Spoken Arabic Dialect ID

Speech & Audio Processing & Recognition

Fadi Biadsy

March 13, 2008
Background

- Modern Standard Arabic (MSA): standard language throughout the Arab world (Literary Arabic)
  - A native Language of Nobody

- Colloquial Arabic: collective term for all dialects of Arabic
Maghrebi, Egyptian, Sudanese, Levantine, Iraqi, Arabian
Dialect ID

- Given a speech segment as short as possible → Dialect ID
Why Study Dialect ID

- Interesting problem
  - Phonetic cues?
  - Prosodic cues? (e.g., intonational contours, phrase accents, durational features...)
  - *Lexical and syntactic features?
Why Study Dialect ID
Why Study Dialect ID

- ASR fails when an Arabic speaker code switches to her regional dialect
Why Study Dialect ID

- ASR fails when an Arabic speaker code switches to her regional dialect
  - Identifying dialects prior to recognition enables the ASR to adapt its:
Why Study Dialect ID

- ASR fails when an Arabic speaker code switches to her regional dialect

  - Identifying dialects prior to recognition enables the ASR to adapt its:
    - Pronunciation Model
    - Acoustic Models
    - Morphological Model
    - Language Model
Why Study Dialect ID

- ASR fails when an Arabic speaker code switches to her regional dialect
  - Identifying dialects prior to recognition enables the ASR to adapt its:
    - Pronunciation Model
    - Acoustic Models
    - Morphological Model
    - Language Model

- Speaker Annotation
Dialect ID – Our Approach

- Phonotactic Modeling
  - Hypothesis: Every Arabic dialect has its own phonetic distribution
  - This approach was successfully used in Language ID
Dialect ID - TRAIN
Dialect ID - TRAIN

- First, train an MSA Arabic “phone” recognizer
Dialect ID - TRAIN

- First, train an MSA Arabic “phone” recognizer
- Now, given K dialects
  - For Dialect $i$

\[
\begin{align*}
\text{dh uw z hh ih n d uw} \\
\text{w ay ey d y aw ao uh} \\
\text{jh y eh k oh aa k v hh aw ao n} \\
\text{f uw v ow z l iy g s m} \\
\text{p l k dh n eh g f e y m} \\
\text{p l ay ae} \\
\text{dh iy jh sh p eh ae ey} \\
\text{d p sh ua r m ey f ay n z}
\end{align*}
\]
First, train an MSA Arabic “phone” recognizer

Now, given K dialects

- For Dialect $i$

Train an n-gram model $\lambda_i$
Given a speech segment $S$ from an unknown dialect:

$$Dialect(PS) = \arg\max_i (P(\lambda_i | PS))$$

$$= \arg\max_i (P(PS | \lambda_i)P(Di))$$
Dialect ID - TEST

- Given a speech segment $S$ from an unknown dialect:

$$P(PS \mid \lambda_i) = \prod_{k=0}^{n-1} P(p_k \mid p_{k-1}, p_{k-2})$$

$Dialect(PS) = \arg \max_i (P(\lambda_i \mid PS))$

$= \arg \max_i (P(PS \mid \lambda_i)P(Di))$
Experiment

- Train an MSA “phone” recognizer on ~37 hours of speech from TDT4 Broadcast News
Corpora – Levantine
Corpora – Levantine

- Arabic CTS Levantine Fisher Training Data Set 1,2,3 Speech
  - 762 Dialogues ➔ 1524 speaker
  - Each dialogue is 10 minutes ➔ 127 hours of speech
  - Annotated: LEB=547, JOR=393, PAL=187, SYR=72
Corpora – Levantine

- Arabic CTS Levantine Fisher Training Data Set 1,2,3 Speech
  - 762 Dialogues ➔ 1524 speaker
  - Each dialogue is 10 minutes ➔ 127 hours of speech
  - Annotated: LEB=547, JOR=393, PAL=187, SYR=72

- Silence based segmentation + remove every segment < 0.5s
Corpora – Egyptian
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
  - 120 Dialogues ➔ 240 speakers
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
  - 120 Dialogues ➔ 240 speakers
  - Each dialogue is 30 minutes ➔ 60 hours of speech
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
  - 120 Dialogues ➔ 240 speakers
  - Each dialogue is 30 minutes ➔ 60 hours of speech
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
  - 120 Dialogues → 240 speakers
  - Each dialogue is 30 minutes → 60 hours of speech
  - Silence based segmentation + remove every segment < 0.5s
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
  - 120 Dialogues  ➔  240 speakers
  - Each dialogue is 30 minutes  ➔  60 hours of speech

- Silence based segmentation + remove every segment < 0.5s
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
  - 120 Dialogues ➞ 240 speakers
  - Each dialogue is 30 minutes ➞ 60 hours of speech

- Silence based segmentation + remove every segment < 0.5s
Corpora – Egyptian

- CALLHOME Egyptian Arabic Speech
  - 120 Dialogues ➔ 240 speakers
  - Each dialogue is 30 minutes ➔ 60 hours of speech

- Silence based segmentation + remove every segment < 0.5s
Experiment
Experiment

- Egyptian corpus: held-out 20/240 speakers
  - Run the Arabic phone recognizer on 220 files:
    - $\sim 18.3$ million phones

- Levantine corpus, held out 757/1524
  - Run the Arabic phone recognizer on 220 files:
    - $\sim 19.4$ million phones
Results on the held out Data

- Levantine: 98.3% 744/757 were correctly classified as Levantine
- Egyptian: 95% 19/20 were correctly classified as Egyptian
Results on a different corpus

- Babylon Levantine corpus
  - Microphone Recordings
  - 164 speakers
  - ~60 hours of speech
  - Accuracy: 96.3% speakers
TODO

- Test on a different corpus for Egyptian
- Try to identify “sub” dialects (from the same corpus)
- Identify Gulf and Iraqi Arabic
- Incorporate English phone recognizer
Important issue (TODO)

- We use all the speech of a speaker
  - avg: ~5 minutes for Lev.
  - avg: ~15 minutes for Egy.

- Will this approach work if we use less than 30s of speech?
Thank you!