Speaker Adaptation

Steve Renals

Automatic Speech Recognition— ASR Lecture 11
January-March 2012

Speaker independent / dependent / adaptive

- **Speaker independent** (SI) systems have long been the focus for research in transcription, dialogue systems, etc.
- **Speaker dependent** (SD) systems can result in word error rates 2–3 times lower than SI systems (given the same amount of training data)
- A **Speaker adaptive** (SA) system... we would like
  - Error rates similar to SD systems
  - Building on an SI system
  - Requiring only a small fraction of the speaker-specific training data used by an SD system

Speaker Adaptation

- Introduction: speaker-specific variation, modes of adaptation
- Model-based adaptation: MAP
- Model-based adaptation: MLLR
- Model-based adaptation: Speaker space models
- Speaker normalization: VTLN

Speaker-specific variation

- **Acoustic model**
  - Speaking styles
  - Accents
  - Speech production anatomy (e.g., length of the vocal tract)
  - Also non-speaker variation, such as channel conditions (telephone, reverberant room, close talking mic) and application domain
  - Speaker adaptation of acoustic models aims to reduce the mismatch between test data and the models
- **Pronunciation model**: speaker-specific, consistent change in pronunciation
- **Language model**: user-specific documents (exploited in personal dictation systems)
Modes of adaptation

- Supervised or unsupervised
  - Supervised: the word level transcription of the adaptation data is known (and HMMs may be constructed)
  - Unsupervised: the transcription must be estimated (e.g. using recognition output)

- Static or dynamic
  - Static: All adaptation data is presented to the system in a block before the final system is estimated (e.g. as used in enrollment in a dictation system)
  - Dynamic: Adaptation data is incrementally available, and models must be adapted before all adaptation data is available (e.g. as used in a spoken dialogue system)

Approaches to adaptation

- Model based: Adapt the parameters of the acoustic models to better match the observed data
  - Maximum a posteriori (MAP) adaptation of HMM/GMM parameters
  - Maximum likelihood linear regression (MLLR) of Gaussian parameters

- Speaker Normalization: Normalize the acoustic data to reduce mismatch with the acoustic models
  - Vocal Tract Length Normalization (VTLN)

- Speaker space: Estimate multiple sets of acoustic models, characterizing new speakers in terms of these model sets
  - Cluster-adaptive training
  - Eigenvoices

Model-based adaptation: The MAP family

- Basic idea: Use the SI models as a prior probability distribution over model parameters when estimating using speaker-specific data
  - Theoretically well-motivated approach to incorporating the knowledge inherent in the SI model parameters
  - If the parameters of the models are denoted \( \lambda \), then maximum likelihood (ML) training chooses them to maximize \( p(X | \lambda) \)
  - Maximum a posteriori (MAP) training maximizes:
    \[
    p(\lambda | X) \propto p(X | \lambda) \rho_0(\lambda)
    \]
    \( \rho_0(\lambda) \) is the prior distribution of the parameters
  - The use of a prior distribution, based on the SI models, means that less data is required to estimate the speaker-specific models: we are not starting from complete ignorance
Refresher: ML estimation of GMM/HMM

- The mean of the \( m \)th Gaussian component of the \( j \)th state is estimated using a weighted average

\[
\mu_{mj} = \frac{\sum_n \gamma_{jm}(n)x_n}{\sum_n \gamma_{jm}(n)}
\]

- Where \( \sum_n \gamma_{jm}(n) \) is the component occupation probability
- The covariance of the Gaussian component is given by:

\[
\Sigma_{mj} = \frac{\sum_n \gamma_{jm}(n)(x_n - \mu_{jm})(x_n - \mu_{jm})^T}{\sum_n \gamma_{jm}(n)}
\]

MAP estimation

- What is \( p_0(\lambda) \)?
- Conjugate prior: the prior distribution has the same form as the posterior. There is no simple conjugate prior for GMMs, but an intuitively understandable approach may be employed.
- If the prior mean is \( \mu_0 \), then the MAP estimate for the adapted mean \( \hat{\mu} \) of Gaussian is given by:

\[
\hat{\mu} = \frac{\tau \mu_0 + \sum_n \gamma(n)x_n}{\tau + \sum_n \gamma(n)}
\]

\( \tau \) is a hyperparameter that controls the balance between the ML estimate of the mean, its prior value. Typically \( \tau \) is in the range 2–20.
- \( x_n \) is the adaptation vector at time \( n \)
- \( \gamma(n) \) the probability of this Gaussian at this time
- As the amount of training data increases, so the MAP estimate converges to the ML estimate

Local estimation

- Basic idea The main drawback to MAP adaptation is that it is local
- Only the parameters belonging to Gaussians of observed states will be adapted
- Large vocabulary speech recognition systems have about \( 10^5 \) Gaussians: most will not be adapted
- Structural MAP (SMAP) approaches have been introduced to share Gaussians
- The MLLR family of adaptation approaches addresses this by assuming that transformations for a specific speaker are systematic across Gaussians, states and models
- MAP adaptation is very useful for domain adaptation:
  - Example: adapting a conversational telephone speech system (100s of hours of data) to multiparty meetings (10s of hours of data) works well with MAP

SMAP: Structural MAP

- Basic idea share Gaussians by organising them in a tree, whose root contains all the Gaussians
- At each node in the tree compute mean offset and diagonal variance scaling term
- For each node, its parent is used as a prior distribution
- This has been shown to speed adaptation compared with standard MAP, while converging to the same solution as standard MAP in the large data limit
The Linear Transform family

- **Basic idea** Rather than directly adapting the model parameters, estimate a transform which may be applied to Gaussian means and covariances.
- Linear transform applied to parameters of a set of Gaussians: adaptation transform parameters are shared across Gaussians.
- This addresses the locality problem arising in MAP adaptation, since each adaptation data point can affect many of (or even all) the Gaussians in the system.
- There are relatively few adaptation parameters, so estimation is robust.
- Maximum Likelihood Linear Regression (MLLR) is the best known linear transform approach to speaker adaptation.

Interim Summary

- Speaker-specific variation
- Adaptation: supervised/unsupervised, static/dynamic
- Model-based adaptation: MAP
- Introduction to model-based adaptation
- Next lecture: MLLR, adaptive training, speaker space models and vocal tract length normalisation.

Overview

Speaker Adaptation

- Introduction: speaker-specific variation, modes of adaptation
- Model-based adaptation: MAP and MLLR
- Adaptive training
- Model-based adaptation: Speaker space models
- Speaker normalization: VTLN
The Linear Transform family

- **Basic idea** Rather than directly adapting the model parameters, estimate a transform which may be applied to the Gaussian means and covariances.
- Linear transform applied to parameters of a set of Gaussians: adaptation transform parameters are shared across Gaussians.
- This addresses the locality problem arising in MAP adaptation, since each adaptation data point can affect many of (or even all) the Gaussians in the system.
- There are relatively few adaptation parameters, so estimation is robust.
- Maximum Likelihood Linear Regression (MLLR) is the best known linear transform approach to speaker adaptation.

**Regression classes**

- The number of transforms may obtained automatically.
- A set of Gaussian components that share a transform is called a regression class.
- Obtain the regression classes by constructing a regression class tree.
- Each node in the tree represents a regression class sharing a transform.
- For an adaptation set, work down the tree until arriving at the most specific set of nodes for which there is sufficient data.
- Regression class tree constructed in a similar way to state clustering tree.
- In practice the number of regression may be very small: one per context-independent phone class, one per broad class, or even just two (speech/non-speech).

**MLLR: Maximum Likelihood Linear Regression**

- MLLR is the best known linear transform approach to speaker adaptation.
- Affine transform of mean parameters
  
  \[ \hat{\mu} = A\mu + b \]

  If the observation vectors are \( d \)-dimension, then \( A \) is a \( d \times d \) matrix and \( b \) is \( d \)-dimension vector.
- If we define \( W = [bA] \) and \( \eta = [1\mu^T]^T \), then we can write:
  
  \[ \hat{\mu} = W\eta \]

- In MLLR, \( W \) is estimated so as to maximize the likelihood of the adaptation data.
- A single transform \( W \) can be shared across a set of Gaussian components (even all of them!).

**Estimating the transforms**

- The linear transformation matrix \( W \) is obtained by finding its setting which optimizes the log likelihood.
- **Mean adaptation**: Log likelihood
  
  \[ L = \sum_r \sum_n \gamma_r(n) \log \left( K_r \exp \left( -\frac{1}{2}(x_n - W\eta_r)^T\Sigma_r^{-1}(x_n - W\eta_r) \right) \right) \]

  where \( r \) ranges over the components belonging to the regression class.
- Differentiating \( L \) and setting to 0 results in an equation for \( W \): there is no closed form solution if \( \Sigma \) is full covariance; can be solved if \( \Sigma \) is diagonal (but requires a matrix inversion).
- Variance adaptation is also possible.
MLLR in practice

- Mean-only MLLR results in 10–15% relative reduction in WER
- Few regression classes and well-estimated transforms work best in practice
- Robust adaptation available with about 1 minute of speech; performance similar to SD models available with 30 minutes of adaptation data
- Such linear transforms can account for any systematic (linear) variation from the speaker independent models, for example those caused by channel effects.

Constrained MLLR (cMLLR)

- Basic idea: use the same linear transform for both mean and covariance
  \[ \hat{\mu} = A'\mu - b' \]
  \[ \hat{\Sigma} = A'\Sigma A'^T \]
- No closed form solution but can be solved iteratively
- Log likelihood for cMLLR
  \[ L = \mathcal{N}(Ax_n + b; \mu, \Sigma) + \log(|A|) \]
  \[ A' = A^{-1}; b' = Ab \]
  Equivalent to applying the linear transform to the data!
- Iterative solution amenable to online/dynamic adaptation, by using just one iteration for each increment
- Similar improvement in accuracy to standard MLLR

Speaker-adaptive training (SAT)

- Basic idea: Rather than SI seed (canonical) models, construct models designed for adaptation
- Estimate parameters of canonical models by training MLLR mean transforms for each training speaker
- Train using the MLLR transform for each speaker; interleave Gaussian parameter estimation and MLLR transform estimation
- SAT results in much higher training likelihoods, and improved recognition results
- But: increased training complexity and storage requirements
- SAT using cMLLR, corresponds to a type of speaker normalization at training time

Speaker Space Methods

- Gender-dependent models: sets of HMMs for male and for female speakers
- Speaker clustering: sets of HMMs for different speaker clusters
- Drawbacks:
  - Hard division of speakers into groups
  - Fragments training data
- Weighted speaker cluster approaches which use an interpolated model to represent the current speaker
  - Cluster-adaptive training
  - Eigenvoices
Cluster-adaptive training

- **Basic idea** Represent a speaker as a weighted sum of speaker cluster models
- Different cluster models have shared variances and mixture weights, but separate means
- For a new speaker, mean is defined as
  \[ \mu = \sum_c \lambda_c \mu_c \]
  
  Given the canonical models, only the \( \lambda_c \) mixing parameters need estimated for each speaker
- Given sets of weights for individual speakers, means of the clusters may be updated
- CAT can reduce WER in large vocabulary tasks by about 4–8% relative
- See Gales (2000) for more

Eigenvoices

- **Basic idea** Construct a speaker space from a set of SD HMMs
- Could regard each canonical model as forming a dimension of speaker space
- Generalize by computing PCA of sets of “supervectors” (concatenated mean vectors), to form speaker space: each dimension is an “eigenvoice”
- Represent a new speaker as a combination of eigenvoices
- Close relation to CAT
- Computationally intensive, does not scale well to large vocabulary systems
- See Kuhn et al (2000) for more

Feature normalization

- **Basic idea**: Transform the features to reduce mismatch between training and test
- **Cepstral Mean Normalization** (CMN): subtract the average feature value from each feature, so each feature has a mean value of 0. makes features robust to some linear filtering of the signal (channel variation)
- **Cepstral Variance Normalization** (CVN): Divide feature vector by standard deviation of feature vectors, so each feature vector element has a variance of 1
- Cepstral mean and variance normalisation, CMN/CVN:
  \[ \hat{x}_i = \frac{x_i - \mu(x)}{\sigma(x)} \]
  
  Compute mean and variance statistics over longest available segments with the same speaker/channel
- Real time normalization: compute a moving average

Vocal Tract Length Normalization (VTLN)

- **Basic idea** Normalize the acoustic data to take account of changes in vocal tract length
- Vocal tract length (VTL):
  - First larynx descent in first 2-3 years of life
  - VTL grows according to body size, and is sex-dependent
  - Puberty: second larynx descent for males
- VTL has large effect on the spectrum
  - Tube acoustic model: formant positions are inversely proportional to VTL
  - Observation: formant frequencies for women are 20% higher than for men (on average)
- **VTLN**: compensate for differences between speakers via a warping of the frequency axis
Warping functions: Piecewise linear

\[ \hat{t} = \alpha f \]

Warping functions: Power function

\[ \hat{t} = \alpha^{3f/8000} f \]

Warping functions: Power function

\[ \hat{t} = f + \arctan \left( \frac{(1 - \alpha) \sin f}{1 - (1 - \alpha) \cos f} \right) \]

Approaches to VTLN

- Classify by frequency warping function
  - Piecewise linear
  - Power function
  - Bilinear transform
- Classify by estimation of warping factor \( \alpha \)
  - Signal-based: estimated directly from the acoustic signal, through explicit estimation of formant positions
  - Model-based: maximize the likelihood of the observed data given acoustic models and a transcription, \( \alpha \) is another parameter set so as to maximize the likelihood
Signal-based VTLN

- **Basic idea**: Estimate the warping factor from the signal without using the speech recognition models.
- Estimate warping factor \( \alpha \) from formant positions; e.g., Eide and Gish (1996) used ratio of median position of 3rd formant for speaker \( s \) (\( F_{3,s} \)) to the median for all speakers (\( F_3 \)):
  \[
  \alpha_s = \frac{F_{3,s}}{F_3}
  \]
- Wegmann et al. (1996) used a generic voiced speech model, estimated using maximum likelihood. During training, estimation of warping factors was alternated with estimating the phone models using the warped data.
- These approaches require an accurate estimation of voiced parts of the speech signal.

Model-based VTLN

- **Basic idea**: Warp the acoustic features (for a speaker) to better fit the models — rather than warping the models to fit the features!
- Estimate the warping factor \( \alpha \) so as to maximise the likelihood of the acoustic models.
- After estimating the warp factors, normalize the acoustic data and re-estimate the models.
- The process may be iterated.
- Model-based VTLN does not directly estimate vocal tract size; rather it estimates an optimal frequency warping, which may be affected by other factors (e.g., F0).
- Exhaustive search for the optimal warping factor would be expensive.
  - Approximate the log likelihood wrt \( \alpha \) as a quadratic, and find the maximum using a line search (Brent’s method).

VTLN: Training

1. **HMM(0)**
2. **ML estimation of warp factors: based on quadratic estimate of log likelihood (Brent search)**
3. **HMM(i)**
4. Repeat until WER is stable on dev data.

Speaker warping factor estimation
Computation of normalised features
HMM Training using normalised features
HMM(0)
HMM(i)
VTLN: Recognition

- Decode using non-normalized features and models
- Preliminary transcription
- Evaluate warping factors using normalized models and preliminary transcription
- Decode using normalized models and features

VTLN: Warp factor estimation, females, non-normalized

- Histogram of warping factors for females calculated using non-normalized model

VTLN: Warp factor estimation, females, pass 1

- Histogram of warping factors for females after first VTLN training pass

VTLN: Warp factor estimation, females, pass 2

- Warping factors histogram 2nd females pass
VTLN: Warp factor estimation, males, pass 3

VTLN: WER (%) on conversational telephone speech

<table>
<thead>
<tr>
<th></th>
<th>Tot</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>F</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adapt</td>
<td>37.2</td>
<td>24.2</td>
<td>8.8</td>
<td>4.2</td>
<td>36.7</td>
<td>37.6</td>
</tr>
<tr>
<td>Test only</td>
<td>36.4</td>
<td>23.6</td>
<td>8.5</td>
<td>4.3</td>
<td>36.1</td>
<td>36.7</td>
</tr>
<tr>
<td>1 pass</td>
<td>35.7</td>
<td>22.9</td>
<td>8.9</td>
<td>3.8</td>
<td>35.0</td>
<td>36.4</td>
</tr>
<tr>
<td>2 pass</td>
<td>35.0</td>
<td>22.5</td>
<td>8.8</td>
<td>3.7</td>
<td>34.2</td>
<td>35.8</td>
</tr>
<tr>
<td>3 pass</td>
<td>34.5</td>
<td>22.0</td>
<td>8.7</td>
<td>3.7</td>
<td>33.6</td>
<td>35.3</td>
</tr>
<tr>
<td>4 pass</td>
<td>34.2</td>
<td>22.0</td>
<td>8.6</td>
<td>3.6</td>
<td>33.3</td>
<td>35.1</td>
</tr>
</tbody>
</table>

- 7–10% relative decrease in WER is typical for VTLN
- VTLN removes the need for *gender-dependent* acoustic models

Summary

Speaker Adaptation

- One of the most intensive areas of speech recognition research since the early 1990s
- Substantial progress, resulting in significant, additive, consistent reductions in word error rate
- Close mathematical links between different approaches
- Linear transforms at the heart of many approaches

Reading

- Gales and Young (2007), sec. 5. Good overview.