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

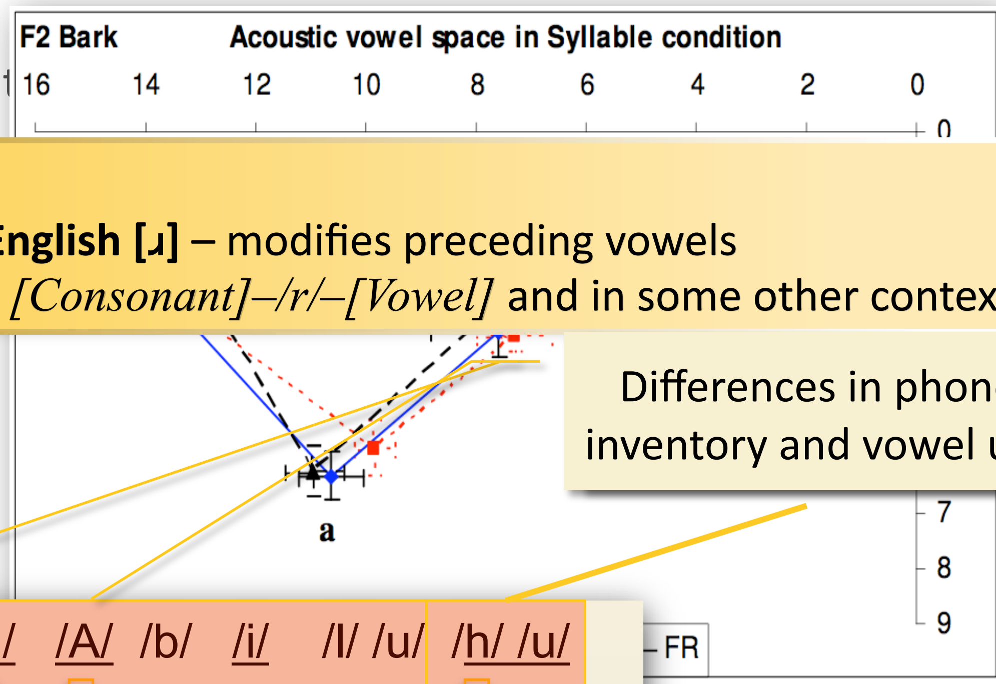
- Similar to language recognition, but use dialects/accents of the same language
- Dialects may differ in any dimension of the linguistic spectrum
  - Differences are likely to be more subtle across dialects than those across languages
  - Thus, more challenging problem than language recognition

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 (cont
- Ph



Example: */r/*

Approximant in American English [ɹ] – modifies preceding vowels

Trilled in Scottish English in *[Consonant]–/r/–[Vowel]* and in some other contexts

Differences in phonetic inventory and vowel usage

"She will meet him"

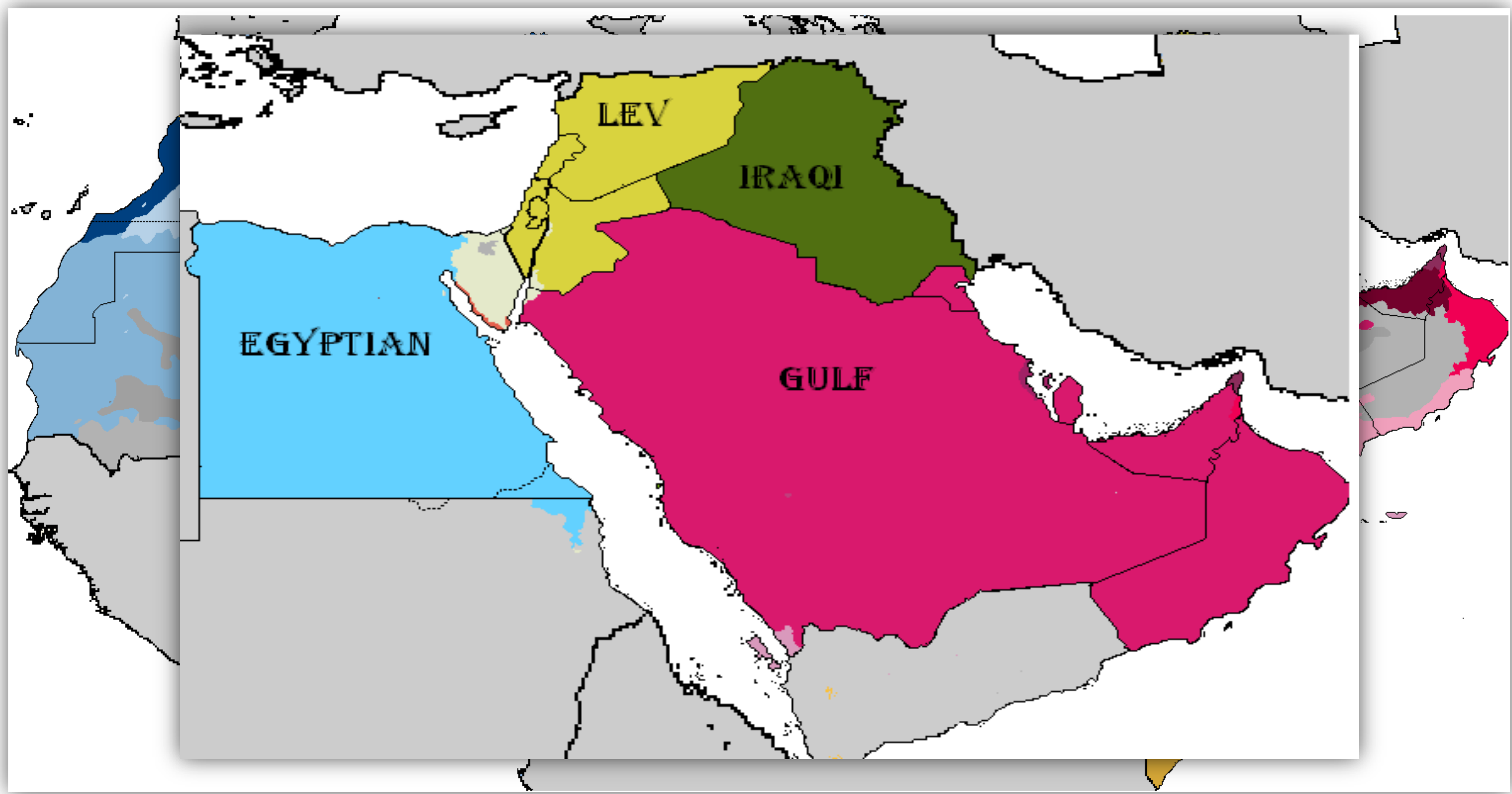
MSA:	/s/ /a/	/t/ /u/ /q/	/A/	/b/	/i/	/l/ /u/	/h/ /u/
Egy:	/H/ /a/	/t/ /?/	/a/	/b/		/l/	/u/
Lev:	/r/ /a/ /H/	/t/ /g/	/A/	/b/		/l/	/u/

2005)

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

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)

*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 → **40D**
  - CMVN
- Speaker Adaptation:
  - **fMLLR followed by MLLR**
- Unigram phone language model trained on MSA

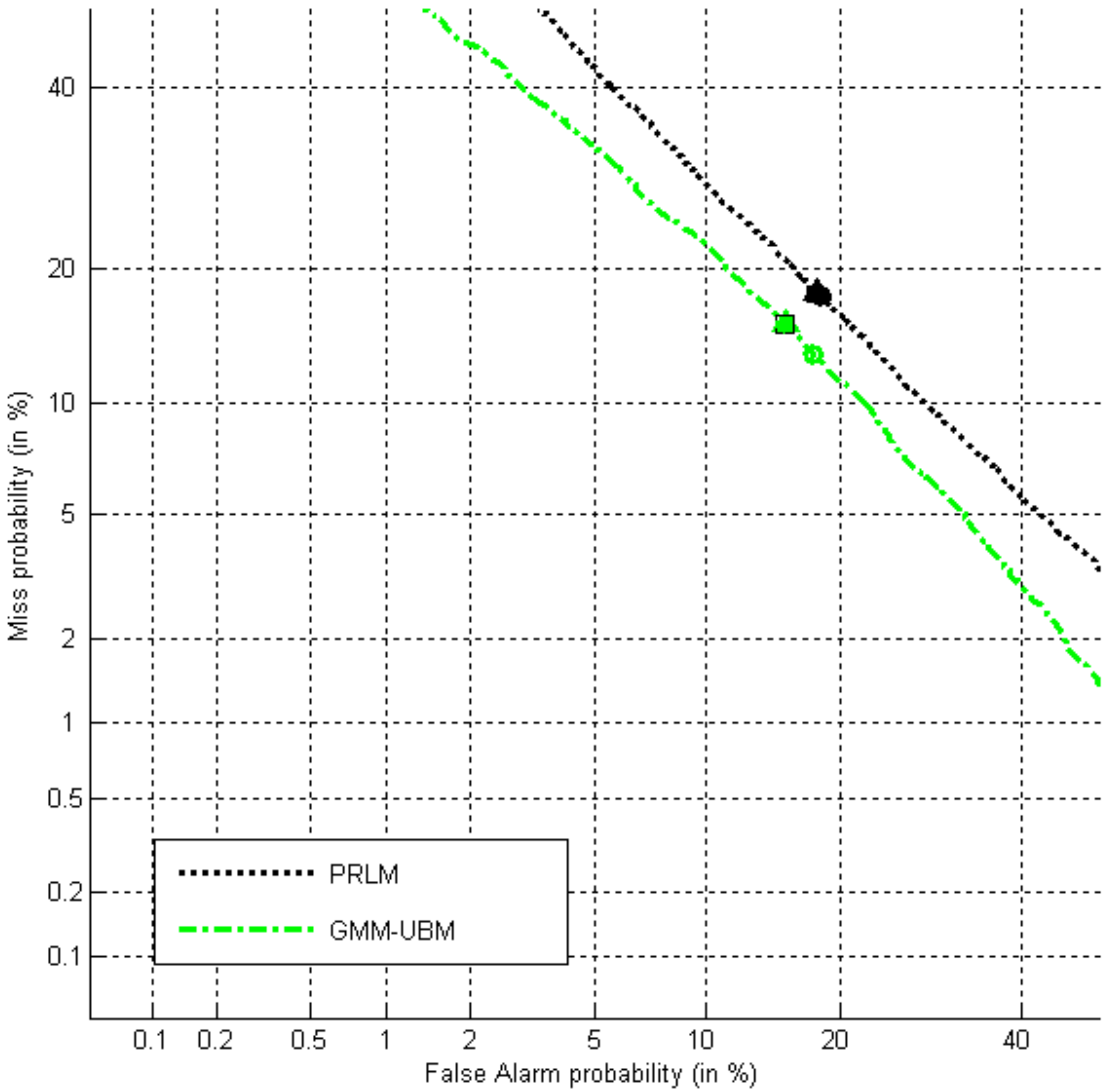
Outline

- 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



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

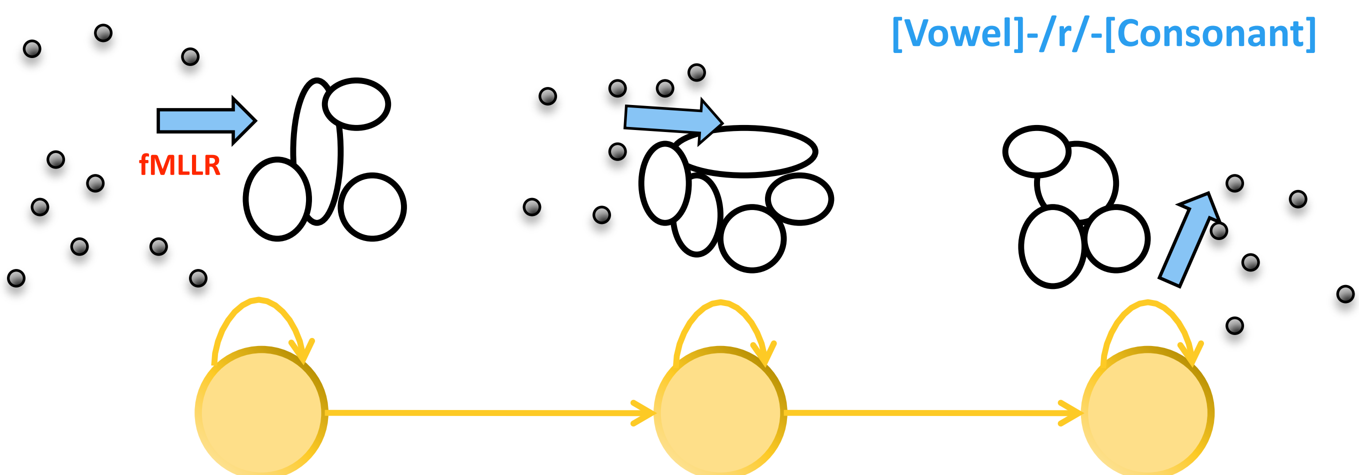
*Comparable to GMM-UBM of (Torres-Carrasquillo et al., 2008) on 3 dialects

Outline

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

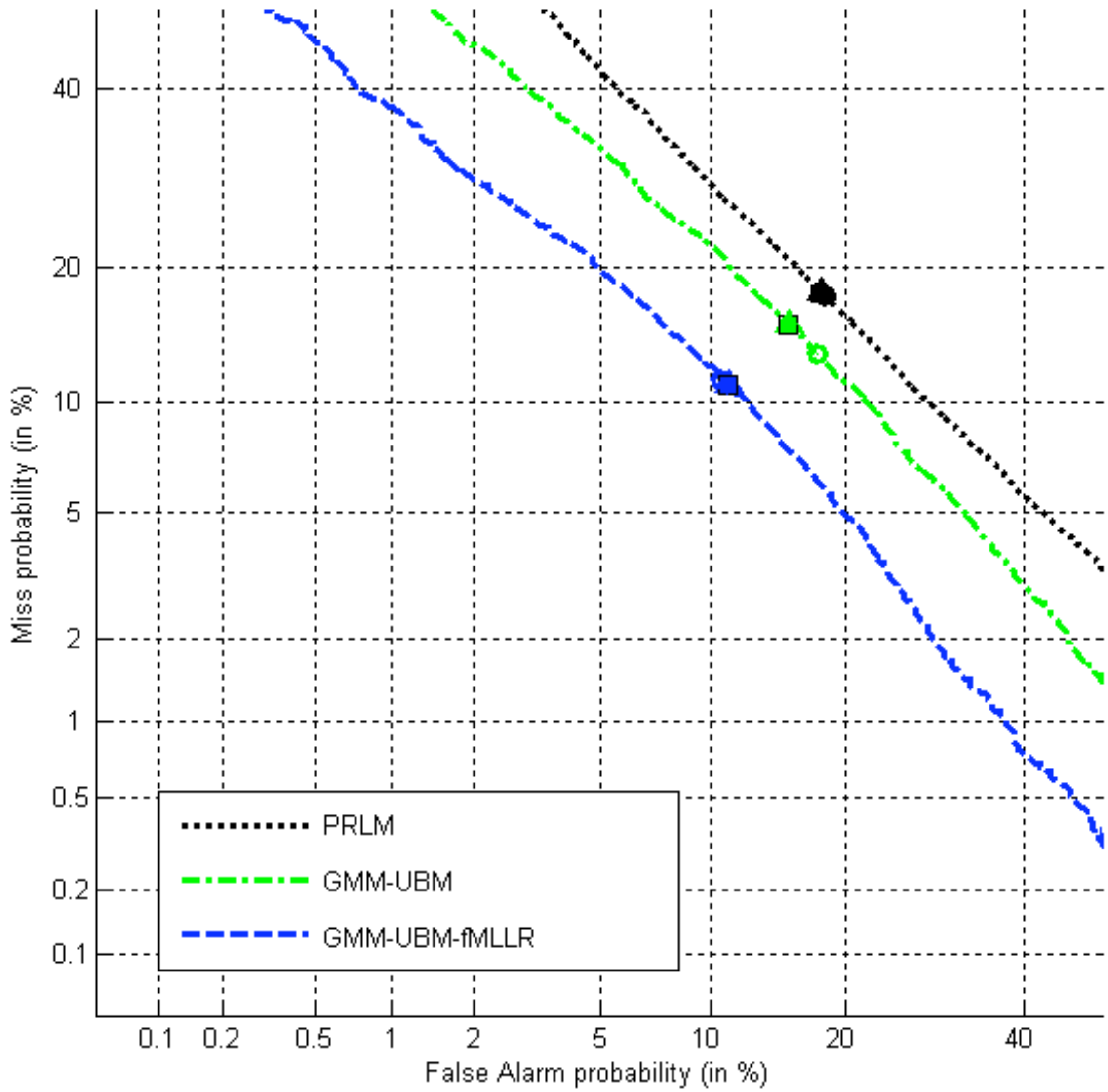
Our GMM-UBM Improved with fMLLR

- 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



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

Outline

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

Discriminative Phonotactics

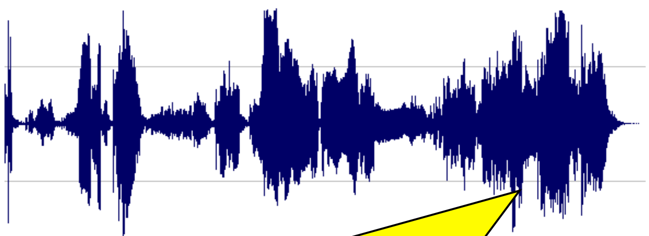
- **Hypothesis:** Dialects differ in their allophones (context-dependent phones) and their phonotactics

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

/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 our CD-phone recognizer

CD-phone sequence

…

[Back vowel]-r-[Central Vowel]

[Plosive]-A-[Voiced Consonant]

[Central Vowel]-b-[High Vowel]

…

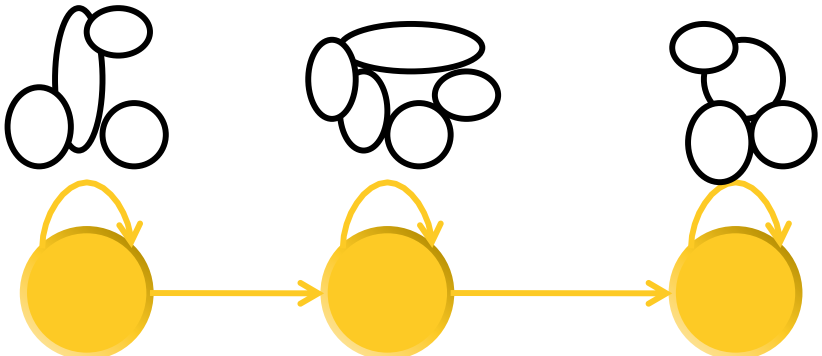
…

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

Do the above for all training data of all dialects

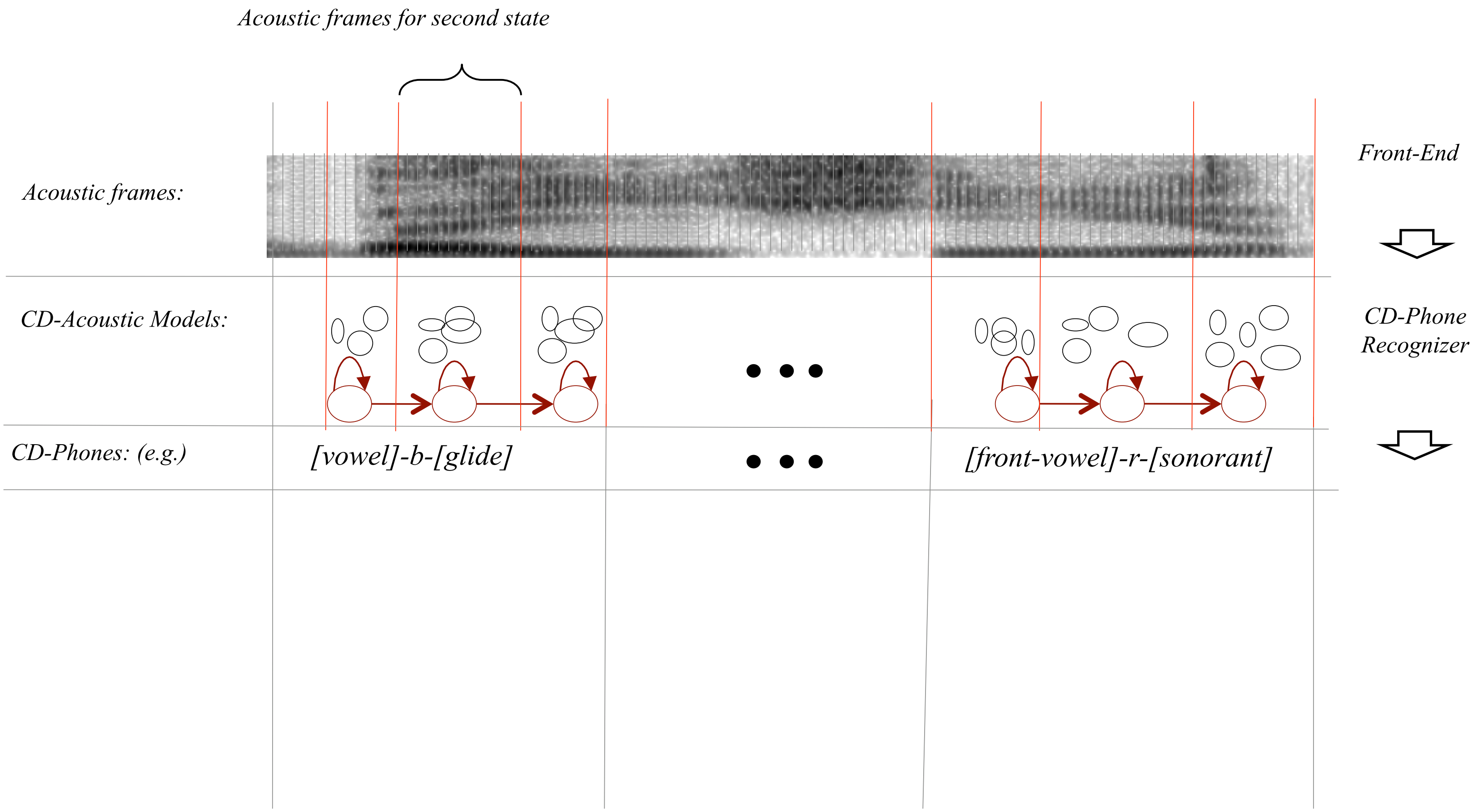
CD-Phone Universal Background Acoustic Model

Each CD phone type has an acoustic model:

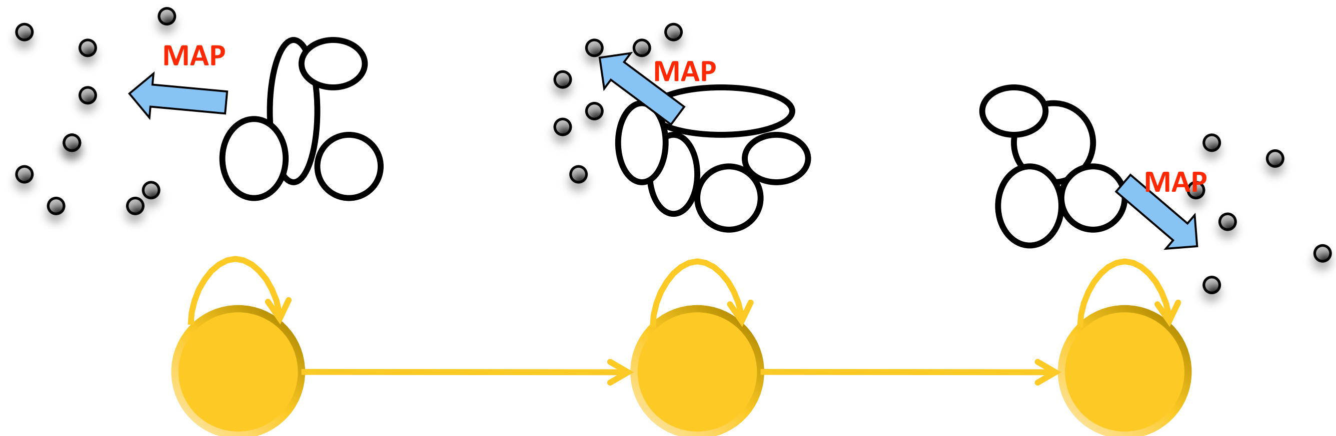


e.g., [Back vowel]-r-[Central Vowel]

Obtaining CD-Phones + Frame Alignment



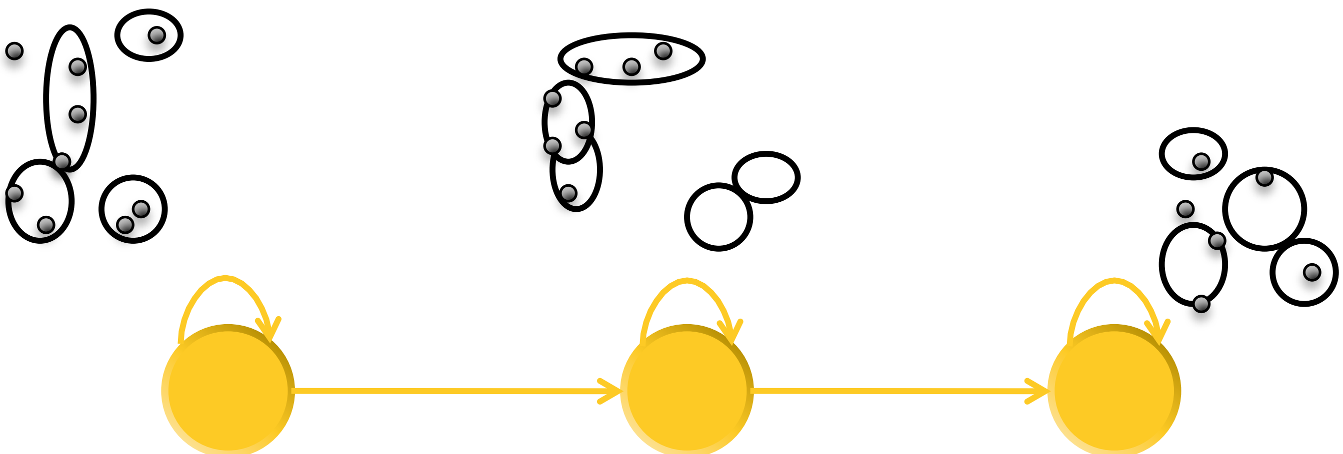
MAP Adaptation of each **CD-Phone Instance**



[Back Vowel]-r-[Central Vowel]

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

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



[Back Vowel]-r-[Central Vowel]

MAP adapt the CD-phone acoustic model GMMs to the corresponding frames*

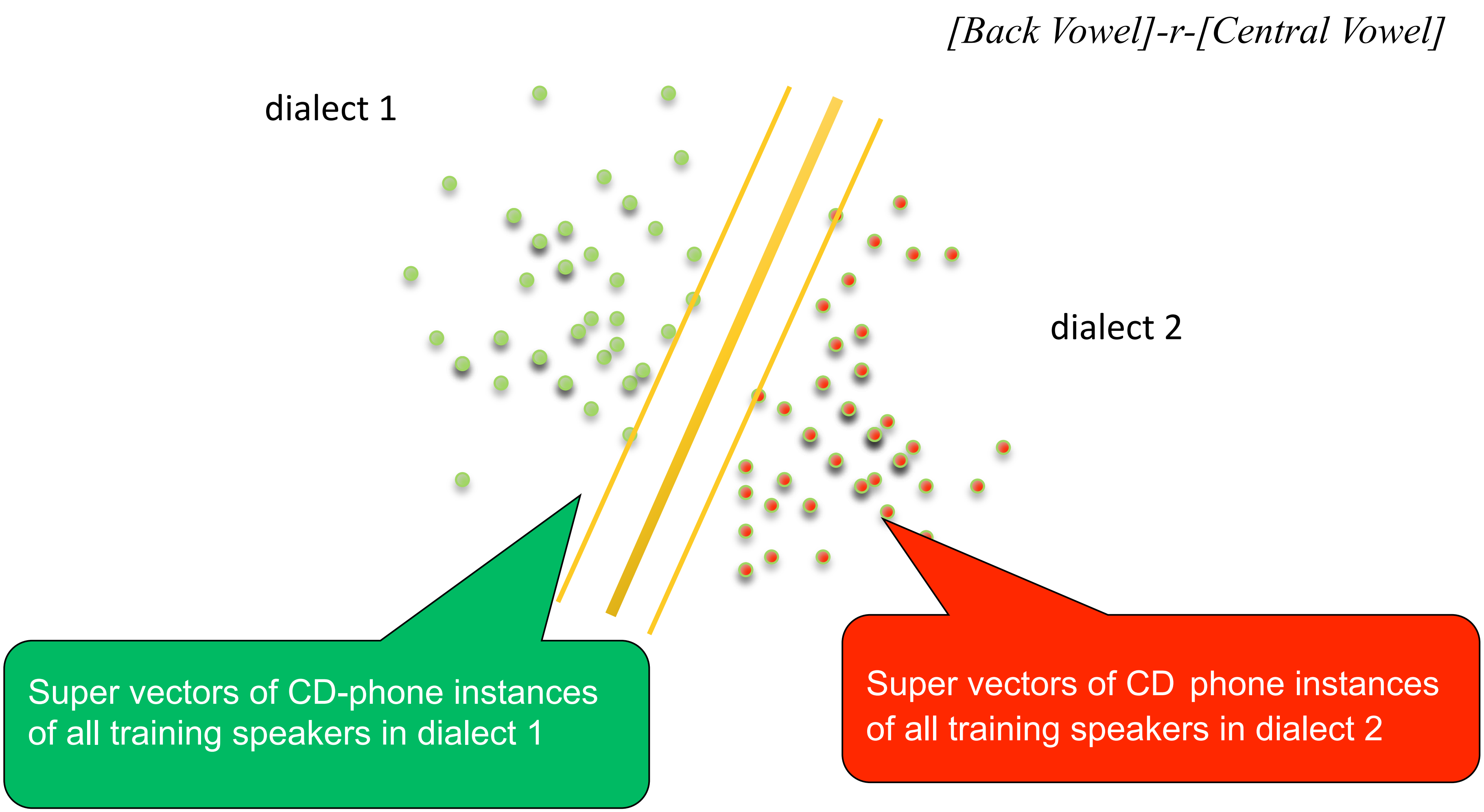
One Super Vector for each CD phone instance:

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

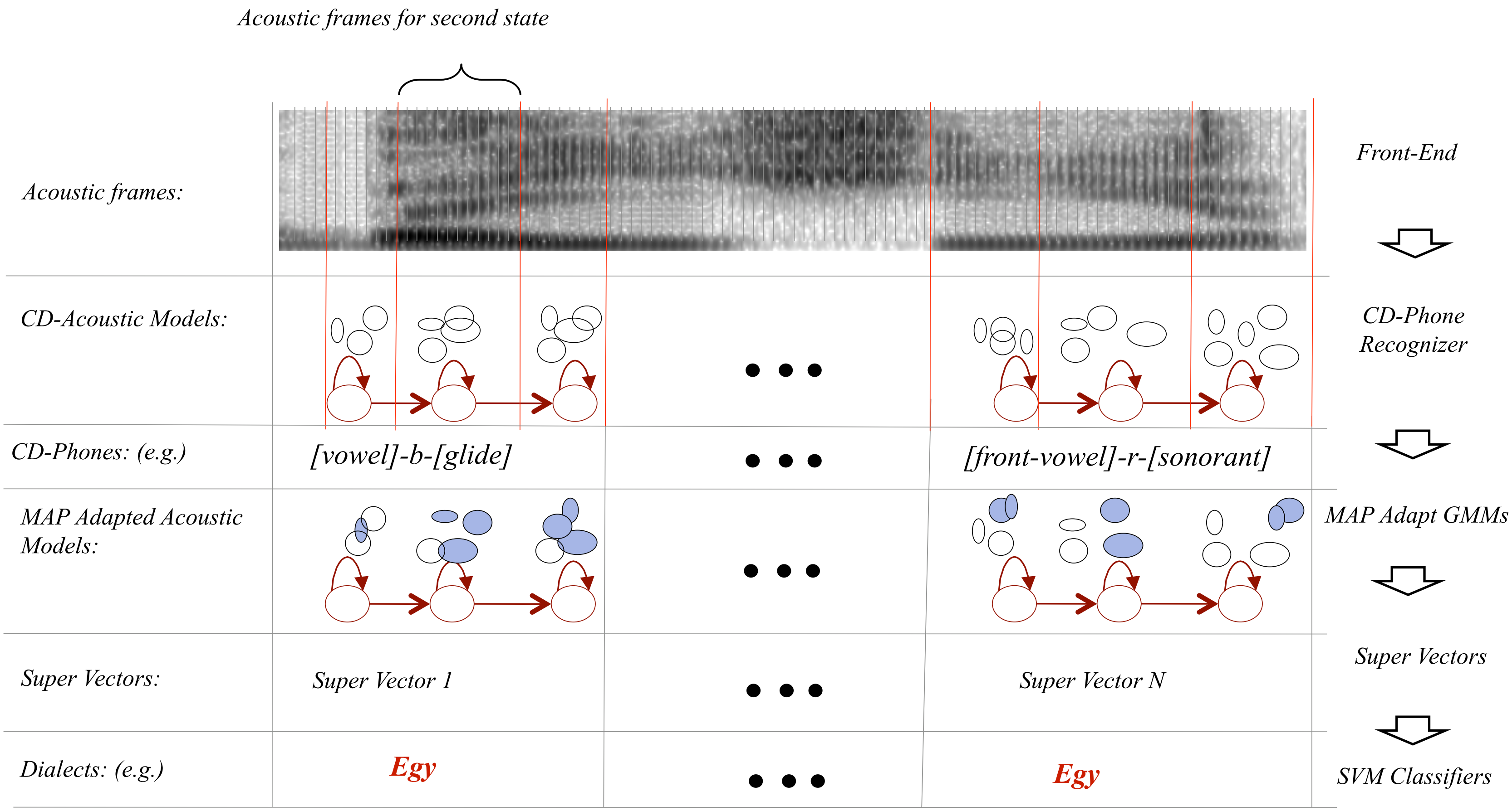
i.e., summarize the acoustic-phonetic features of each CD-phone in one vector

*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



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

* performed significantly better than chance (50%)

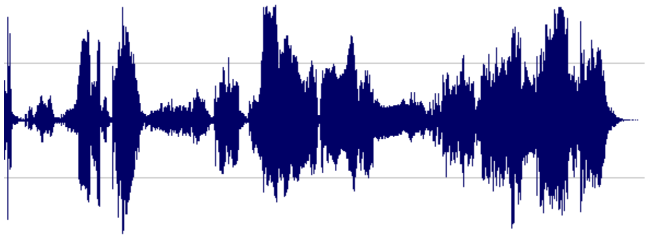
Extraction of Linguistic Knowledge

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

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

...

[Back vowel]-r-[Central Vowel]

[Plosive]-A-[Voiced Consonant]

[Central Vowel]-b-[High Vowel]

...

...

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

...

[Back vowel]-r-[Central Vowel] ***Egyptian***

[Plosive]-A-[Voiced Consonant] ***Egyptian***

[Central Vowel]-b-[High Vowel] ***Levantine***

...

...

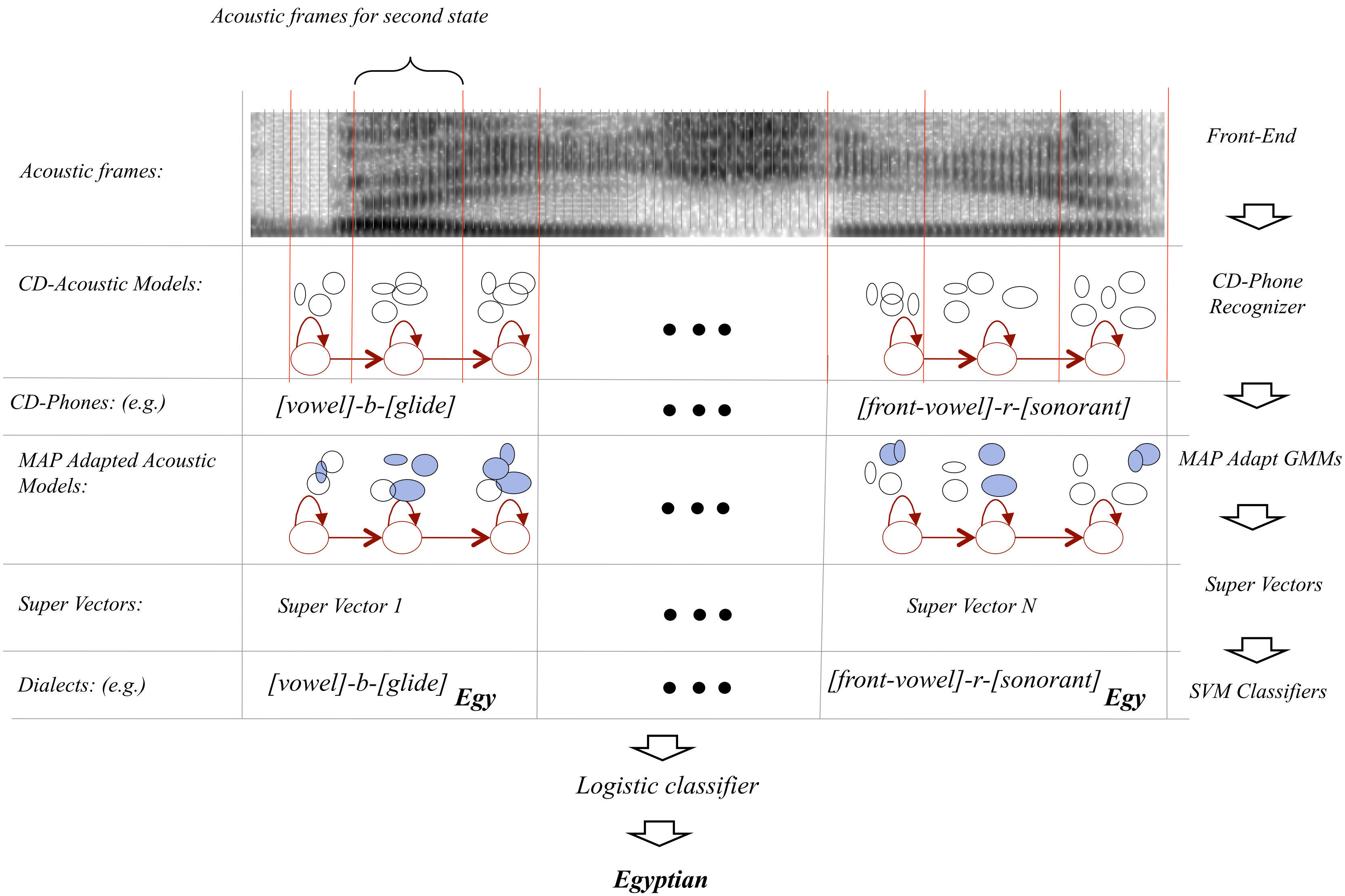
Textual Feature Extraction for Discriminative Phonotactics

- Extract the following textual features from each pair of dialects
  - Frequency of annotated CD-Phone bigrams, e.g.,
    "[Nasal]–***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 2nd half
- Train the logistic classifier on the annotated sequences

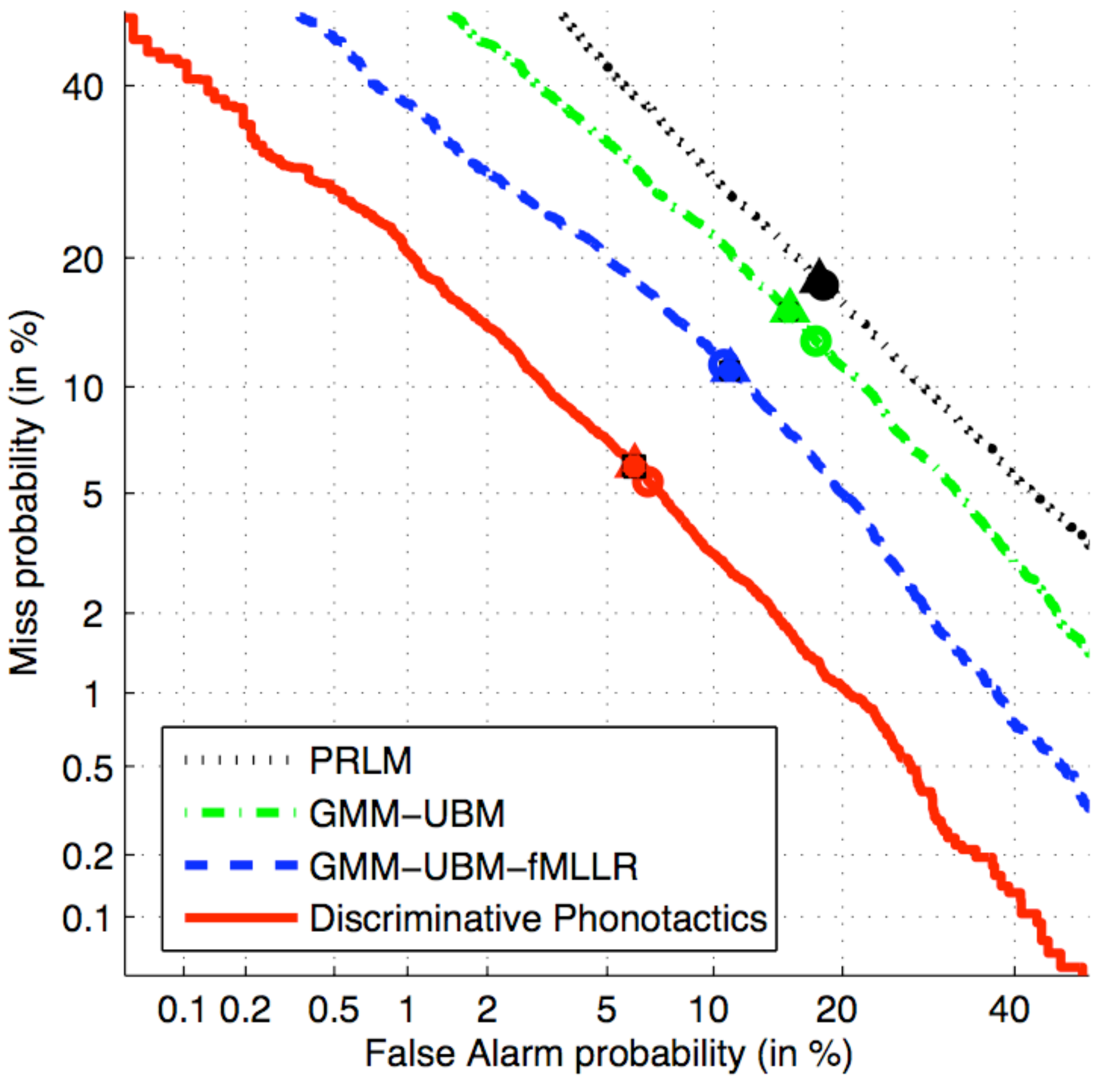
Discriminative Phonotactics – Dialect Recognition



Baselines

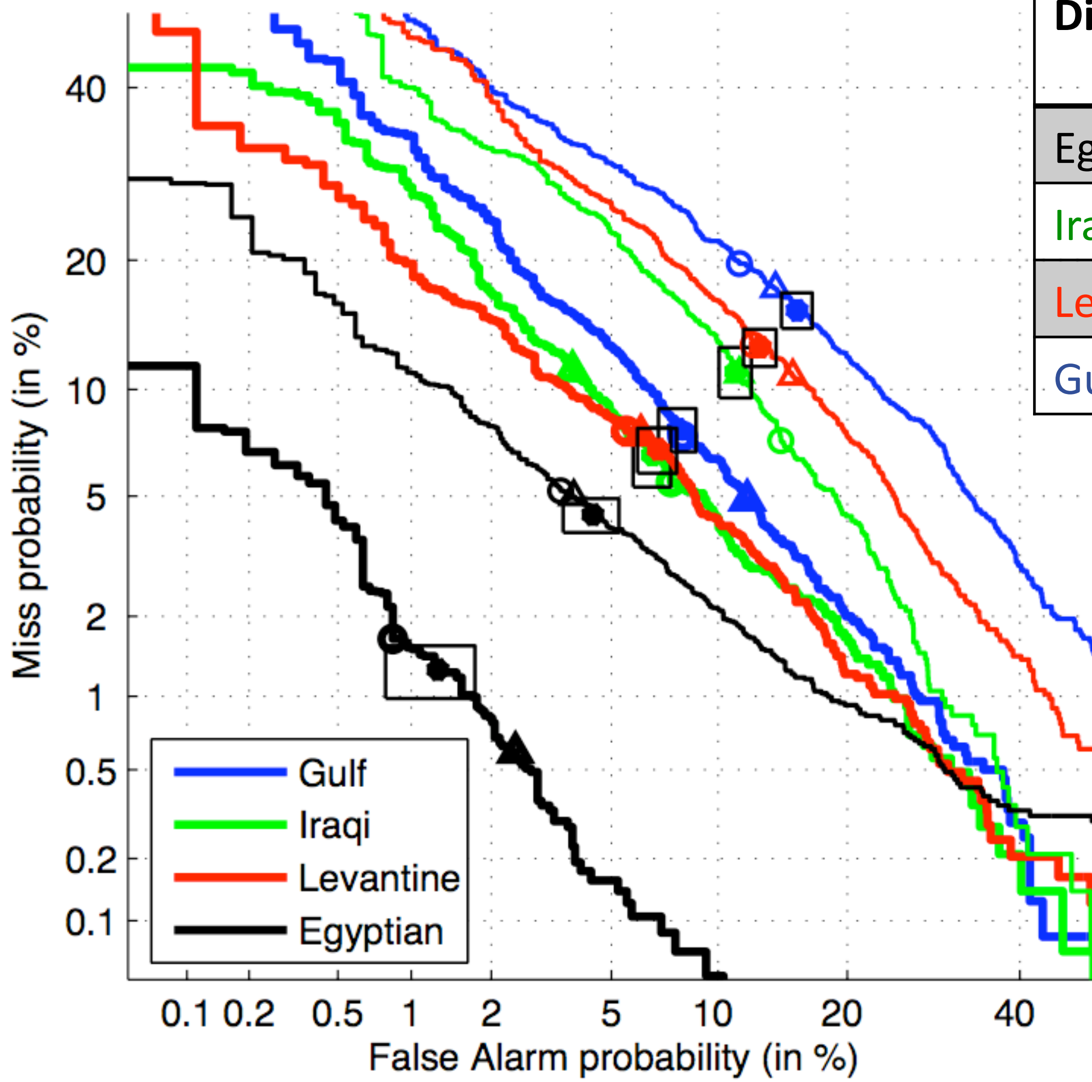
- Standard PRLM: a trigram phonotactic model per dialect
- Standard GMM-UBM:
  - Front-End:
    - 13D PLP features from 9 frames followed by LDA → 40D
    - CMVN
  - 2048 Gaussians – ML trained on equal number of frames from each dialect
  - Dialect Models are MAP adapted with 5 iterations (similar to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

Results – Discriminative Phonotactics



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%

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 super-vectors 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