Speech Processing

Using Speech with Computers
Overview

- **Speech vs Text**
  - Same but different

- **Core Speech Technologies**
  - Speech Recognition
  - Speech Synthesis
  - Dialog Systems
  - Other Speech Processing
The vocal tract

alveolar ridge

hard palate

soft palate

tongue tip

uvula

tongue blade

pharynx

tongue root

epiglottis

tongue body

larynx
From meat to voice

- **Blow air through lungs**
  - Vibrate larynx
  - Vocal tract shape defines resonance
  - Obstructions modify sound
    - Tongue, teeth, lips, velum (nasal passage)
The ear
Sound waves

- Vibrate ear drum
- Cause fluid in cochlear to vibrate
- Spiral cochlear
  - Vibrate hairs inside cochlear
  - Different frequencies vibrate different hairs
  - Converts time domain to frequency domain
Phonemes

- Defined as fundamental units of speech
  - If you change it, it (can) change the meaning

  “pat” to “bat”
  “pat” to “pam”
Vowel Space

- One or two banded frequencies (formants)
# English (US) Vowels

<table>
<thead>
<tr>
<th>AA</th>
<th>wAshington</th>
<th>AE</th>
<th>fAt, bAd</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH</td>
<td>bUt, hUsh</td>
<td>AO</td>
<td>IAWn, mAll</td>
</tr>
<tr>
<td>AW</td>
<td>hOW, sOUth</td>
<td>AX</td>
<td>About, cAnoe</td>
</tr>
<tr>
<td>AY</td>
<td>hlde, bUY</td>
<td>EH</td>
<td>gEt, fEAther</td>
</tr>
<tr>
<td>ER</td>
<td>makER, sEARch</td>
<td>EY</td>
<td>gAte, Elght</td>
</tr>
<tr>
<td>IH</td>
<td>blt, shlp</td>
<td>IY</td>
<td>bEAt, shEEp</td>
</tr>
<tr>
<td>OW</td>
<td>lOne, nOse</td>
<td>OY</td>
<td>tOY, OYster</td>
</tr>
<tr>
<td>UH</td>
<td>fUll</td>
<td>UW</td>
<td>fOOl</td>
</tr>
</tbody>
</table>
English Consonants

- **Stops:** P, B, T, D, K, G
- **Fricatives:** F, V, HH, S, Z, SH, ZH
- **Affricatives:** CH, JH
- **Nasals:** N, M, NG
- **Glides:** L, R, Y, W

**Note:** voiced vs unvoiced:
- P vs B, F vs V
Not all variation is Phonetic

- Phonology: linguistically discrete units
  - May be a number of different ways to say them
  - /r/ trill (Scottish or Spanish) vs US way

- Phonetics vs Phonemics
  - Phonetics: discrete units
  - Phonemics: all sounds

- /t/ in US English: becomes “flap”
  - “water” / w ao t er /
  - “water” / w ao dx er /
Dialect and Idiolect

- **Variation within language (and speakers)**

- **Phonetic**
  - “Don” vs “Dawn”, “Cot” vs “Caught”
  - R deletion (Haavaad vs Harvard)

- **Word choice:**
  - Y’all, Yins
  - Politeness levels
Not all languages are the same

- Asperated stops (Korean, Hindi)
  - \( P \) vs \( PH \)
  - English uses both, but doesn’t care
  - \( \text{Pot} \) vs \( \text{sPot} \) (place hand over mouth)

- L-R in Japanese not phonological

- US English dialects:
  - Mary, Merry, Marry

- Scottish English vs US English
  - No distinction between “pull” and “pool”
  - Distinction between: “for” and “four”
Different language dimensions

- **Vowel length**
  - *Bit vs beat*
  - *Japanese: shujin (husband) vs shuujin (prisoner)*

- **Tones**
  - *F0 (tune) used phonetically*
  - *Chinese, Thai, Burmese*

- **Clicks**
  - *Xhosa*
Prosody

- **Intonation**
  - *Tune*

- **Duration**
  - *How long/short of each phoneme*

- **Phrasing**
  - *Where the breaks are*

- **Used for:**
  - *Style, emphasis, confidence etc*
Intonation Contour
Intonation Information

- Large pitch range (female)
- Authoritive since goes down at the end
  - News reader
- Emphasis for Finance H*
- Final has a raise – more information to come

- Female American newsreader from WBUR
- (Boston University Radio)
Words

- The things with space around them (sort of)
- Chinese, Thai, Japanese doesn’t use spaces

Words aren’t always what they seem

- Can you pass the salt?
- Boston. Boston! Boston?
- Yeah, right

Multiple ways to say the same thing:

- I want to go to Boston.
- Yes
Speech Recognition

- **Two major components**
  - Acoustic Models
  - Language Models

- **Accuracy various with**
  - Speaker, language, dialect
  - Microphone type, environment
  - Speaking style:
    - Good Recognition:
      - Head mounted mike, controlled language, careful speaker
    - Not so good recognition:
      - Remote mike, chatting between friends, in open cafe
But not just acoustics

- But not all phones are equi-probable
- Find word sequences that maximizes $P(W \mid O)$
- Using Bayes’ Law
  \[ \frac{P(W)P(O \mid W)}{P(O)} \]
- Combine models
  - Use HMMs to provide $P(O \mid W)$
  - Use language model to provide $P(W)$
Speech Synthesis

- **Three Levels**
  - **Text analysis**
    - From characters to words
  - **Prosody and Pronunciation**
    - From words to phonemes and intonation
  - **Waveform generation**
    - From phonemes to waveforms
• This is a pen.
• My cat who lives dangerously has nine lives.
• He stole $100 from the bank.
• He stole 1996 cattle on 25 Nov 1996.
• He stole $100 million from the bank.
• It's 13 St. Andrew St. near the bank.
• Its a PIII 1.5Ghz, 512MB RAM, 160Gb SATA, (no IDE) 24x cdrom and 19" LCD.
• My home pgae is http://www.geocities.com/awb/. 
Waveform Generation

- Formant synthesis
- Random word/phrase concatenation
- Phone concatenation
- Diphone concatenation
- Sub-word unit selection
- Cluster based unit selection
- Statistical Parametric Synthesis
- Wavenet Neural Synthesis
**Pronunciation Lexicon**

- **List of words and their pronunciation**
  - (“pencil” n (p eh1 n s ih l))
  - (“table” n (t ey1 b ax l))
- **Need the right phoneme set**
- **Need other information**
  - Part of speech
  - Lexical stress
  - Other information (Tone, Lexical accent …)
  - Syllable boundaries
Homograph Representation

- **Must distinguish different pronunciations**
  - (“project” n (p r aa1 jh eh k t))
  - (“project” v (p r ax jh eh1 k t))
  - (“bass” n_music (b ey1 s))
  - (“bass” n_fish (b ae1 s))

- **ASR multiple pronunciations**
  - (“route” n (r uw t))
  - (“route(2)” n (r aw t))
Pronunciation of Unknown Words

- How do you pronounce new words
- 4% of tokens (in news) are new
- You can’t synthesis them without pronunciations
- You can’t recognize them without pronunciations
- Letter-to-Sounds rules
- Grapheme-to-Phoneme rules
Hand written rules

• \([\text{LeftContext}] \, X \, [\text{RightContext}] \rightarrow Y\)
• e.g.
• \(c \, [h \, r] \rightarrow k\)
• \(c \, [h] \rightarrow ch\)
• \(c \, [i] \rightarrow s\)
• \(c \rightarrow k\)
• Need an existing lexicon
  • Pronunciations: words and phones
  • But different number of letters and phones

• Need an alignment
  • Between letters and phones
  • checked -> ch eh k t
LTS: alignment

- **checked** -&gt; **ch eh k t**

<table>
<thead>
<tr>
<th>c</th>
<th>h</th>
<th>e</th>
<th>c</th>
<th>k</th>
<th>e</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch</td>
<td>_</td>
<td>eh</td>
<td>k</td>
<td>_</td>
<td>_</td>
<td>t</td>
</tr>
</tbody>
</table>

- **Some letters go to nothing**

- **Some letters go to two phones**
  - **box** -&gt; **b aa k-s**
  - **table** -&gt; **t ey b ax-l -**
Find alignment automatically

- **Epsilon scattering**
  - Find all possible alignments
  - Estimate $p(L,P)$ on each alignment
  - Find most probable alignment

- **Hand seed**
  - Hand specify allowable pairs
  - Estimate $p(L,P)$ on each possible alignment
  - Find most probable alignment

- **Statistical Machine Translation (IBM model 1)**
  - Estimate $p(L,P)$ on each possible alignment
  - Find most probable alignment
Not everything aligns

- **0, 1, and 2 letter cases**
  - e -> epsilon "moved"
  - x -> k-s, g-z "box" "example"
  - e -> y-uw "askew"

- **Some alignments aren’t sensible**
  - dept -> d ih p aa r t m ax n t
  - cmu -> s iy eh m y uw
Training LTS models

- **Use CART trees**
  - One model for each letter
- **Predict phone (epsilon, phone, dual phone)**
  - From letter 3-context (and POS)
- # # # c h e c c -> ch
- # # c h e c k k -> _
- # c h e c k e e -> eh
- c h e c k e d -> k
## LTS results

- **Split lexicon into train/test 90%/10%**
  - i.e. every tenth entry is extracted for testing

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Letter Acc</th>
<th>Word Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD</td>
<td>95.80%</td>
<td>75.56%</td>
</tr>
<tr>
<td>CMUDICT</td>
<td>91.99%</td>
<td>57.80%</td>
</tr>
<tr>
<td>BRULEX</td>
<td>99.00%</td>
<td>93.03%</td>
</tr>
<tr>
<td>DE-CELEX</td>
<td>98.79%</td>
<td>89.38%</td>
</tr>
<tr>
<td>Thai</td>
<td>95.60%</td>
<td>68.76%</td>
</tr>
</tbody>
</table>
For letter V:
if (n.name is v)
    return _
    if (n.name is ≠)
        if (p.p.name is t)
            return f
        return v
    if (n.name is s)
        if (p.p.p.name is n)
            return f
        return v
return v
But we need more than phones

- **What about lexical stress**
  - `$p r a a 1 j e h k t$ -> `$p r a a j e h1 k t$`

- **Two possibilities**
  - A separate prediction model
  - Join model – introduce `eh/eh1` (BETTER)

<table>
<thead>
<tr>
<th></th>
<th>LTP+S</th>
<th>LTPS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L no S</strong></td>
<td>96.36%</td>
<td>96.27%</td>
</tr>
<tr>
<td><strong>Letter</strong></td>
<td>---</td>
<td>95.80%</td>
</tr>
<tr>
<td><strong>W no S</strong></td>
<td>76.92%</td>
<td>74.69%</td>
</tr>
<tr>
<td><strong>Word</strong></td>
<td>63.68%</td>
<td>74.56%</td>
</tr>
</tbody>
</table>
Does it really work

- 40K words from Time Magazine
  - 1775 (4.6%) not in OALD
  - LTS gets 70% correct (test set was 74%)

<table>
<thead>
<tr>
<th></th>
<th>Occurs</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names</td>
<td>1360</td>
<td>76.6</td>
</tr>
<tr>
<td>Unknown</td>
<td>351</td>
<td>19.8</td>
</tr>
<tr>
<td>US Spelling</td>
<td>57</td>
<td>3.2</td>
</tr>
<tr>
<td>Typos</td>
<td>7</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Speech Synthesis Techniques

- **Unit selection**
- **Statistical parameter synthesis**
- **Automated voice building**
  - *Database design*
  - *Language portability*
- **Voice conversion**
• Target cost and Join cost [Hunt and Black 96]
  – Target cost is distance from desired unit to actual unit in the databases
    • Based on phonetic, prosodic metrical context
  – Join cost is how well the selected units join
Clustering Units

- Cluster units [Donovan et al 96, Black et al 97]

\[
Adist(U, V) = \begin{cases} 
\frac{WD*|U|}{|V|} * \sum_{i=1}^{n} \sum_{j=1}^{\frac{|U|}{|V|}} W_j \cdot (abs(F_{ij}(U) - F_{i*|V|/|U|j}(V))) \\
Adist(V, U) 
\end{cases} 
\]

- \( |U| \) = number of frames in \( U \)
- \( F_{xy}(U) \) = parameter \( y \) of frame \( x \) of unit \( U \)
- \( SD_j \) = standard deviation of parameter \( j \)
- \( W_j \) = weight for parameter \( j \)
- \( WD \) = duration penalty
Unit Selection Issues

- Cost metrics
  - Finding best weights, best techniques etc
- Database design
  - Best database coverage
- Automatic labeling accuracy
  - Finding errors/confidence
- Limited domain:
  - Target the databases to a particular application
  - Talking clocks
  - Targeted domain synthesis
Old vs New

Unit Selection:
- large carefully labelled database
- quality good when good examples available
- quality will sometimes be bad
- no control of prosody

Parametric Synthesis:
- smaller less carefully labelled database
- quality consistent
- resynthesis requires vocoder, (buzzy)
- can (must) control prosody
- model size much smaller than Unit DB
Parametric Synthesis

• Probabilistic Models
  \[ \text{argmax}(P(O|W)) \]

• Simplification
  \[ \text{argmax}(P(o_0|W), P(o_1|W), \ldots, P(o_n|W)) \]

• Generative model
  – Predict acoustic frames from text