Speech Processing 11-492/18-492

Speech Synthesis
Signal Processing
Signal Manipulation

- **Signal Parameterization**
  - Joining
  - LPC
  - PSOLA: pitch and duration modification

- **Statistical Parameterization**
  - MELCEP/MLSA
  - LSF, STRAIGHT, HNM, HSM
TTS Signal Processing

- Join together pieces of speech
- Prosodic modification
  - Pitch (F0)
  - Duration
  - Power
- Change spectral properties
  - Stress/unstress
  - Spectral tilt
  - Speaking style
Joining

- **Just put them together**
  - Gets clicks at join points
- **Join them at zero crossings**
- **Window them and overlap them**
  - WSOLA
- **Join them at pitch periods**
Prosodic Modification

- **Modify pitch and duration** independently
- **Changing sample rate** changes both
  - “chipmunk” style speech
- **Duration**
  - Duplicate/delete parts of the signal
- **Pitch**
  - “resample” to change pitch
Speech and Short Term Signals
Duration Modification
Pitch Modification
Modify pitch and duration

- Find ideal pitch periods and duration
- Find closest actual periods from units
- End with
  - Pitch period (short term signals)
  - Distances between them
Signal Reconstruction

- **TD-PSOLA™**
  - Time domain pitch synchronous overlap and add
- **Patented by France Telecom**
  - Expired 2004
- **Very efficient:**
  - No FFT (or inverse FFT)
- **Can modify Hz * 2.0 (or 0.5)**
- **The reason no one publishes algorithms**
- **The (partial) reason unit selection typically doesn’t do pitch/duration modification**
LPC: Linear predictive coding

- Linear predictive coding
  - Predict next sample point from previous
  - Weighted sum of previous points
  - Filter of order p.

\[ s_n = \sum_{i=1}^{p} a_i s_{n-i} \]

- Residual excited LPC

\[ s_n = \sum_{i=1}^{p} a_i s_{n-i} + r_n \]
LPC

- Works well but can be buzzy
- Can be very compact
- Can be pitch synchronous
- Excited
  - Pulse
  - Triangular pulse
  - Multi-pulse
  - Full residual
- Used in standard speech coding
  - LPC10: 2.4kps
  - CELP: codebook excited LPC
Other Parametric Representations

- Typically split spectral and residual
- **MBROLA:**
  - Multi-band overlap and add
- **HNM/HSM:**
  - Harmonic plus (noise/stochastic) modeling
- **STRAIGHT**
- **MELCEP/MLSA**
  - Often used in HMM synthesis
- **Sinusoidal (HARMONIC)**
- Wavelet
- LSF/LPC
Predict the time domain signal directly

Deepmind’s Wavenet (van den Oord et al 2016)

Cf of PixelRNN and PixelCNN models
- Predict sequences of quantized PCM
- 16,000 times a second
- Sort of unit selection at the very very local signal level
- Has a strong “Language Model” (it can “babble”)
- Similar quality to unit selection
- Some properties of SPSS though
- Very very expensive to train
- Expensive to run (or maybe not any more)
Choosing the right unit type

- **Diphones**
  - Phone-phone
  - Joins at stable portions, not transitions
- **Half phone (AT&T Natural Voices)**
- **Hybrid systems (Hadifix – Bonn systems)**
- **Other selection systems:**
  - Syllable, phone, HMM state
  - Even frame level
Acoustically Derived Units

- E.g Bacchiani 99 or Rita Singh CMU
- From some waveforms
  - Find N most diverse unit types
  - Varied in length
- Still need to map letters to units
Acoustic Phonetic Clustering

- Parameterize database
  - Melcep plus power
- K-means
  - Euclidean distance measure
  - 100 clusters
- Label DB with best cluster
- Build clunits synthesizer
  - Can’t predict APC cluster directly
  - Use held out data for testing
Acoustic Phonetic Clustering
Grapheme Based Synthesis

- **Synthesis without a phoneme set**
  - “End-to-End” synthesis
- **Use the letters as phonemes**
  - (“alan” nil (a l a n))
  - (“black” nil (b l a c k))
- **Spanish (easier ?)**
  - 419 utterances
  - HMM training to label databases
  - Simple pronunciation rules
  - Polici’a -> p o l i c i’ a
  - Cuatro -> c u a t r o
<table>
<thead>
<tr>
<th>Word</th>
<th>Castillian</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>casa</td>
<td>/k a s a/</td>
<td>house</td>
</tr>
<tr>
<td>cesa</td>
<td>/th e s a/</td>
<td>stop</td>
</tr>
<tr>
<td>cine</td>
<td>/th i n e/</td>
<td>cinema</td>
</tr>
<tr>
<td>cosa</td>
<td>/k o s a/</td>
<td>thing</td>
</tr>
<tr>
<td>cuna</td>
<td>/k u n a/</td>
<td>cradle</td>
</tr>
<tr>
<td>hechizo</td>
<td>/e ch i th o/</td>
<td>charm, spell</td>
</tr>
</tbody>
</table>

In Spanish the letter “c” may be pronounced /k/, /ch/ and /th/ or /s/ (depending on dialect). The choice of phone is determined by the letter context.
English Grapheme Synthesis

- Use Letters are phones
- 26 “phonemes”
  - ( “alan” n (a l a n))
  - ( “black” n (b l a c k))
- Build HMM acoustic models for labeling
- For English
  - “This is a pen”
  - “We went to the church at Christmas”
  - Festival intro
  - “do eight meat”
- Requires method to fix errors
  - Letter to letter mapping
Signal Processing for TTS

- Pitch and duration modification
- LPC
- Finding the right unit type
- Grapheme-based Synthesis
HW2: TTS

- **Due 3:30pm Mon October 16th and 23rd**
  - Like the website says
- **Install Festival and Festvox**
- **Find 10 errors in each of two different synthesizers**
- **Build a voice**
  - A Talking Clock
  - A general voice
  - (or both)