Speech and Speaker Recognition

Acoustic Modeling: Other techniques

KTH/CSC: DT2118

Mats Blomberg

Other Techniques

• In addition to standard HMM
  – Neural Networks (Course book)
  – Segment Models (Course book)
  – 2D HMM
  – Bayesian networks
  – Multi-stream
  – Articulatory oriented representation
  – Prosody and duration
  – (Long range dependencies)
  – Eigenvoices
  – Model transformation using speaker characteristic properties
9.8.1 Artificial Neural Networks (ANN)

\[
Y = f \left( \sum_{i=0}^{N} w_i x_i - \theta \right)
\]

Simple Model of a nerve cell

![Simple Model of a nerve cell diagram]

Figure 1: Computation performed in a single node. Three representative nonlinearities are shown.

ANN – an example

Compute phonetic feature activations (~likelihoods) from a filter bank section

- **Output layer**
  Activation of each feature

- **Hidden layer**

- **Input layer**
  Filter amplitudes
9.8.1 Artificial Neural Networks (ANN)

- Good performance for phoneme classification and isolated, small-vocabulary recognition
- Problem
  - Basic neural nets have trouble handling patterns with timing variability (such as speech)
    - Alignment, training, decoding
- Approaches
  - Integration with HMM (Hybrid system)
    - The ANN replaces the Gaussian mixture densities
  - Recurrent neural networks
    - Memory of previous outputs and internal states
  - Time Delay Neural Networks (TDNN)
    - A time sequence of input features and hidden layers

Recurrent network

```
Input history and feedback from the previous hidden and output layers are used to classify the current frame.
```

Figure 9.14 A recurrent network with contextual inputs, hidden vector feedback, and output vector feedback.
9.8.2 Segment Models

- **Problem**
  - The HMM output-independence assumption results in a stationary process (constant mean and variance) in each state
    - Bad model, speech is non-stationary
    - Delta and acceleration features help, but the problem remains
  - Phantom trajectories can occur
    - Trajectories that did not exist in the training data

- **Approach**
  - An interval trajectory rather than a single frame value is matched
  - Parametric Trajectory Models
  - Heavily increased computational complexity
  - Modest improvement over HMM
  - Unified Frame- and Segment-Based Models
Phantom trajectories

Mixture component sequences that never occurred in the same utterance during training, are allowed during recognition.

Standard HMMs allow every frame in an utterance to come from a different speaker.

Can be prevented by Parametric Trajectory Models.

Unified Frame- and Segment-Based Models

- HMM and segment model (SM) approaches are complementary
  - HMM: detailed modeling but quasi-stationary
  - SM: models transitions and longer-range dynamics but coarse detail
- Combine HMM and SM
  \[ p(X \mid \text{Unified model}) = p(X \mid \text{HMM})p(X \mid \text{SM})^a \]
- 8% WER reduction compared to HMM Whisper
- (Method developed by course book co-author)
2-dimensional HMM

• The speech spectrum is viewed as a Markov process (Weber et al, 2000)

[Diagram showing 2-dimensional HMM]

Articulatory Inspired Modeling

• Estimating articulatory features theoretically interesting but difficult
• Phonemes share articulatory features
  – E.g. voiced phonemes share the characteristics of the glottal source spectrum
  – Using the features as basic unit would reduce the number of parameters
  But complex process to combine into phoneme models
• Linear trajectories in the articulatory domain are transformed to non-linearity in the spectral/cepstral domain
  – Should be better to interpolate coarticulation trajectories in the articulatory domain
• Variation in the synchrony between articulators cause large acoustic variability
  – If more than one articulator position is changed in a phoneme boundary, variability in their relative timing will produce different intermediate segments
  – Ex. Transition region in boundary vowel - unvoiced fricative
  – Devoicing before closure: Aspiration Closure before devoicing: voiced fricative
  – Can be modeled by a single delay parameter

[Diagram showing articulatory inspired modeling]
Multi-stream systems

- Dupont, Bourlard (1997)
- Separate decoding for feature subsets
- Accounts for timing asynchrony within the feature vector
  - E.g. between different frequency regions

Recent approaches to articulatory oriented recognition

- Dynamic Bayesian Networks (Livescu, 2005)
  - HMM is a special case of DBN
- Factored Conditional Random Fields (Prabhavalkar et al, 2011)
- Structural training on corpora with speech and articulatory measurements (Neiberg, Ananthakrishnan, Blomberg, 2009)
Bayesian network for modeling articulatory asynchrony

The observed output $Y_t$ is a function of underlying independent Markov chains (~articulators)
Factorised CRF:s can model asynchrony explicitly

Use of prosody and duration

- Carries semantic, stress, and non-linguistic information
  - Several information sources are superimposed
- Not fully synchronized to the articulation
  - Multi-stream technique would help
- Small improvement reported
  - 1% (Chen et al, 2003)
Fast speaker adaptation

- Eigenvoices
- Predict speaker characteristics not observed in the training data

Eigenvoices

- Positions the new speaker in a speaker space
- The dimensions of this space are the mean elements of each training speaker’s models (supervector)
- The dimensionality is reduced by PCA
- Speaker adaptation consists of finding the test speaker’s position in this space
  - Very little adaptation data needed
Speaker characteristics for modeling new speakers

- The speaker property space has fewer dimensions than the acoustic space
  - Smaller search space
  - All positions in this space represent a realistic speaker model
    - Not the case for the spectral space
  - Objective: Find property-specific transforms to move a conventionally trained model to all speaker positions without any adaptation data
    - E.g. VTLN (Vocal Tract Length Normalization) can transform models from a long to a short vocal tract.
    - Other properties with known transformation such as voice source, speaker variability, speech rate, clarity, etc.

Transformation: two search procedures

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Model transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test utterance</td>
<td>Test utterance</td>
</tr>
<tr>
<td>Transformation</td>
<td>Transformation</td>
</tr>
<tr>
<td>Transformed utterance</td>
<td>Transformed model</td>
</tr>
<tr>
<td>Recognition</td>
<td>Recognition</td>
</tr>
<tr>
<td>0.9 1.0 1.1</td>
<td>0.9 1.0 1.1</td>
</tr>
<tr>
<td>8-64 steps</td>
<td></td>
</tr>
</tbody>
</table>
Feature transformation at KTH/TMH

- Speaker properties tested
  - Vocal tract length
  - Voice source spectrum
  - Speaker variability
  - Size of the speaker’s model space (ongoing)
  - Speech rate

- The search space is still too large for exhaustive multi-dimensional search
  - Efficient tree-based technique for fast search developed

- Recognition of children’s speech using models trained by adult speech
- 45% WER reduced to 3% for children’s digit recognition using male adult models (Blomberg & Elenius D., 2009)

Example: Vowel space adaptation

Vowel distribution in F1-F2 for two male, adult speakers

Speech data: Swedish digits

A single parameter, the size of the vowel space controls the properties of a large number of vowels
Difficult to find with conventional techniques
Further reading


Possible Master Thesis project

- Children’s speech recognition
- Intended application: Early screening for dyslexia
- Mainly algorithm development and offline experiments