Speech Processing 11-492/18-492

Speech Synthesis
Waveform generation
Speech Synthesis

- **Text Analysis**
  - Chunking, tokenization, token expansion

- **Linguistic Analysis**
  - Pronunciations
  - Prosody

- **Waveform generation**
  - From phones and prosody to waveforms
Physical Models

• **Blowing air through tubes…**
  – von Kempen’s synthesizer 1791

• **Synthesis by physical models**
  – Homer Dudley’s Voder. 1939
More Computation – More Data

- **Formant synthesis (60s-80s)**
  - Waveform construction from components

- **Diphone synthesis (80s-90s)**
  - Waveform by concatenation of small number of instances of speech

- **Unit selection (90s-00s)**
  - Waveform by concatenation of very large number of instances of speech

- **Statistical Parametric Synthesis (00s-..)**
  - Waveform construction from parametric models
Waveform Generation

- Formant synthesis
- Random word/phrase concatenation
- Phone concatenation
- Diphone concatenation
- Sub-word unit selection
- Cluster based unit selection
- Statistical Parametric Synthesis
Concatenative Synthesis

- Use human speech
- Need to design database
- Need to carefully label it
- Need to impose prosody on selections
- Results depend on DB contents
  - You get good synthesis
  - But style is like the databases
Diphone Synthesis

- Use databases of natural speech
- From mid-phone to mid-phone
  - Requires phones squared – diphones
- Needs very good definition of phoneset
  - Dialect of speaker becomes important
Diphone Databases

Collect nonsense carrier words

- $t\text{ aa } b\text{ aa } b\text{ aa}$
- $t\text{ aa } m\text{ aa } m\text{ aa}$
- $t\text{ iy } b\text{ iy } b\text{ iy}$
- Good for coverage, consistent
- Not very natural

Collect from “natural” words

- Quebecois arguments (19)
- Arkansas arranging (11)
- Good for naturalness, but maybe not consistent
Recording Databases

- Do recording in best conditions possible
  - Recording studio
  - Head mounted mike
  - Repeatable conditions

- Explain to voice talent
  - Get *signed* permission
  - You are going to steal their voice!
**Diphone Limitations**

- **Only get fixed inventory**
  - Need more than phone-phone
  - Need stressed, positional examples
  - What about consonant clusters?
- **Get more representative samples**
  - Larger databases
  - More natural
  - Harder to ensure it is correct
Database Design

- **Requires:**
  - Good phonetic coverage
  - Good prosodic coverage
  - Easy to read sentences (few mistakes)
  - Consistent delivery
Database Design

- From large databases of text
  - E.g. out-of-copyright books

- Find “nice” sentences
  - Contain only high frequency, easy to pronounce words
  - 5-15 words long
  - No homographs

- Greedily select “nice” sentences with
  - Best phone/diphone/triphone coverage
  - Best characters/dicharacter/tricharacter coverage

- Consider multiple genres
  - Novels, news, bus stops (domain dependent)
CMU ARCTIC Databases

- **1132 sentences (about an hour of speech)**
  - *Author of the danger trail, Phillip Steels etc.*
  - *
- **12 different speakers**
  - *Different English accents*

- **Technique used for other languages**
Select appropriate sub-word units from databases of natural speech

Not simply word concatenation

Not simply longest phrase

Balance

- Appropriate unit
- Good join costs
• **Target cost / 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]

\[
A\text{dist}(U, V) = \begin{cases} 
\text{if } |V| > |U| & A\text{dist}(V, U) \\
\frac{W_D \cdot |U|}{|V|} \cdot \frac{n}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{W_j \cdot (\text{abs}(F_{ij}(U) - F_{ij(\text{abs}|V|/|U|))(V)))}{SD_j \cdot n \cdot |U|} 
\end{cases}
\]

|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
Unit Selection vs Parametric

**Unit Selection**
- The “standard” method
  “Select appropriate sub-word units from large databases of natural speech”

**Parametric Synthesis: [NITECH: Tokuda et al]**
- HMM-generation based synthesis
- Cluster units to form models
- Generate from the models
  “Take ‘average’ of units”
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 Sel DB
Example CG Voices

7 Arctic databases:

1200 utterances, 43K segs, 1hr speech

awb  bdl
clb  jmk
ksp  rms
slt
## Database size vs. Quality

### slt_arctic data size

<table>
<thead>
<tr>
<th>Utts</th>
<th>Clusters</th>
<th>RMS F0</th>
<th>MCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>230</td>
<td>24.29</td>
<td>6.761</td>
</tr>
<tr>
<td>100</td>
<td>435</td>
<td>19.47</td>
<td>6.278</td>
</tr>
<tr>
<td>200</td>
<td>824</td>
<td>17.41</td>
<td>6.047</td>
</tr>
<tr>
<td>500</td>
<td>2227</td>
<td>15.02</td>
<td>5.755</td>
</tr>
<tr>
<td>1100</td>
<td>4597</td>
<td>14.55</td>
<td>5.685</td>
</tr>
</tbody>
</table>
Database size vs. Quality

- **SPS**
  - $rms_{100}$
  - $rms_{1132}$

- **Unit selection**
  - $rms_{100}$
  - $rms_{1132}$
Advantages of SPS

- **Statistical Parameter Synthesis**
  - More robust to errors in data
  - Requires less data
  - Models are smaller (< 2MB vs > 1GB)
  - Parametric models allows further processing
Disadvantages of SPS

- Statistical Parametric Synthesis
  - “buzziness” of resynthesized speech
  - Doesn’t sound as good as the best unit selection
  - Still experimental