Accent Reduction for Computer-Aided Language Learning

Zhao Sixuan, IMI PhD Student
Prof Koh Soo Ngee, Supervisor
School of Electrical and Electronic Engineering, NTU
Prof Luke Kang Kwong, Co-supervisor
School of Humanities and Social Sciences, NTU
Outline

- Introduction
- Accent Reduction Methods
- Evaluation on Converted Speech
- Voice Conversion to Reduce Accent
- Summary
Introduction

- **Spoken English of Non-native English Learners:**
  1. Influenced by accents and knowledge of the learner’s native language.
  2. Failure to pronounce English in an appropriate way.
  3. Failure to manipulate stress and rhythm in English.
  4. From the viewpoint of language acquisition, prosody is the hardest and the last to be acquired.
Introduction

- Proposed System Structure
Introduction

Feedback Utterances in CALL System: utterances pronounced by the teacher with native and natural prosody, not considering the differences between the learner’s and the teacher’s voice features.

- Current Feedback:

- Proposed Feedback:
Motivation for Accent Reduction:

1. A “golden speaker” whose voice is highly similar to that of the learner, can enable the learner to more concentrate on the pronunciation and prosody issues in learning process (K. Probst, Y. Ke, and M. Eskenazi, “Enhancing foreign language tutors-In search of the golden speaker,” Speech Communication, vol. 37, no. 3-4, pp. 161-173, 2002.)


3. Accent reduction can be used to significantly improve the intelligibility of cross-language speech synthesis system. (K. Yanagisawa, and M. Huckvale, “Accent morphing as a technique to improve the intelligibility of foreign-accented speech,” in International Congress of Phonetics Sciences, Saarbrücken, Germany, 2007.)
**Rule-based Method:** The transformation of perceived accents is based on *a thorough study of the different English accents*. Formants & prosodic transformations are performed to convert from the source accent to the target accent.

**Reference-based Method:** Accent reduction is performed with reference utterances given by native speakers. The reference utterances provide linguistic cues to reduce prosody & pronunciation errors of foreign English learners.


**Accent Reduction for Synthesized Speech:** Reference-based or rule-based method is performed on synthesized speech to extend the ability of text-to-speech (TTS) systems, e.g., using TTS system to generate foreign languages.


(K. Yanagisawa, and M. Huckvale, “Accent morphing as a technique to improve the intelligibility of foreign-accented speech,” in International Congress of Phonetics Sciences, Saarbrücken, Germany, 2007.)
Accent Reduction Scheme: Modify the learner’s speech according to the teacher’s speech to obtain accent-free feedback with the learner’s speaker identity:
Prosodic Feature Modification:
1. Forced alignment: to obtain corresponding speech pairs at phoneme level.
2. Obtain the excitation of each phoneme segment.
3. Pitch and duration calculation.
4. Pitch-synchronous overlapping and adding (PSOLA) method: to reconstruct each segment of the learner’s utterance according to the teacher’s pitch profile and duration.
Accent Reduction Methods

- Pitch-synchronous overlapping and adding (PSOLA)[1][2]: Decompose speech into pitch-synchronous frames for analysis and synthesis.

Accent Reduction Methods

- **Selective Pitch Modification**: The pitch contour of a phoneme of the student will be modified if its shape is different from that of the teacher’s phoneme.

<table>
<thead>
<tr>
<th>L-H</th>
<th>H-L</th>
<th>H-L-H</th>
<th>L-H-L</th>
<th>Others</th>
</tr>
</thead>
</table>

- **Pitch Modification Factor**: The pitch contour is substituted by the teacher’s pitch contour whereas *mean pitch value is preserved*

\[
\beta = \frac{(\phi(P^T(t)) - \overline{P^T(t)}) + \overline{P^L(t)}}{P^L(t)}
\]

\[(1)\]
**Accent Reduction Methods**

- **Segmental Feature Modification:** The vocal tract filters (spectral envelope) of the learner in each phoneme segment is modified according to the envelope of the teacher’s segment.
**Accent Reduction Methods**

- **Restoration of the learner’s speaker identity:** After the transformation of segmental feature, the learner’s formant frequencies may also change due to the spectrum modification. **Vocal Tract Length Normalization (VTLN)** [1] is used to *normalize the formant frequency of the modified speech* to that of the learner.

---

Accent Reduction Methods

- **Spectral Interpolation:** The substitution of spectral envelope of the learner’s speech by that of the teacher’s speech may create *discontinuity at phoneme boundaries*. Thus, spectral envelope interpolation should be used to smooth the boundaries and to reduce the distortions.
Accent Reduction Methods

- **Segmental Modification:**

  ![Graphs showing Learner's Spectrum, Flattened Spectrum (Excitation), Reconstructed Spectrum, and Teacher's Spectral Envelope.]

- **Graphs:**
  - Learner's ST-spectrum vs. Learner's Spectral Envelope.
  - Flattened Spectrum (Excitation).
  - Reconstructed ST-spectrum vs. Original Spectral Envelope.
  - Teacher's Spectral Envelope.

- **Institute for Media Innovation**
  - [Logo]
  - [Nanyang Technological University]
Evaluation on Converted Speech

- **Evaluation on Converted Speech**: 200 utterances from non-native students are recorded to test the accent reduction method.

- **Database and Subjective Evaluation:**

<table>
<thead>
<tr>
<th>Database</th>
<th>Boston University Radio News Corpus (BURNC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcriptions</td>
<td>20 unique sentences selected from BURNC</td>
</tr>
<tr>
<td>Students’ Utterances</td>
<td>Total 200 students’ sentences including Chinese, Indians, Vietnamese and Singaporeans.</td>
</tr>
<tr>
<td>Teachers’ Utterances</td>
<td>Speaker M1B and F2B in BURNC with from BURNC.</td>
</tr>
</tbody>
</table>
Evaluation on Converted Speech

- **Samples of Accent Reduced Utterances (male)**:
  
  **Transcription**: What's going to make the *difference this* time around?
  
  1. The Learner’s Speech
  2. The Teacher’s Speech
  3. Prosodic Features Modified
  4. Segmental Features Modified
  5. Both Features Modified
Evaluation on Converted Speech

- **Samples of Accent Reduced Utterances (male)**:
  
  **Transcription:** This last young *mother*-to-be is *luckier* than many.

1. The Learner’s Speech

2. The Teacher’s Speech

3. Prosodic Features Modified

4. Segmental Features Modified

5. Both Features Modified
Evaluation on Converted Speech

- **Evaluation of Converted Speech**: objective measurements are performed to test the "accentedness" as well as the acoustic quality of converted speech.

- **1. Accentedness**: speech recognition software HTK is used to obtain the posterior score of an utterance:

\[
S_{\text{accent}} = -\text{mean}\{\log \frac{p(o_j | \lambda_j)}{p(o_j | \lambda_{\text{max}})} ; j=1, 2, ..., n\}
\]

accent score is calculated for each sentence.

where \( o_j \) is the observation, \( \lambda_j \) is the HMM model, \( S_{\text{accent}} \) is the sentence probability score representing the accentedness of the input speech.

2. **Acoustic Quality**: ITU standard P.563 is used to evaluate the acoustic quality of modified speech. The range of mean opinion score (MOS) is from 1 to 5 (worst to best). The general diagram [1] is shown as follows:

![Diagram of Acoustic Quality Evaluation](image)

Evaluation on Converted Speech

- Accentedness Scores of 200 Utterances for Each Kind Stimuli:

Accentedness: Student > Segmental Modification > Prosodic Modification > Combined Modification > Teacher
Evaluation on Converted Speech

- Acoustic Quality MOS:

MOS: Student > Prosodic Modification > Segmental Modification > Combined Modification
Outline

Introduction

Accent Reduction Methods

Evaluation on Converted Speech

Voice Conversion to Reduce Accent

Summary
Voice Conversion to Reduce Accent

- **Feedback Based on Voice Conversion:** An alternative way to obtain feedback utterances with correct prosody & pronunciation as well as the student’s speaker identity may result from voice conversion methods.

- **Teacher-to-Learner Conversion:** The teacher’s voice is converted to the learner’s voice. The converted speech will reserve the teacher’s prosody and pronunciation (as only speaker identity feature is converted), while obtaining the learner’s voice features.
Voice Conversion to Reduce Accent

- **Voice Conversion Scheme:** The technique to transform the voice of the source speaker to that of the target speaker while preserving the content of speech.
Voice Conversion to Reduce Accent

- **Training Scheme:**

  - Source Speech
  - Analysis
  - Target Frames
  - Source Frames
  - Target Speech
  - Alignment
  - Synthesis Parameters
  - LSF Parameters
  - Training Function (GMMs)

- **Conversion Scheme:**

  - Source Speech
  - Analysis
  - Parameters Conversion
  - Conversion
  - Parameters Conversion
  - Synthesis
  - Converted Speech
  - Synthesis Parameters
  - LSF Parameters
  - LSF Parameters
  - Synthesis Parameters
Voice Conversion to Reduce Accent

- **Samples of Voice Converted Utterances (male):**
  
  **Transcription:** What's going to make the *difference this* time around?

1. The Teacher’s Speech (source)

2. The Learner’s Speech (target)

3. Converted Speech
Voice Conversion to Reduce Accent

- **Samples of Voice Converted Utterances (male)**:
  
  **Transcription**: But state officials aren't *convinced* that's a problem.

1. The Teacher’s Speech (source)
2. The Learner’s Speech (target)
3. Converted Speech
# Voice Conversion to Reduce Accent

## Comparison of Accent Reduction Method and Voice Conversion Method:

<table>
<thead>
<tr>
<th>Method</th>
<th>Voice Quality</th>
<th>Speaker Identity</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent Reduction</td>
<td>Lower quality (due to phoneme-level modification)</td>
<td>Similar to the student (the original excitation is manipulated)</td>
<td>One-to-one mapping, no database required</td>
</tr>
<tr>
<td>Voice Conversion</td>
<td>Smoother but muffled (over-smoothness introduced by GMM)</td>
<td>Contains some of teacher’s voice features (excitation &amp; vocal tract are converted from that of the teacher)</td>
<td>Parallel utterances are required for model training</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Accent Reduction Methods
- Evaluation on Converted Speech
- Voice Conversion to Reduce Accent
- Summary
Summary

- **Accent Reduction**: Accent reduction and speech synthesis techniques can be used to generate effective feedback utterances for speaking skills training by modifying the learner’s utterance according to that of the teacher.

- **Evaluation of Converted Speech**: Accentedness is effectively reduced by accent reduced. Prosodic modification is the primary factor to reduce accentedness. The degradation of acoustic quality mainly results from segmental modification.

- **Voice Conversion Method**: Voice conversion techniques can also be used to obtain the desired feedback utterances. It outperforms accent conversion in terms of smoothness and acoustic quality, but reduces the learner’s speaker identity.
Study the tradeoff between accent reduction and voice conversion methods and try to combine the benefits of the two.

Perform accent reduction based on other different speech synthesis models like HSM (Harmonic Stochastic Model) and STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weighted spectrogram) to further improve the converted speech.

Study the influence of various factors on the quality and the nativeness of converted speech, like database, gender, and nationality of speakers. Preliminary results show that reference corpus can affect the selection of accent conversion methods (i.e., segmental or prosodic modification).

Perform pedagogical experiments to study the effectiveness of the proposed schemes.


Final Slide

Thanks!

Q&A