

Recognizing Arabic Speakers With English Phones

Andreas Stolcke Sachin Kajarekar

*Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA*



Overview

- Background
- Arabic Data Set Development
- Phone-based SID
- Results
 - English vs. Arabic SID difficulty
 - English vs. Arabic phone recognition
 - Effects of Arabic dialects in background data
- Conclusions



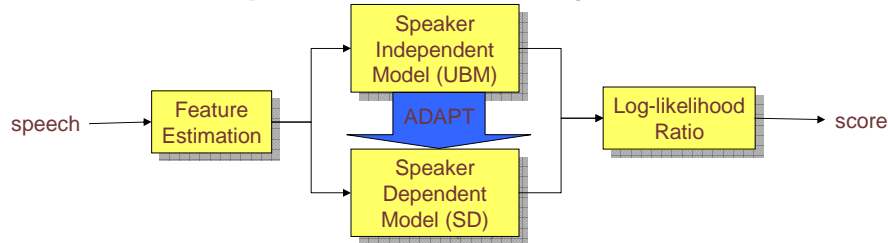
Arabic SID Data and Dialects

- ❑ We used data from 4 Arabic dialects to train background (speaker independent) model for GMM system
 - Modern Standard Arabic (MSA) – broadcaster speech
 - Levantine Arabic (LVA) – telephone conversations
 - Egyptian Arabic (ECA) – telephone conversations
 - Iraqi Arabic (IRA) – interview scenario
- ❑ A separate data set from NIST 2004 and 2005 speaker recognition evaluation (SRE) was used for the evaluation of the system
 - Mixer collection (dialects unknown) – telephone convs.

Experimental Setup

- ❑ Data from NIST 2004 and 2005 SREs
 - Mixer collection, conversational speech, average length of 5 mins
- ❑ Data divided into two sets of speakers
 1. One conversation per speaker
 - 35 conversations used in the speaker independent model training
 2. Two or more conversations per speaker
 - 43 speakers, average 5 conversations per speaker
- ❑ Second set used for evaluation
 - 594 target trials and 26,000 impostor trials
 - Impostor distribution is sampled uniformly to get 5940 impostor trials (to maintain the ratio 1:10)

Cepstral GMM System



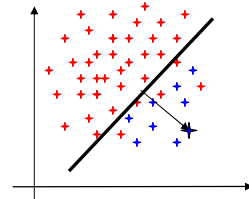
- Gaussian mixture model
 - Universal background model (UBM) trained on population sample, mix of dialects
 - UBM represents generic impostor speaker
 - Speaker dependent models created by MAP adaptation of UBM
 - Moves Gaussian means of UBM towards target speaker parameters
 - Log-likelihood ratio is SID score
 - Gives ratio of probabilities that test data is from target vs. impostor
 - More positive score means more likely to be target speaker

Phone-Based Speaker Recognition

- Motivation: phone based speech modeling gives
 - Portability to different languages (unlike word recognition)
 - More detail, longer-term modeling than cepstral system
- Here we tried two models previously used in English:
 - Phoneloop MLLR-SVM
 - Models speaker-specific translation/rotation of Gaussian means of phone recognition models
 - Phone N-gram SVM
 - Models relative frequencies of phone sequences from unconstrained recognition
- Initially: use existing phone models trained on *English* speech

Speaker Verification with SVMs

- ❑ Each conversation side = one point in feature space
- ❑ SVM trained to separate target from background samples (with maximal margin)
- ❑ SID score = distance from test sample to hyperplane
- ❑ SVM cost function weighted to compensate for training data imbalance
- ❑ Linear kernel functions work well for most features tried to date.



- + Target training sample(s)
- + Background samples
- + Test sample

3/5/2008

Odyssey 2008

7



MLLR Speaker Adaptation

- ❑ Speech recognizer adapts speaker-independent model to best fit test speaker



- ❑ Adaptation transform estimated by Maximum Likelihood Linear Regression (MLLR)
 - Maximizes likelihood of test data under recognition hypothesis or (in our case) phone-loop model
- ❑ Transform rotates and shifts Gaussian means (= matrix + vector)
- ❑ Separate transforms for different phone classes
 - non-speech, obstruents, nonobstruents

3/5/2008

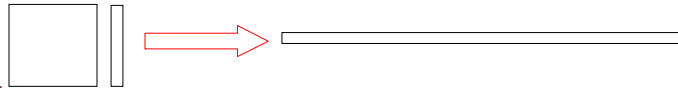
Odyssey 2008

8



MLLR-SVM Speaker Recognition

- ❑ Idea: MLLR transform encapsulates what makes target speaker different from the “average speaker”
- ❑ Transforms are based on detailed, sequential speech models (unlike cepstral SID models)
- ❑ Use normed transform coefficients as feature vector



- ❑ Refinements:
 - Combine transforms for different phone classes, discarding non-speech transform
 - Combine transforms relative to different recognition models
- ❑ Model feature vectors with SVMs

Phone N-grams

- ❑ Idea (originally by Zissman for language ID):
 - Convert continuous acoustic signal into a stream of phone labels
 - Allows modeling of long-range, sequential acoustic patterns (unlike frame-level cepstra)
- ❑ Implementation:
 - Run phone recognizer without language constraints
 - Extract phone N-gram frequencies from lattices
 - Model frequency feature vector with SVMs

Results

System	Bkg. data	%EER
Cepstral GMM	English	10.27
Cepstral GMM	Arabic	9.09
Phone N-gram SVM	Arabic	11.11
Phoneloop MLLR SVM	Arabic	8.41
Ph. N-gram + MLLR	Arabic	7.74
Ph. N-gram + MLLR + cep. GMM	Arabic	7.45

- ❑ Matched background data is important
- ❑ English-based phone-level system are competitive with GMM *without using Arabic-specific models*, and beat it when combined
- ❑ Large improvements over baseline by combining multiple systems

Arabic vs. English SID

- ❑ Compare similar models on different tasks

System	Test data language	
	Arabic	English
Cepstral GMM	9.09	7.16
Phone N-gram SVM	11.11	12.75
Phone-loop MLLR SVM	8.41	7.91

- ❑ Performance generally comparable
- ❑ Phone N-gram worse on English, others better
- ❑ Note: still using English phone models for both lang.

Effect of Recognition Models

- ❑ First experiments used English-trained phone models (mostly for expedience)
- ❑ Can we improve with Arabic phone models ?
- ❑ Tried using MSA recognition models (from GALE)
- ❑ Differences from English models:
 - Broadcast genre (versus telephone conversations)
 - Less training data (100h versus 300h)
 - Gender-independent
 - Smaller phone set (33 versus 46 phones)
 - Less sophisticated training algorithms



Arabic Phone Models: Results

System	Phone models trained on	
	Arabic	English
Phone N-gram SVM	19.70	11.11
MLLR SVM (m+f)	n/a	8.41
MLLR SVM (unisex/female)	10.44	9.60

- ❑ English model work better than Arabic
 - Good news for language-independent speaker recognition!
- ❑ Arabic models might suffer from modeling differences given earlier
- ❑ To do: Train Arabic conversational speech model comparable to English models



Effect of Background Data

- ❑ Fundamental Problems
 - Arabic is mix of dialects
 - Only minimal Mixer background data available
 - Therefore, background data is mismatched to test data
- ❑ Use combination of different conversational dialectal Arabic corpora
- ❑ Used here (all telephone data)
 - Egyptian (ECA) -- 238 conversation sides
 - Levantine (LVA) – 880 conv sides
 - Mixer (MM) – 35 conv sides
 - Iraqi (ICA) – 478 conv sides [recently added]
 - Gulf (GCA) – 526 conv sides [recently added]



Background Data: Results

Background data	MLLR-SVM	Phone N-g
ECA+LEV	8.42	11.45
ECA+LEV+MM	8.42	11.11
ECA+LEV+MM+ICA	8.24	10.94
ECA+LEV+MM+ICA+GCA	8.92	10.94

- ❑ Mix-of-dialects approach seems to work
- ❑ The more dialects, the better -- up to a point
- ❑ Gulf Arabic data seems to be mismatched to the rest
 - Need to investigate acoustic, language properties
 - We don't know if Gulf Arabic even occurs in Mixer test data



Conclusions

- ❑ Initial results for Arabic SID using phone-based models
- ❑ English-trained models work very well – better than Arabic models tried so far
- ❑ Much to do in future work:
 - Can we improve Arabic phone models?
 - Multi-language phone models?
 - Multi-language model combination