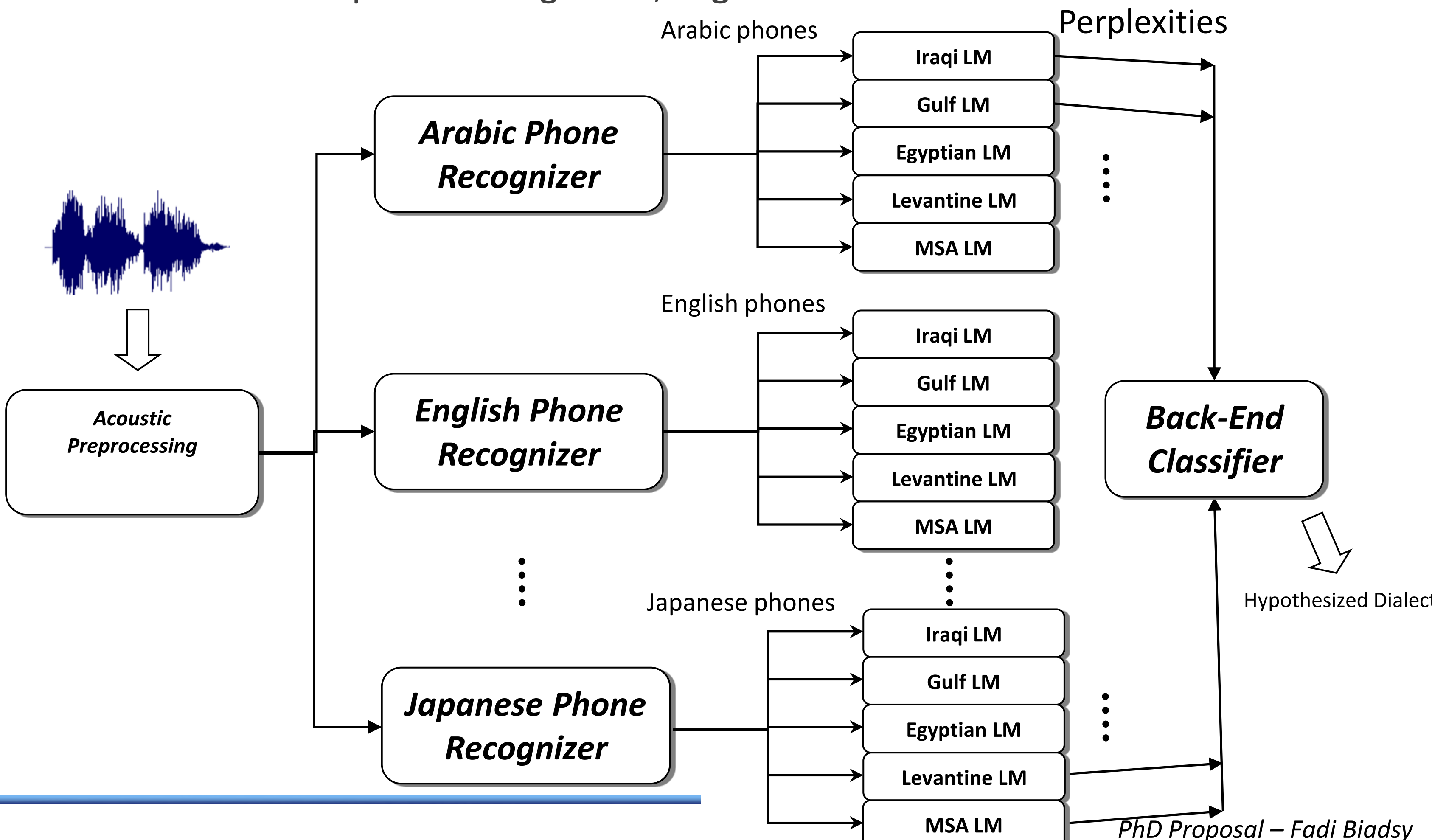
uw hh ih n d uw w ay ey uh jh y eh k oh v hh …

C

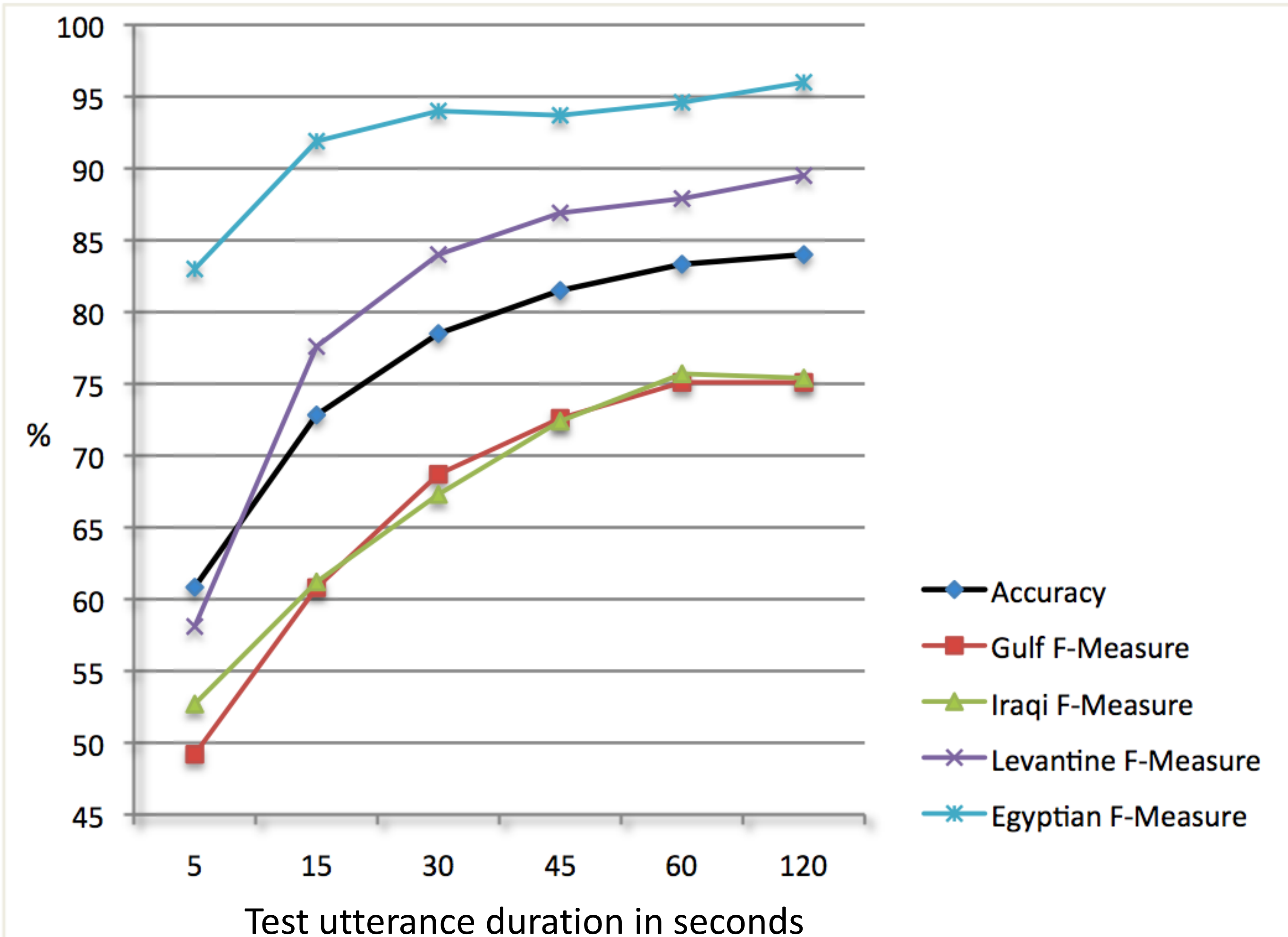
$$\underset{i}{\operatorname{argmax}}\, P(C = c_1, .., c_T; \lambda_i)$$

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



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

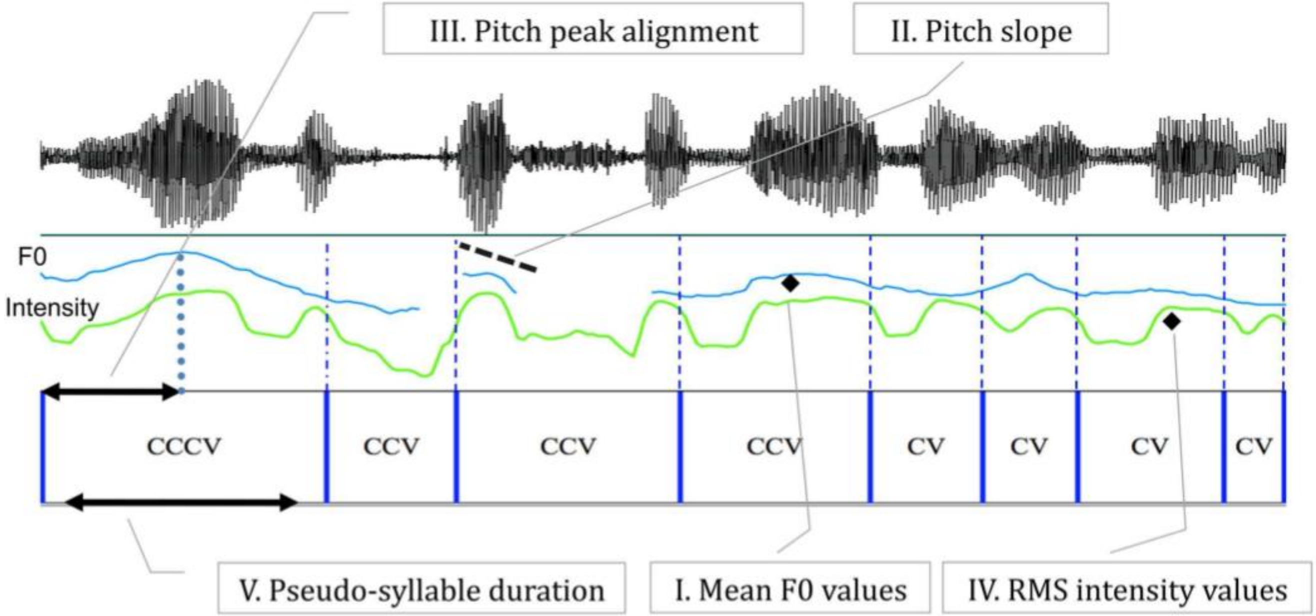
Prosodic Differences Across Dialects

- **Hypothesis:** Dialects differ in their prosodic structure
  - What are these differences?
- Global Features
  - Pitch: Range and Register, Peak Alignment, STDV
  - Intensity
  - Rhythmic features: ΔC, ΔV, $\%V$ (using pseudo syllables)
  - Speaking Rate
  - Vowel duration statistics
- Compare dialects using descriptive statistics

New Approach: Prosodic Modeling

$$\underset{i}{\text{argmax}}\ \cancel{P(D_i)}\ \cancel{P(C|D_i)}\ P(S, \vec{f}|C, D_i)\ \cancel{P(\vec{a}|C, S, \vec{f}, D_i)}$$

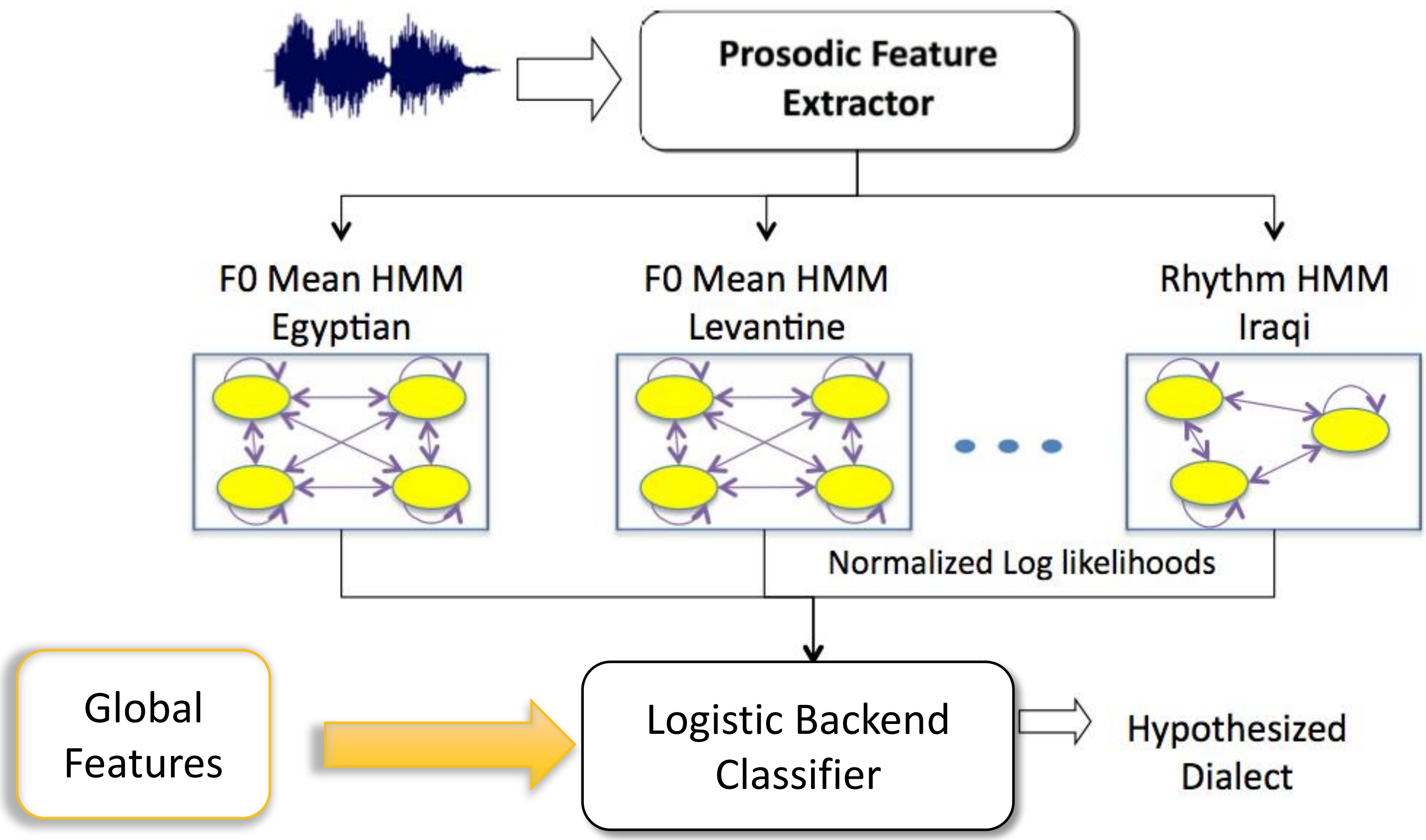
- Pseudo-syllabification
- Sequential local features at the level of pseudo-syllables:



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

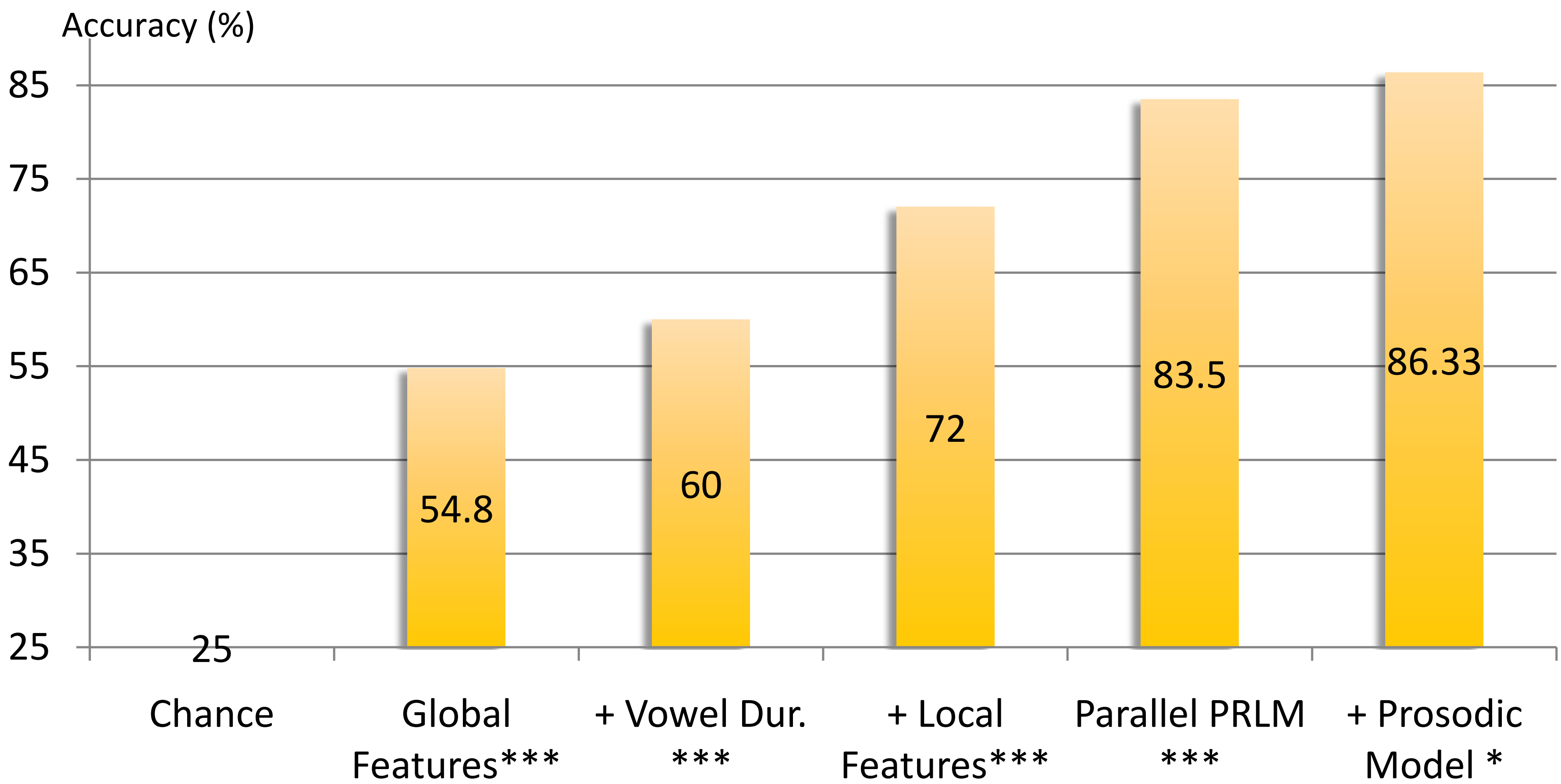
New Approach for Prosodic Modeling

- Dialect Recognition System:



Prosodic Modeling – Results (2m test utterances)

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



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

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

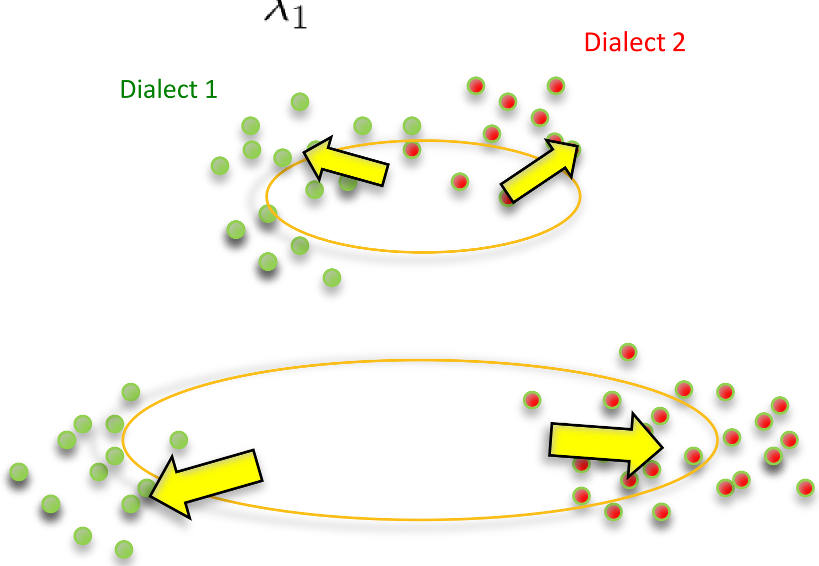
Baseline: Acoustic Modeling

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

$$\arg\max_i \cancel{P(D_i)\ P(C|D_i)\ P(S,\vec{f}|C,D_i)}\ P(\vec{a}|\cancel{C,S,\vec{f}},D_i)$$

- 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

$$\arg\max_i P(\vec{a};\lambda_i)$$

Corpora – Four Dialects – DATA II

Dialect	# Speakers	Test 20% – 30s test cuts	Corpus
Gulf	976	801	(Appen Pty Ltd, 2006a)
Iraqi	478	477	(Appen Pty Ltd, 2006b)
Levantine	985	818	(Appen Pty Ltd, 2007)

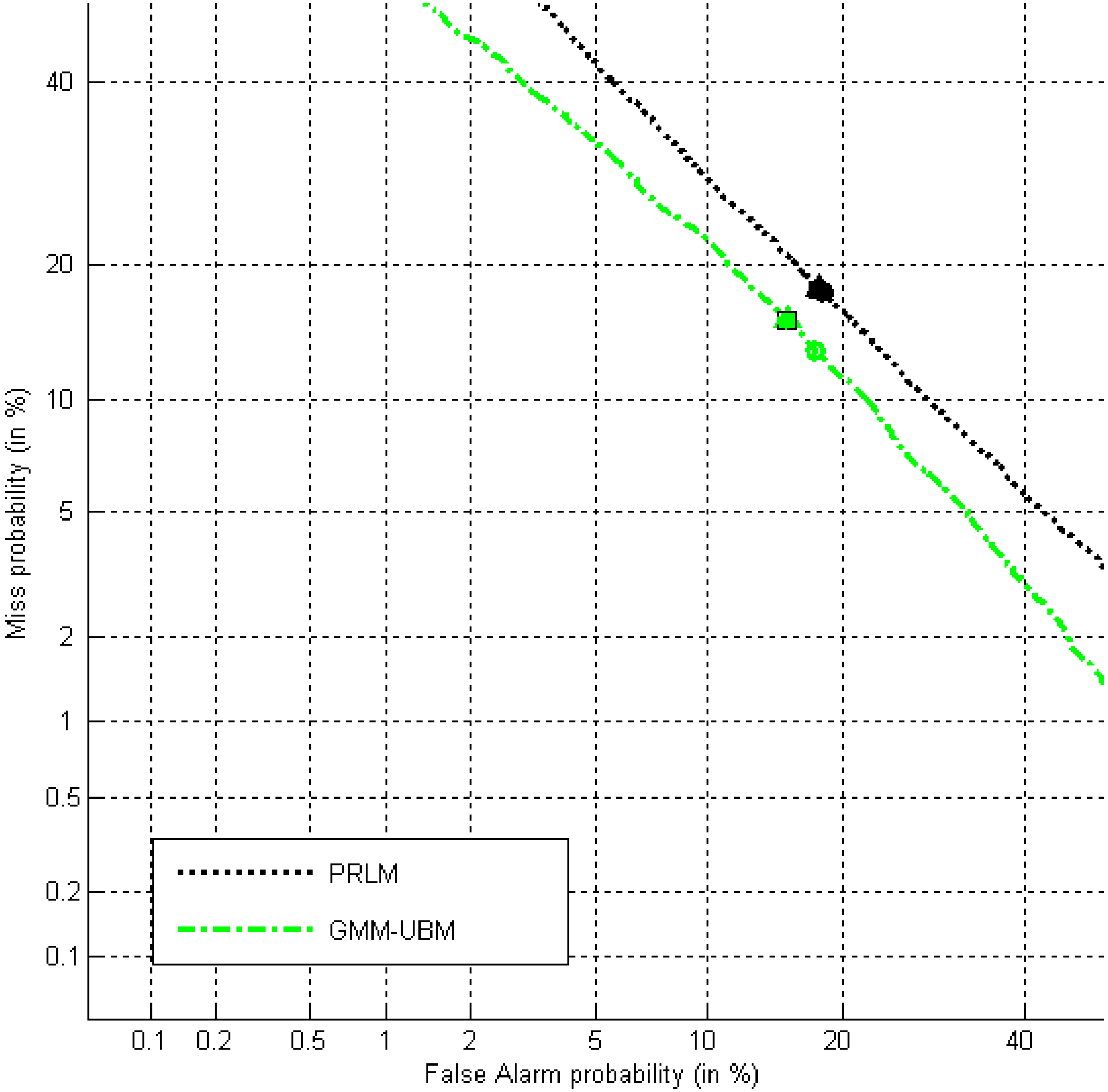
- For testing:
  - (25% female – mobile, 25% female – landline, 25% male – mobile, 25 % male – landline)
- Egyptian: Training: CallHome Egyptian, Testing: CallFriend Egyptian

Dialect	# Training Speakers	# 120 speakers 30s cuts	Corpora
Egyptian	280	1912	(Canavan and Zipperlen, 1996) (Canavan et al., 1997)

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)

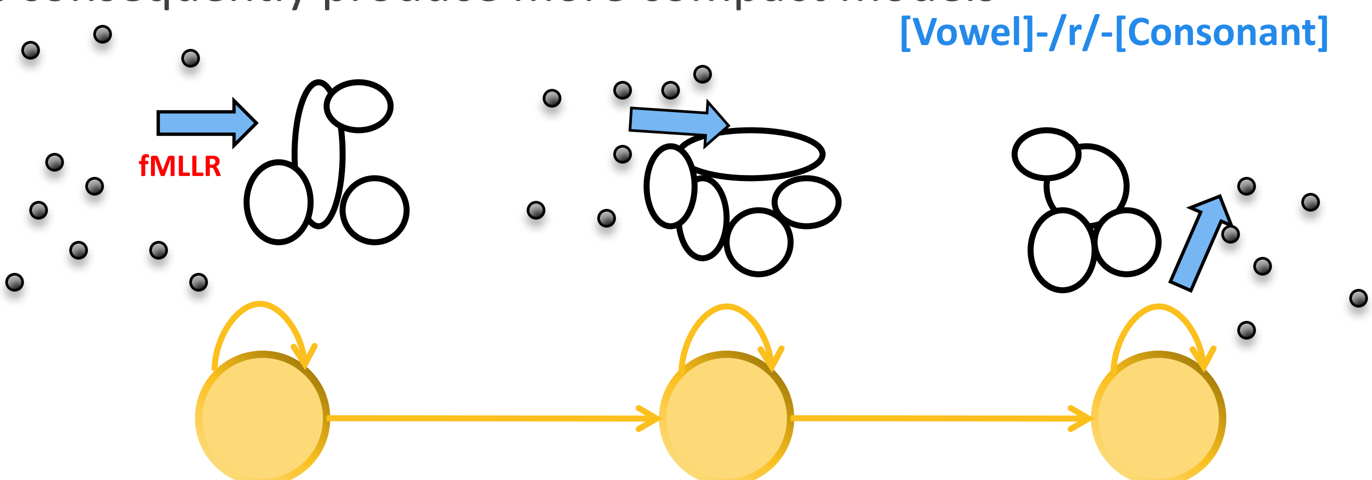
Results (DET curves of PRLM and GMM-UBM) – 30s Cuts (Data II)



Approach	EER (%)
PRLM	17.7
GMM-UBM	15.3

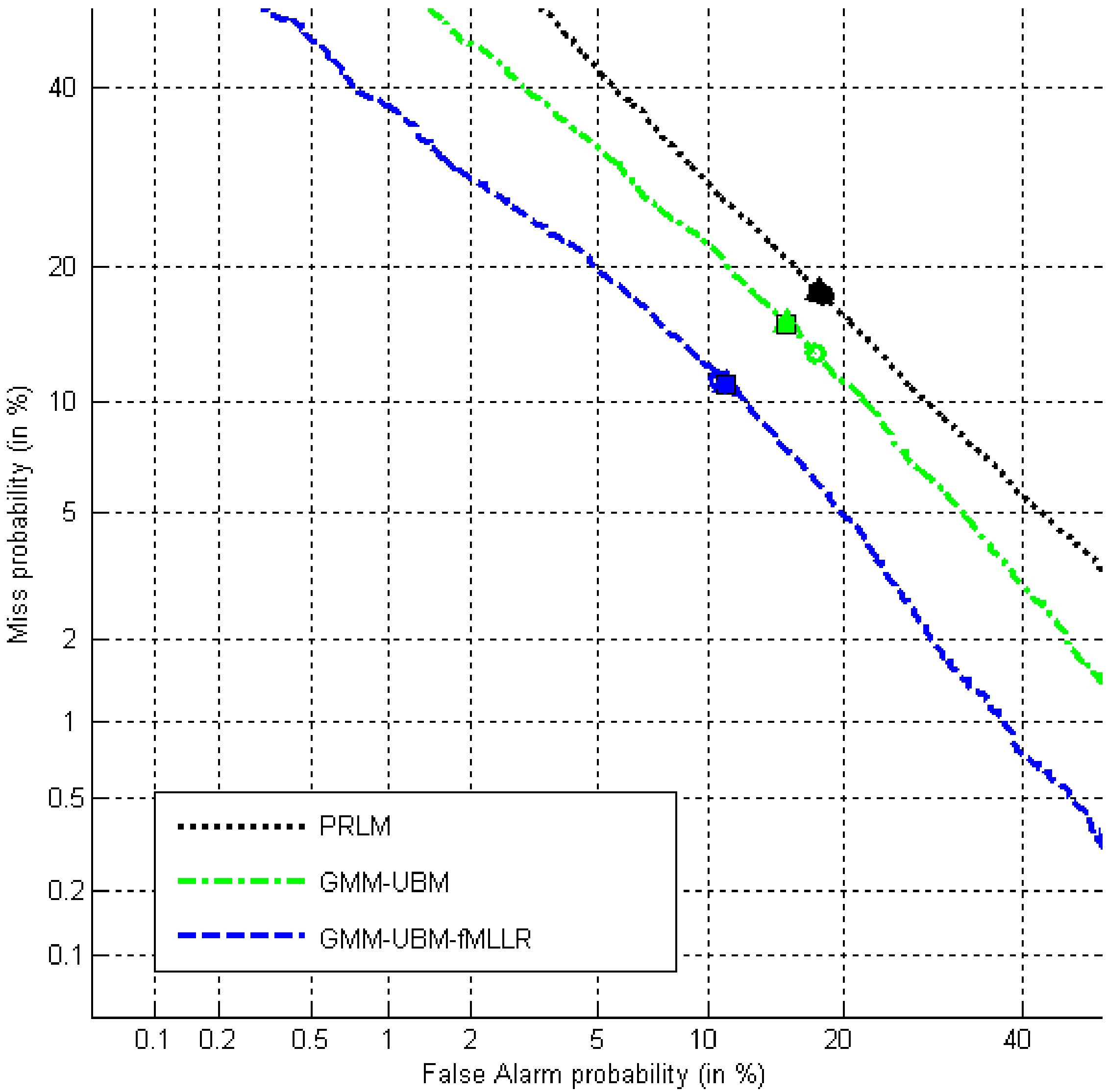
Our GMM-UBM Improved with fMLLR

- 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



Approach	EER (%)
PRLM	17.7
GMM-UBM	15.3
GMM-UBM-fMLLR	**11.0%**

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

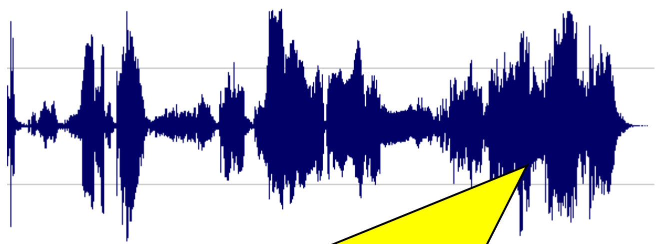
- **Hypothesis:** Dialects differ in their allophones (context-dependent phones) and their phonotactics

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

> */r/* is Approximant in American English [ɹ] and trilled in Scottish in *[Consonant] – /r/ – [Vowel]*

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 Attila context-dependent phone recognizer **<u>(trained on MSA)</u>**

Context-dependent (CD) phone sequence

...

[Back vowel]-r-[Central Vowel]

[Plosive]-A-[Voiced Consonant]

[VCentral]-b-[High Vowel]

...

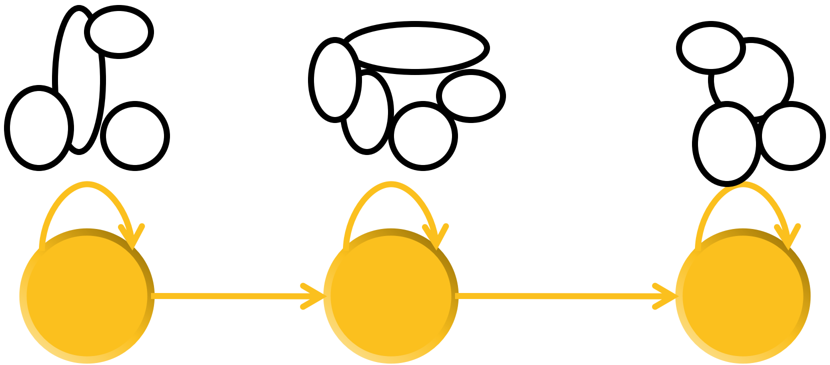
...

** not just /r/ /A/ /b/*

Do the above for all training data of all dialects

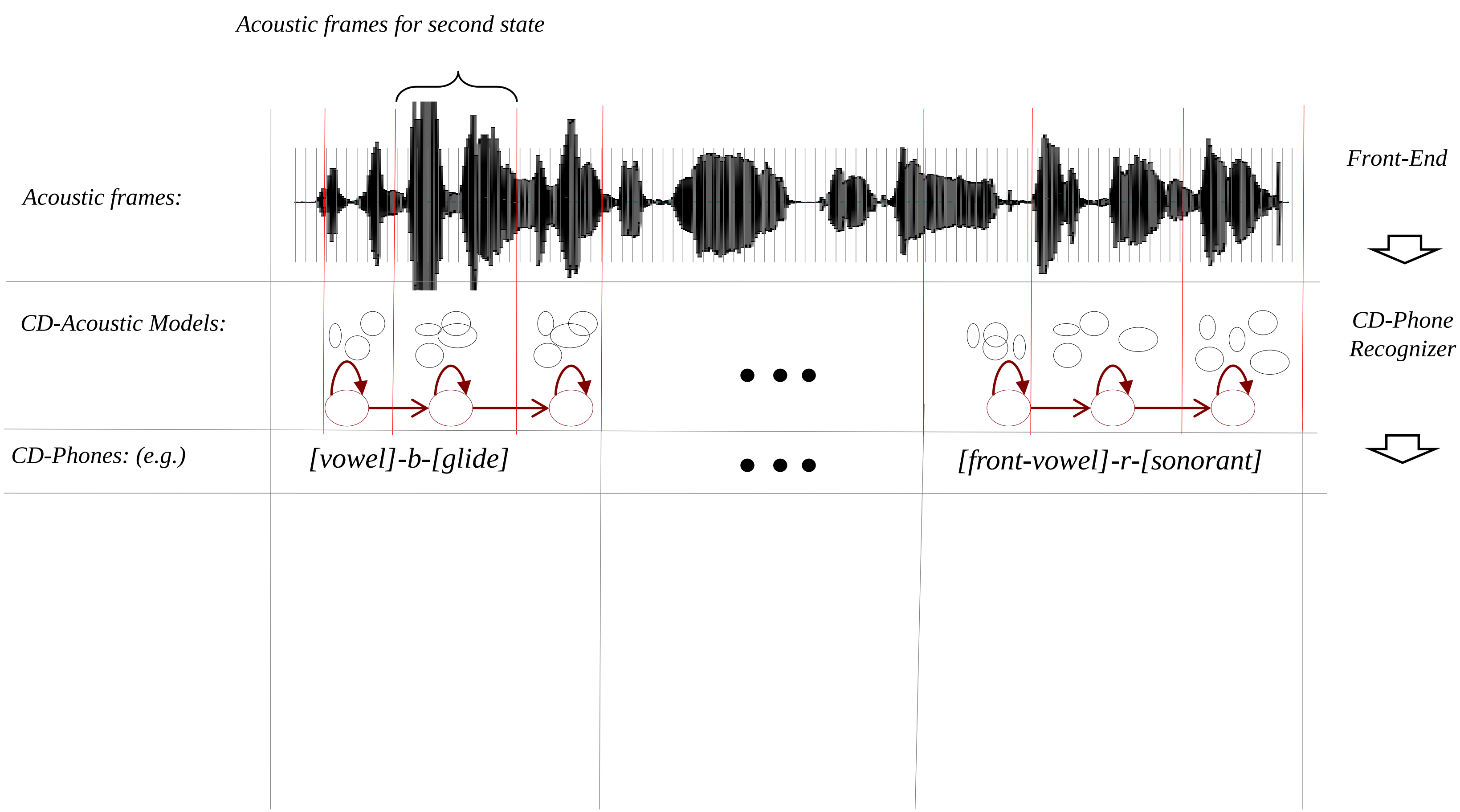
CD-Phone Universal Background Acoustic Model

Each CD phone type has an acoustic model:

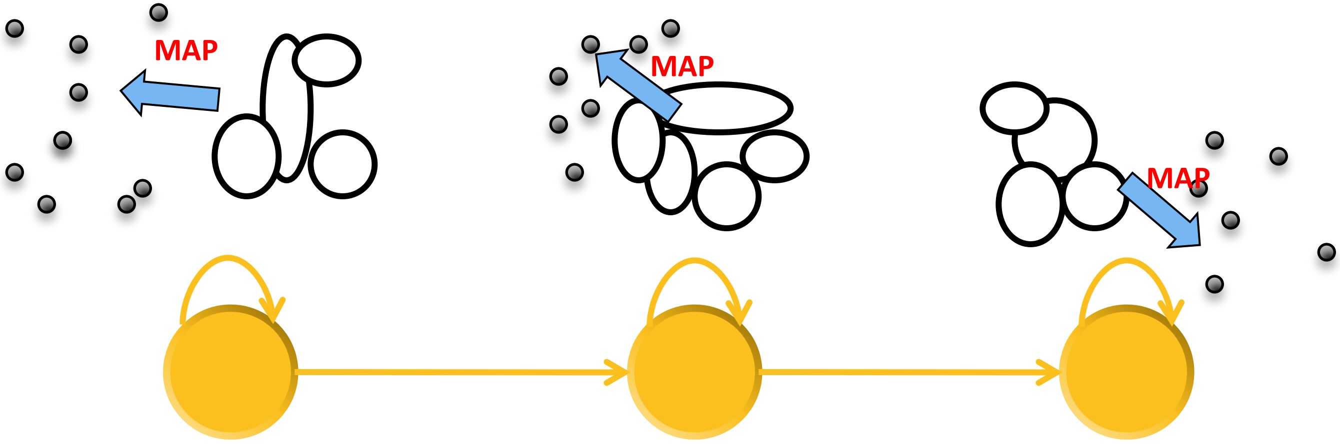


e.g., [Back vowel]-r-[Central Vowel]

Obtaining CD-Phones + Frame Alignment



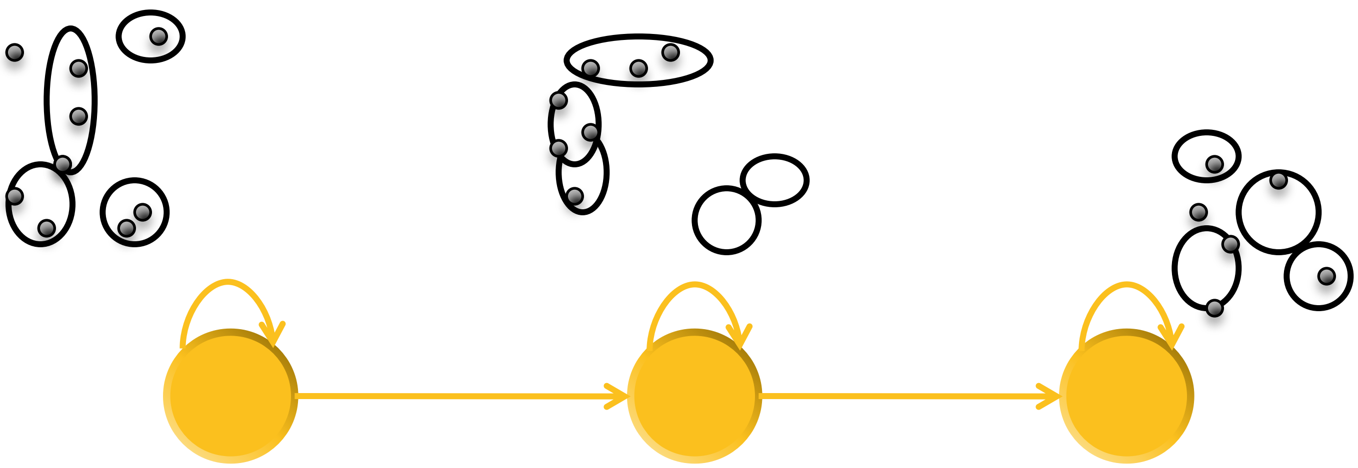
MAP Adaptation of each **CD-Phone Instance**



[Back Vowel]-r-[Central Vowel]

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

MAP Adaptation of each **CD-Phone Instance**



[Back Vowel]-r-[Central Vowel]

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

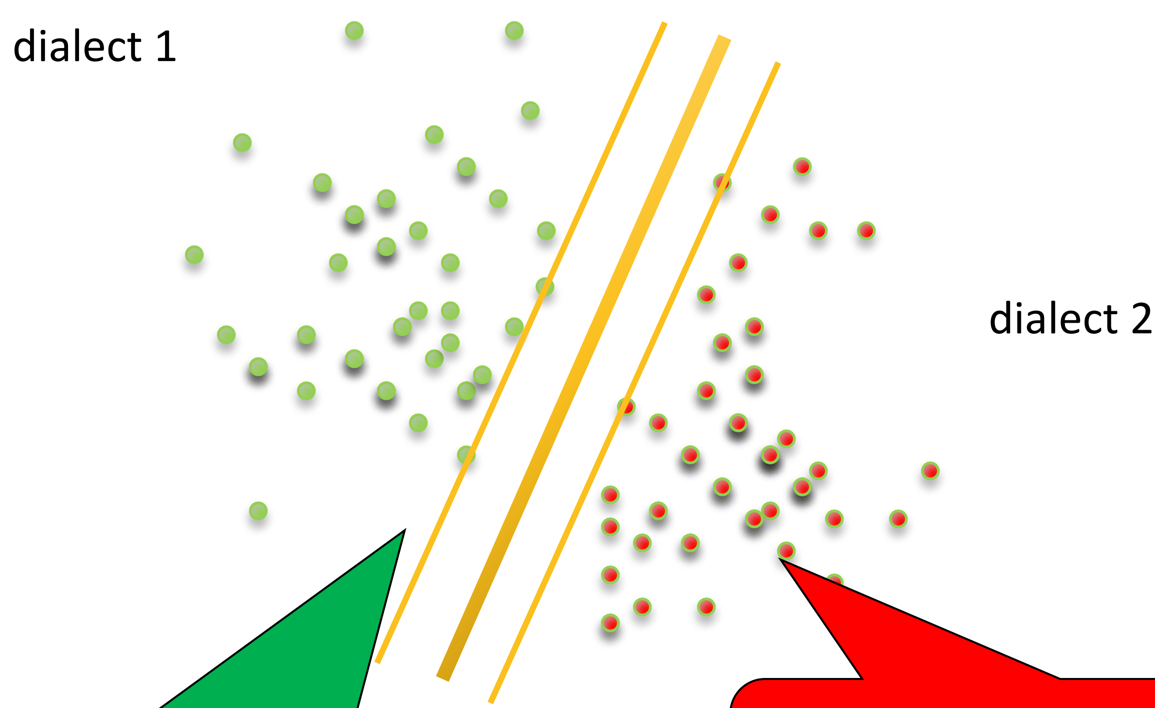
One Super Vector for each CD phone instance:

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

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

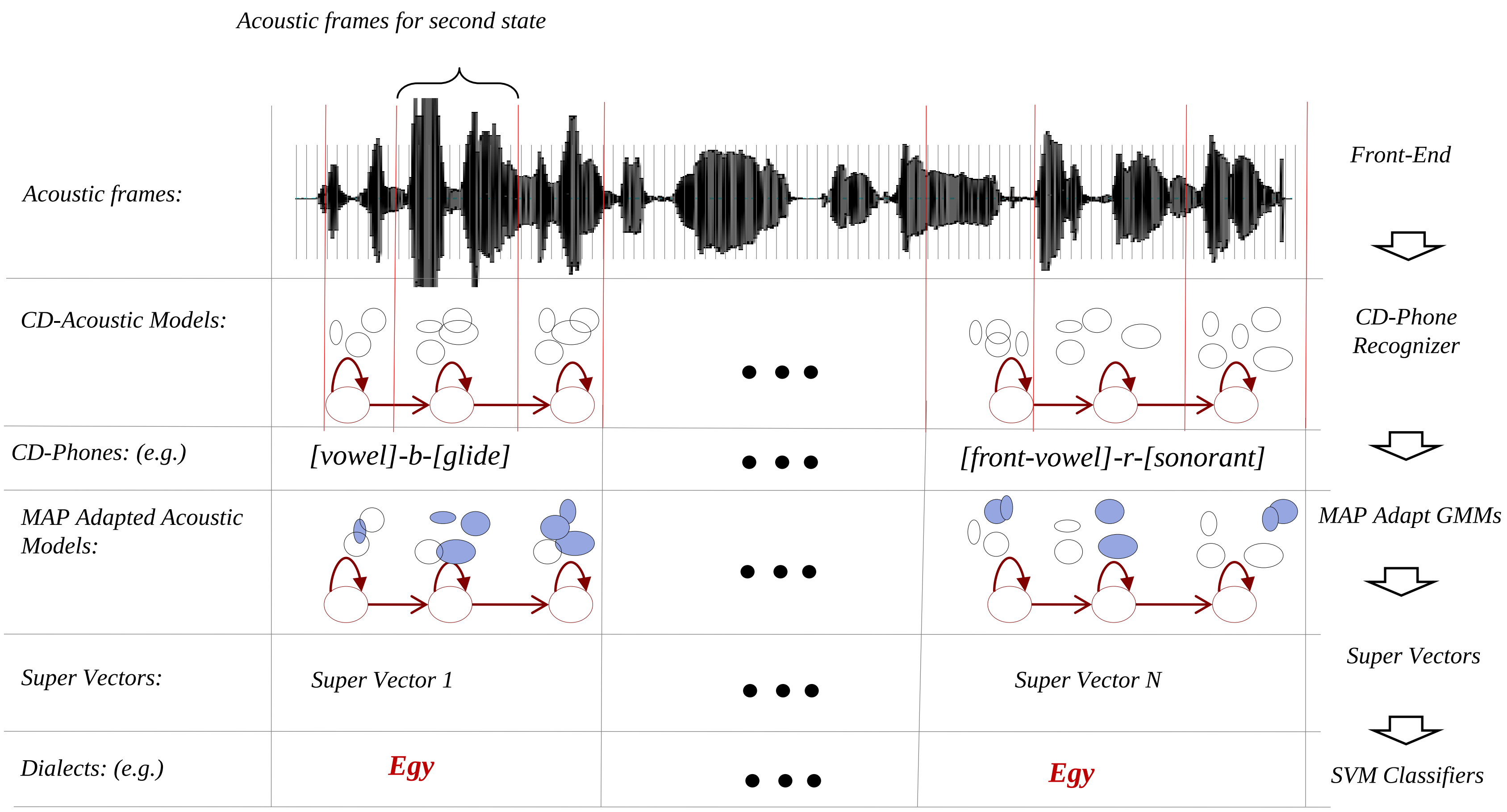
[Back Vowel]-r-[Central Vowel]



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

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

Discriminative Phonotactics – CD-Phone Classification



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 1st half and test on 2nd

Dialect Pair	Num. of * classifiers	Weighted accuracy (%)
Egyptian/Iraqi	195	70.9
Egyptian/Gulf	196	69.1
Egyptian/Levantine	199	68.6
Levantine/Iraqi	172	63.96
Gulf/Iraqi	166	61.77
Levantine/Gulf	179	61.53

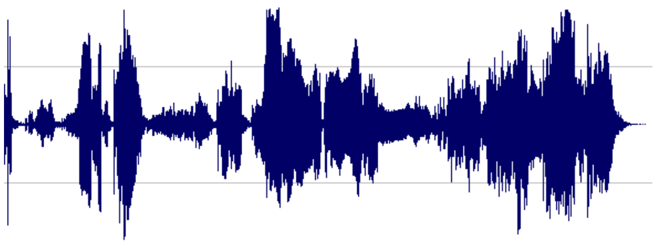
Extraction of Linguistic Knowledge

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

CD-Phone ([l-context]–phone–[r-context])	Accuracy	#
[*]–***sh***–[*]	71.1	6302
[SIL]–***a***–[*]	70.3	3935
[SIL]–***?***–[Central Vowel]	68.7	1323
[*]–***j***–[*]	68.5	3722
[! Central Vowel]–***s***–[! High Vowel]	68.5	1975
[Nasal]–***A***–[Anterior]	68.1	5459
[!SIL & ! Central Vowel]–***E***–[!Central Vowel]	67.8	3687
[Central Vowel]–***m***–[Central Vowel]	66.7	2639
[!Voiced Cons. & !Glottal & !Pharyngeal & !Nasal & !Trill & !w & !Emphatic]–***A***–[Anterior]	66.4	11857
[*]–***k***–[Central Vowel]	66.4	1433
...	...	...
[!SIL & !Central Vowel]–***G***–[!Central Vowel]	57.5	852
[!A]–***h***–[Back Vowel]	57.0	409
[!Vowel & !SIL]–***m***–[!Central Vowel & !Back Vowel]	56.2	300

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***

…

…

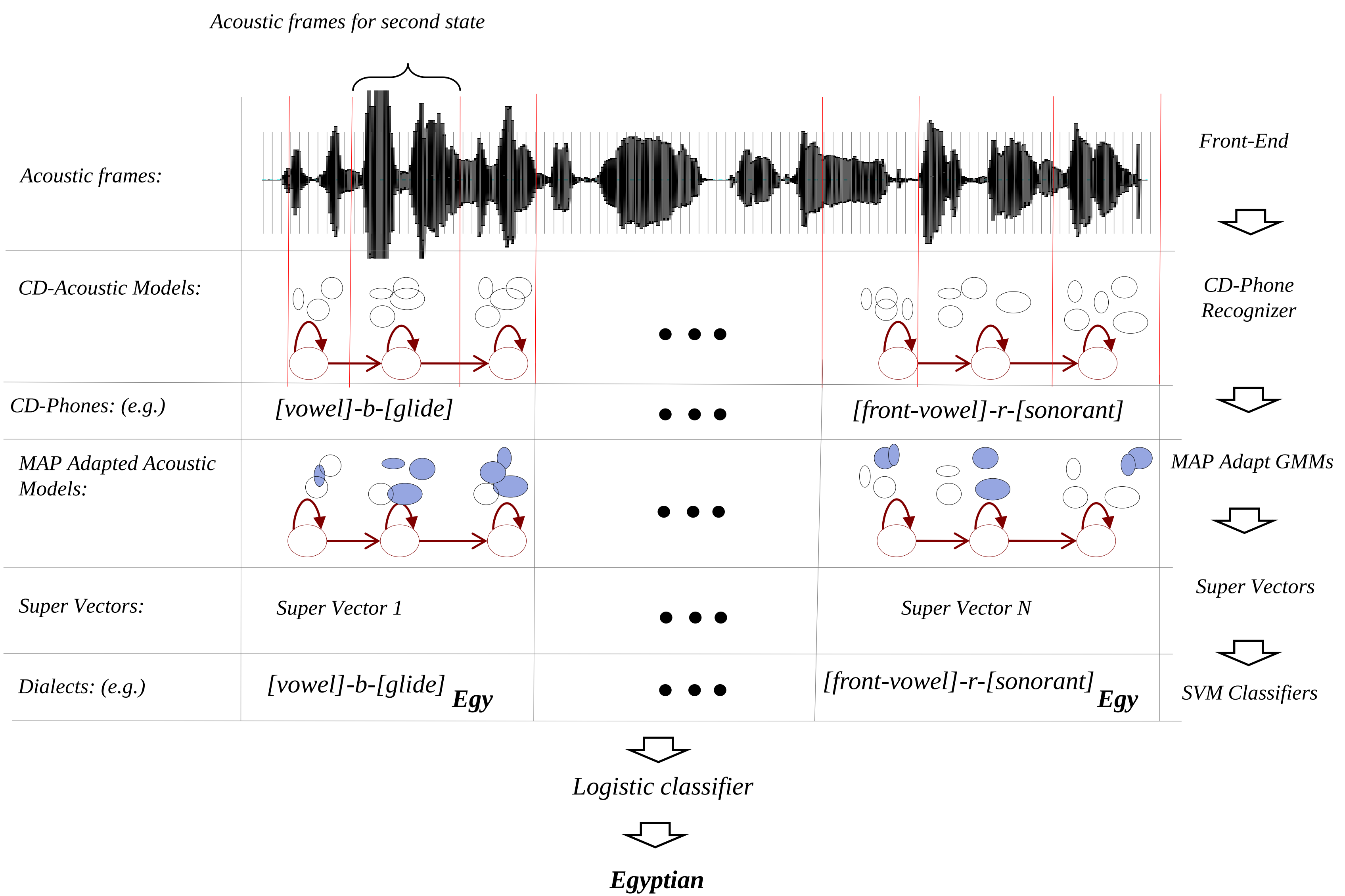
Textual Feature Extraction for Discriminative Phonotactics

- Extract the following textual features from each pair of dialects
  - Frequency of annotated CD-Phone bigrams, e.g.,
    "[Nasal]–$\boldsymbol{r}$–[Vowel]$_{Iraqi}$ [Voiced Cons.]–$\boldsymbol{a}$–[Liquid]$_{Gulf}$"
  - Frequency of bigrams with only one annotated CD-Phone, e.g.,
    "[Nasal]–$\boldsymbol{r}$–[Vowel] [Voiced Cons.]–$\boldsymbol{a}$–[Liquid]$_{Gulf}$"
  - Frequency of annotated unigrams, e.g.,
    [!Central Vowel]–$\boldsymbol{E}$–[Central Vowel]$_{Gulf}$
  - Frequency of not annotated CD-Phone unigrams and bigrams, e.g.,
    "[Nasal]–$\boldsymbol{r}$–[Vowel] [Voiced Cons.]–$\boldsymbol{a}$–[Liquid]"
  - Frequency of context *independent* phone *trigrams*, e.g.,
    "s A l"
- Normalize vector by its norm
- Train a logistic regression with L2 regularizer

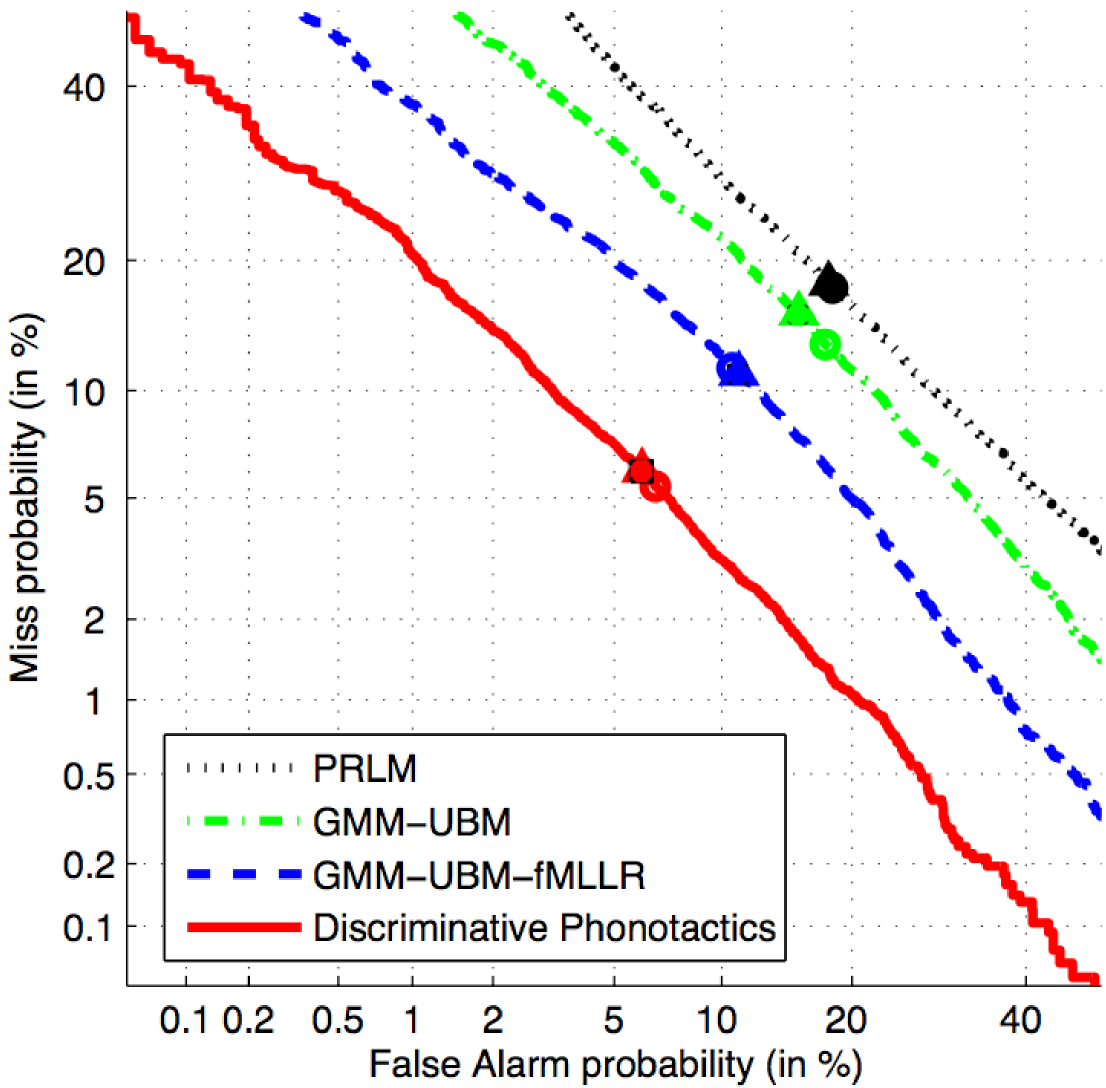
Experiments – Training Two Models

- 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 2nd half
- Train the logistic classifier on the annotated sequences

Discriminative Phonotactics – Dialect Recognition

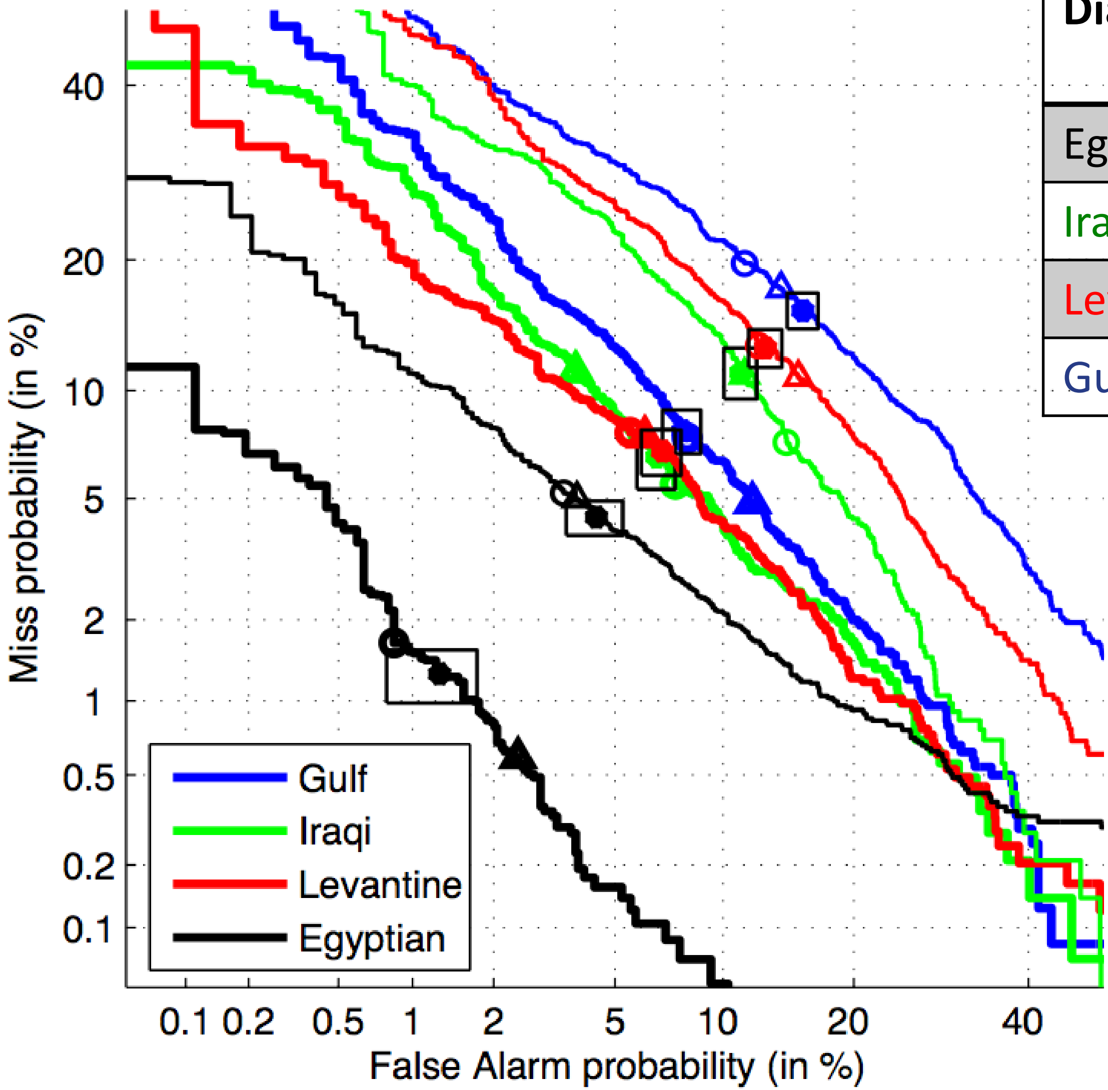


Results – Discriminative Phonotactics



Approach	EER (%)
PRLM	17.7
GMM-UBM	15.3
GMM-UBM-fMLLR	11.0%
Disc. Phonotactics	**6.0%**

Results per Dialect



Dialect	GMM fMLLR	Disc. Pho.
Egyptian	4.4%	1.3%
Iraqi	11.1%	6.6%
Levantine	12.8%	6.9%
Gulf	15.6%	7.8%

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>
<tr><th>Month</th><th colspan="2">Tasks</th></tr>
<tr><td>Mar 2010</td><td>Further analyses of the disciminative 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 colspan="2">Build the new Bi-phone system using HTK</td></tr>
<tr><td rowspan="2">Jun 2010</td><td colspan="2">Test different techniqiues for biphone acoustic models on Arabic dialects</td></tr>
<tr><td colspan="2" rowspan="2">Experiment with different languages: Chinese, Spanish, American vs. Indian English, and American English Dialects</td></tr>
<tr><td rowspan="2">July 2010</td></tr>
<tr><td colspan="2" rowspan="3">Work with IBM to Improve Arabic ASR using the best approach for dialect ID</td></tr>
<tr><td>Aug 2010</td></tr>
<tr><td>Sep 2010</td></tr>
<tr><td>Oct 2010</td><td colspan="2" rowspan="2">Write Dissertation</td></tr>
<tr><td>Nov 2010</td></tr>
<tr><td rowspan="2">Dec 2010</td><td colspan="2">Prepare Dissertation Defense</td></tr>
<tr><td colspan="2">Defend Dissertation</td></tr>
</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

Frame Alignment

For each CD phone sequence:

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