AUTOMATIC ERROR DETECTION IN PRONUNCIATION TRAINING: WHERE WE ARE AND WHERE WE NEED TO GO

Silke Witt
June 2012
Outline

- A short History of CAPT
- What is Pronunciation?
- Existing Research Areas
- Commercial Systems
- Challenges in CAPT
- Conclusions: Where we need to go
Why ‘ADEPT’?

- Automatic Detection of Errors in Pronunciation Training:
  - Increasing global need for foreign language skills
  - Enable cheap access to powerful training methods
  - Availability: Anywhere and Anytime

Why now?

- Increased globalization has increased demand for foreign language learning
- Mobile devices and tablets are proving to be powerful application platforms
Outline

A short History of CAPT

What is Pronunciation?

Existing Research Areas

Commercial Systems

Challenges in CAPT

Conclusions: Where we need to go
Many contribution disciplines

- Pedagogy
- Second language learning
- Language acquisition theories
- Psycholinguistics
- Auditory processing
- Speech recognition
- Signal processing
- Phonetics
- Linguistics: syntax, semantics
- HCI: Human – computer interaction
A Short History of CAPT

Proliferation on Publications: > 100!

1990 2000 2010

CMU
SRI

KTH

Univ. of Nijmegen
Japan

Entropic
RosettaStone
Aleho

ETS

Ordinate/Pearson

CU HongKong
Erlangen
TU Dresden

Proliferation on Publications: > 100!
Outline

A short History of CAPT

What is Pronunciation?

Existing Research Areas

Commercial Systems

Challenges in CAPT

Conclusions: Where we need to go
Pronunciation Error Sources

**L₁ induced errors:**
- Pronunciation learning natural for young children up to about 10 years
  - Later brain loses flexibility, adaptability
- L₁ to L₂ language transfer
- L₂ phonemes that do not exist in L₁

**Spelling induced errors:**
- Unknown spelling rules or letter combinations
- Transfer of L₁ letter-to-sound rules
Pronunciation Error Types

Phonemic Error Types

- Phoneme Distortion Errors
- Phoneme Insertion
- Phoneme Substitution
- Phoneme Deletion

Syllable-level / diphone / Coarticulation Errors

Prosodic Error Types

- Stress
- Rhythm
- Intonation
“Native-like” versus “intelligible” pronunciation

- Intelligibility is an essential component of communicative competence.
- Errors related to prosodic features, such as vowel insertion, impact intelligibility more than segmental errors (Raux et al. 2002).
- A model of the human auditory system can be used to identify those pronunciation errors that are most noticeable to native speakers (Koniaris et al., 2011).
Distortion Errors, i.e. Mispronunciations

- Phonemic pronunciation error detection appropriate for language learning

- For advanced learners, focus needs to be on detecting subtleties.

Native speaker

```
h_{L2} \rightarrow ae \rightarrow l \rightarrow o
```

advanced non-native speaker

```
h \rightarrow ae \rightarrow l \rightarrow o
```

beginning non-native speaker

```
h_{L1} \rightarrow ae \rightarrow l \rightarrow o
```
Measuring Pronunciation

How do we go from this....... .. to “something” like this
How to score? How to measure?

- Many performance metrics:
  - Correlation, Pearson’ Coefficient, F1 score, labeling accuracy
  - Most metrics compare to human labeling
  - But human labeling is far from consistent...

- Maybe need something like a Bleu score for pronunciation?
  - Is there a way to establish a golden reference?
# Examples of Pronunciation Features

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Feature Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phonemic</strong></td>
<td>Phone-level log-likelihood scores, GOP</td>
</tr>
<tr>
<td></td>
<td>Vowel durations, duration trigrams</td>
</tr>
<tr>
<td></td>
<td>Phoneme pair classifiers</td>
</tr>
<tr>
<td></td>
<td>spectral features (formants)</td>
</tr>
<tr>
<td></td>
<td>Articulatory-acoustic features</td>
</tr>
<tr>
<td><strong>Prosodic</strong></td>
<td>distances between stressed and unstressed syllables</td>
</tr>
<tr>
<td><em>(Intonation, Stress, fluency)</em></td>
<td>Mean, max, min power per word (energy)</td>
</tr>
<tr>
<td></td>
<td>Fo contours (slope and maximum)</td>
</tr>
<tr>
<td></td>
<td>rate of speech (words per second/minute)</td>
</tr>
<tr>
<td></td>
<td>Trigram models to model phoneme duration in context</td>
</tr>
<tr>
<td></td>
<td>Phonation/time ratio, mean phoneme duration</td>
</tr>
<tr>
<td></td>
<td>Articulation Rate (phonemes/sec)</td>
</tr>
<tr>
<td></td>
<td>Mean and standard deviation of long silence duration</td>
</tr>
<tr>
<td></td>
<td>Silences per second</td>
</tr>
<tr>
<td></td>
<td>Frequency of disfluencies (pauses, fillers etc)</td>
</tr>
<tr>
<td></td>
<td>Total and mean pause time (i.e. duration of interword pauses)</td>
</tr>
</tbody>
</table>
Pronunciation Assessment

- Assessment assigns an overall pronunciation score to audio from a test participant.
- Allows averaging over several minutes of data.
- Allows averaging many instances of the same phoneme.
- Allows averaging over a number of intonation patterns.

- Extensive work has been done by Pearson and ETS.
- Comparing automated test results to human judgement showed high correlation in the range of 0.7 - 0.92 (Bernstein et al. 2010).
- Shows effectiveness of combining many phonemic as well as prosodic features.
Outline

- A short History of CAPT
- What is Pronunciation?
- Existing Research Areas
- Commercial Systems
- Challenges in CAPT
- Conclusions: Where we need to go
Likelihood-based Scoring

- Basic scoring method at phoneme level (Franco et al. 2000, Witt, 1999)
- Currently used in many systems as a baseline or in combination with other features Requires knowledge of spoken text
- Limited accuracy, particularly for partial mispronunciations
- At speaker or sentence level high correlation with human ratings
- Detection accuracy decreases for syllable and phoneme-level
- Detection accuracy also decreases for increased fluency levels
L1 Dependent Approaches

- Requires annotated non-native corpora -> expensive
  - Many existing corpora, but most task specific
  - New corpus required for each L1/L2 pair

Automatic L1/L2 mapping generation:
- Manual mapping derivation:
  - L1-L2 map (Husby et al.)
  - Manual error group detection (Neri et al.)
- Machine-learning based derivation:
  - Extended recognition networks (Harrison et al. 2008, 2009)
  - Statistical Machine-translation applied to learn errors (Stanley et al.)
L1 Independent Approaches

- Historically less accurate than L1-dependent scoring

- Improvements possible if focusing on common errors independent of L1 (Cucchiarini et al. 2011)

- Improvements possible if combining multiple features in classifiers (Cincarek et al. 2009)
Classifier-based Scoring

- Acoustic-phonetic LDA classifiers for selected set of phoneme pairs (Truong et al. 2004, Strik et al. 2007):
  - Focus on common phonetic substitution errors
  - L1-independent
  - Outperforms GOP

- Landmark*-based SVM classifier combined with confidence scores. Good results but only for specific phonemes. (Yoon et al. 2010)

- LDA-based classifier for loglikelihood, loglikelihood ratio, energy each for consonants and vowels (Amdal et al. 2009)

- Combination of confidence metrics and classifiers (Doremalen et al. 2009)

- Main challenge: Classifier training data need, limited set of errors address

*Landmark = sudden signal change
Non-native acoustic modeling

- Standard adaptation algorithms like MAP or MLLR adaptations give substantial accuracy gains (Hui et al., 2005)

- Even unsupervised adaptation to error-inflicted reduces error rate is better than no adaptation (Saz et al., 2009)
Text-independence

- One of most efficient ways for students to learn is via immersion in speaking environments.
- Early systems text-depending
- Non-native acoustic modeling makes text independence pronunciation assessment feasible

- Two-step approach (Moustroufas et al. 2007, Chen et al., 2009):
  1. Text recognition with adapted acoustic models
  2. Pronunciation assessment with native acoustic models
Prosody error detection

- Linear relationship between fluency measures and human judgments of prosody (Bernstein et al., 2011)
- Pitch recognition with SVM-based classifier (Levow et al. 2009)
- Extension of this approach:
  - Prosodic multi-feature set to detect word accents (Hönig et al. 2009)
  - Discriminative training with even larger set of prosodic features (Hönig et al. 2012)
Prosody Error Detection II

- Prosody feedback idea: Use user speech to create target user-specific pitch pattern (Bonneau et al. 2011)

- Tones as a special component of prosody, particularly intonation (Mixdorff et al. 2009, Hussein et al. 2011)

- Rhythm features added to existing features improved correlation with human ratings (Chen et al. 2011)
Auditory modeling

- Audiovisual articulatory feature inversion (Engwall et al. 2009 and 2012)
  - Estimate learner’s articulation from audio data
  - Compare native and learner articulation
  - Show tongue, mouth, lip movement with computer animation (both of target and student)
- L1 independent
- Initial evaluation results are promising
Corrective Feedback

- First challenge of L2 learners
  - can’t discern sound variations that don’t exist in L1
  - requires perception training
  - (Sonu et al. 2011 showed effectiveness of word and sentence-level perception training)

- Second challenge of L2 learners
  - Which parts of jaw, tongue, lips and vocal cords to move
  - how to produce the right sounds
  - If a score is given, what does the score mean?
Interactive CAPT system design

- Early immersive, interactive system: Subarashii (Bernstein et al. 1999)
- Dialog system design currently manual
- Needs interaction creation framework that can be used by teachers (see Alelo’s authoring framework)

- More evaluation, effectiveness and usability testing needed. (For example see Neri et al. 2006, Strik et al. 2009)
- Quantity and types of error feedback needs to be appropriate to student’s skill level
Outline

A short History of CAPT

What is Pronunciation?

Existing Research Areas

Commercial Systems

Challenges in CAPT

Conclusions: Where we need to go
Commercial Pronunciation Systems

- About 10 existing, established systems with complete lesson programs
- Several upcoming mobile or tablet applications, very much in infancy
- Pronunciation assessment to various degrees of effectiveness and detail
- Limited corrective feedback
- Three main challenges:
  - Ease of Lesson creation
  - “appropriateness” of feedback
  - Combination of all possible features
Example commercial system
Example commercial system
### Commercial Systems Overview

<table>
<thead>
<tr>
<th>Product</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Versant and VersantPro</td>
<td>Automated pronunciation assessment, measures speaking as well as listening</td>
</tr>
<tr>
<td>SpeechRater Engine (<a href="http://www.ets.org/research/topics/as_nlp/speech">http://www.ets.org/research/topics/as_nlp/speech</a>)</td>
<td>Automated pronunciation assessment as part of standardized tests &lt;br&gt;Pronunciation learning via AMEnglish.com includes training on stress, rhythm, intonation &lt;br&gt;Part of TOELF since 2006</td>
</tr>
<tr>
<td>EnglishCentral</td>
<td>English learning website, Assigns pronunciation score at sentence level, Tracks progress over time</td>
</tr>
</tbody>
</table>
## Commercial Systems Overview II

<table>
<thead>
<tr>
<th>Product</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CarnegieSpeech Assessment Climb Level 4 NativeAccent SpeakRussian SpeakFarsi</td>
<td>Pronunciation assessment as well as pronunciation teaching. Feedback at phone and sentence level Prosody Measuring pausing and duration</td>
</tr>
<tr>
<td>EduSpeak</td>
<td>Acoustic modeling of childrens’ speech</td>
</tr>
<tr>
<td>RosettaStone Totale</td>
<td>Immersion approach, all teaching in target language</td>
</tr>
</tbody>
</table>
## Commercial Systems Overview III

<table>
<thead>
<tr>
<th>Product Name &amp; Link</th>
<th>Company</th>
<th>Languages</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TellMeMore v10.0 (<a href="http://www.tellmemore.com">www.tellmemore.com</a>)</strong></td>
<td>Auralog</td>
<td>Spanish, French, German, Italian, English, Dutch, Chinese, Japanese, Arabic</td>
<td>Front &amp; sideview visualization of words, audio and Fo tracking.</td>
</tr>
<tr>
<td><strong>EyeSpeak (<a href="http://www.eyespeakenglish.com">www.eyespeakenglish.com</a>)</strong></td>
<td>EyeSpeak</td>
<td>US and British English</td>
<td>Audio comparison, measures each phoneme, timing, loudness. Student can listen to each phoneme segment, visual cross-section of mouth for each sound, pitch tracking</td>
</tr>
<tr>
<td><strong>Tactical Iraqi, Dari and Pashto</strong></td>
<td>Aleho (alelo.com)</td>
<td>Iraqi, Dari and Pashto</td>
<td>Pronunciation teaching and immediate corrective feedback embedded in interactive, 3D video games.</td>
</tr>
</tbody>
</table>
Outline

A short History of CAPT

What is Pronunciation?

Existing Research Areas

Commercial Systems

Challenges in CAPT

Conclusions: Where we need to go
Challenges I

1. Reliable phoneme-level error detection
   - Human agreement on phone-level errors ~0.7
   - Computer should have higher agreement while minimizing false positives

2. Distortion error assessment
   - Existing metrics do poorly on accented speech (Müller et al. 2009)
   - Discriminative training can help but only for some, pre-selected phoneme pairs (Yan et al., 2011)
   - How to measure accent that is close to native-like?
Challenges II

3. Text independence
   - Conservational, language immersion requires pronunciation assessment on unknown text

4. L1 independence
   - Using L1 knowledge good for increasing accuracy and corrective feedback, but too expensive and time consuming
   - Need ways to measure distance of student to native target
Challenges III

4. Corrective, audiovisual feedback
   • Needs knowledge of exact nature of error
   • Needs better auditory imaging to display instruction

5. Robust, interactive system design
   • Limited availability of flexible architectures to implement lessons
   • Challenges lie in integrating many possible features in an efficient manner
An imaginary, best-of-all worlds system

- Capability to map spoken audio to vocal-track, mouth movements and positions
- Depending on student’s level, focus is on intelligibility training versus native-like training
- Learning is embedded in natural conversations with a virtual agent (mimicking how a child learns)
- Dependent student’s level, virtual agent provide corrective, audio-visual feedback
- Tracking student’s progress over time
- System also does perception training
- System can detect and teach pronunciation for any L1
Conclusions

Much has been done, much is left to be done

Hand-crafted systems for a couple L1/L2 combinations are quite powerful, but more generic systems are too expensive to build.

Overarching challenge: How to effectively combine the many different approaches into a system

Many different features contribute to pronunciation. Error detection ideally utilize as many as possible

Big gap between and efficient creation of systems for any L1/L2
Thank you

Contact:
switt@fluentialinc.com