Candidacy Exam

Fadi Biady

September 24th, 2008
Topics

1. Language Identification
2. Arabic Dialect Modeling and Analysis
3. Disfluency Detection
• Given a speech segment from an unknown language:

Why Study Language ID?

• For multi-lingual NLP tasks, we need to identify the language spoken first
  • QA and Information Retrieval from multilingual data
  • Automatic Dialogue systems: “the International airport of the future” (Hazen and Zue, 1993)

  • Call Centers: Route an incoming telephone call to a human switchboard operator (crucial in emergency situations)

  • We can learn about differences between languages
Automatic language identification using a segment-based approach
(Hazen and Zue, 1993)

- **Goal:** Design a general probabilistic framework for LID to combine:
  - Phonotactic (LM), Prosodic, and Acoustic models

- The most general expression to describe the LID problem, mathematically:
  \[
  \arg\max_i \Pr(L_i \mid \tilde{a}, \tilde{f})
  \]

- Using chain and conditioning rules, we can get this framework:
  \[
  \arg\max_i \Pr(\tilde{a} \mid C_b, S_b, \tilde{f}, L_i) \Pr(S_b, \tilde{f} \mid C_b, L_i) \Pr(C_b \mid L_i) \, P(L_i)
  \]

  - Acoustic model
  - Prosodic model
  - LM
  - Prior
  (duration \(\mid\mid\) F0)
Acoustic Models

\[
\text{argmax}_{i} \Pr(\vec{a} | C_b, S_b, \vec{f}, L_i) \Pr(S_b, \vec{f} | C_b, L_i) \Pr(C_b | L_i) \Pr(L_i)
\]

- **Hypothesis:** Languages differ in their spectral distributions

- **Two Approaches to acoustic modeling:**
  - (Hazen and Zue, 1993)
  - (Zissman, 1996)
Prosodic Models

\[
\arg\max_i \Pr(\tilde{a} \mid C_b, S_b, \tilde{f}, L_i) \Pr(S_b, \tilde{f} \mid C_b, L_i) \Pr(C_b \mid L_i) \Pr(L_i)
\]

- **Hypothesis:** Languages differ in their prosodic structure
  - Duration, F0 patterns, energy, speaking rate, and rhythm

- **4 Approaches to Prosodic Modeling:**
  - (Hazen and Zue, 1993)
  - (Zissman, 1996)
  - (Rouas, 2005)
  - (Timoshenko and Hoge, 2007)
Language Models

$$\text{argmax}_i \Pr(\tilde{a} \mid C_b, S_o, \hat{f}, L_i) \Pr(S_b, \hat{f} \mid C_b, L_i) \Pr(C_b \mid L_i) \Pr(L_i)$$

Acoustic model  Prosodic model  LM  Prior

**Hypothesis:** Languages differ in their phonetic constraints and inventory

*For each language i:*

- Train a language model $\lambda_i$
- Run a phone Recognizer

**Languages:**

- dh uw z hh ih n d uw w ay ey d y aw ao uh jh y eh k oh aa k v hh aw ao n
- f uw v ow z l iy g s m p l k dh n eh g f ey m p l ay ae
- dh iy jh sh p eh ae ey d p sh ua r m ey f ay n z
Language Models

$$\text{argmax}_i \Pr(\bar{a} | C_b, S_o, f, L_i) \Pr(S_b, f | C_b, L_i) \Pr(C_b | L_i) P(L_i)$$

Test utterance:

Run the phone Recognizer

- **4 Approaches to Language Modeling:**
  - (Hazen and Zue, 1993)
  - (Zissman, 1996)
  - (Kirchhoff and Parandekar, 2001)
  - (Torres-Carrasquillo, et al., 2002)
Automatic language identification using a segment-based approach
(Hazen and Zue, 1993)

- **OGI-TS**: Oregon Graduate Institute Multi-Language Telephone Speech database
  - 10 languages, 90 speakers

- **Acoustic Model**:
  - Trained a Gaussian for each of 23 phone classes for each language
  - Features: 14 MFCC coefficients + 14 delta cepstra

- **Prosodic Model**:
  - A duration model: a distribution for each phone
  - F0 model: two histogram for each language (F0 normalized)

- **Phonotactic Model**:
  - 23 phone-class recognizer
  - Trigram LM

- **Results (on 10s test utterances)**:
  - Combining all models: 47.7% (accuracy)
Comparison of Four Approaches to Automatic Language Identification of Telephone Speech
(Zissman, 1996)

- **Goal**: Compare the performance of 4 LID approaches evaluated on common corpora (OGI-TS):

  I. **GMM Acoustics**: Gaussian Mixture Modeling
     - Train 2 GMMs (of 40 mixtures) for each language
     - 1st GMM: 12 MFCC coefficients
     - 2nd GMM: 13 deltas

  II. **Three Phonotactic Approaches**
      1. PRLM: Single-language phone recognition followed by language dependent n-gram model
      2. Parallel PRLM: Use multiple phone recognizers, each trained on a different language
      3. PPR: Language dependent parallel phone recognitions
          - A phone recognizer for each language to be identified
          - The LM is embedded in the HMM

- **Results (comparing GMM to LMs)**:
  - On 3 languages: PPR performs as well as Parallel PRLM, and achieve best accuracy (85%)
  - On 10 languages: Parallel PRLM achieves best accuracy, 63%, PRLM: 54%, GMMs, 50%

- **Improvements**:
  - Prosodic model (duration only: short vs. long phones)
  - Gender Model
    - ➤ 79% on 11 languages
Language Identification using Gaussian Mixture Model Tokenization
(Torres-Carrasquillo et al., 2002)

- **Goal:** Instead of using a phone recognizer in the front-end (PRLM approach), use cluster indexes

- **Approach**
  - Train a GMM on the acoustic data of one language only
  - **Tokenizer:** for each frame, output the index of the Gaussian scoring highest in the GMM
  - Train a bigram model for each language over these indexes

- **Advantages:**
  1. No need of manually transcribed data
     - Avoid mislabeled data
     - More data can be added easily
  2. GMM is less expensive than phone recognizers, faster processing during recognition
  3. Can be combined with the phone language model score to further boost performance

- **Results:**
  - On 12 languages (test 30s)
  - This approach did not outperform the baseline
  - But combination of 3 approaches is the best (83%)
  - Backend classifier always improve accuracy

<table>
<thead>
<tr>
<th>Systems used</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-PRLM</td>
<td>22.0</td>
</tr>
<tr>
<td>GMM tokenizers</td>
<td>36.3</td>
</tr>
<tr>
<td>GMM acoustics</td>
<td>35.5</td>
</tr>
<tr>
<td>GMM tokenizers + GMM acoustics</td>
<td>26.7</td>
</tr>
<tr>
<td>P-PRLM + GMM acoustics</td>
<td>19.5</td>
</tr>
<tr>
<td>P-PRLM+ GMM tokenizers + GMM acoustics</td>
<td>17.0</td>
</tr>
</tbody>
</table>
Hierarchical Language Identification based on Automatic Language Clustering
(Yin et al., 2007)

- **Goal:** Which feature type is most important to distinguish one language from another
- Instead of fusing all features in one classifier, use a multi-level hierarchical classifier that splits the languages based on one feature type at a time
- Use Agglomerative clustering technique.
  1. Init: one cluster for each language
  2. Iteratively train classifiers on each language cluster pair:
     - merge the pair with MAX feature \(\text{min accuracy of the classifier on dev-set}\)
  3. Repeat the process till all languages are in one cluster
- **Feature types:**
  - MFCC with 7 coefficients (denoted as M)
  - Prosodic features: Pitch and Intensity (P)
  - Concatenation of both (M+P)
- **Results on OGI:** (10 languages, 10s)
  - The system outperforms significantly all baselines
  - Adds 1.1% to baseline
  - Accuracy: 91.3%

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>30s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>9.7%</td>
<td>12.9%</td>
</tr>
<tr>
<td>Prosodic</td>
<td>18.6%</td>
<td>25.7%</td>
</tr>
<tr>
<td>MFCC + Prosodic</td>
<td>7.1%</td>
<td>9.8%</td>
</tr>
<tr>
<td>GMM Fusion</td>
<td>7.3%</td>
<td>9.9%</td>
</tr>
<tr>
<td>Hierarchical LID</td>
<td>6.2%</td>
<td>8.7%</td>
</tr>
</tbody>
</table>
Phone>c
knowledge,
phonotac>cs
and
perceptual
valida>on
for
automa>c
language
iden>fica>on
(Adda
‐Decker
et
al.,
2003)

• **Goal:** Estimate the upper bound of the phonotactic approach by discarding linguistic noise due to recognition errors
  1. Compare language ID performance trained on phonetic hand-labeled data vs. automatic phone recognizer
  2. See how well humans perform in identifying languages

• **Experiments:** *(BN 8 languages, 3h each)*
  o C/V sequences, 10 classes, 19 megaphones from 70 multilingual phones
  o Train 5-gram phonotactic models

• **Results on hand labeled:**
  o Test utterances of 10 phones
  o Test utterances of **20 phones**, **100%** accuracy is obtained

<table>
<thead>
<tr>
<th>class set:</th>
<th>CV</th>
<th>10-class</th>
<th>19-class</th>
</tr>
</thead>
<tbody>
<tr>
<td>%LID</td>
<td>35.5</td>
<td>78.6</td>
<td>96.0</td>
</tr>
</tbody>
</table>

• **Results on automatic phonetic annotations:**
  o Using PR with 19 classes:
    • <51.9%, 10 phones, 0.7s> <83.7%, 40 phones, 3s> <93.7%, 80 phones, 6s>
  o Using 70 classes, accuracy increased (63.2% for 10 phones)

• **Human LID performance on 8 languages**
  o Subjects were trained on 20 seconds per language
  o 14 native French speaking academics listened to 3 tokens per language of 1.5 to 2 seconds
  o A combined correct identification is **87.6%**. Spanish was the hardest (63%)
Multi-Stream Statistical N-Gram Modeling With Application to Automatic Language ID
(Kirchhoff and Parandekar, 2001)

- **Goal:** Use multiple parallel sequences of articulatory phonetic features
  - Voicing, manner of articulation, consonantal place of articulation, nasality, etc.
  - E.g., one stream: <glide, vowel, plosive, vowel, fricative, vowel, plosive, affricative...>

- **Approach**
  - Acoustic model for each feature
  - Decode each feature group $F_i$ independently
  - Classification:
    \[
    \arg\max_i P(F_1, ..., F_K | L_i)
    \]
  - Stream selection $\Rightarrow \{\text{manner, consonantal place, vowel place, front-back, and rounding}\}$

- **Results:**
  - OGI-TS corpus, including < 3s utterances
  - The phone based approach performs better than independent streams
  - With some dependency, feature model $\gg$ phone model on
    - On short utterances $50.8\% \rightarrow 54.8\%$ (<15s) and $33.3\% \rightarrow 48\%$ (< 3s) only

- **Advantages:**
  - Unseen phone context in the test data
  - 47 feature models, but 126 phone models $\Rightarrow$ more robust n-gram models
  - Training data for phonetic features can be shared across phones $\Rightarrow$ more robust acoustic models
  - Language independent nature of phonetic features
Language identification with suprasegmental cues: A study based on speech resynthesis (Ramus and Mehler, 1999)

• **Goal:** How do newborns separate input utterances from two languages in a bilingual environment?

• **Rhythm hypothesis:** Newborns are able to discriminate languages which have different rhythmic structure

• **Intonation Hypothesis:** The discrimination is on basis of intonation and not rhythm

• **Stimuli:** 20 Japanese and 20 English sentences read by 4 native speakers

• **Resynthesis Experiments:**
  1. **saltanaj:** Intonation, rhythm, and broad phonetic categories were preserved
     • All non-prosodic, lexical and syntactic information was lost (replace all phonemes by 6 broad categories)
  2. **sasasa:** Only intonation and rhythm were preserved
  3. **aaaa:** Only the intonation of the original sentences was preserved. Interpolate F0 over unvoiced frames
  4. **Flat sasasa:** Syllabic rhythm only is preserved (constant fundamental frequency)
Language identification with suprasegmental cues: A study based on speech resynthesis (Ramus and Mehler, 1999)

- **Perceptual Experiment:**
  - 64 Adult French subjects were told that the utterances were from acoustically modified Sahatu and Moltec
  - Subjects (after passing a training session) were asked to answer S or M

- **Results:**
  - *saltanaj, sasasa, and flat sasasa* were identified significantly above chance but NOT *aaaa*
  - When the stimuli presented to 16 English subjects, who were told that one of the languages is English, they could significantly discriminate between *aaaa* English and *aaaa* Sahatu

- **Conclusion**
  - Syllabic rhythm was a robust cue for discrimination
  - Intonations can be of greater interest of native speakers
Using Speech Rhythm for Acoustic Language Identification
(Timoshenko and Hoge, 2007)

- **Goal**: Use rhythm as a feature for Language ID
  - Speech rhythm is modeled using the durations of two successive syllables

- Automatic syllabification of the acoustic signal is hard (it’s a language dependent task). Instead, use pseudo-syllables:
  - \( C^N V \), where \( N \geq 0 \).
  - duration = \(|C^N| + |V|\)

- **Approach**:
  - A speech utterance is modeled as a sequence of pseudo-syllable durations \( D=d_1d_2...d_N \)
  - Learn a bigram model over these sequences for each language
  - Given a test utterance, get the duration sequence, and test which bigram model provides the highest likelihood

- **Results (on 7 languages, 7s)**
  - Rhythm alone with ANN provide \( \sim 32\% \) accuracy
  - The acoustic system alone: 92.08% accuracy
  - The fused system 92.9% accuracy
Modeling Long and Short-term prosody for language identification (Rouas, 2005)

- **Goal:** Investigate the efficiency of prosodic features for language ID at the level of pseudo syllables

- Modeling short and long term prosody
  - **Long term prosody models prosodic movements over several pseudo-syllables**
  - **Short-term prosody represents prosodic movement inside a pseudo-syllable**

- **F0** is modeled using Fujisaki model (phrase and local accentuation)
  - The baseline of the F0 contour is computed by connecting all the local minima (*mark the slopes with ups/downs/silence*)
  - Subtract it from the original F0 contour. The residue is then approximated using linear regression in each unit (long/short) and then mark the slopes

- **Energy:** for each pseudo-syllable segment, compute the linear regression of the energy and then mark the slopes

- **Duration:** mark the units as short/long for each (what is short/long?)

- They use a N-multigram model to model these sequences

- **Results (on 7 languages, 20s)**
  - The long-term prosodic model provides: 41%
  - The short-term prosodic model provides: 63%
  - Merging both models: 71.2%
<table>
<thead>
<tr>
<th>Paper</th>
<th>Approach</th>
<th>Accuracy</th>
<th># Languages</th>
<th>Dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Hazen and Zue, 1993)</td>
<td>A + P + PRLM</td>
<td>47.7%</td>
<td>10</td>
<td>10s</td>
</tr>
<tr>
<td>(Zissman, 1996)</td>
<td>Parallal PRLM + Duration + Gender dependent models</td>
<td>79%</td>
<td>11</td>
<td>10s</td>
</tr>
<tr>
<td>(Kirchhoff and Par., 2001)</td>
<td>PRLM of articulatory phonetic features</td>
<td>48%</td>
<td>10</td>
<td>&lt;3s</td>
</tr>
<tr>
<td>(Torres et al., 2002)</td>
<td>Parallel PRLM + GMM-A + tokenizer</td>
<td>83%</td>
<td>12</td>
<td>30s</td>
</tr>
<tr>
<td>(Adda et al., 2003)</td>
<td><strong>LM on manual phones</strong></td>
<td>(100%)</td>
<td>8 BN</td>
<td>&lt;3s</td>
</tr>
<tr>
<td>(Rouas, 2005)</td>
<td>Short+Long term prosody</td>
<td>71.2%</td>
<td>7</td>
<td>20s</td>
</tr>
<tr>
<td>(Timosh. and Hoge, 2007)</td>
<td>GMM-A + rhythm</td>
<td>92.9%</td>
<td>7</td>
<td>7s</td>
</tr>
<tr>
<td>(Yin et al., 2007)</td>
<td>A-GMM + prosodic features in a hierarchical classifier + Recent SP techiques</td>
<td>91.3%</td>
<td>10</td>
<td>10s</td>
</tr>
</tbody>
</table>
Arabic Dialect Modeling and Analysis -- Motivation

- Arabic: Modern Standard Arabic (formal) and colloquial Arabic (informal or casual)
- Typically, not MSA ➔ Spontaneous, unconstrained, not well studied, so lack of NLP tools
- One of the main challenges of Speech Recognition and NLP tasks is to deal with informal data
- **Goal:** What are the speech cues that make Arabic dialects different?
  - Dialect ID
  - Code Switching and Mixing

Maghrebi, Egyptian, Sudanese, Levantine, Iraqi, Gulf
1. **Rhythmic and Syllabic Structure**
   1. (Ramus, 2002)
   2. (Hamdi et al., 2004)
   3. (Hamdi et al., 2005)

2. **Intonation and Stress**
   1. (Barkat, 1999)
   2. (De Jong and Zawaydeh, 1999)
   3. (Hellmuth and El Zarka, 2007)

3. **Handling ASR problems for Dialects**
   1. (Kirchhoff and Vergyri, 2004)
   2. (Vergyri, et al., 2005)

4. **Morphology**
   1. (Habash and Rambow, 2006)
Acoustic correlates of linguistic rhythm: Perspectives (Ramus, 2002)

(Ramus, 2002) – rhythm-based measure to classify:

- Stressed-Timed Languages
- Syllable-Timed Languages
- Mora-Timed Languages
• **Goals:** Compare two recent measures to distinguish rhythmic structure of languages
  1. Duration and variability of vocalic and intervocalic intervals -- %V, ΔC, and ΔV (Ramus et al. 1999)
  2. Pairwise Variability Index (rPVI and nPVI) of vocalic and non-vocalic intervals (Grabe and Low, 2002)

• **Issue:** Rhythm is, at least in part, a matter of duration, and duration are affected by speaking rate (across languages)

• **Conclusion:** nPVI method is robust to variability due to variations of speaking rate

(Ramus, 2002) – rhythm-based measure to classify:

- Stressed-Timed Languages
- Syllable-Timed Languages
- Mora-Timed Languages

(Hamdi et al., 2004) Rhythm across Dialects Using Ramus metric

Languages: English, Dutch, Arabic
Speech timing and rhythmic structure in Arabic dialects: A comparison of two approaches

*(Hamdi et al., 2004)*

- **Goal:** Explore the differences between the rhythmic structure of Arabic dialects
  - Subjects in (Barakat et al., 1999): “Western Arabic sounded faster and jerkier than Eastern Arabic” ↔ speech rhythm

- **Question:** Are there systematic rhythmic distinctions between dialects?

- **Dialects and Languages:** 6 Arabic dialects (3 western, and 3 eastern) and 3 other languages
  - 30 sentences per language/dialect of 2.5s duration on avg (3 male speakers per language)

- **Method:** Compute %V, and ΔC for each language/dialect and compare

- **Results:**
  - A gradual increase of %V as one moves from West to East (*)
  - ΔC decreases from West to East (**)
  - French has larger vocalic intervals than the other languages and Arabic dialects
  - ΔC of French is similar to that of the eastern dialects
  - Significance differences between regions (West vs. East) but not between dialects within the same group
  - The average values of %V of all western dialect is significantly higher than the eastern dialects
  - The average values of ΔC of EA is significantly higher than WA

- **Comparing rhythm methods:** high correlation between ΔC and rPVI-C and ΔV and rPVI V
  - In the LID framework ➔ prosodic model
Syllable Structure in Spoken Arabic: a comparative investigation (Hamdi et al., 2005)

(Ramus, 2002) – rhythm-based measure to classify:

- Stressed-Timed Languages
- Syllable-Timed Languages
- Mora-Timed Languages

(Syllable/rhythm relationship)

(English, Dutch, Arabic)

(Hamdi et al., 2004) Rhythm across Dialects Using Ramus measures

(Hamdi et al., 2005) Syllabic Structure of Arabic Dialects
Syllable Structure in Spoken Arabic: a comparative investigation (Hamdi et al., 2005)

- **Goal:** compare in detail the syllabic structure of three Arabic dialects: Moroccan, Tunisian, and Lebanese Arabic to understand their rhythmic tendency

- **Data:** 8-10 minutes of spontaneous speech at normal speaking rate for each dialect

- **Analysis:**
  - Consonant clusters are more frequent in Western dialects, especially in Moroccan Arabic
  - CV and CVC are the two dominant types, together: 55% in Moroccan, 65% in Tunisian and 76% in Lebanese
    -- CV is the most frequent
  - CV syllables are much more frequent in Lebanese than in western dialects
  - Moroccan dialect may include up to 3 consonants in onset position and 2 in the coda
  - Tunisian Arabic syllable complexity is between Moroccan and the Lebanese dialects

- **The findings of this paper support those of (Hamdi et al. 2004)**
  - Vowel reduction and short vowel deletion in western dialects $\Rightarrow$ lower %V (*)
  - More complex syllabic structure $\Rightarrow$ higher $\Delta C$ (**)

- **Conclusion:**
  - Within a language, dialects exhibit detectable differences in rhythmic/syllabic characteristics

- In the LID framework $\Rightarrow$ (broad) phonotactic model + prosodic model
1. **Rhythmic and Syllabic Structure**
   1. (Ramus, 2002)
   2. (Hamdi et al., 2004)
   3. (Hamdi et al., 2005)

2. **Intonation and Stress of Arabic Dialects**
   1. (Barkat, 1999)
   2. (De Jong and Zawaydeh, 1999)
   3. (Hellmuth and El Zarka, 2007)

3. **Handling ASR problems for Dialects**
   1. (Kirchhoff and Vergyri, 2004)
   2. (Vergyri, et al., 2005)

4. **Morphology**
   1. (Habash and Rambow, 2006)
Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects
(Barakat et al., 1999)

(Barakat, 1999) F0 and energy differences between E and W Arabic
Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects
(Barakat et al., 1999)

- **Goal:** Test if prosodic patterns are reliable cues for perceptually discriminating Arabic dialects
  - 4 Dialects: Western Arabic (Morocco and Algerian) and Eastern Arabic (Syria and Jordan)

- **Data:** Six “passages” spoken by 4 male speakers ➔ 24 total for each dialect

- **Subjects:** 19 Native western Arabic; and 19 non-Arabic

1. **Baseline perceptual experiment:** natural speech to evaluate the subjects' knowledge and perception of dialects
   - **Results:** 97% of correct identification by the Arabic subjects and 56% (significant) by the non-Arabic

2. **Masking perceptual experiment** (buzz sounds) to evaluate the reliability of prosodic information on discrimination
   - **Results:** 58% of correct identification (significant) by Arabic subjects and 49% (not significant) by non-Arabic

- **Problem:** the two experiments were presented to the same subjects in a row
Stress, duration, and intonation in Arabic word-level prosody (de Jong and Zawaydeh, 1999)

Acoustic-prosodic correlates of lexical stress

(De Jong and Zawaydeh, 1999)
Stress, duration, and intonation in Arabic word-level prosody
(de Jong and Zawaydeh, 1999)

- **Goals:** Examine the prosodic correlates of Ammani-Jordanian Arabic lexical stress
  - Does Arabic have similar cues as in English? Increase in duration, extreme formant values, increased intensity, and F0

- **Data:** 10 types of words spoken in the five prosodic conditions spoken in 5 conditions by 4 speakers (target syllables had /d/ onsets and /a/ in the nucleus)

- **Duration**
  - Duration of vowels in stressed syllables >> unstressed syllables in antepenultimate (word) position only
  - Syllable position more consistent determiner of duration across subjects than is the stress

- **Formant patterns:**
  - Stressed /a/ has a systematically higher F1

- **F0 patterns**
  - Stress associated with an increase in F0
  - F0 in penultimate significantly greater than in antepenultimate syllables

- **Relationships between higher-level prosody and word-level effects:**
  - Vowel durations for syllable followed by break index 4 is significantly longer from 2,3, but no significant difference between 2, 3
  - Most speakers use L-L% contours for statements and L-H% for questions
  - High pitch accents commonly occur in statements

- **Conclusion:** Arabic word-level prosody is remarkably like that of English
  - The expression of stress, linkage of pitch accents to stressed syllables, and in the occurrence of pre-boundary lengthening
Variation in phonetic realization or in phonological categories? Intonational pitch accents in Egyptian Colloquial Arabic and Egyptian Formal Arabic (Hellmuth and El Zarka, 2007)

(Hellmuth and El Zarka 2007) Egyptian Colloquial vs. Formal Registers

$F_0$, $Word_i$, $Lexical stress$, $2^{nd} Register F_0$
Variation in phonetic realization or in phonological categories? Intonational pitch accents in Egyptian Colloquial Arabic and Egyptian Formal Arabic (Hellmuth and El Zarka, 2007)

• **Goal:** Explore the assumption that formal Arabic will have the intonational characteristics of the speaker’s colloquial variety

• **Material:** 2 Egyptian speakers read ECA and EFA words that share same stressed syllables. Words were put in sentences and read 3 times (total 72 target words)

• **Qualitative Analysis:**
  - Similarities between ECA and EFA
    1. Pitch accent on almost every content word in both registers
    2. Accent shape for a content word is mostly the same for both registers
    3. Low plateau between successive H peaks
  - Differences:
    • EFA contains greater proportion of phrase boundaries

• **Quantitative Analysis:** Test pitch event to stressed syllable:
  - Relative peak delay: (distance of H peak from the stressed syllable onset) / (stressed syllable duration) varies significantly between registers for speaker A only and not speaker B.
  - For both speakers, in CVV and CVC syllables, H is aligned within accented syllable in both registers

• **Conclusion:**
  • speakers carry their prosodic events of the mother’s tongue dialect

![Figure 1: Schematised peak alignment in CV syllables in EFA and ECA, as observed in prior studies.](image)
Prosodic differences among Arabic dialects

(Hellmuth and El Zarka, 2007) Egyptian Colloquial vs. Formal Registers

(Barakat et al., 1999) F0 and energy differences between E and W Arabic

(De Jong and Zawaydeh, 1999) Acoustic-prosodic correlates of lexical stress

Word \(_i\)

Lexical stress

2\(^{nd}\) Register F0

F0
1. **Rhythmic and Syllabic Structure**
   1. (Ramus, 2002)
   2. (Hamdi et al., 2004)
   3. (Hamdi et al., 2005)

2. **Intonation and Stress of Arabic Dialects**
   1. (De Jong and Zawaydeh, 1999)
   2. (Barkat, 1999)
   3. (Hellmuth and El Zarka, 2007)

3. **Handling ASR problems for Dialects**
   1. (Kirchhoff and Vergyri, 2004)
   2. (Vergyri, et al., 2005)

4. **Morphology**
   1. (Habash and Rambow, 2006)
ASR for Arabic

- Typically Arabic transcripts lack short vowels (small diacritics in Arabic script)
- Multiple valid vocalizations => pronunciation for most words ➔ challenge for ASR

*qbl, qabla, qabli, qablu* (before)

*qabila* (accept)

*qab ala* (to kiss)

*....*
Cross-Dialectal Acoustic Data Sharing for Arabic Speech Recognition
(Kirchhoff and Vergyri, 2004)

• **Goal:** Use unvocalized MSA data to improve ASR for Egyptian Conversational Arabic (ECA)

• **Motivation:** Not enough training data for ECA, especially for triphone acoustic models
  - 40% of the CallHome (ECA corpus) triphones also occur in the FBIS (MSA corpus)

• **Automatic Diacritization**
  1. Generate all possible diacritized variants for each word, along with their morphological analyses
  2. Train an unsupervised trigram tagger to assign probabilities to sequences of morphological tags
  3. Use the trained tagger to assign probabilities to all possible diacritizations for a given utterance
  4. Use the weighted diacritizations as pronunciation networks and use acoustic models trained on ECA to find the most likely diacritization

• **Results:**
  - Training a system with the pooled data (CallHome + FBIS) ➔ did not outperform baseline
  - When training two independent systems, ROVER combination outperforms CallHome-Only ➔ 0.8% absolute improvement on dev set and 1.0% improvement on eval set (0.1 significance level) *(accuracy 58.3%)*
Development of a Conversational Telephone Speech Recognizer for Levantine Arabic (Vergyri et al., 2005)

- **Goal:** Describe the development of Levantine Speech Recognizer, and discuss:
  - **Grapheme Acoustic Models**
    - Each acoustic model implicitly models either a long vowel or a consonant with optional short vowel (obtained by simple orthographic rules)
  - **Modeling of Short Vowels**
    1. **Generic Vowel:** Add one optional generic vowel phone in all possible positions in pronunciation
    2. **Auto-Vowel:**
       1. Annotate subset of training data
       2. Train 4-gram language model with hidden events to predict the vowels in all training data
       \(\Rightarrow\) (30\% of the words have at least one wrong character)
  
- **Morphological language modeling**
  - Affixations and a subset of POS are identified using “a simple script and knowledge of Levantine”
  - Use MSA morphological analyzer
    \(\Rightarrow\) train factored LM

- **Results:**
  - Combination of **Auto-vowelized AM + LM** outperform the generic and grapheme models
  - **Factored LM** improves the accuracy but not significantly.
  - ROVER combination of systems (grapheme + generic vowel + auto-vowel): accuracy 53.5\% – the best
1. **Rhythmic and Syllabic Structure**
   1. (Ramus, 2002)
   2. (Hamdi et al., 2004)
   3. (Hamdi et al., 2005)

2. **Intonation and Stress of Arabic Dialects**
   1. (De Jong and Zawaydeh, 1999)
   2. (Barkat, 1999)
   3. (Hellmuth and El Zarka, 2007)

3. **Handling ASR problems for Dialects**
   1. (Kirchhoff and Vergyri, 2004)
   2. (Vergyri, et al., 2005)

4. **Morphology**
   1. (Habash and Rambow, 2006)
• **Goal:** Develop a general framework for morphological analyzer and generator for dialects of one language family

![Diagram of MAGEAD](image)

- **MAGEAD Advantages:**
  - Very general framework, supporting a new dialect requires specifying concrete morphemes and orthographic and phonological rules for this dialect
  - It can be used without a lexicon or with a partial lexicon
  - It can be used as analyzer and generator
  - It adds short vowels to the analyzed words (good for ASR)

- **Evaluation (on verbs of 3 radicals):**
  - The system outperforms Buckwalter analyzer on MSA (*on mbc verb list*)
    - token precision: 94.9%; recall: 95.8%
  - On Levantine (on *all*)
    - Context token recall: MSA system on Levantine data: 60.4% → Levantine system: 94.2%
Disfluency Detection – Motivation

- ~10% of spontaneous utterances contain disfluencies (Hindle, 1982)
- One Disfluency per 4.6 seconds for radio talk shows (Blackmer and Mitton, 1991).

**Disfluency types:**
- **Hesitations:** “Ch* Change Strategy”
- **Restarts (or false starts):** “It’s also * I like it”
- **Fillers: (filled and unfilled):** “um * Baltimore”
- **Self repairs (or self-corrections):** “I think that you get * it’s more strict in Catholic”

**Disfluencies are useful!**
- Disfluencies may facilitate language acquisition by highlighting equivalent classes
- Sometimes reduce the mental and memory load to digest information
- Getting or keeping the floor
- For dialogue systems (to pretend real-time performance, keep the turn)

**Disfluencies are an obstacles for NLP tasks:**
- ASR, Speech Understanding, Parsing, QA, and Summarization
1. Disfluency Correction and Identification
   1. (Hindle, 1982)
   2. (Nakatani and Hirschberg, 1994)
   3. (Liu et al., 2003)
   4. (Snover et al., 2004)

2. Modeling Disfluency to improve ASR
   1. (Stolcke et al., 1999)
   2. (Stouten and Martens, 2004)

3. Human and Disfluency
   1. (Bard and Lickley, 1997)
Give me airlines flying at uh flying to Boston from San Francisco next...

1. Text: (Nakatani and Hirschberg, 1994)
2. A-P: (Liu et al., 2003)

IP Identification in repairs

Interruption Point

Give me airlines flying at -- uh flying to Boston from San Francisco next...

Textual Input with Edit Signal annotated

(Hindle, 1982)

Disfluency Correction

Give me airlines flying to Boston from San Francisco next...
Deterministic parsing of Syntactic non-fluencies (Hindle, 1982)

- **Goal**: Expunge self-repairs \(\Rightarrow\) produce well-formed syntactic structure that is consistent with the intended meaning
- **Editing Signal**: Minimal non-lexical material that self-repair might insert
- **Assumption**: phonetically recognizable and equivalent

- **Method**: Integrate correction rules in a parser to specify how much, if anything, to expunge when an edit signal is detected
- **Rules**:
  - **Surface Copy Editor**: search for exact repetition separated by edit signal; expunge one
  - **Category Copy Editor**: search for exact repetition with same category separated by an edit signal, expunge the first
  - **Stack Copy Editor**: search for exact repetition with similar syntactic constituent separated by an edit signal, expunge the first
- **Results on one interview**:
  - 1512. 27% of the sentences had edit signal, 73% of the sentences had no edit signal
  - Surface copy: 29% | Category Copy: 9% | Stack Copy: 27%
  - Removing edit signal only: 24%
  - **Failures**: 3% | Remaining unclear and ungrammatical: 2%
A corpus-based study of repair cues in spontaneous speech  
(Nakatani and Hirschberg, 1994)

**Goals:**

1. Propose a framework to investigate repairs that divides the repair event into three temporal intervals (RIM)
2. Identify robust acoustic-prosodic cues in each of these intervals to detect repairs with no reliance upon sophisticated understanding of the text
3. Build a repair IS detector

**Data:** 6414 utterances from the ARPA Airline Travel and Information System  
122 speakers 346 utterances contained at least one repair (5.4%)
A corpus-based study of repair cues in spontaneous speech
(Nakatani and Hirschberg, 1994)

- **Reparandum**
  - 73.3% of all reparanda end in word fragment
  - Majority of fragment words are content words and rarely more than one syllable long, sometimes glottalized and sometimes exhibit coarticulatory effects.
  - #words in reparandum: (non-fragment repairs: 1, 52% 2, 32%) (fragment: 1, 65% 2, 23%)

- **Disfluency Interval**
  - Filled pauses and cue phrases occur in DI (9.4%) – significantly more often in non-fragment repairs than in fragment repairs
  - Speakers take less time to initiate the production of the repairing in fragmented repairs
  - DI duration for fragment repairs is significantly shorter than for non-fragment repairs
  - Small but reliable increases in F0 and amplitude from the end of the reparandum to the beginning of the repair

- **The Repair Interval**
  - Phrase boundaries can serve to identify the repair region
  - For 43% of the repairs, the repair offset coincides with phrase boundary
  - 70% of the remaining have the first phrase boundary after the repair onset at the right edge of a syntactic constituent
A corpus-based study of repair cues in spontaneous speech
(Nakatani and Hirschberg, 1994)

- Predicting repairs from acoustic and Prosodic Cues
  - Distinguish \{IS, fluent-phrase boundary, non-repair disfluency, simple word boundary\}
  - They considered every word boundary to be a potential repair site
  - Utterances in the test data have at least one IS
  - Feature examples:
    - The duration of pause between $w_i$ and $w_j$
    - The occurrence of one or more word fragments within $w_i$ and $w_j$

- Recall: 86.1% Precision: 91.2%
Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources (Liu et al., 2003)

- **Goal:** Investigate multiple knowledge sources of identifying reparanda
  1. Decision Tree classifier that uses the acoustic prosodic features $\rightarrow$ posterior probability: **IP vs. non-IP** between each pair of words
     - Duration + F0 features + voice quality features
     - POS/Word LM with hidden event “<IP>”

- **Results:**
  - Prosody model only $\gg$ chance performance on downsampled data. **Recall 77.5%; precision 77.6%** (baseline is 50%)
  - On non-downsampled data: **Word-LM & POS-LM & Prosody** $\gg$ baseline (96.62%)
    - Accuracy: 98.1%
    - Recall: 56.76%
    - Precision: 81.25%
    - Some degradation on ASR output (using only Word-LM 97.05 vs. 98.01)
    - More repetition (IPs) are identified by the pattern LM than by the word-based LM.
A Lexically-Driven Algorithm for Disfluency Detection
(Snover et al., 2004)

- **Goal**: Design transformation-based learning approach to disfluency detection using primarily lexical features without the use of extensive prosodic cues
  - Rule example: “change the label of word with POS X from L1 to L2 if followed by word with POS Y”

- **Task**: Tag each word in either a reference or ASR sentence with {filler, edit, fluent}

- **Features**:
  - Lexemes, speaker identity, and whether the word is followed by a silence

- **Training**:
  - Input: Time aligned transcript: speaker id, sentence boundaries, edits, fillers and interruption points are annotated
  - Rules are created by expanding rule templates, which are given as input to the learner, for example:
    - Change the label of word X from L1 to L2
  - The algorithm greedily selects the rule that reduces the error rate the most
    → 106 rules were learned

- **Results** (using lexeme error rate):
  - 2 Baselines: both systems using prosodic and lexical features
  - No system performs well on ASR transcripts (86-96% for edits).
  - Filled pause identification: REF: ~18%, ASR: 48-57% (comparable to acoustic/prosodic system)
  - Edits: Acoustic system (59%) significantly outperforms the lexical system (68%) only in CTS.
1. Disfluency Correction
   1. (Hindle, 1982)
   2. (Nakatani and Hirschberg, 1994)
   3. (Liu et al., 2003)
   4. (Snover et al., 2004)

2. Modeling Disfluency to improve ASR
   1. (Stolcke et al., 1999)
   2. (Stouten and Martens, 2004)

3. Human and Disfluency
   1. (Bard and Lickley, 1997)
Modeling Disfluency to Improve ASR

(Stouten and Martens, 2004) FP detector as a front-end

(Stolcke et al., 1999) Model prosody → disfluency

ASR
Modeling the Prosody of Hidden Events for Improved Word Recognition (Stolcke et al., 1999)

- **Goal:** Model prosody to improve speech recognition by modifying the language model to represent hidden events:
  - sentence boundaries and **various forms of disfluencies**

```
Right <S> I <REP> I don’t <DEL> uh <FP> I’m not really sure
```

\[
W^* = \arg \max_{W} P(W \mid A, F) \approx \arg \max_{W} \sum_{S} P(W, S, F) P(A \mid W)
\]

\[
P(W, S, F) \Leftrightarrow \text{HMM}:
\]

- **States:** \langle \text{word, event} \rangle \text{ pairs}
- **Observations:** prosodic features
- **Transition probabilities:** n-gram probabilities
- **Emission probabilities:** posterior probabilities from a decision tree

\[
P(W, S) P(F \mid W, S) \approx \prod_{i} P(F_{i} \mid E_{i}, W)
\]

Fi computed from window i

Standard Acoustic Models
Modeling the Prosody of Hidden Events for Improved Word Recognition (Stolcke et al., 1999)

- **Prosodic features**
  - Durations of pauses, of final vowels, and of syllable rhymes

- **0.9%** significant absolute reduction of word error rate

**Error Analysis:**
- Fewer substitution and insertion but more deletion.
- Prosodic model reduces errors
  - of high-frequency words that tend to occur at sentence boundaries
    - ... that at church to <s> ➔ ....that at church too <s>
  - occur around filled pauses
    - ...to perform in and col weather ➔ ....to perform in UH cold weather
Coping with Disfluencies in Spontaneous Speech Recognition
(Stouten and Martens, 2004)

- **Goal**: Detect first simple disfluencies and then change the behavior of the search engine and LM (for Dutch)

- If a FP is detected,
  - **FP frame dropping**: discard the frames in FP interval
    - Significant reduction of WER when using reference only
  
  - **FP probability adaptation**: Locally raise the probability of entering FP state
    - Change the probability to the FP arc in the LM when more than 50% of the frames consumed by this arc fall inside a detected FP interval
    - ➔ for each FP, **1.04 (auto) - 1.7 (ref)** words were corrected FP; traditional model (modeling FP as a word): **0.75**

- If a word repetition is identified,
  - **WR frame dropping**: Drop the frames of the repeated word
    - ➔ Corrects 0.6% words per WR (ref)
  
  - **WR probability adaptation**: Raise the probability of paths that include WRs
    - Change the probability of reentering the word after detecting repetition
    - ➔ Corrects 1.03% words per WR (ref)
1. Disfluency Correction and Identification
   1. (Hindle, 1982)
   2. (Nakatani and Hirschberg, 1994)
   3. (Liu et al., 2003)
   4. (Snover et al., 2004)

2. Modeling Disfluency to improve ASR
   1. (Stolcke et al., 1999)
   2. (Stouten and Martens, 2004)

3. Human and Disfluency
   1. (Bard and Lickley, 1997)
On Not Remembering Disfluencies  
(Bard and Lickley, 1997)

• **Goal:** Examine evidence that failures of memory and perception are involved in human ability to miss disfluencies

• **Stimuli:** Spontaneous speech utterances with 80 simplex and 16 complex disfluencies
  • 30 recast (no words from reparandum is repeated in repair) and 50 with repeats

• **Subjects** were instructed to transcribe everything they heard into real words in standard orthography and to be as accurate as possible.

• **Results:**
  1. **Listeners had great difficulty in reporting words from reparanda**
  2. Recall of words in reparanda was worse in the longest strings
  3. All fluent outcomes were significantly better than any disfluent outcomes
  4. **The longer the repair the less recall of words in reparanda**
  5. **Report rate falls more sharply in repetition disfluencies than in others**
  6. Repetition disfluencies are significantly more forgettable than others when they occur in utterances which are already difficult to process because of multiple false starts

  ➔ Repetition deafness helps to expunge disfluencies

• **Using Multiple regression**, repetition and recast disfluencies were subject to somewhat different influences
Thank You!

Many Thanks to Andrew Rosenberg, Bob Coyne, Agustin Gravano, Frank Enos, David Elson, Sasha Caskey, Kapil Thadani, Kristen Parton, and Sara Stolbach