Speech Processing 11-492/18-492

Speech Synthesis
Prosody
Speech Synthesis

- **Linguistic Analysis**
  - Pronunciations
  - Prosody
Prosody

- How the phonemes will be said
- Four aspects of prosody
  - Phrasing: where the breaks will be
  - Intonation: pitch accents and F0 generation
  - Duration: how long the phonemes will be
  - Power: energy in signal
Phrase Breaks

- Need to take a breath
- Need to chunk relevant parts together
  - Sub-sentential
  - Supra-word
- First approximation
  - At punctuation (comma, semicolon, etc.)
  - Too little
- Second approximation
  - At each (or some) of the content/function words
  - Too much
Next week, some inmates released early from the Hampton County jail in Springfield, will be wearing a wristband that hooks up to a special jack on their home phones.
Phrasing

- **Bachenko and Fitzpatrick 90**
  - Rule driven with punctuation, POS and syntax
  - Balanced phrasing
  - (the boy saw) (the girl in the park)
  - (the boy in the park) (saw the girl)

- **Hirschberg and Prieto 94**
  - CART trees (similar features)

- **Ostendorf and Veilleux 94**
  - Hierarchical statistical model
  - Multilevel breaks
Balance length of phrases

- Predict probability of break with CART (use POS)
- Use n-gram of B/NB to keep balance

\[
\prod_{k=1}^{n} \frac{P(B_k \mid B_{k-1}, \ldots, B_{k-N+1})P(T_{k-N,\ldots,k+1} \mid B_k)}{P(T_{k-N,\ldots,k+1})}
\]

Trained on BBC Radio 4 (NPR-like)

- 31,707 words, 6,346 breaks
- 91% correct with 6-gram
- Still makes errors – especially around “I”
Phrasing

- **What is correct?**
  - Lots of answers are correct.
  - But some are definitely bad.

- **Ostendorf and Vielleux 94**
  - Multiple people read same paragraphs
  - If your method matches any single person’s version it is correct.
The fundamental tune

- Accents (highlighting important parts)
- F0 generation (the tune itself)
Intonation Contour
Intonation Information

- Large pitch range (female)
- Authoritative 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 Public Radio)
Intonation Examples

- Fixed durations, flat F0.
- Declining F0
- “hat” accents on stressed syllables
- Accents and end tones
- Statistically trained
Accents and Boundaries

- Where are the important changes in F0?

Accents on syllables

- Identifies “important” words
  - It will be RAINY today in Boston
  - It will be rainy TODAY in Boston
  - It will BE rainy today IN Boston (strange)
Where do the accents go?

- On important words
- First approximation
  - On stressed syllables in content words
    - It WILL be RAINY TODAY in BOSTON
  - About 80% correct on news reader speech
- CART training on more features
  - Content, proper nouns, POS, position in text
  - (not semantic information)
Tones and Break Indices
- A labeling for intonation (English)

Different accent types
- $H^*$, $!H$, $L^*$, $L+H^*$

Different boundary types
- $L+L^\%$, $L+H^\%$, $H+H^\%$, 
ToBI examples

Marianna made the marmelade.

- H* H* L-L default reading
- H* L-L% emphasis on Marianna
- L+H* L-L% contrastive reading
- L* H-H% incredulous
- L* L* H-H% doubly incredulous
- L+H* L-H% L* H* L-L% (2 intonation phrases)
F0 Generation

- Contour from accents (and durations)
- Piece together shapes of different accents
- Generated
  - By rule
  - Trained from data
Using real contours

- From a data base of different contours
  - Select most appropriate one
- Record lots of different intonation examples
  - He DID then KNOW what HAD occurred
  - TARZAN and JANE raised THEIR heads
  - ...
- Label them and select the contours when you want emphasis
Emphasis Synthesis

- This is a short example
- THIS is a short example
- This IS a short example
- This is A short example
- This is a SHORT example
- This is a short EXAMPLE
Duration Prediction

- Each phone needs a duration
  - Make it 80ms
- Vowels are typically longer than consonants
- Emphasis/accent/stress lengthens them
- Initial and final phones are longer
Prediction Models

- **By rule**
  - Klatt rules

- **By training (using Klatt features)**
  - CART / linear regression
  - Easy to get reasonable durations
  - Hard to get very good durations
Fast and Slow Speech

- Speaking fast: not uniformly shorter durations
  - Have less prosodic breaks
  - Reduce syllables
  - Make consonants shorter
  - Make vowels a little shorter

- Speaking slow: not uniformly longer durations
  - Add more prosodic breaks
  - Small increases in vowel duration (?)
Prosody

- Phrasing
- Intonation
  - Accents + F0 generation
- Duration
- Power