CS626-460: Lecture 34

Pronunciation Scoring For Language Learners Using A Phone Recognition System

Presented by,
K.L Srinivas (M.Tech 2nd year)

Guided by,
Prof. Preeti Rao (Elect. Dept)

Department of Electrical Engineering, IIT Bombay
Mumbai, India
Introduction

Pronunciation refers to the manner in which a particular word of a language is uttered.

Motivation

- Accurate pronunciation or articulation is a vital component of a language acquisition process.
- Fluency in speech of a non-native speaker of a language can be judged by pronunciation and prosody.
- Non availability of a classroom environment for learners.

Subjective Evaluation

Word spoken: Kaleidoscopic

Speaker 1  Speaker 2
Problem statement

- Developing computer based automatic pronunciation scoring system.
- Accessing the closeness of language learner pronunciation to that of reference speaker (already stored in system).
- To provide language learner with pronunciation score and feedback.
A brief on Automatic Speech Recognition
Introduction

- Automatic speech recognition (ASR) is a process by which an acoustic speech signal is converted into a set of words.

- Getting a computer to understand spoken language.

- Approaches to ASR
  - Template matching
  - Knowledge-based (rule based approach)
  - Statistical approach (machine learning)
Statistical based approach:

- Collect a large corpus of transcribed speech recordings.
- Train the computer to learn the corresponding instances (Machine learning).
- At run time, apply statistical processes to search through the space of all possible solutions and pick the statistically most likely one.
Speech recognition tool kits:

- Sphinx and HTK are two widely accepted and used speech recognition tools.
  - CMU sphinx: Carnegie Mellon University (CMU)
  - HTK: Cambridge University

- Both the frameworks are used for developing, training and testing a speech model from existing corpus speech data.

- Both use Hidden Markov Modeling techniques.
MFCC feature vector:

- The *Mel-Frequency Cepstrum Coefficients* (MFCC) is a popular choice
- Frame size: 25 msec
- Hop size: 10 msec
- 39 feature per 10ms frame
  - Absolute: Log Frame Energy (1) and MFCCs (12)
  - Delta: First-order derivatives of the 13 absolute coefficients
  - Delta-Delta: Second-order derivatives of the 13 absolute coefficients
Sphinx 3:

Training:

Testing / Decoding:
Decoder output:

Recognition Hypothesis:
- This gives the single best recognition result for each utterance processed.
- Linear word sequence with their time segmentation and their scores.

Output format:
<word> <start frame> <end frame> <AScr> <LM Score> <AScore +LM Score> <Ascale>
Automatic Speech Recognition for non-native speech

- Non-native speech characters:
  - Phone substitutions: S in word ‘she’ pronounced as s
  - Phonotactic constraints: Stop cluster sk in school pronounced as iskU

- Use of language model masks out the non-nativeness during recognition.
- Accuracy of state-of-the-art phone recognition systems as low as 50%-70%
- Traditional ASR techniques cannot be used for non-native speech
- Phone recognition to be carried out in constrained mode
Back to pronunciation scoring
Pronunciation Scoring System

- Canonical transcription of the utterance
- Generation of Pronunciation Variants
- Constrained Phone Decoder
- Variant Selection
- Boundary Refinement and Prosodic analysis
- Prosody Score
- Articulation Score
- Pronunciation Score

Input speech signal

Department of Electrical Engineering, IIT Bombay
Pronunciation Variants

**Word**: Fundamentals

**Canonical form**: SIL f aa n d aa m ee n’ clt t aa l s SIL

**Variant_1**: SIL f aa n d a m ee n’ clt t aa l s SIL
**Variant_2**: SIL f aa n d ee m ee n’ clt t aa l s SIL

➢ Challenges

• No ready database of speakers of Indian English
• Multiple L1s for Indian speakers poses further challenges.
• Native Hindi and native English databases are available
Constrained Phone Decoding
HMM based recognizers

- HTK 3.4
- Sphinx 3

Decoding

- Input Speech Utterance
- Extraction of MFCC Feature Vectors
- Decoder in Forced Alignment Mode

Acoustic Models from training
Variants

Aligned Phone Sequence with likelihood for each variant
Variant Selection

Aligned Phone Sequence with likelihood for each variant

Select Variant with the highest likelihood

Visual Feedback and Articulation Score

Past Work

- Strik and Cucchiarini (2000): Pronunciation variations and modeling
- Goronzy, Rapp and Komp (2004): Non-native pronunciation variations and generation (native English speakers speaking German)
- Witt and Young (2000): Presented likelihood based goodness of pronunciation scheme
Databases

- **TIMIT database**
  - 630 speakers of 8 major dialects of American English.
  - Each speaking 10 phonetically rich sentences.

- **TIFR database**
  - 100 native speakers of Hindi.
  - Each speaking 10 phonetically rich sentences.

- **Indian English database - Testing**
  - 30 Indian college students each speaking the 2 common sentences from TIMIT database.

Acoustic Phone Models

- **47 class TIMIT models**: Entire phone set from TIMIT.
- **52 class Union models**: Entire TIMIT phone set (47 phones) and 5 additional phones from the TIFR phone set making a total of 52 phones.
- **48 class Union models**: Entire TIFR Hindi phone set (36 phones) and 12 phones from TIMIT.
Experiments and Evaluation

The focus of this work is to investigate the effect of selection of phone models from one of 47, 52-union and 48-union phone models.

Evaluation Measures

- **Method I**: The number of instances in which the surface transcription is within the top N decoded sequences in terms of likelihood score.

- **Method II**: The edit distance between the most likely phone sequence and the surface transcription in terms of %correct and %accuracy.

- **Method III**: Normalized likelihood error. A value of “0” for this measure indicates the best achievable performance.
Performances of Method I and II

Tabulation of Method I and Method II of evaluation for HTK 3.4 and Sphinx 3

<table>
<thead>
<tr>
<th>Decoder models</th>
<th># of Unique variants</th>
<th>HTK 3.4</th>
<th></th>
<th></th>
<th></th>
<th>Sphinx3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Method I</td>
<td>Method II</td>
<td>%Corr</td>
<td>%Acc</td>
<td>Method I</td>
<td>Method II</td>
<td>%Corr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Top 1</td>
<td>Top 5</td>
<td></td>
<td></td>
<td>Top 1</td>
<td>Top 5</td>
<td></td>
</tr>
<tr>
<td>SA1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47class</td>
<td>636</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52class</td>
<td>1263</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48class</td>
<td>763</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47class</td>
<td>1026</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52class</td>
<td>1026</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48class</td>
<td>1026</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Department of Electrical Engineering, IIT Bombay
Performance of Method III

Distribution of the likelihood scores across the 60 utterances

48-phone class has average likelihood error closest to zero of the three phone sets.
Articulation Scoring methods:

- Articulation score indicates the closeness of language learner’s pronunciation with native speaker (of target language) pronunciation.
- Detects phoneme level mispronunciation and extent to which phoneme has been mispronounced.
- Algorithm uses speech models derived from speech database of native speakers.
- Uses forced align tools in the background to get acoustic scores (quantitative measure indicating acoustic fit for that particular speech segment).

- Two methods investigated:
  - GOP (Goodness of Pronunciation) score [2].
  - Method by Sunil K. Gupta [9].
GOP scoring method:

- Confidence with which particular phone has been recognized.
- Also called as Goodness Of Pronunciation (GOP) score.
- GOP score is given by normalized log posterior probability

\[
GOP(p) = \log \left( P\left( p \mid O^{(p)} \right) \right) / NF(p)
\]

\[
GOP(p) = \log \left( \frac{P\left( O^{(p)} \mid p \right) P(p)}{\max_{q\in Q} P\left( O^{(p)} \mid q \right) P(q)} \right) / NF(p)
\]
GOP scoring method (cont.):

\[
GOP(p) = \frac{\log \left( P\left( O^{(p)} \mid p \right) \right)}{NF(p)} - \frac{\log \left( \max_{q\in Q} P\left( O^{(p)} \mid q \right) \right)}{NF(p)} = l_p - l_g
\]

Block diagram for Articulation scoring
Method by Sunil K. Gupta:

- Shortcoming of GOP score:
  - Threshold selection was based on subjective rating of human judges.
  - Not providing any quantitative measure to measure extent of mispronunciation.
  - Free decoder not accurate enough leading to alignment errors.

- In this method two speech models have to be prepared:
  - 48 class phone models (36 TIFR Hindi + 12 TIMIT English)
  - Garbage model (all phonemes of speech data combined to get one speech model)
Garbage model:

- A single speech model combining all the phonemes of speech data.
- Entire speech corpus trained with garbage transcription.
Methodology (cont.) :

- Utterance is force aligned using Sphinx3_align with the reference transcription using 48 class phone models.
  
  - Each phoneme of the transcription will have its own acoustic score.
  
  - These log-likelihood scores are duration normalized given by
    \[ l_q \text{ for } q_i \]

- Similarly, utterance is force aligned with garbage transcription using Garbage model.
  
  \[ l_g \text{ for } q_i \]

- Difference between these two likelihood is current phoneme likelihood
  \[ d = l_q - l_g \text{ for } q_i \]

- This difference score \((d)\) is used for coming up with phone articulation score using lookup score table.(explained in next slide)
Formation of score table:

- For each utterance “In-grammar” and “Out-grammar” is formed
  - In-grammar: When the transcription is conforming to target acoustic waveform.
  - Out-grammar: Transcription selected is some random phrase from training database not conforming to target acoustic waveform.

- In-grammar and Out-grammar transcriptions are force aligned to come up with log-likelihood scores:
  - In-grammar: $l^i_q \text{ and } l^i_g \text{ for } q_i$.
    $$d^i = l^i_q - l^i_g$$
  - Out-grammar: $l^o_q \text{ and } l^o_g \text{ for } q_i$.
    $$d^o = l^o_q - l^o_g$$
Score table (cont.):

- Using all the in-grammar points and out-grammar points, pdf is formed for each phoneme.

\[ f^i = N(\mu^i, \sigma^i) \]
\[ f^o = N(\mu^o, \sigma^o) \]

- Using these probability density functions are used for coming up with score table. (table shown in results section)
Results:

- Histograms and Gaussian pdf (approximating data points) for both In-grammar and Out-grammar for phoneme “aa”:
Results (cont.):

- Histograms and Gaussian pdf (approximating data points) for both In-grammar and Out-grammar for phoneme “ee”:

![Histograms and Gaussian pdf](image)
Results (cont.):

- Combined PDF of In-grammar and Out-grammar for “aa”:
Results (cont.):

- Combined PDF of In-grammar and Out-grammar for “ee”:
Results (score table):

- Below calculations and table is for phoneme “aa”:

- \( f \) denotes probability density function.

\[ h(x) = \log f^i(x) - \log f^o(x) \]

- \( \mu^i \) and \( \mu^o \) are In-grammar and Out-grammar mean respectively.

- For In-grammar and Out-grammar points:

\[ h(\mu^i) = \log f^i(\mu^i) - \log f^o(\mu^i) = 1.242 \]
\[ h(\mu^o) = \log f^i(\mu^o) - \log f^o(\mu^o) = -0.272 \]

- Score table for phoneme “aa” in next slide:
Score table (phoneme “aa”):

<table>
<thead>
<tr>
<th>D</th>
<th>$h(x)$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(\mu^i)$</td>
<td>1.242</td>
<td>100</td>
</tr>
<tr>
<td>[ \log(90/10) \log(10) h(\mu^i) ]</td>
<td>1.185</td>
<td>90</td>
</tr>
<tr>
<td>[ \log(80/20) \log(10) h(\mu^i) ]</td>
<td>0.748</td>
<td>80</td>
</tr>
<tr>
<td>[ \log(70/30) \log(10) h(\mu^i) ]</td>
<td>0.457</td>
<td>70</td>
</tr>
<tr>
<td>[ \log(60/40) \log(10) h(\mu^i) ]</td>
<td>0.219</td>
<td>60</td>
</tr>
<tr>
<td>[ \log(50/50) \log(10) h(\mu^i) = 0 ]</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>
Score table (phoneme “aa”):

<table>
<thead>
<tr>
<th>D</th>
<th>$h(x)$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log(50/50)/\log(10) h(\mu^o) = 0$</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>$\log(90/10)/\log(10) h(\mu^o)$</td>
<td>-0.048</td>
<td>40</td>
</tr>
<tr>
<td>$\log(80/20)/\log(10) h(\mu^o)$</td>
<td>-0.1001</td>
<td>30</td>
</tr>
<tr>
<td>$\log(70/30)/\log(10) h(\mu^o)$</td>
<td>-0.164</td>
<td>20</td>
</tr>
<tr>
<td>$\log(60/40)/\log(10) h(\mu^o)$</td>
<td>-0.259</td>
<td>10</td>
</tr>
<tr>
<td>$h(\mu^o)$</td>
<td>-0.272</td>
<td>0</td>
</tr>
</tbody>
</table>
Result (Speaker 1 : Fundamentals) :

<table>
<thead>
<tr>
<th>Phone</th>
<th>Correct Pronun.</th>
<th>Incorrect Pronun.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d Score</td>
<td>d Score</td>
</tr>
<tr>
<td>h</td>
<td>-8647</td>
<td>-90232</td>
</tr>
<tr>
<td>aa</td>
<td>-14397 60%</td>
<td>-2990 80%</td>
</tr>
<tr>
<td>n</td>
<td>-99.5</td>
<td>-8224</td>
</tr>
<tr>
<td>d</td>
<td>-23238</td>
<td>-27363</td>
</tr>
<tr>
<td>aa</td>
<td>-15280 60%</td>
<td>-40304 0%</td>
</tr>
<tr>
<td>m</td>
<td>-756</td>
<td>-813</td>
</tr>
<tr>
<td>ee</td>
<td>-19837</td>
<td>-19767</td>
</tr>
<tr>
<td>n’</td>
<td>-7571</td>
<td>-5941</td>
</tr>
<tr>
<td>SI</td>
<td>-12570</td>
<td>-5250</td>
</tr>
<tr>
<td>t</td>
<td>-2023</td>
<td>-6451</td>
</tr>
<tr>
<td>aa</td>
<td>-8659.5 80%</td>
<td>-10531 70%</td>
</tr>
<tr>
<td>l</td>
<td>-10920</td>
<td>-2014</td>
</tr>
</tbody>
</table>

Department of Electrical Engineering, IIT Bombay
### Result (Speaker 2 : Fundamentals):

<table>
<thead>
<tr>
<th>Phone</th>
<th>Correct Pronun.</th>
<th>Incorrect Pronun.</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>Score</td>
<td>d</td>
</tr>
<tr>
<td>h</td>
<td>370</td>
<td>-9969</td>
</tr>
<tr>
<td>aa</td>
<td>-10269 70%</td>
<td>-10499 70%</td>
</tr>
<tr>
<td>n</td>
<td>-9675</td>
<td>-4115</td>
</tr>
<tr>
<td>d</td>
<td>-25162</td>
<td>-17556</td>
</tr>
<tr>
<td>aa</td>
<td>-2295 80%</td>
<td>-46778 0%</td>
</tr>
<tr>
<td>m</td>
<td>-4100</td>
<td>-11144</td>
</tr>
<tr>
<td>ee</td>
<td>-24043</td>
<td>-34685</td>
</tr>
<tr>
<td>n’</td>
<td>4284</td>
<td>-5253</td>
</tr>
<tr>
<td>SI</td>
<td>-4595</td>
<td>-10971</td>
</tr>
<tr>
<td>t</td>
<td>-10534</td>
<td>-5146</td>
</tr>
<tr>
<td>aa</td>
<td>-14271 50%</td>
<td>-7271 80%</td>
</tr>
<tr>
<td>l</td>
<td>-10917</td>
<td>-18839</td>
</tr>
</tbody>
</table>
Duration scoring:

- Duration score provides feedback on normalized relative duration difference between language learner speech and reference speaker speech.
- Denotes whether a particular syllable is stressed or not.
- If $L_i$ and $R_i$ are their respective durations corresponding to phoneme $q_i$ then, utterance consisting of $N$ phones can be denoted by:
  
  $L = (L_1, L_2, \ldots \ldots L_N)$ for language learner speech
  
  $R = (R_1, R_2, \ldots \ldots R_N)$ for reference speaker speech

- Normalized durations given by:
  
  $\hat{L}_i = \frac{L_i}{\sum_{i=1}^{N} L_i}$ and $\hat{R}_i = \frac{R_i}{\sum_{i=1}^{N} R_i}$
Duration scoring (cont.):

- Overall duration score given by:

\[ D = \max \left\{ 0, 1 - \sum_{i=1}^{N} \left| L_i - R_i \right| \right\} \]

- Maximum duration score is ‘1’ and minimum is ‘0’.
Duration scoring (Results):

- Speaker_1 was taken as reference and duration scores were calculated for other speakers.

- Speaker_1

---

Department of Electrical Engineering, IIT Bombay
Duration scoring (Results):

- Speaker_1 🎤 Vs Speaker_2 🎤
- Duration score = 0.573 (low due to differences in ‘f’, ’a’ and ‘s’ duration)
Duration scoring (Results):

- **Speaker_1** 🗣️ **Vs** **Speaker_3** 🗣️
- **Duration score = 0.485** (low due to differences in ‘f’, ’a’ and ‘s’ duration)
Feedback, Articulation and Duration Score

Speaker 1
Canonical Transcription (Reference speaker)
SIL f aa n d aa m ee n’ clt t aa l s SIL

Speaker 2
Transcription
SIL f aa n d ee m ee n’ clt t aa l s SIL

Articulation Score: 72% Duration Score: 0.573

Feedback
SIL f aa n d ee m ee n’ clt t aa l s SIL
References


