Computer-Assisted Language Learning (CALL) Systems

Tatsuya Kawahara (Kyoto University, Japan)
Nobuaki Minematsu (University of Tokyo, Japan)

Outline

- Introduction (TK)
  - Segmental Aspect & Speech Recognition Tech. (TK)
    - Pronunciation Structure Model (NM)
  - Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL Systems (TK)
- Database for CALL (NM)
(Traditional) LL → CALL

- (Traditional) LL: magnetic audio tape
  - Single media, Sequential access

- Computer-Assisted LL
  - Multi-media, Random access
    - Easier comparison of learner’s speech and model speech
  - Speech technology can be incorporated
    - Partly replace rater’s or teacher’s jobs

Speech Technology for LL

- Automate assessment of proficiency
  - PhonePass → Versant
  - ETS-TOEFL
  - PSC (Putonghua Shuiping Ceshi)

- Assist LL
  - With light supervision...CALL classroom
  - Self-learning
    - Need to keep motivating...Edutainment
    - Need to avoid enhancement of errors
Target Population of CALL

• **Non-native speakers**
  - Particular L1 (ex.) English LL for Japanese people
    - Still diverse in proficiency level, but L1 knowledge useful
    - Unlimited L1 (ex.) Japanese LL for people in the world

• **Children (native) [Russel 1996]**

• **Handicapped (Hearing or Articulation-impaired) people [Bernstein 1977]**

• **Accented (dialect) people**
  - Putonghua [Liu 2006]
  - Operators at Call Centers

Target Skill of CALL

• **Reading**

• **Writing**

• **Listening**

• **Speaking...including sentence composition**
  - Vocabulary, Grammar
  - Pragmatic Dialog (Communication)
    - travel-shopping, business-negotiation

• **Pronunciation...of given sentences**
  - Phone...segmental only
  - Word....+accent
  - Sentence...+intonation
  - Paragraph....+prominence
Importance of Pronunciation Training
[Bernstein 2003]

\[ \text{Comm} = \text{pron} \times \text{lex} \times (1 + \text{syn} + \text{rhet} + \text{prag} + \text{soc}) \]

- comm. = communication skills
- pron. = pronunciation
- lex. = lexical control and vocabulary
- syn. = syntax
- rhet. = rhetorical form
- prag. = pragmatics
- soc. = sociolinguistics

- Pronunciation skill affects entire communicative performance
- Native-sounding pronunciation may not be needed, but acceptable (intelligible enough) pronunciation is desired for smooth communication

Articulation \(\rightarrow\) Speech

- Students must learn how to control articulators (vocal tract)

- But it is not easy to observe the movement of these organs

- Observation is feasible for acoustic aspect of speech
Visual Presentation of Articulation

- Talking Head showing correct articulation [Massaro 2006]
- Acoustic-to-articulatory inversion to estimate the articulatory movements [Badin 2010]

Segmental and Prosodic Aspects

- Segmental Pronunciation  
  - Phonemes (Sub-words)
  - Features: spectrum envelop-based

- Prosody  
  - Tones
  - Lexical accents
  - Intonation and rhythm patterns
  - Features: fundamental frequency, power, and duration

(cf.) speech recognition

(cf.) speech synthesis
OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL Systems (TK)
- Database for CALL (NM)

Speech Technology used in CALL

- Speech analysis
  - spectrum, pitch, power
  - Feature normalization required for objective comparison with model speaker
- (Constrained) speech recognition (ASR)
  - Speech segmentation-alignment
  - Error detection
  - Scoring
  - Need to model non-native speech and handle erroneous input
  - Not only segmental aspect, but also prosodic aspects
- Speech synthesis (Minematsu)
Flowchart of Pronunciation Error Detection and Scoring

Speech Analysis

Segmentation

Error Detection

Scoring

Acoustic Model

Pronunciation Model

Speech Recognition

Formant and Articulatory Features

- Potentially useful for effective diagnosis and feedback
  - Direct relationship with articulation
- Not easy to make reliable and robust estimation
  - Not used in ASR
Classification of Vowels

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>front</th>
<th>middle</th>
<th>back</th>
</tr>
</thead>
<tbody>
<tr>
<td>narrow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>open</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“bat” vs. “but”</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relationship between Articulation and Formants

Articulation Chart

Formant Chart

This still requires expert knowledge on phonetics!
### Classification of Consonants (Japanese)

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Alveolar</th>
<th>Palatal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>voiced</td>
<td>unvoiced</td>
<td>voiced</td>
<td>unvoiced</td>
</tr>
<tr>
<td>Fricative</td>
<td>f*)</td>
<td>z</td>
<td>s</td>
<td>ʒ</td>
</tr>
<tr>
<td>Affricate</td>
<td>dz</td>
<td>ts</td>
<td>dʒ</td>
<td>tʃ</td>
</tr>
<tr>
<td>Stop</td>
<td>b</td>
<td>p</td>
<td>d</td>
<td>t</td>
</tr>
<tr>
<td>Semi-vowel</td>
<td>w</td>
<td>r**`)</td>
<td>j</td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“sea” vs. “she”

### MFCC: Mel-Frequency Cepstrum Coefficient

- Most widely-used spectral feature
  - Mel-bandwidth ← human perception
  - Cepstrum → spectrum envelope
    - orthogonal & less correlated → appropriate for statistical model

1. DFT(FFT) → power spectrum
2. Mel-conversion (Mel-band filter bank)
3. Logarithm + Cosine Transform (IDFT) → cepstrum
4. Extract low quefrency (12) coefficients
Feature Normalization in Speech Analysis

- Feature normalization
  - for objective comparison with model speaker
  - for score calculation via speech recognition
  - against speakers (native/non-native)
  - against acoustic channels (database/users)

- Normalization methods for MFCC
  - Cepstrum Mean Normalization (CMN)
  - Cepstrum Variance Normalization (CVN)
  - Histogram Equalization

Speaker Normalization in Speech Analysis

- Vocal-Tract Length Normalization (VTLN)
  - Warping spectral dimension
  - Based on acoustic model likelihood

- Pronunciation Structure (by Minematsu)
  - Invariant-feature (F-divergence)
Speech Recognition for CALL

- Tasks
  - Speech segmentation-alignment
  - Error detection
  - Scoring
- Challenges
  - Modeling non-native speech
  - Handling erroneous speech input
- Constraint
  - Target word or sentence is given

ASR vs. CALL

X: speech input,  W: phone label ↔ word sequence (target)

- ASR
  - For given X, find W that maximizes p(W|X)
  - Solved by max p(W)*p(X|W)
  - Each phone model p(x|w) is trained
- CALL
  - W (oracle) and X (not reliable) given,
  - Segmentation: Viterbi forced alignment
  - Error detection: find W’ such that p(X|W’)>p(X|W)
  - Scoring: evaluate p(X|W)?? How to train the model??
Flowchart of Pronunciation Error Detection and Scoring

- Pre-process for scoring
- Viterbi forced alignment with HMM representing W
- In fact, there may be pronunciation errors in X
  - Insertion & deletion seriously affect alignment
  - Error prediction/detection may be necessary
Segmentation

Speech Analysis

Segmentation

Error Detection

Scoring

Acoustic Model

Pronunciation Model

Flowchart of Pronunciation Error Detection and Scoring
Error Detection

- Find W' such that p(X | W') > p(X | W)
- Compute scores p(x | w') for alternative phones w' for each segmented region x
- When we take into account insertions and deletions, we need to generate a network of possible errors
- Error prediction can be done with prior knowledge, such as L1
  - Alternative phones w' can be taken from L1

Error Prediction in Pronunciation Model

- No equivalent syllable in L1 (ex.) sea → she
- No equivalent phoneme in L1 (ex.) l → r, v → b
- Vowel insertions (ex.) b-r → b-uh-r
Error Detection based on Classification Approach

- Not necessarily compute $p(x|w')$, but test if $w'$ is more likely than $w$

- Explicit classifier (verifier) learning
  - Incorporate many features
  - Focus on error detection by assuming segmentation
  - Linear Discriminant Analysis (LDA)
  - Support Vector Machines (SVM)

Other Issues in Error Detection

- Filter and prioritize many (possible) individual phone errors
- error miss $>>$ false alarm
  - Not to discourage learners
- Feedback
  - How to correct errors
Flowchart of Pronunciation Error Detection and Scoring

Scoring: Standpoints

- **Native-likeness**
  “How close to golden native speakers?”
  \[ P(X|W, \lambda_G) \]
  - What is the “golden” model? British? American?...
  - Impossible to free from L1 effect, speaker characteristic

- **Intelligibility**
  “How distinguishable (less confusable) from other phones?”
  \[ p(W|X) \]
  - Some pronunciation may not be recognized as anything
  - Need to consider L1 phones as well \( \rightarrow \) assume L1
Scoring based on Native-Likeness

- How close to golden native speakers?
  - Defined by \( p(X|W, \lambda_G) \) \( \lambda_G \): golden model
  - Normalized by \( p(X|W, \lambda_N) \) \( \lambda_N \): non-native model
  - In summary, likelihood ratio

\[
\frac{p(X|W, \lambda_G)}{p(X|W, \lambda_N)} \approx \prod_{w} \frac{p(x_i| w, \lambda_G)}{p(x_i| w, \lambda_N)} \approx \prod_{t} \frac{p(x_{t}| w, \lambda_G)}{p(x_{t}| w, \lambda_N)}
\]

Mean w.r.t. phones Mean w.r.t. time-frame

\( \Pi \): geometric mean= arithmetic mean in logarithm

Scoring based on Intelligibility

- How distinguishable (less confusable) from other phones?
  - Measured by \( p(W|X) \)

\[
\frac{p(X|W)}{\sum_{W'} p(X|W')} \approx \prod \frac{p(x| w)}{\sum_{w'} p(x| w')} \approx \prod_{t} \frac{p(x_{t}| w)}{\max_{w'} p(x_{t}| w')}
\]

\( \text{posterior prob.} \) \( \text{forced alignment (segmentation)} \) \( \text{Viterbi score} \)

- Often called GOP (Goodness Of Pronunciation)
- Similar to confidence measure in ASR...becomes 1 if best \( w'=w \)
Extension of GOP

- Better to limit w’ to confusing phones, rather than all possible phones [Wei 2009]
  - (ex.) /v/ → /b/
- Better to set different thresholds depending on phones (or phone clusters) for error detection [Ito 2005]
  - Any error can be detected if we use phone-loop model

- Acoustic & Pronunciation Model
  - Need to adapt to non-native speakers
  - Better to consider L1 phones

Scoring to Assessment

- Other factors
  - Duration modeling & evaluation
  - Other prosodic aspects...accent, intonation
  - Speech rate
- Score mapping
  - Linear regression to fit to human rater’s evaluation
Flowchart of Pronunciation Error Detection and Scoring

Acoustic Modeling: Native vs. Non-native

- Native speech
  - “Gold standard”, but does not match
- Non-native speech
  - Matched, but error-prone
  - There is not large database available
- Adaptation from native to non-native
- Phone model of L1 is used for the same phone (in the IPA inventory)
Context-Independent Modeling

- Context-dependent (e.g. triphone) models are widely used in ASR

- Context-independent (monophone) model works well, even better, in CALL
  - Phonetic context is not reliable in non-native speech \(\leftarrow\) insertion of vowels
  - Better segmentation accuracy even in native speech

Speaker Adaptation of Acoustic Model to Non-native Speech

- Pronunciation of adaptation data may not be correct
- Compare baseform label (automatic but error prone) and hand label (correct but costly)
- Phone accuracy: measured based on hand-label including errors

[Tsubota 2004]

<table>
<thead>
<tr>
<th>Acoustic model (native model)</th>
<th>Phone accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adaptation</td>
<td>75.4</td>
</tr>
<tr>
<td>Hand label</td>
<td>81.0</td>
</tr>
<tr>
<td>Baseform label</td>
<td>80.6</td>
</tr>
</tbody>
</table>

Lexicon baseform label is sufficient
Acoustic Model:
Native model vs. Non-native (L2) model

- Non-native speech database (MEXT project)
  - 13129 utterances by 178 speakers
  - Pronunciation errors are not annotated (too costly)
  - Dictionary label vs. automatic label with ASR
    - Both are error prone

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>baseline</th>
<th>speaker adapt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native English model</td>
<td>75.4</td>
<td>80.6</td>
</tr>
<tr>
<td>Non-native model (baseform)</td>
<td>78.0</td>
<td>81.8</td>
</tr>
<tr>
<td>Non-native model (ASR)</td>
<td>77.1</td>
<td>81.5</td>
</tr>
</tbody>
</table>

- Non-native model is more effective, even with dictionary label
- The superiority is reduced with speaker adaptation

Acoustic Model:
Native/L2 model + L1 model

- Many phones are shared by target L2 and L1
  - (ex.) most of consonants shared by Japanese and English
- Incorporate L1 acoustic model for shared phones in parallel

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>baseline</th>
<th>speaker adapt</th>
</tr>
</thead>
<tbody>
<tr>
<td>English native model + L1 model</td>
<td>78.9</td>
<td>81.3</td>
</tr>
<tr>
<td>Non-native model + L1 model</td>
<td>78.7</td>
<td>81.5</td>
</tr>
</tbody>
</table>

- Incorporating L1 model in parallel is very effective
Flowchart of Pronunciation Error Detection and Scoring

- Standard baseform → possible errors
- Constraint of L1 is effective
- Linguistic knowledge
  - /v/ → /b/, /θ/ → /s/
  - Substitution with similar phone of L1
  - Insertion of vowels

- For GOP score computation, simple phone loop model (=no pronunciation model) is used
  - Any errors can be detected, with many false alarms
**Error Prediction in Pronunciation Model**

- No equivalent syllable in L1
  (ex.) sea → she
- No equivalent phoneme in L1
  (ex.) l → r, v → b
- Vowel insertions
  (ex.) b-r → b-uh-r

**Pronunciation Model Training**

- Hand-craft phonological rules
  - Expert knowledge needed
  - Too many rules cause false alarms, degrading recognition performance
  - Tradeoff between coverage and perplexity

- Machine learning from annotated data
  - Statistical learning of rewriting rules [Meng 2011]
  - Decision tree to find critical rule set [Wang 2009]
OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL Systems (TK)
- Database for CALL (NM)

English CALL System: HUGO @Kyoto Univ. [Tsubota, Imoto, Raux 2002]

- For Japanese college students, so that they can introduce Japanese cultures
- Dedicated acoustic model & error prediction scheme for Japanese students
- Deployed and used in classrooms
English CALL System: HUGO @Kyoto Univ.

• Goal: Pinpointing the pronunciation errors which degrade intelligibility and providing effective feedback

• Practice consists of two phases
  1. Dialogue-based skit (for natural conversation)
  2. Training on specific errors detected in the first phase (using a phrase or a word)

• Pronunciation error detection
  • Segmental pronunciation $\leftarrow$ hand-crafted phonological rules
  • Accent (Primary & Secondary Stress) $\leftarrow$ multiple prosodic features
/R-L/ substitution

Select an item: Period

You
Model
Words

Evaluating segmentals
Performing phoneme recognition...

R and L
To pronounce an L, your tongue must touch the back of your upper teeth.
For R, concentrate on progressively rounding your lips.

Function words

Select an item: During this period.

You
Model
Words

Please record the given item

Function words
Words such as prepositions (after, under...), pronouns (he, them...), or articles (the, her...) are called function words. They don't carry any essential meaning but have a grammatical role. Such words are usually not stressed in English. Concentrate on shortening their vowels and keeping intonation and pace as flat as possible on these words.
List of Pronunciation Errors

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>W/Y deletion (would)</td>
<td>Final vowel insertion (let)</td>
</tr>
<tr>
<td>SH/CH substitution (choose)</td>
<td>CCV-cluster insertion (active)</td>
</tr>
<tr>
<td>R/L substitution (road)</td>
<td>VCC-cluster insertion (study)</td>
</tr>
<tr>
<td>ER/A substitution (paper)</td>
<td>H/F substitution (fire)</td>
</tr>
<tr>
<td>Non-reduction (student)</td>
<td></td>
</tr>
</tbody>
</table>

- Built from literature in ESL
- Remove error patterns with low detection rate

Intelligibility Assessment based on Error Statistics

- Estimate
- Find critical errors
Priority of Training on Specific Errors according to Intelligibility Level

NativeAccent [Eskenazi 2007]

- Product of Fluency Project of CMU
- English learning
  - Error detection and feedback on articulation
  - Up to 28 L1: Japanese, Russian, French...
  - 800 exercises
English CALL System @ CUHK [Meng 2010]

- For Chinese learners of English
- Corpus: 100 Cantonese and 111 Mandarin L1
  - Reading a paragraph, words
- Pronunciation error model
  - Hand-crafted phonological rules
  - Data-driven patterns
- GOP score
- Pre-filtering based on duration models
- Synthesizing expressive speech to convey emphasis in feedback generation
- Synthesizing visual speech with articulator animation

Shadowing Exercise [Luo 2009]

- listening and repetition of native utterances, online
  - Simultaneous training of listening and speaking skills
- High correlation between GOP and TOEIC scores (= 0.90)
  - Higher than simple reading
ETS SpeechRater for TOEFL
[Zechner 2007]

- Assessment of unconstrained English speech
  - TOEFL iBT Practice Online (TPO)
  - iBT Field Study
- Based on LVCSR methodology
  - Acoustic model: non-native speech (30hours)
  - Language model: non-native speech + broadcast news
- Features: ASR results (word ID, confidence), speech rate, pause length... 40 in total
- Scoring: linear regression model
- Correlation with human rater: 0.67
  - Inter-human correlation 0.94

Dialog-Based English CALL @POSTECH
[Lee 2010]

- Situated dialog...(ex.) shopping
- Based on SDS (Spoken Dialog System) methodology
  - Task-dependent ASR+SLU
- Example-Based Dialog Management
- Corrective feedback based on example selection

- Field trial on elementary school
Japanese CALL system: CALLJ @Kyoto Univ. [Wang 2009]

- Exercise of basic sentence production (text & speech), given a image scene
- Key features
  - Dynamic generation of questions & ASR grammar network with error prediction
  - Interactive hints


Japanese CALL system: CALLJ @Kyoto Univ. How to Try

- Windows only.

0. (Unzip CALLJ1.5.zip).
1. Move to the directory CALLJ.
2. Click “StartCALLJ”.
3. Create your account by clicking “New” in login window for the first time.

- You need some knowledge on Japanese.
## References

References


