Automatic Speech Recognition Part 2

Training & adaptation

Mats Blomberg

Speech, Music and Hearing
KTH

Acoustic and language model training

- Large training corpora (speech and text) are required to estimate the statistical distribution of the acoustic and linguistic models
- Automatic training techniques are necessary

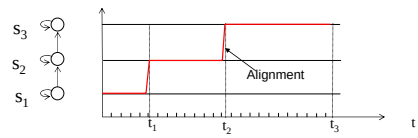
Language model training

- N-grams are estimated by counting word sequence frequency in text corpus
  - Unigram $P(w_i) = \frac{C(w_i)}{\sum_j C(w_j)}$
  - Bigram $P(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$
  - Trigram $P(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}$

Acoustic model training

Simple approach (Viterbi)

1. An initial model is created (existing model, from hand labeled data, or flat start)
2. Align (Viterbi) the training utterances with their phonetic transcriptions using the current model. (The alignment is found by backtracking the search lattice.) Measure likelihood.
3. The statistical distribution of each model state is set to that of all frames aligned to it.
4. Repeat 2 until likelihood convergence

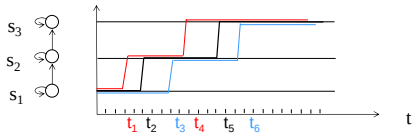


Example: New average $E(s_2) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_{2-1}} X(t)$

Problem: If the Viterbi alignment is incorrect, the new estimate is bad.

Account for alignment uncertainty

When estimating the parametric distribution of a state, the contribution of each frame is weighted by the probability that it is assigned to this state. **Expectation Maximization (EM) algorithm**



Example for three possible paths with probabilities P_{red}, P_{black} and P_{blue}:

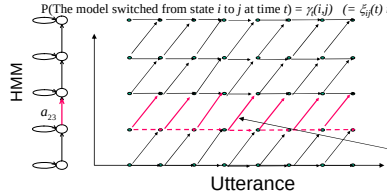
$$E(S_2) = \frac{\sum_{t=1}^{T} P(s=S_2,t)X(t)}{\sum_{t=1}^{T} P(s=S_2,t)} = \frac{\sum_{t=1}^{T} (P_{red}(s=S_2,t) + P_{black}(s=S_2,t) + P_{blue}(s=S_2,t))X(t)}{\sum_{t=1}^{T} P_{red}(s=S_2,t) + P_{black}(s=S_2,t) + P_{blue}(s=S_2,t)}$$

Not as simple as it may look. Many individual paths with parts in common.
=> The Forward-Backward (Baum-Welch) algorithm

The Forward-Backward algorithm

Also known as the Baum-Welch algorithm.
Compute probabilities for all state transitions between two adjacent frames.
This is a summation for all individual alignment curves passing through this point.

P(The model switched from state *i* to *j* at time *t*) = $\gamma_t(i,j)$ (= $\xi_{ij}(t)$ *in Holmes & Holmes)*



Example: Transition probability:

$$a_{23} = \frac{\sum_{t=1}^{8} \gamma_t(2,3)}{\sum_{t=1}^{8} \sum_{k=2}^{3} \gamma_t(2,k)}$$

Iteratively estimate new model parameters until convergence. Guaranteed probability increase.

The γ values are computed by searching both forward and backward in time. Hence the name.

Notation from Huang, Acero and Hon (2001)

The Forward-Backward Algorithm (cont.)

New model estimates:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{k=1}^{N} \gamma_t(i,k)} \quad (8.40)$$

The ratio between the expected number of transitions from state i to j and the expected number of all transitions from state i

$$\hat{b}_j(k) = \frac{\sum_{t=1,\, o_t = o_k}^{T} \sum_{i} \gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{i} \gamma_t(i,j)} \quad (8.41)$$

Discrete model:
The ratio between the expected number of times the observation data emitted from state j is o_k and the expected number of times the model is in state j

Quite intuitive equations! Continuous observations are more tricky though.

Notation from Huang, Acero and Hon (2001) $\gamma_t(i,j) = \xi_{ij}(t)$ *in Holmes & Holmes*

Swedish training corpora at TMH

- Acoustic
  - Telephone speech - SpeechDat
    - 5000 speakers over fixed telephone network
      - 350 hour recordings, ca 4 minute recordings per speaker
    - 1000 speakers over mobile telephones
    - Speakers are balanced according to
      - dialect, age and gender
  - Wideband speech – Speecon
    - 600 speakers incl. 50 children
    - 82 hour recordings
    - 6 dialect regions
  - Children's speech – PF-Star
    - 200 children 4 – 8 years old, Stockholm area
- Text
  - Totally ca 150 million words, mostly from newspapers and books
  - 1,9 million unique words
  - ca 1 million words occur only once

Parameter Smoothing
Compensate for insufficient training data

- Increase the data *("There is no data like more data")*
- Reduce the number of free parameters
- Backing off
  - Use more general models if specific models don't exist
- Deleted interpolation
  - Weighted combination of models with different context dependence
- Parameter flooring to avoid small probability values
- Tying parameters (SCHMM)
- Covariance matrix
  - Interpolate via MAP
  - Tie matrices (= some states have common covariance matrix)
  - Use diagonal covariance matrices

Adaptation and Normalisation
Compensate for mismatch between training and test data

- Normalisation
  - Adjust the *input signal* to remove non-phonetic information
- Adaptation
  - Adjust the *trained models* to minimize the mismatch with the calibration data
  - Supervised
    - The word sequence of the utterance is known
  - Unsupervised
    - Use the recognition result

Normalisation to speaker and environment

- Vocal Tract Length Normalisation (VTLN)
  - Expand/Compress the frequency scale to adjust for mismatch in vocal tract length
- Environmental noise compensation
  - Estimate the noise in non-speech segments
  - Remove noise from the input signal – Spectral subtraction
  - (Insert noise in the models - adaptation)
- Normalisation of channel frequency response
  - Cepstral Mean Subtraction (CMS)
    - Subtract the utterance average from each frame
  - RASTA
    - Hearing-inspired bandpass filtering of amplitude envelopes (4 – 10 Hz).

Adaptation Techniques

- Maximum A Posteriori (MAP)
  - Interpolation between an old and a new model
- Maximum Likelihood Linear Regression (MLLR)
  - Transformation of groups of phonemes (regression classes)
- Eigenvoices
  - Positions the new speaker in a speaker space
  - Very little adaptation data needed
- Speaker Adaptive Training (SAT)
  - Adaptation during training to reduce model variance

Maximum a Posteriori (MAP)

- A new model is estimated using the training data interpolated with old information about the model

$$\hat{\mu} = \tau\mu_{obs} + (1-\tau)\mu_{prior}$$

- τ is a balancing factor between the prior mean and the ML estimate. Depends on the likelihood of the adaptation data
- Limitations
  - The prior model needs to be accurate
  - Needs observations for all models

Maximum Likelihood Linear Regression (MLLR)

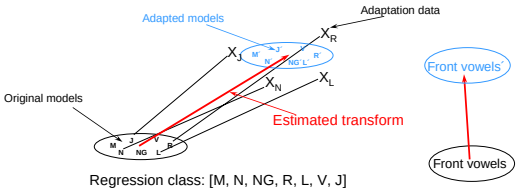
- Linear regression functions transform mean and covariance for maximizing the likelihood of the adaptation data

$$\overline{\boldsymbol{\mu}}_{ik} = \mathbf{A}_c\boldsymbol{\mu}_{ik} + \mathbf{b}_c$$

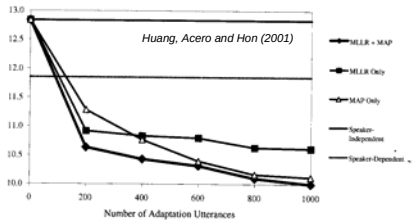
- $\mathbf{A}_c$ is a regression matrix, $\mathbf{b}_c$ is an additive vector for regression class *c*, $\mathbf{u}_{ik}$ is mean vector for mixture *k* in state *i*
- Adapts means and variances, but not transition probabilities
- **A** and **b** are estimated by the EM algorithm on the adaptation data
- All models in the same regression class have the same transform. Also models not in the adaptation data are updated
- If little training data, use few regression classes

MLLR adaptation illustration

- The transform for a class is optimized to maximize the likelihood of the adapted models to generate the adaptation data



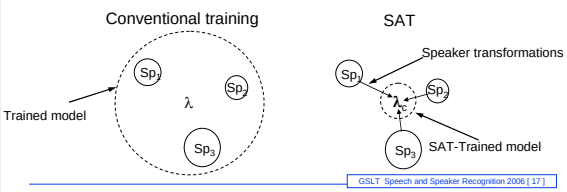
MLLR and MAP Comparison



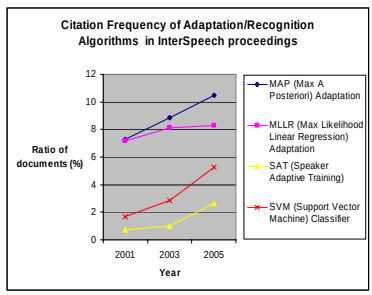
- MLLR is better for small adaptation data, MAP is better when the adaptation data is large. Combined MLLR+MAP best in both cases

Speaker-Adaptive Training (SAT)

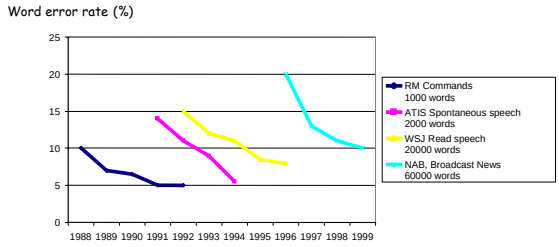
- Problem in speaker independent models
  - Large model variances due to inter-speaker differences
- Solution: Speaker Adaptive Training
  - MLLR adaptation "transforms" every speaker to an average position before training
  - The model variances are decreased, reducing errors 5-10% vs MLLR alone
  - Requires adaptation during recognition



Trends



Performance progress
ARPA evaluations 1988-1999

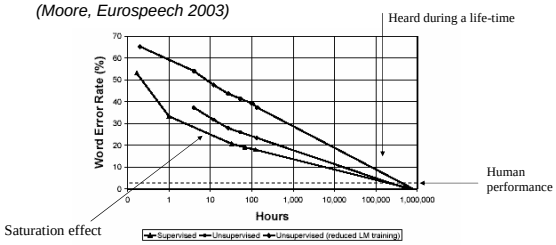


Results 2005
(previous results in parenthesis)

Corpus	Speech type	Lexicon size	Word Error Rate (%)	Human Error Rate (%)
Digit string (phone)	spontaneous	10	0.3	0.009
Resource Management	read	1000	3.6	0.1
ATIS	spontaneous	2000	2	-
Wall Street Journal	read	64000	6.6	1
Radio News	mixed	64000	13.5	-
Switchboard (phone)	conversation	10000	19.3	4
Call Home (phone)	conversation	10000	30	-

How large training data to reach human listening performance?

Extrapolated word error rates for increasing quantities of training data
(Moore, Eurospeech 2003)



END