Dialect Recognition Using a Phone-GMM-Supervector-Based SVM Kernel

Fadi Biadsy, *Julia Hirschberg, Michael Collins*

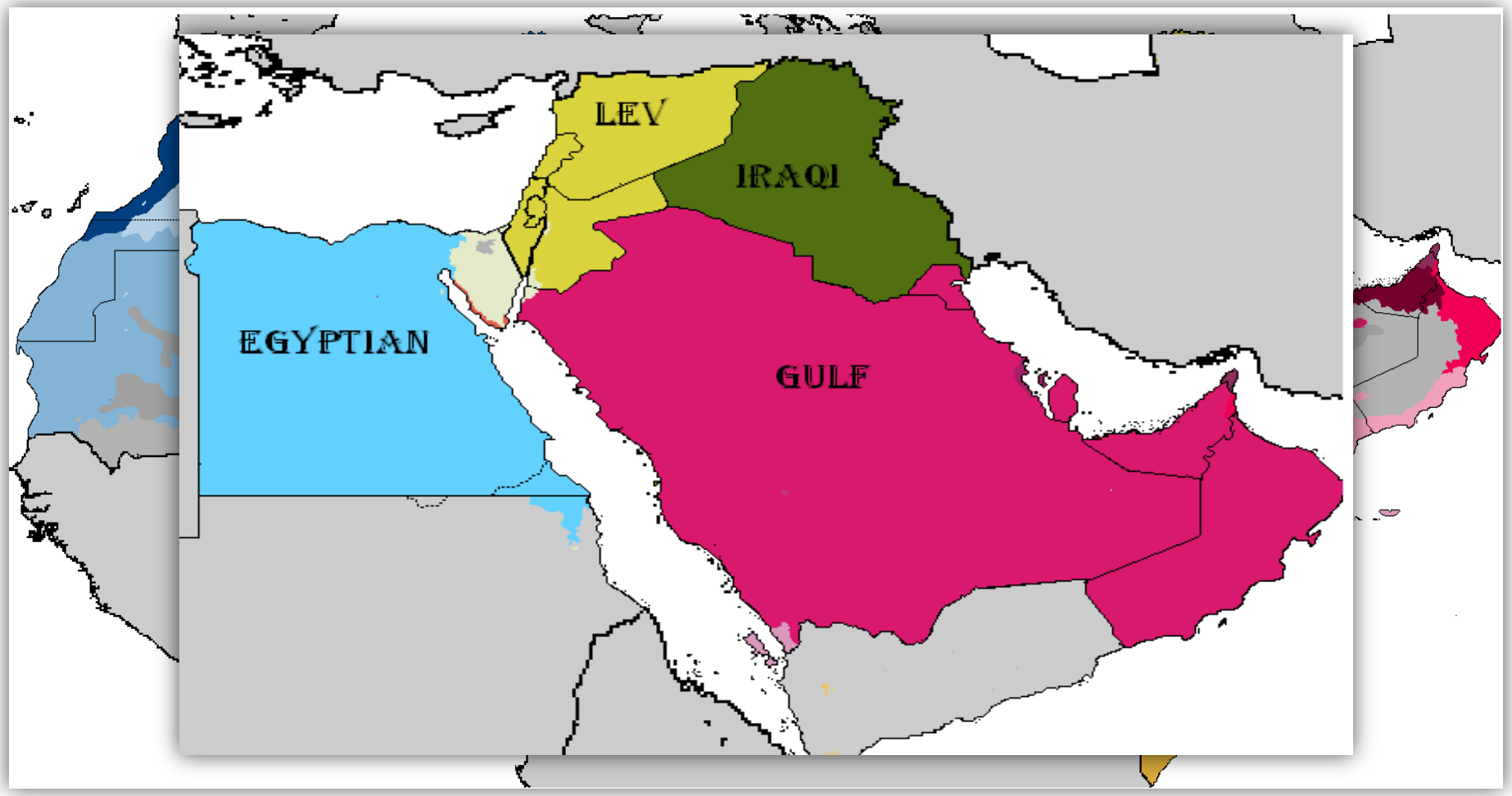
Columbia University, NY, USA

Sep 28st, 2010

Motivation: Why Study Dialect Recognition?

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

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

Baselines

I. Standard PRLM

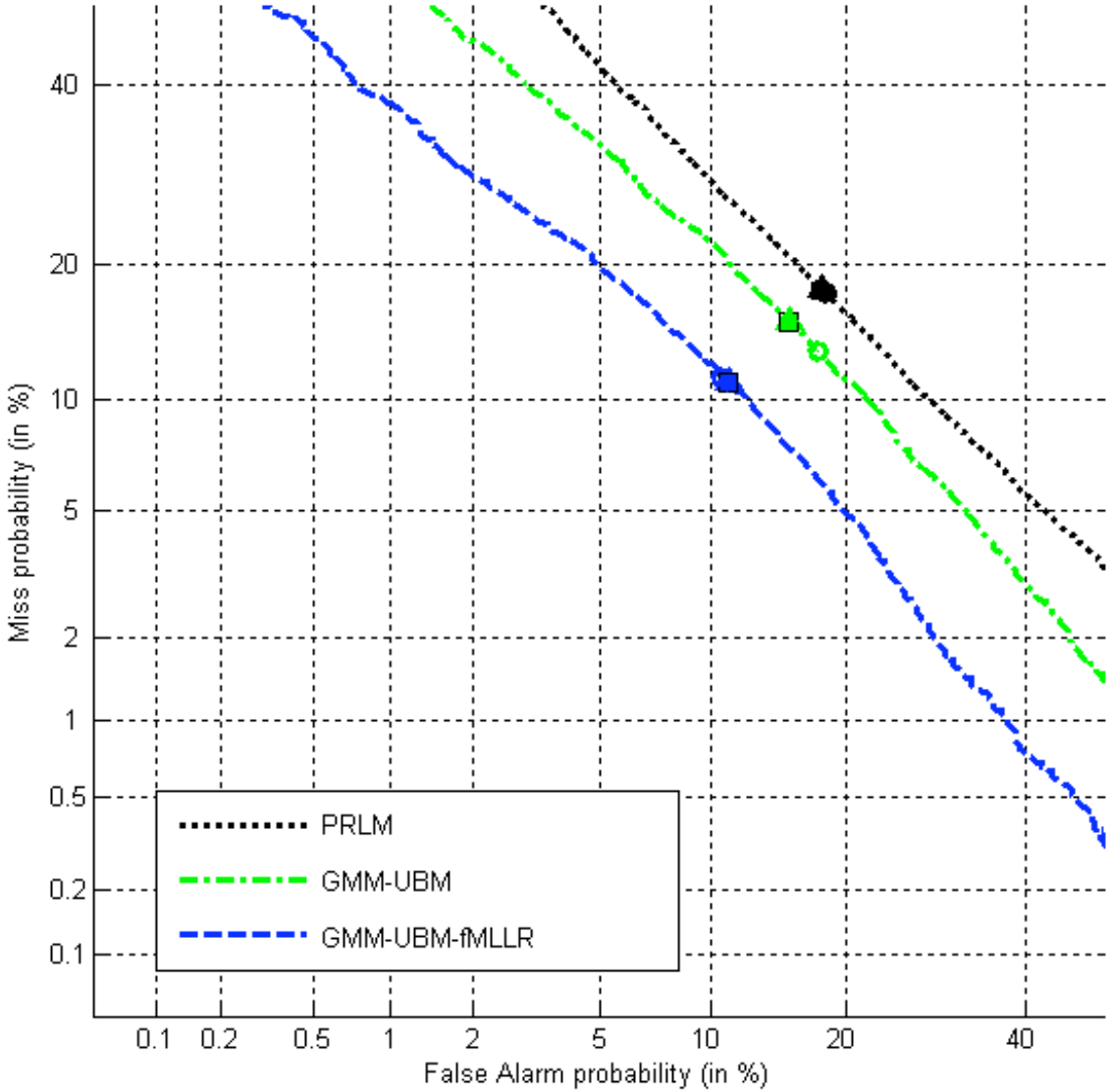
- A trigram phonotactic model per dialect

II. Standard GMM-UBM:

- Front-End:
  - 13D PLP features per frame
  - Each frame is spliced together with four preceding and four succeeding 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)

III. GMM-UBM with fMLLR (Biadsy et al., 2010)

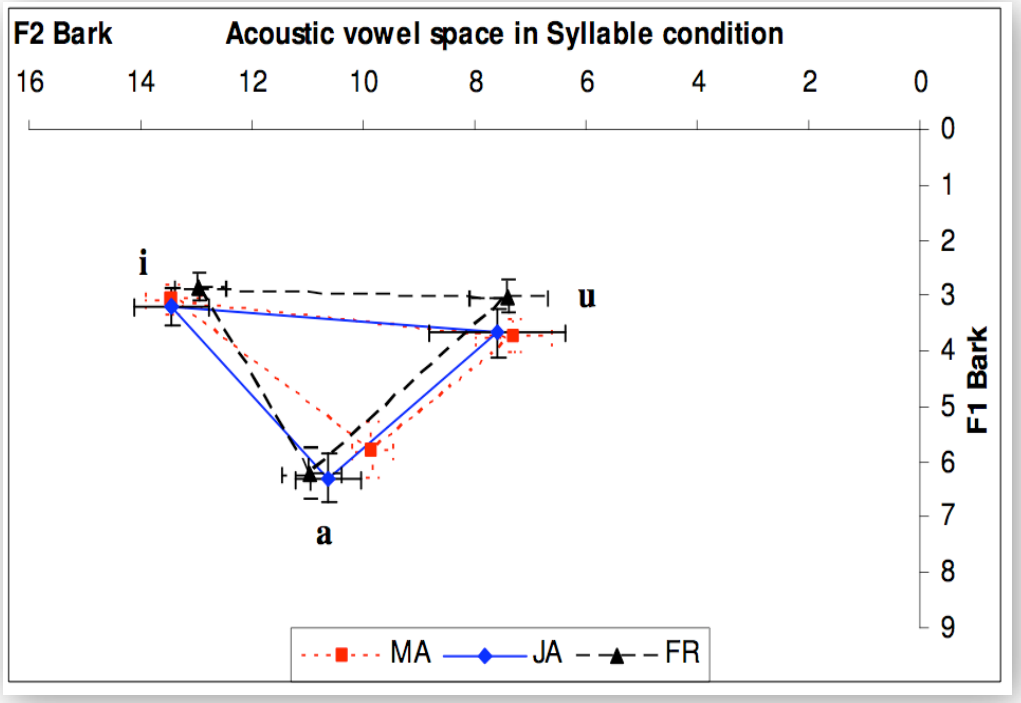
Baselines



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

Hypothesis in current work

- Rely on the hypothesis that dialects differ in the realization of certain phonemes



(*Al-Tamimi & Ferragne, 2005*)

General Idea

- Compare utterances at the phonetic level

Current Approach

- Build a GMM-UBM for each phone type
- Extract GMM-Supervectors at the level of phones
- Design a kernel function that computes similarity between pairs of utterances
- Train an SVM classifier for each pair of dialects

Current Approach

- Build a GMM-UBM for each phone type
- Extract GMM-Supervectors at the level of phones
- Design a kernel function that computes similarity between pairs of utterances
- Train an SVM classifier for each pair of dialects

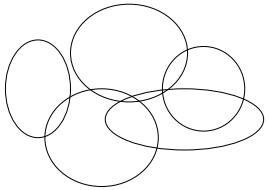
Front-End

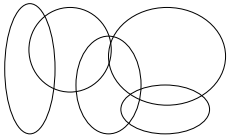
- Using IBM's Attila System (Soltau et al. 2009):
  - 13D PLP features per frame
  - Each frame is spliced together with four preceding and four succeeding frames followed by LDA ➔ **40D**
  - CMVN
  - fMLLR adaptation using hypothesized CD-phones

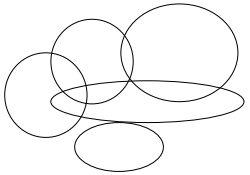
Phone GMM-UBM

- Run a phone recognizer on all data
- Extract frames aligned to each phone type
- Train a GMM-UBM for **every phone type**
  - Using frames from all dialects

Phone GMM-UBM

/aa/ 

/b/ 

/z/ 

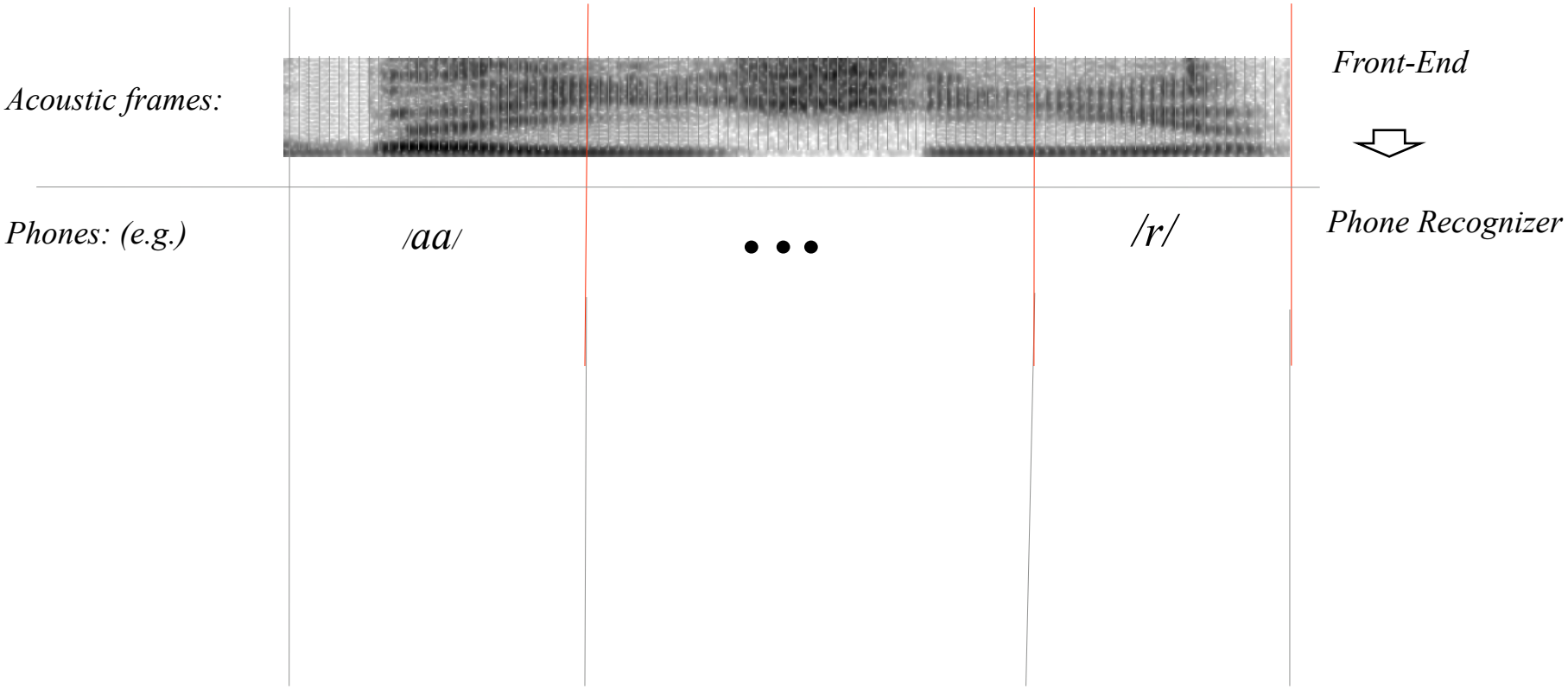
34 Arabic phones ➔ 34 Phone GMM-UBMs

Current Approach

- Build a GMM-UBM for each phone type
- Extract GMM-Supervectors at the level of phones
- Design a kernel function that computes similarity between pairs of utterances
- Train an SVM classifier for each pair of dialects

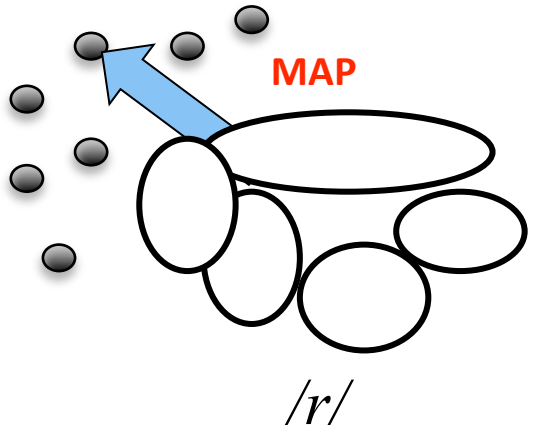
Step 1

Given an utterance U:



Step 2 - MAP Adaptation of each **Phone Instance**

- Given a phone instance acoustic frames:



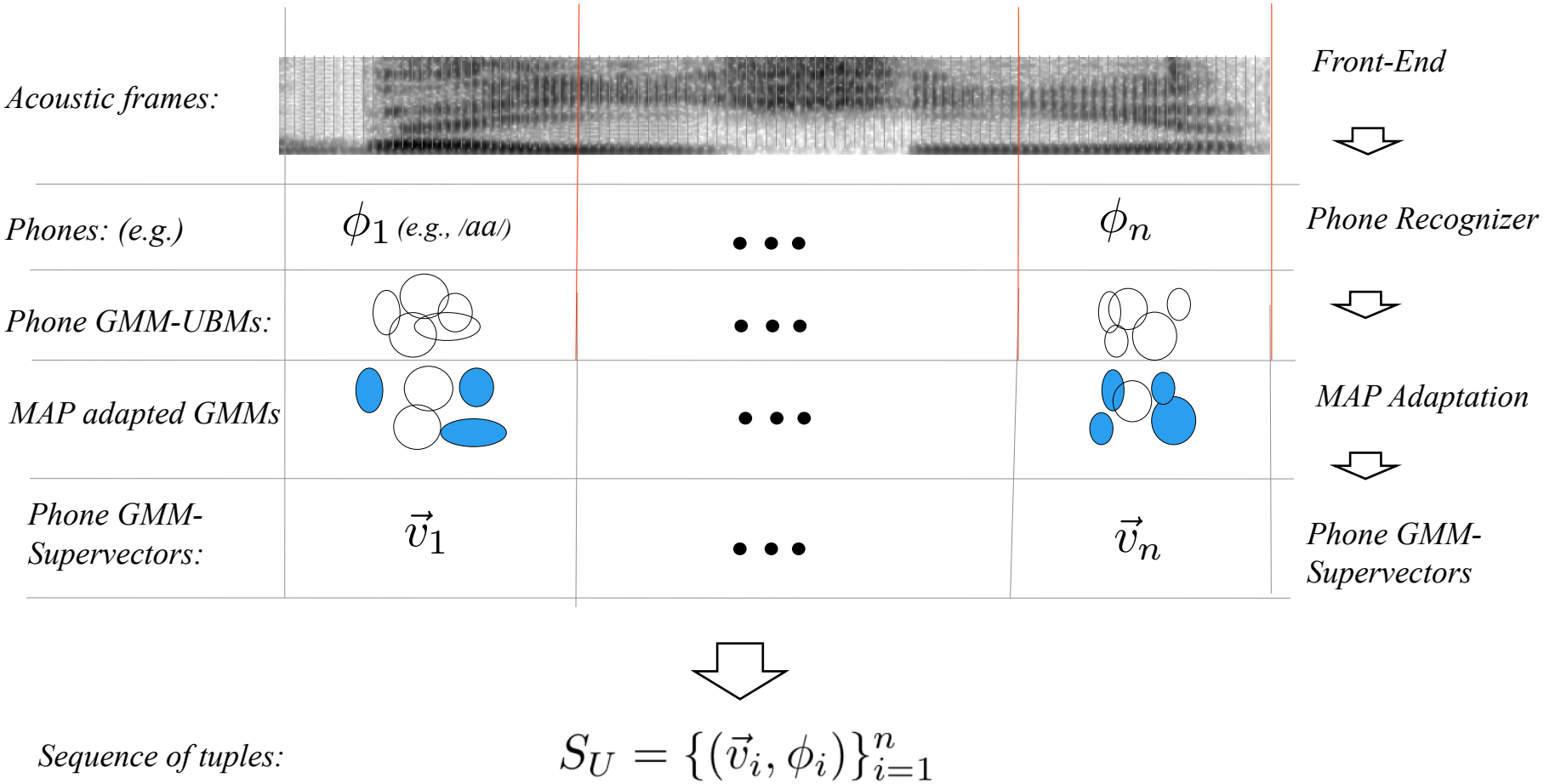
I. MAP adapt the phone GMM-UBM using the phone acoustic frames
II. Stack all the **Gaussian means and phone duration** $V_k = [\mu_1, \mu_2, \ldots, \mu_N, duration]$ in one supervector

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

*Similar to (Campbell et al., 2006) but at the phone level

Steps

Given an utterance *U*:



Classification Task

- Distinguish between pairs of dialects given a sequence of tuples

- Classifier choice:
  - SVM has been shown to model well supervector-like representation (e.g., Campbell et al., 2006)

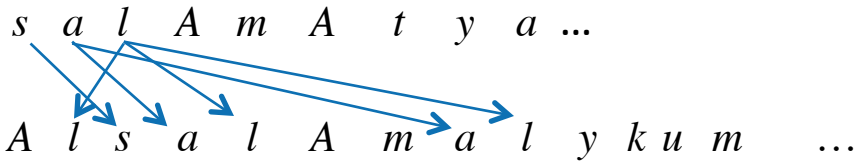
- We need a kernel function that computes the similarity between a pair of utterances U_a and U_b

Phone-GMM-Supervector-Based Kernel

- Let S_{U_a} be the sequence of tuples of utterance U_a
- Let S_{U_b} be the sequence of tuples of utterance U_b

$$K(S_{U_a}, S_{U_b}) = \sum_{i,j:\phi_i=\psi_j} e^{-\|\vec{v}_i - \vec{u}_j\|^2/2\sigma^2}$$

- Sum of RBF kernels between every pair of Supervectors of phone instances with the same type across the two utterances



Dialect Recognition

- Compute a kernel matrix using our kernel function for each pair of dialects
- Train an SVM classifier using this kernel matrix for the pair of dialects
- During testing, given an utterance U:
  1. Construct the sequence of tuples S_U
  2. Compute the kernel value with every support vector

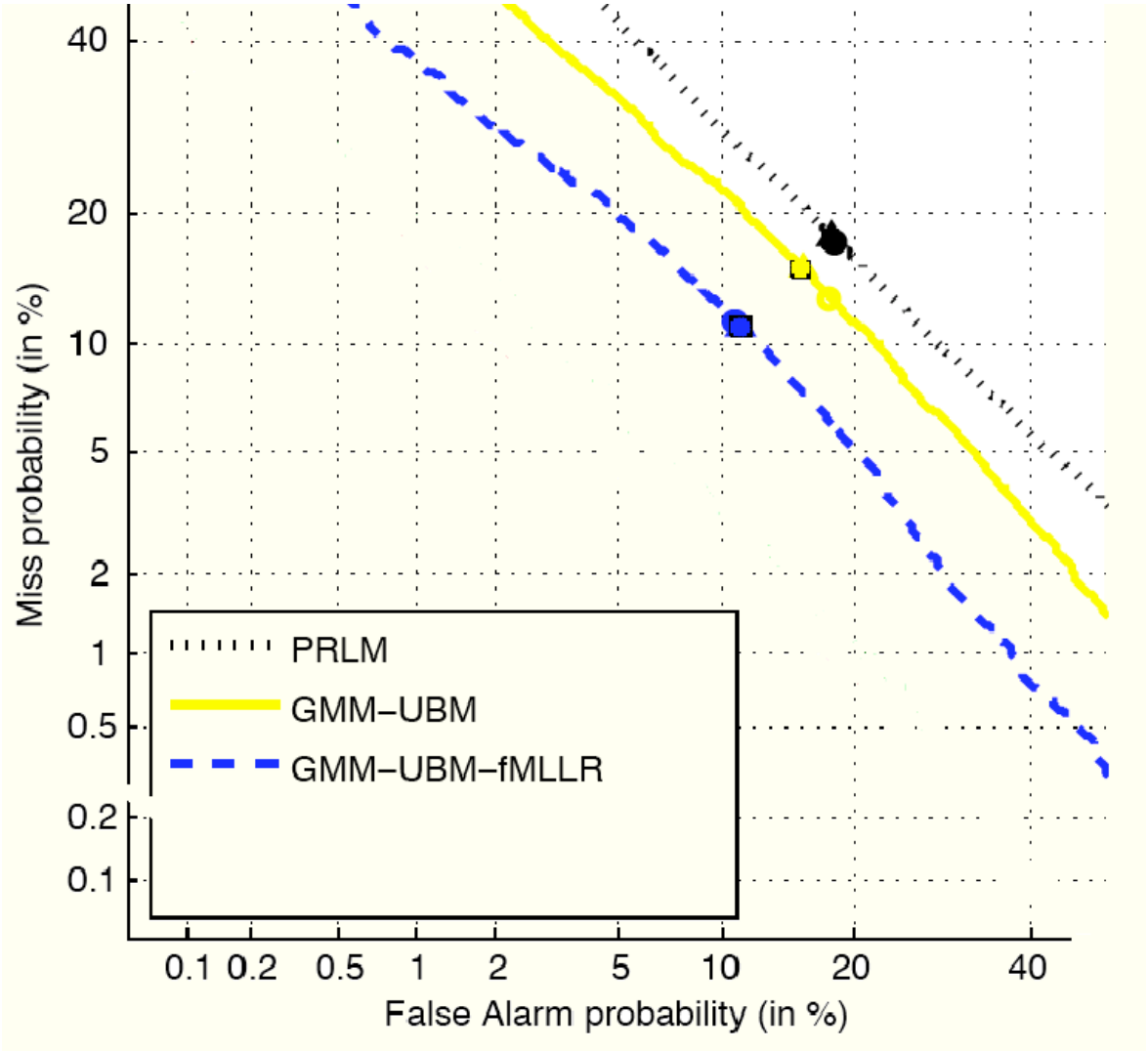
$$f(S_U) = \sum_{i=1}^{N} \alpha_i y_i K(S_U, x_i) + b$$

  3. The sign of this function is our hypothesized dialect class

Evaluation

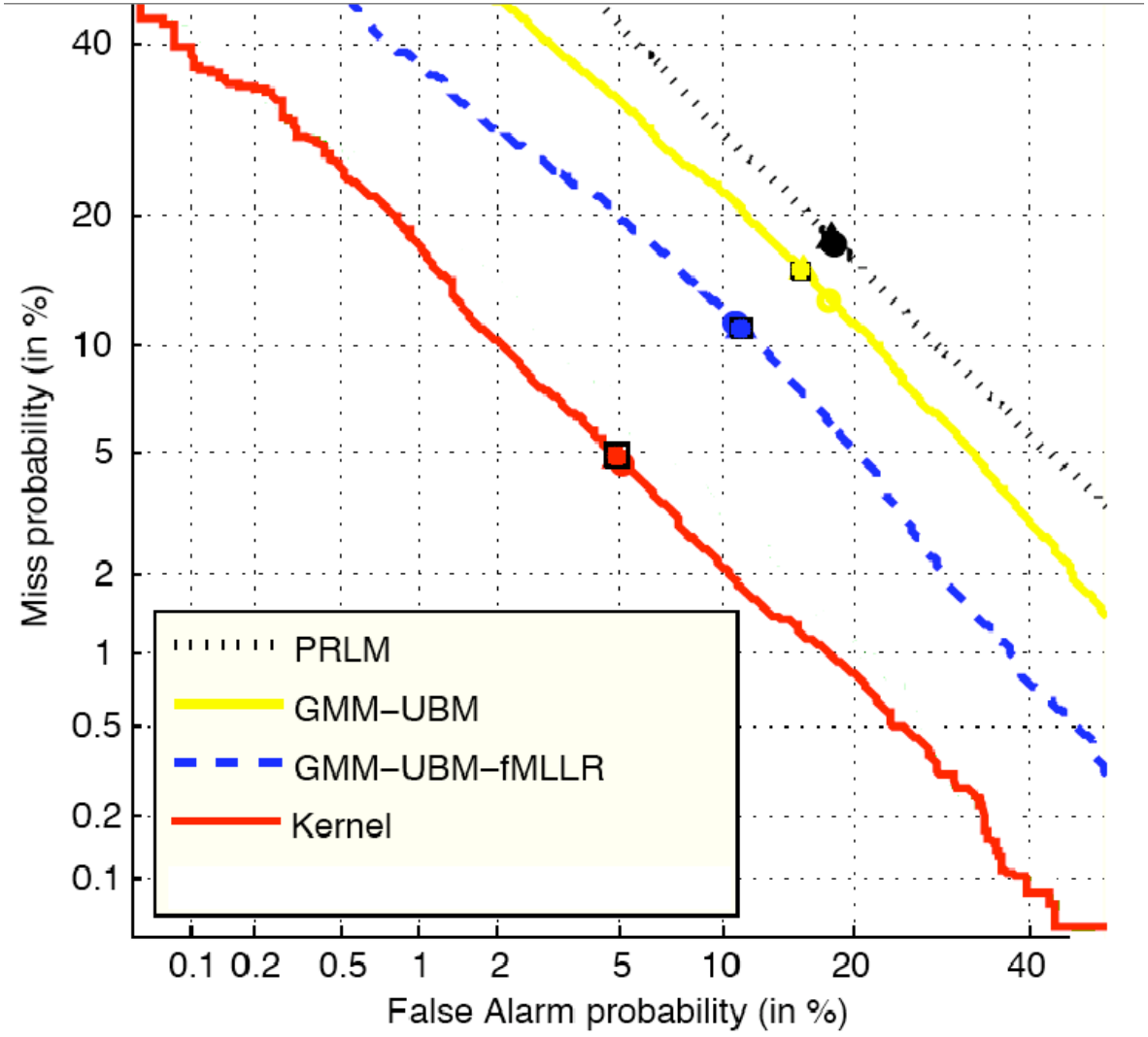
- 34 phone GMM-UBMs are Maximum-Likelihood trained with 100 Gaussians each
- We segment the training speakers in our corpora to 30s cuts
- Using all training cuts, train an SVM classifier for each pair of the 4 dialects (6 classifiers)
  - Use SVMs that estimate posterior probabilities (Wu et al., 2004)
- Use the posterior as the detection score to plot DET curves

Results and Baseline comparison



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

Results and Baseline comparison



Approach	EER (%)
PRLM	17.7
GMM-UBM	15.3
GMM-UBM-fMLLR	11.0%
Kernel	**4.9%**

Comparison to (Torres-Carrasquillo et al., 2008)

- GMM-UBM-based model discriminatively trained with SDC features
- Eigen-channel compensation and VTLN
- Back-end classifier

→ **EER 7.0%** on 3 Arabic dialects

- Our approach on exactly the same segments as in (Torres-Carrasquillo et al., 2008)

→ **EER 6.4%**

Conclusions

- Modeling the differences between dialects at the phonetic level is very effective
- New Approach:
  - Supervector representation at the phone level
  - New Phone GMM-UBM-Supervector-based Kernel function
- Significantly outperforms: PRLM, GMM-UBM, GMM-UBM-fMllR
- To our knowledge, represents new state-of-the-art performance for Arabic

Future Work

- 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
  - Portuguese Dialects
- Missing components:
  - VTLN
  - NAP channel compensation (need to modify to accommodate for short context supervectors)

Thank You!

- Acknowledgments:
  - P. Torres-Carrasquillo and N. Chen for providing us with the segmentation
  - IBM T. J. Watson Speech Team