Computer-Aided Evaluation of Prosody for Language Learning

Supervisor: Prof Koh Soo Ngee
School of Electrical and Electronic Engineering, NTU

Co-Supervisor: Prof Luke Kang Kwong
School of Humanity and Social Science, NTU

PhD Student: Zhao Sixuan
School of Electrical and Electronic Engineering, NTU
General Objective

- General Objective: To develop a robust and effective CALL (Computer Aided Language Learning) system based on prosody evaluation, which mainly targets at non-native English speakers.

One of the top 14 engineering challenges identified by the National Academy of Engineering is “Advance Personalized Learning” and to move towards this goal, more effective CALL systems must be developed.

- Advantages of CALL System:
  1. Self-adapting Learning
  2. Anxiety-free Practice
  3. Personalized Learning
General Objective

- **Current CALL Systems:** Current systems are mainly based on pronunciation evaluation of each uttered word, syllable or phoneme rather than prosody.
- **Pronunciation Patterns**
- **Eyespeak**
- **Speak to me**
- **Proposed CALL Approach:** We propose to use the concepts of prosodic analysis to identify appropriate prosodic units for matching and evaluation rather than concentrating on individual syllable or word.
Prosody Problems for Learners

- Prosody Problems of Non-native English Learners:
  1. Pronunciation influenced by their accents and knowledge of their native language.
  2. Failure to manipulate stress and phrasing in English.
  3. Inappropriate usage of prosody in different situations.
  4. From the viewpoint of language acquisition, prosody is the hardest and the last to be acquired.
As it is quite a new research topic, prosody evaluation system can only be found from very few recent papers:


List of Recent Papers


Roles of Engineering and Linguistics

- **Role of Engineering:**
  1. to extract features from speech signals for computer processing.
  2. to develop evaluation method for extracted features.
  3. to establish machine scoring based on evaluation results.

- **Role of Linguistics:**
  1. to define important features of prosody.
  2. to define prosodic units for engineering modeling.
  3. to provide human scoring for comparison.
Important Linguistic Features

- **Definition of Prosody:** The supra-segmental feature which combines stress, rhythm and intonation of speech, reflecting emphasis, contrasts, emotional states and so forth.

- **Prosody Assessment in a CALL system:**
  1. feature extraction from input speech utterances.
  2. modeling of features to describe the prosody.
  3. assessment on modeled features.
Important Linguistic Features

- Fundamental Components of Prosody: prosodic phrasing and prominence.

1. Prosodic Phrasing: The grouping of adjacent words in speech utterances according to the prosodic features.

2. Prominence: The particular stress or emphasis on syllable expressing either prosodic or lexical meanings.
Prosodic Phrasing

- **Intonation Group**: A continuous pitch contour, which may always be a clause or something less than a clause. *E.g.* Please do this for me/ since I am busy now.

- **Factors to Determine Intonation Group**:
  1. The average length. The average length for an intonation group is five words.
  2. Syntactic correlation. A group of words with syntactic cohesion is more likely to be an intonation group.
  3. Boundary markers in the utterance.
Prosodic Phrasing

- **Boundary Marker**: The linguistic phenomenon to divide intonation groups.

1. **External Boundary**: the phonetic cues existing at the actual boundary, i.e. pause.

- **Two Categories of Pause**:
  1. the unfilled pause (silence).
  2. filled pause, e.g. like the vowel in "gate" and "play".

2. **Internal Boundary**:

1. Change of pitch level or duration of unstressed syllables. E.g. My *Ma-ma* lives in the *mountain*.
2. In-sentence pause (pauses between clauses).
Prominence

- **Prominence**: Two categories of prominences exist, as shown below:

1. **Lexical Stress**: In lexical stressed language, the syllables in one word are not pronounced equally, the emphasis on some syllables can affect the meaning of the word. E.g. *present, present*.

2. **Phrasal Prominence**: The relatively stronger syllables above the level of the phrase due to the phrasing or context requirement. E.g. *Your behavior is wrong!"*
Prominence

- **Concentration on Prominence:** Current research on prominence mainly concentrates on the nucleus of prominence.
- **Nucleus:** The prominence which stands as the most prominent in an intonation-group.

- **Features of English Nucleus:**
  1. The initial movement from the nucleus: fall, rise or level.
  2. The start point of this initial movement: high or low.
  3. A second change of pitch direction following the nucleus.
1. General Objective

2. Important linguistic Features

3. Engineering Methods

4. Proposed Research Directions

5. Conclusion
Prosody Evaluation Methods

- Mainstream Prosody Evaluation Method:

  1. **Deterministic Method:** To directly compare learners’ prosody with that of the pre-recorded experts.
  2. **Statistical method:** To build statistical model for prosodic components based on linguistic theory and thus to evaluate the prosody.

- **Choice:** Deterministic model.
  2. Convenience for modification.
  3. Avoidance of training process.
Flowchart of Deterministic Method

1. Learners' Utterances
2. Feature Extraction
3. Alignment
4. Comparison
5. Machine Score Calculation
6. Teachers' Utterance
7. Feature Extraction
Feature Extraction: Prosodic features can be basically divided into two main groups, rhythm and intonation, which mainly provides prominence and phrasing information, respectively.

**Rhythm**: Rhythm is the variation of the length and accentuation of a series of sounds or other events.

**Intonation**: Intonation is the variation of pitch while speaking which is not used to distinguish words.
1. **Rhythm Feature**: Speech duration and stress pattern.

   - **Speech Duration**: The lasting time of parts of speech utterances.

     The comparison between the speech durations of the leaner and that of the teacher can be described by “duration ratio”:

     $$ R_L(k) = \frac{\max(L(k), L^s(k))}{\min(L(k), L^s(k))} $$

     Here $L(k)$ is the duration of the learner’s utterance and $L^s(k)$ stands as the duration of the teacher’s utterance.
**Rhythm Feature**

- **Stress Pattern**: The intensity contour of the speech utterance.

  With DTW (Dynamic Time Warping) algorithm, the stress comparison result can be defined as:

  \[
  X_{DP}(k) = \frac{1}{I_k + J_K} g(I_k, J_K)
  \]

  \(I\) and \(J\) are the frame numbers of teachers’ and learners’ speech. \(g\) is the accumulated distance between the stress patterns of two sequences. \(K\) is the number of utterance.
2. Intonation Feature: 4-dimension feature vector with pitch, derivative pitch, power, derivative power. Due to the different lengths of teacher’s and learner’s speech utterances, DTW is used in this section to align the teachers’ and learners’ utterances as the same length, as shown below:
Outline

1. General Objective
2. Important linguistic Features
3. Engineering Methods
4. Proposed Research Directions
5. Conclusion
Proposed Research Directions

1. Involvement of Different Linguistic Units:
   Prosodic features involve various hierarchies, including:
   - Syllable
   - Word
   - Foot: delimited by the stress of sentences. E.g. /My/ Ma-ma lives in the/ mountain/.
   - Intonation Unit: a continuous pitch contour.
   The concentration on new prosodic hierarchies can further improve the evaluation result.
2. Modeling of Linguistic Features:

More linguistic features can be identified and modeled to improve the evaluation on the prosody:

- Boundary markers of intonation group
- Prominence information
- Nucleus

Acoustic features such as pitch, intensity and duration can be modeled to describe above features. The aim is to find the best modeling method.
3. Alignment Between Two Speech Utterances:
   In dynamic programming, there are various methods for aligning two utterance sequences:
   - alignment based on intonation contour
   - alignment based on linear or nonlinear combination of different acoustic features
   - alignment based on MFCC (mel frequency cepstrum coefficients)
   The selection of alignment methods can greatly influence the comparison results.
4. Appropriate Combination with Other Evaluation Method:

Language learning system is not only based on prosody evaluation, but also includes the following techniques:

- Phoneme Detection
- Face Detection
- Spectrogram Comparison

Combination with above systems can further improve the human-machine correlation of the evaluation results.
Outline

1. General Objective
2. Important linguistic Features
3. Engineering Methods
4. Proposed Research Directions
5. Conclusion
Conclusion

- **Proposed Evaluation Method:** Deterministic method which compares the reference speech features with those of the learner.

- **Evaluation Features:** The foot and intonation unit evaluation will be included, also combined with current word and syllable level evaluation.

- **Feature Extraction:** Using acoustic features to describe linguistic features, along with MFCC for alignment. Besides, spectrogram comparison will be included.
Combination with Other System: We will first include phoneme detection system. The combination of the two can describe both lexical and prosodic errors made by English learners.
Final Slide

Thanks!

Q&A