Speech Processing 15-492/18-492

Speech Recognition
Template matching
Speech Recognition by Templates

- A little history …
- Matching Templates
- DTW (Dynamic Time Warping)
- Beyond template matching
Radio Rex (1922)

- Toys always lead technology …
- Call “Rex” and he comes out of his kennel

(Crystalradio.com and Rhys Jones)
Toy ASR “Tricks”

- **Radio Rex**
  - Recognizes vowel formants in “EH”
- **Voice activated toy train**
  - Multilingual stop/go hashire/tomate
- **Toys “pets” don’t need perfect ASR**
Template Matching

- Record templates from user
  - Store in library
- Record ASR example
  - Compare against each library template
- Select closest example
- For example ...
  - On a voice dialing system
Voice Dialing System

- Library
  - Mom
  - Dad
  - Bob
  - Mario’s Pizza

- Let’s Go Bus Information System
Matching in Time Domain

- **Duration**
  - Will discriminate some examples
  - But Mom, Bob and Dad will be confused
- **What about spectral properties**
Matching in Frequency Domain

Mom

Bob
Different deliveries

- *We change durations*
  - Two utterances are never the same
- *When it fails we change our delivery*
  - Become more articular
  - “clearer”
Dynamic Time Warping

Template

Sample Speech
DTW algorithm

*For each square*

- \( \text{Dist}(\text{template}[i], \text{sample}[j]) + \) \( \text{smallest}_{\text{of}} \) \( (\text{Dist}(\text{template}[i-1], \text{sample}[j]) \)
  \( \text{Dist}(\text{template}[i], \text{sample}[j-1]) \)
  \( \text{Dist}(\text{template}[i-1], \text{sample}[j-1]) \)

*Remember which choice your took (count path)*
Multiple Templates

- **Compare against each**
- **Find closest**
- **Need to normalize scores**
  - *(divide by length of matches)*
Matching Templates

Template Library

Sample

Word0
Word1
Word2
...

For Word in Templates
    Score = dtw(Template[Word], Sample);
    if (Score < BestScore)
        BestWord = Word;
    DoAction(Action[BestWord])
DTW issues

- **What happens with no-matches**
  - Need to deal with none of the above

- **What happens with more templates**
  - Harder to choose between
  - Once variance greater than differences

- **Choose templates that are very different**
DTW/Template Applications

- Voice dialer
- Simple command and control
- Speaker ID
For Speaker in Templates
    Score = dtw(Template[Speaker], Sample);
    if (Score < BestScore)
        BestSpeaker = Speaker;
DTW

**Advantages**
- Works well for small number of templates (<20)
- Language independent
- Speaker specific
- Easy to train (end user controls it)

**Disadvantages**
- Limited number of templates
- Speaker specific
- Need actual training examples
More reliable matching

• Distance metric
  – Euclidean \[ \sqrt{\frac{1}{N} \sum_{i=0}^{N} (T_i - S_i)^2} \]

• But some distances are bigger than others
  – Silence is pretty similar
  – Fricatives are quite larger
    • A longer fricative might give large score
    • A longer vowel might give smaller score
More reliable matching

• Having multiple template examples
  – Individual matches or
  – Average them together
• DTW align all of the examples
• Collect statistics as a Gaussian
  – Mean and standard deviation for each coeff

\[ \{ \mu_0, \sigma_0, \mu_1, \sigma_1, \mu_2, \sigma_2, \ldots \} \]
More reliable distances

- Instead of Euclidean distance
  - Doesn’t care about the standard deviation
    \[ \sqrt{\frac{N}{\sum_{i=0}^{N} (T_i - S_i)^2}} \]

- Use Mahalanobis distance
  - Care about means and standard deviation
    \[ \sqrt{\frac{N}{\sum_{i=0}^{N} \left( \frac{\mu_i - S_i}{\sigma_i} \right)^2}} \]
Extending Template matching

- **String word templates together**
  - Need to find word segmentation

- **But there are many words ...**
Extending template model

- **String phoneme templates together**
  - A template model for each phoneme

Sample

- k
- ae
- t

Phoneme Templates

- Phone0
- Phone1
- Phone2
- ...
Summary

- **Speech Recognition by Templates**
  - Good for simple small vocabulary tasks
- **Dynamic Time Warping (DTW)**
  - Can match different durational examples
- **Averaging over multiple models**
- **Distance metrics**
  - Euclidean vs Mahalanobis