Fundamental Equations

Bayes' decision rule:

$$\widehat{\omega} = \arg\max_{\omega} \{P(\omega|O)\} = \arg\max_{\omega} \{P(\omega)P_{\omega}(O)\}$$

- $P_{\omega}(O)$ — acoustic model.
  - For word sequence ω, how likely are features O?
- $P(\omega)$ — language model.
  - How likely is word sequence ω?

Lecture 9

Speaker Adaptation

Michael Picheny, Bhuvana Ramabhadran, Stanley F. Chen, Markus Nussbaum-Thom

Watson Group
IBM T.J. Watson Research Center
Yorktown Heights, New York, USA
{picheny,bhuvana,stanchen,nussbaum}@us.ibm.com

8 April 2016

Where Are We?

[0]

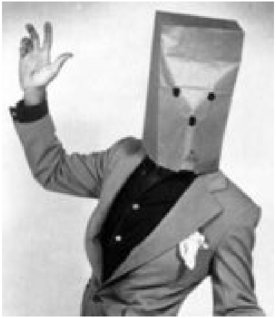
Problem: Sources of Variability

- gender: male / female
- age: young / old
- accents: Texas, South-Carolina
- environment noise: office, car, shopping mall
- different types of microphone
- channel characteristics: high-quality, telephone, mobile phone

Question: Are all these effects covered in training ?

Changing Conditions (I)

- Training data: Should represent test data adequately.
- Problem: There will always be new speakers or conditions.
- Consequence: What will happen ?



- Recognition performance drops.

Changing Conditions (II)

- Why does the performance drop ?
- The features are different from training.
- Situation in training: Larger amount data for a specific set of speakers.
- Situation in recognition: Small amount of data from a target speaker.
- What can we do ?

Adaptation vs. Normalization

- What can we do to overcome the mismatch between training and recognition ?
- Change features O or the acoustic model $P(O|\omega, \theta)$.
  - O: feature sequence.
  - ω: word sequence.
  - θ: free model parameters.

- Model-based – Feature-based:
  - Modify model to better fit the features $\Rightarrow$ adaptation.
  - Transform features to better fit model $\Rightarrow$ normalization.

Terminology

- Speaker: Rather a concept for different signal conditions.
- Speaker-independent (SI) system: trained on complete data.
- Speaker-dependent (SD) system: trained on all the data per speaker.
- Speaker-adaptive (SA) system: adapted SI system using the speaker dependent data.

Adaptation/Normalization Types

- Supervised vs. Unsupervised: Is correct transcription of utterance available at test time ?
- Batch vs. Incremental adaptation/normalization: Whole vs. small (time critical) portion of the test data is available.
- Online vs. Offline system: Real-time demand vs. No time restriction.

Question answer (I)

- What is the concept of a speaker ?
- Are all speaker covered in the training data ? What happens ?
- How do we approach the problem of unseen speaker ?
- Supervised vs. unsupervised ?
- Batch vs. Incremental ?

Strategies Summary

Goal: Fit acoustic model/features to speaker.

- Use new acoustic adaptation data from the current speaker.
- Model-based – Feature-based:
  - Modify model to better fit the features $\Rightarrow$ adaptation.
  - Transform features to better fit model $\Rightarrow$ normalization.
- Supervised – Unsupervised:
  - Transcription is available for adaptation $\Rightarrow$ supervised.
  - No transcription is available $\Rightarrow$ unsupervised.
- Training:
  - Normalization/Adaptation also in training $\Rightarrow$ Speaker Adaptive Training (SAT).
- Incremental – Batch:
  - adaptation only on small parts $\Rightarrow$ incremental.
  - adaptation on all data $\Rightarrow$ batch.

Transformation of a Random Variable

- Consider a random variable O:
  - with density $P(O)$
  - and transform $O^{'} = f(O)$ (assume f can be inverted)
- Then the density $P(O^{'})$ is:

$$P(O^{'}) = \frac{1}{\left|\frac{d\,f(O)}{d\,O}\right|} P(f^{-1}(O^{'})) = \frac{1}{\left|\frac{d\,f(O)}{d\,O}\right|} P(O)$$

- with Jacobian determinant:

$$\left|\frac{d\,f(O)}{d\,O}\right|$$

- or equivalent:

$$P(O) = \left|\frac{d\,f(O)}{d\,O}\right| P(O^{'})$$

Supervised Normalization and Adaptation

Estimation of adaptation parameters

- Model based: $\theta' = f(\theta, \Phi)$

$$\hat{\Phi} = \arg\max_{\Phi} P(O|\omega, \theta')$$

- Feature based: $O' = f(O, \Phi)$ $\qquad$ ($|\cdot|$ is the Jacobi matrix)

$$\hat{\Phi} = \arg\max_{\Phi} \left| \frac{d\, f(O, \Phi)}{d\, O} \right| P(O'|\omega, \theta)$$

- Correct transcript ω of adaptation data is given.

Unsupervised Normalization and Adaptation

Estimation of adaptation parameters and generation of adaptation word sequence.

- Model based: $\theta' = f(\theta, \Phi)$

$$(\hat{\omega}, \hat{\Phi}) = \arg\max_{\omega,\Phi} P(O|\omega, \theta').$$

- Feature based: $O' = f(O, \Phi)$

$$(\hat{\omega}, \hat{\Phi}) = \arg\max_{\omega,\Phi} \left| \frac{d\, f(O, \Phi)}{d\, O} \right| P(O'|\omega, \theta).$$

- In practice infeasable.
- In practice transcript ω of adaptation data is approximated.

Unsupervised Normalization and Adaptation: First Best Approximation

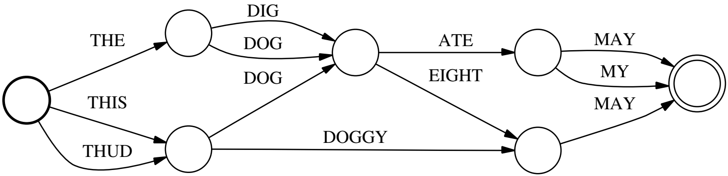
First-best approximation

- A speaker independent system generates the first best output.
- Estimation is performed exactly as in the supervised case, but use first pass output as transcription.
- Most popular method.

Unsupervised Normalization and Adaptation: Word Graph Approximation

Word graph based approximation

- A first pass recognition generates a word graph.
- Use the forward-backward algorithm as in the supervised case based on the word graph.
- Weighted accumulation.



Question answer (II)

- Superprvised vs. Unsupervised adaptation ?
- Approximations ?

Where Are We?

[0]

Online System

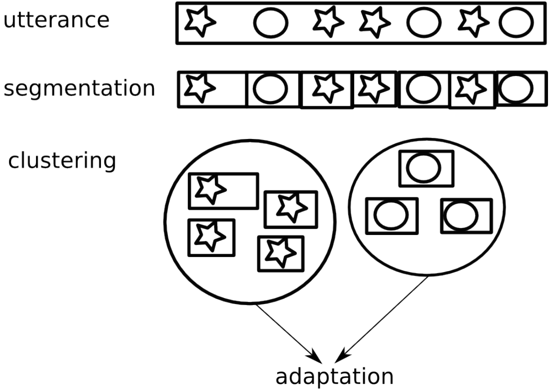
Objective: generate transcription from audio in real time.

- The audio has multiple unknown speakers and conditions.
- Requires fast adaptation with very little data (a couple of seconds).
- Can benefit from incremental adaptation which continuously updates the adaptation for new data.

Offline System (I)

Objective: generate transcription from audio. Multiple passed over the data are allowed.

- The audio has multiple unknown speakers and conditions.
- Segmentation and Clustering:



Offline System (II)

- Where are the speakers and conditions ? $\Rightarrow$
- Segmentation and Clustering
  - Segmentation: Partitioning of the audio into homogenous areas, ideally one speaker/condition per segment.
- What are the speakers are conditions ?
  - Clustering: Clustering into similar speakers/conditions.
- The speakers unknown/no transcribed audio data $\Rightarrow$ unsupervised adaptation.

Audio Segmentation(I)

- Objective: split audio stream in homogeneous regions
- Properties:
  - speaker identity,
  - recording condition (e.g. background noise, telephone channel),
  - signal type (e.g. speech. music, noise, silence), and
  - spoken word sequence.

Audio Segmentation (II)

- Segmentation affects speech recognition performance:
  - speaker adaptation and speaker clustering assume one speaker per segment,
  - language model assumes sentence boundaries at segment end,
  - non-speech regions cause insertion errors,
  - overlapping speech is not recognized correctly, causes errors at sorrounding regions,

Audio Segmentation: Methods

- Metric based
  - Compute distance between adjacent regions.
  - Segment at maxima of the distances.
  - Distances: Kullback-Leibler distance, Bayesian information criterion.
- Model based
  - Classify regions using precomputed models for music, speech, etc.
  - Segment changes in acoustic class.
- Decoder guided
  - Apply speech recognition to input audio stream.
  - Segment at silence regions.
  - Other decoder output useful too.

Audio Segmentation: Bayesian Information Criterion

Bayesian Information Criterion (BIC):

- Likelihood criterion for a model Θ given observations O:

$$\mathrm{BIC}(\Theta, O) = \log p(O|\Theta) - \frac{\lambda}{2} \cdot d(\Theta) \cdot \log(N)$$

  $d(\Theta)$: number of parameters in Θ, λ: penalty weight for model complexity.

- used for model selection: choose model maximizing BIC.

Change Point Detection: Modeling

Change point detection using BIC:

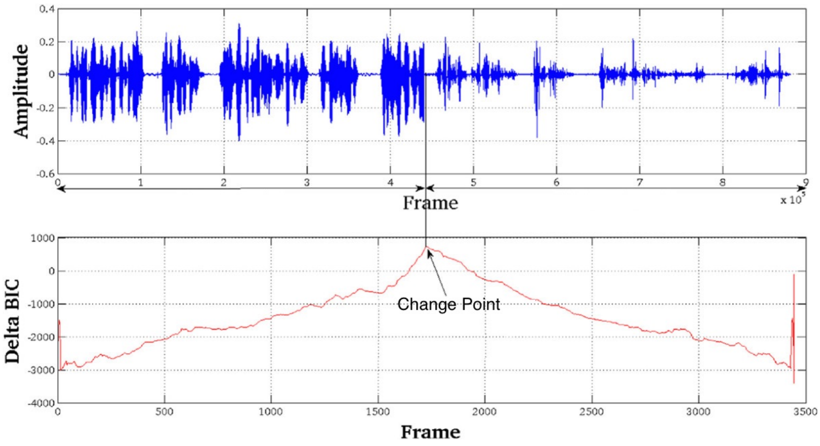
- Input stream is modeled as Gaussian process in the cepstral domain.
- Feature vectors of one segment: drawn from multivariate Gaussian: $O_i^j := O_i \dots O_j \sim \mathcal{N}(\mu, \Sigma)$
- For hypothesized segment boundary t in O_1^T decide between
  - $O_1^T \in \mathcal{N}(\mu, \Sigma)$ and
  - $O_1^t \in \mathcal{N}(\mu_1, \Sigma_1)$ $O_{t+1}^T \in \mathcal{N}(\mu_2, \Sigma_2)$
- Use difference of BIC values:

$$\Delta\mathrm{BIC}(t) = \mathrm{BIC}(\mu, \Sigma, O_1^T) - \mathrm{BIC}(\mu_1, \Sigma_1, O_1^t) - \mathrm{BIC}(\mu_2, \Sigma_2, O_{t+1}^T)$$

Change Point Detection: Criterion

- Detect single change point in $O_1 \ldots O_T$:

$$\hat{t} = \arg\max_t \{\Delta\mathrm{BIC}(t)\}$$



Change Point Detection: Example

- $\Delta\text{BIC}(t)$ can be simplified to:

$$\Delta\text{BIC}(t) = T\log|\Sigma| - t\log|\Sigma_1| - (T-t)\log|\Sigma_2| - \lambda P$$

  - number of parameters $P = \frac{(D+\frac{1}{2}D(D+1))}{2}\log T$,
  - D : dimensionality.

t	1	2	3	4	5	6
O_1^6	4	3	2	9	5	7
Σ	5.67					
$\Sigma_{1,t}$	0	0.25	0.67	7.25	5.84	5.67
$\Sigma_{2,t}$	6.56	6.69	2.67	1	0	0
$\Delta\text{BIC}(t)$	-0.89	5.57	8.68	2.48	-0.17	0

Question answer (II)

- What should a segmentation ideally do ?
- What problems can occur due to segmenting (LM, Non-Speech, Overlap) ?
- What methods exist for segmentation ?
- What ist the Bayesian Information criterion ?
- How does change point detection work ?

Speaker Clustering: Introduction

- Objective: Group speech segments into clusters for adaptation.
- Segments from same or similar speakers should be grouped.

BIC Method

- Uses acoustic features only.
- Greedy, bottom up clustering.
- BIC used to control number of clusters.

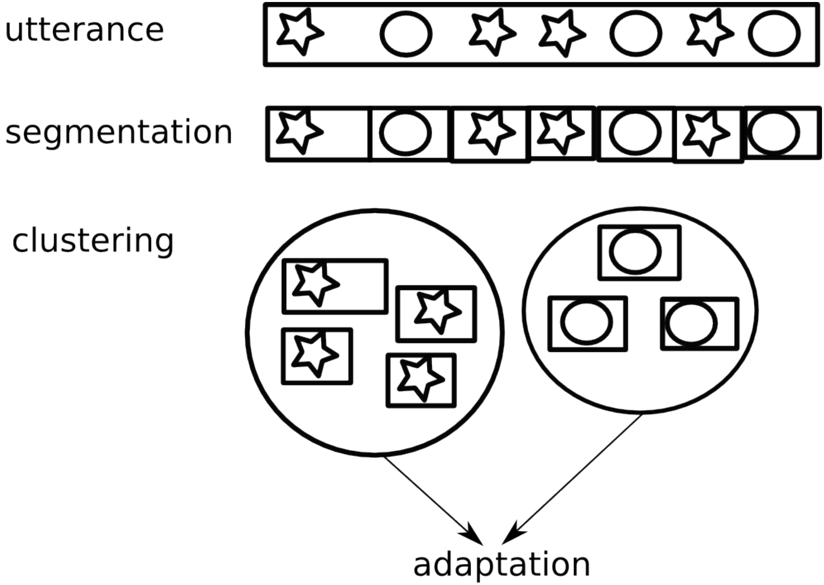
Speaker Clustering: BIC Clustering

- Greedy, bottom up, BIC clustering method.
- Each cluster is modeled using single Gaussian, full covariance.
- BIC criterion, Requirement: Clustering should give lowest possible adaptation WER.
- Algorithm:
    1. Start with one cluster for each segment.
    2. Try all possible pairwise cluster merges.
    3. Merge the pair that gives the largest increase in BIC.
    4. Iterate from 1, until BIC starts to decrease.

Where Are We?

[0]

Remember ? Batch Adaptation



Maximum Likelihood Linear Regression (MLLR)

Goal: Modify speaker indpendent model to better fit featues.

- Speaker dependent transform:

$$f(\mu, (A, b)) = A \cdot \mu + b$$

  Simplified to $f(\mu, A) = A \cdot \mu$

- Maximum likelihood:

$$(\hat{A}, \hat{\omega}) = \arg\max_{A,\omega} \{P(\omega)P(O|\omega, \mu, \sigma, A)\}$$

- Is a simultaneous optimization practical ?

Maximum Likelihood Linear Regression (MLLR)

- $\hat{W}$ is the result from a speaker independent system.

$$\hat{A} = \arg\max_{A} \left\{ P(W) P(O|\hat{W}, A) \right\}$$

- EM-Algorithm:
  1. Find best state sequence for $\hat{W}$ for given $\hat{A}$.
  2. Estimate new parameters $\hat{A}$ based on given $\hat{W}$.
  3. Iterate 1.
- EM Estimate: Maximum Likelihood or Forward backward.

Remember ? Gaussian Mixture Models

The Speaker Dependent Model is a Gaussian Mixture Model (GMM).

- Probability of an utterance given a hypothesized word sequence:

$$P(O|\omega,\mu,\sigma) = \prod_{t=1}^{N} \sum_{k=1,\dots,K} \frac{p_k}{\sqrt{2\pi}\sigma_k} e^{-\frac{(O_t-\mu_k)^2}{2\sigma_k^2}}$$

- Log-likelihhood is just as good:

$$\log P(O|\omega,\mu,\sigma) = \sum_{t=1}^{T} \ln\left[\sum_{k=1,\dots,K} \frac{p_k}{\sqrt{2\pi}\sigma_k} e^{-\frac{(O_t-\mu_k)^2}{2\sigma_k^2}}\right]$$

Remember ? Maximum Approximation

- For simplification: Maxium approximation

$$\log P(O|\omega, \mu, \sigma) = \sum_{t=1}^{T} \ln \left[\max_{k=1,\ldots,K} \frac{p_k}{\sqrt{2\pi}\sigma_k} \; e^{-\frac{(O_t - \mu_k)^2}{2\sigma_k^2}} \right]$$

- Path:

$$s_1^t = s_1, \ldots, s_t = \arg\max_{k_1^T} \ln \left[\sum_{t=1}^{T} \ln \frac{p_{k_t}}{\sqrt{2\pi}\sigma_{k_t}} \; e^{-\frac{(O_t - \mu_{k_t})^2}{2\sigma_{k_t}^2}} \right]$$

Simple Linear Regression - Review (I)

Say we have a set of points $(O_1, \mu_{s_1}), (O_2, \mu_{s_2}), \ldots, (O_N, \mu_{s_T})$ and we want to find coefficients A so that

$$\sum_{t=1}^{T} (O_t - (A\mu_{s_t}))^2$$

is minimized.

Taking derivatives with respect to A we get

$$\sum_{t=1}^{T} 2\mu_{s_t}(O_t - \mu_{s_t}^T A) = 0$$

Simple Linear Regression - Review (II)

Taking derivatives with respect to A we get

$$\sum_{t=1}^{T} 2\mu_{s_t}(O_t - \mu_{s_t}^T A) = 0$$

$$\Leftrightarrow \sum_{t=1}^{T} 2\mu_{s_t} O_t^T = \sum_{t=1}^{T} 2\mu_{s_t}\mu_{s_t}^T A$$

$$\Leftrightarrow A = \left[\sum_{t=1}^{T} \mu_{s_t}\mu_{s_t}^T\right]^{-1} \sum_{t=1}^{T} \mu_{s_t} O_t^T$$

so collecting terms we get

$$A = \left[\sum_{t=1}^{T} \mu_{s_t}\mu_{s_t}^T\right]^{-1} \sum_{t=1}^{T} \mu_{s_t} O_t^T$$

MLLR: Estimation

- Minimize the speaker transformed log-likelihood:

$$\sum_{t=1}^{T} \ln \left[\max_{k=1,\ldots,K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(O_t - A\mu_k^T)^2}{2\sigma^2}} \right]$$

- Consider observations $O_1, \ldots, O_T$ and path $s_1, \ldots, s_T$:

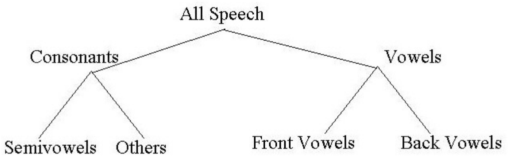
$$\frac{d}{dA} \left\{ \sum_{t=1}^{T} \frac{(O_t - A\overline{\mu}_{s_t})^2}{\sigma^2} \right\} = 0 \qquad (1)$$

- Compare with linear regression:

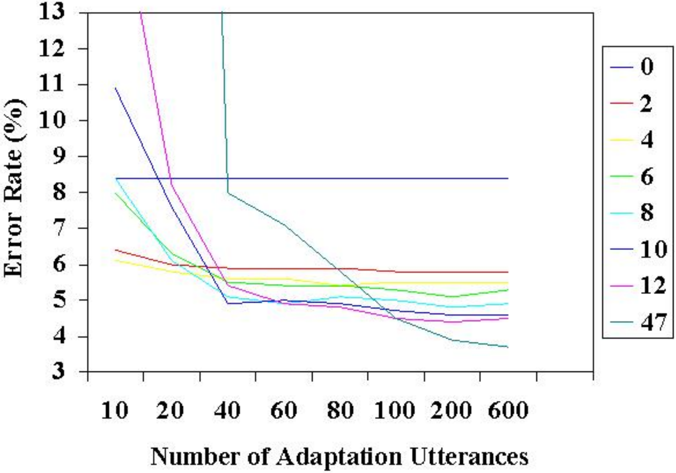
$$A = \left[\sum_{t=1}^{T} \mu_{s_t} \mu_{s_t}^T \right]^{-1} \sum_{t=1}^{T} \mu_{s_t} O_t^T$$

MLLR - Multiple Transforms

- Single MLLR transform for all of speech is very restrictive.
- Multiple transforms can be created having state dependent transforms.
- Arrange states in form of tree
- If there are enough frames at a node, a separate transform is estimated for all the phones at the node.



MLLR - Performance



Where Are We?

[0]

Feature based Maximum Likelihod Linear Regression (fMLLR)

Goal: Normalize features to better fit speaker.

- Speaker dependent transform:

$$O_t' = A \cdot O_t + b$$

- Transformation of Gaussian:
  $P(O_t') = \mathcal{N}(O_t' | \mu_k, \sigma_k) \Leftrightarrow P(O_t) = |A| \mathcal{N}(AO_t + b | \mu_k, \sigma_k)$
- Pure feature transform $\Rightarrow$ no changes to decoder necessary.
- Speaker adaptive training easy to implement.

Where Are We?

[0]

Speaker Adaptive Training

- Introduction:
  - Adaptation compensates for speaker differences in recognition.
  - But: We also have speaker differences in training corpus.
  - Question: How can we compensate for both these differences?
- Speaker-Adaptive Normalization:
  - Apply transform on training data also.
  - Model training using transformed acoustic features.
- Speaker-Adaptive Adaptation:
  - Interaction between model and transform requires simultaneous model and transform parameter training.
  - Cannot simply retrain model modified acoustic model training necessary.

Speaker Adaptive Training: Training

- Training Procedure:
  1. Estimate speaker independent model.
  2. Compute viterbi path using a simple target model.
  3. Use simple viterbi path to estimate fMLLR adaptation supervised for each speaker in training.
  4. Transform features using the estimated fMLLR adaptation.
  5. Train speaker adaptive model MSAT on transformed features, starting from the speaker independent system.

Speaker Adaptive Training: Recognition

- Recognition Procedure:
  1. First pass recognition using speaker independent model.
  2. Estimate fMLLR adaptation unsupervised using simple target model.
  3. Transform features using the estimated fMLLR adaptation.
  4. Second pass using the speaker adaptive model using the transformed features.

Performance of MLLR and fMLLR

	Test1	Test2
BASE	9.57	9.20
MLLR	8.39	8.21
fMLLR	9.07	7.97
SAT	8.26	7.26

- Task is Broadcast News with a 65K vocabulary.
- 15-26% relative improvement.