
Adaptation and Adaptive Training

A Discriminative and Bayesian Approach

C. K. Raut



Cambridge University
Engineering Department

August 2009

Speech Recognition Research - Background

- ▶ **PhD (Information Engineering) – continuing**
Discriminative Adaptive Training and Bayesian Adaptation
 Supervisor: Dr. Mark Gales



- ▶ Bayesian Discriminative Adaptation and Inference (ICASSP'09)
- ▶ Discriminative MAP Adaptation (ICASSP'09)
- ▶ Expectation Propagation for Bayesian Inference
- ▶ Discriminative Adaptive Training (Interspeech'08)

- ▶ **Master's in Information Science and Technology**
Robust Speech Recognition with Noise and Long Reverberation
 Supervisor: Prof. Shigeki Sagayama



- ▶ ML State Regression for Reverberant Speech (ASRU'05, ICASSP'06)
- ▶ HMM State Splitting for Long Reverberation (Interspeech'05)
- ▶ Noise-Driven Spectral Filtering (ASJ'04)
- ▶ Polynomial Approximation for Model Composition (ICSLP'04)

- ▶ **Bachelor in Electronics Engineering**
Speech Based System using Artificial Intelligence

- ▶ Connectionist Speech Recognition System
- ▶ Speech Processing



Outline

Adaptation and Adaptive Training

A Discriminative and Bayesian Approach

Introduction

Adaptation to Speaker or Environment

- Maximum Likelihood and Discriminative Transforms
- Discriminative Mapping Transforms (DMTs)

Adaptive Training

- Maximum Likelihood SAT
- ML-transforms Based Discriminative SAT
- DMT-based Discriminative SAT

Bayesian Approaches to Adaptation and Adaptive Training

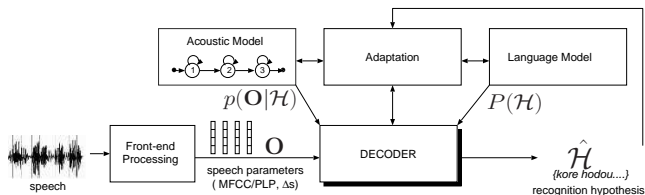
- Maximum Likelihood Bayesian Approaches
- Discriminative Bayesian Approaches



Automatic Speech Recognition Systems

- ▶ statistical speech recognition systems

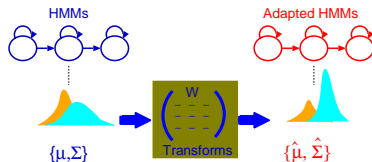
$$\hat{\mathcal{H}} = \arg \max_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}) = \arg \max_{\mathcal{H}} \underbrace{\left\{ \overbrace{p(\mathbf{O}|\mathcal{H})}^{\text{acoustic score}} \overbrace{P(\mathcal{H})}^{\text{LM score}} \right\}}_{\text{inference evidence}}$$



- ▶ HMMs used as acoustic models
 - ▶ trained from large amount of speech data
- ▶ testing speakers/environment may differ from training
 - ▶ adaptation of models – important to reduce mismatch

Adaptation to Speaker or Environment

- ▶ mismatch coming from
 - ▶ difference between speakers
 - ▶ physiological: age, gender, vocal tract, stress-level
 - ▶ linguistic: accent, non-native speaker
 - ▶ difference in acoustic environment
 - ▶ background noise, reverberation, microphone
- ▶ HMM parameters – adapted to match test speaker/environment
- ▶ linear transforms commonly used
 - ▶ transform mean and covariance of state output distributions of HMMs



- ▶ mean transform

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b}$$

$$\mathbf{W} = [\mathbf{A} \ \mathbf{b}] : \text{affine transform}$$



Transforms Estimation

- ▶ Maximum-likelihood Linear Regression (MLLR) (Leggetter & Woodland [4]; Gales [5])

$$\mathbf{W}_{\text{ml}}^{(s)} = \arg \max_{\mathbf{W}} \left\{ \log p(\mathbf{O}^{(s)} | \mathcal{H}^{(s)}; \mathbf{W}, \mathcal{M}) \right\}$$

- ▶ Discriminative Linear Transforms (DLT) (Tsakalidis et al. [15]; Wang [16])
 - ▶ use discriminative criteria

$$\mathbf{W}_{\text{d}}^{(s)} = \arg \min_{\mathbf{W}} \left\{ \underbrace{\sum_{\mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}, \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)})}_{\text{expected loss}} \right\}$$

$\mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)})$: raw phone error – minimum phone error (MPE) ✓
 (a form of general minimum Bayes risk, MBR criterion)

- ▶ needs supervision hypothesis for generating transforms
 - ▶ usually hypothesis from initial decoding with SI model or other systems
 - **unsupervised adaptation**



The Problems of Unsupervised Adaptation

- ▶ CTS task with the standard MPE-SAT system (described later)
- ▶ **using reference as supervision**
 - ▶ significant improvement: 4.9% with MLLR, 7.5% absolute with DLT !!
 - ▶ **discriminative transforms much better than MLLR** ✓
 - 2.6% absolute gain compared to MLLR

Adaptation	Supervision		
	Reference	1-best Hyp.	Lattice
-	29.2		
MLLR	24.3	27.0	26.7
DLT	21.7	26.8	26.6

- ▶ **using 1-best hypothesis (unsupervised adaptation)**
 - ▶ much reduced improvement
 - ▶ discriminative transforms affected more – only 0.2% gain over MLLR
 - **hypothesis bias problem** -
 - ▶ transform estimation biased to supervision hypothesis
 - ▶ transforms much sensitive to any errors in supervision
 - lattice-based or confidence-based adaptation: only little improvement

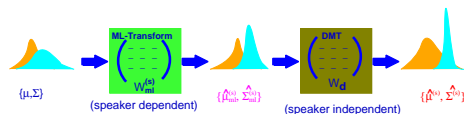


Discriminative Mapping Functions

- ▶ one strategy to obtain overall discriminative transforms: decomposition
 - ▶ train speaker-specific parameters with maximum-likelihood

$$\mathbf{W}_{m1}^{(s)} = [\mathbf{A}_d^{(s)} \quad \mathbf{b}_d^{(s)}] \quad \text{– speaker-specific}$$
 - ▶ train speaker-independent parameters discriminatively

$$\mathbf{W}_d = [\mathbf{A}_d \quad \mathbf{b}_d] \quad \text{– speaker-independent}$$



- ▶ mean adapted first by $\mathbf{W}_{m1}^{(s)}$, followed by \mathbf{W}_d

$$\hat{\boldsymbol{\mu}}_{m1}^{(s)} = \mathbf{A}_{m1}^{(s)} \boldsymbol{\mu} + \mathbf{b}_{m1}^{(s)}$$

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_d \hat{\boldsymbol{\mu}}_{m1}^{(s)} + \mathbf{b}_d$$

- ▶ ML-transforms basically mapped into discriminative-like transforms

$$\mathbf{A}_d^{(s)} = \mathbf{A}_d \mathbf{A}_{m1}^{(s)}; \quad \mathbf{b}_d^{(s)} = \mathbf{A}_d \mathbf{b}_{m1}^{(s)} + \mathbf{b}_d$$

- ▶ $\mathbf{W}_d = [\mathbf{A}_d \quad \mathbf{b}_d]$ called discriminative mapping transforms (DMT)

Discriminative Mapping Transforms (DMTs)

- ▶ DMT estimation same as DLT, but uses training data from all speakers

$$\mathbf{W}_d = \arg \min_{\mathbf{W}} \left\{ \sum_s \sum_{\mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}, \mathbf{W}_{ml}^{(s)}, \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}$$

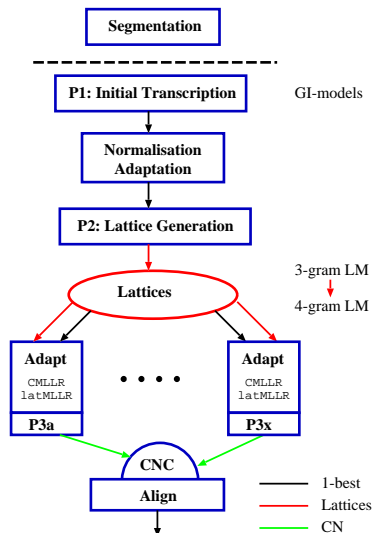
- ▶ DMT characteristics
 - ▶ estimated during training, same used for testing
 - ▶ not sensitive to errors in the test supervision hypothesis
 - ▶ possible to estimate large number of transforms
- ▶ English CTS task
 - ▶ ML Systems: 2.3% gain for SI, 1.8% gain for SAT over MLLR
 - ▶ MPE Systems: 0.8% gain over standard MLLR adaptation (Yu et al. [8])



AGILE Chinese BC/BN Transcription Task

- ▶ investigated DMT in a multipass framework with structured transforms
- ▶ CU S32 System
 - ▶ AM: 1670 hours of acoustic data
 - ▶ LM: about 2.9G words
 - ▶ manual segmentation
 - ▶ P1: use GI models
 - ▶ generate 1-best supervision
 - ▶ P2: use GD models
 - ▶ gender alignment, clustering, adaptation
 - ▶ lattice generation and expansion 3-gram to 4-gram
 - ▶ P3: GD or SAT models
 - ▶ CMLLR and latMLLR adaptation
 - ▶ rescoreing
- ▶ detailed description in AGILE QPR

(Oct 08 - Dec 08)



AGILE Chinese BC/BN Transcription – DMTs Performance

- ▶ P3b branch: CMLLR+latMLLR adaptation
- ▶ P3b branch modified for DMT:
 - ▶ covariance adaptation for latMLLR turned off
 - ▶ GI model used instead of GD models
 - as DMT estimated using GI models
 - ▶ DMT uses 1750 base classes

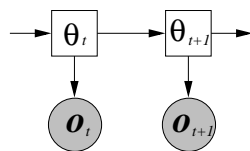
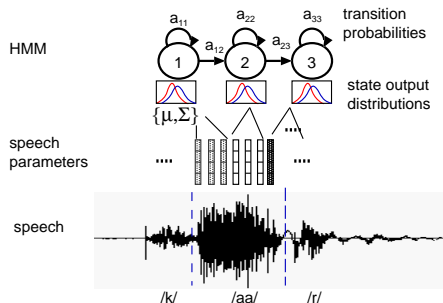
DMT	WER%				
	bcmdev05	bnmdev06	dev07	evrt07ns	dev08
×	15.9	6.7	9.4	9.0	9.2
✓	15.5	6.7	9.2	9.0	9.0

- ▶ 0 – 0.4% absolute gain with DMT
- ▶ directly using CMLLR based DMT may improve performance
 - ▶ implementation issues only



Acoustic Model Training

- ▶ traditional systems use multi-style trained HMMs
 - ▶ single set of models estimated using data from all speakers/environment
 - ▶ observations 'assumed' to be generated from speech state distributions

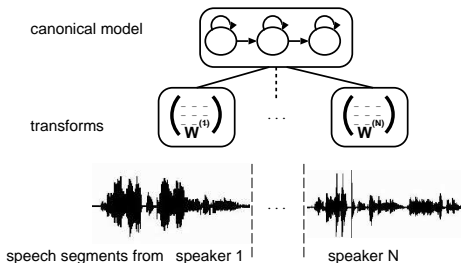


A DBN for a standard HMM
(generative model)

- ▶ large variabilities in training data
 - ▶ wide “spread” or coverage required
 - ▶ inherently less discriminatory models

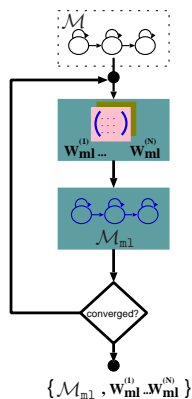
Adaptive Training

- ▶ speech and non-speech variabilities separately modelled
 - ▶ speech → canonical models \mathcal{M}_c
 - speaker/environment independent in nature
 - ▶ non-speech → set of transforms $\{\mathbf{W}^{(1)}, \mathbf{W}^{(2)} \dots \mathbf{W}^{(N)}\}$
 - specific to each speaker/homogeneous block



Maximum-Likelihood Speaker Adaptive Training (ML-SAT)

(Anastasakos et al. [6]; Matsoukas et al. [7]; Gales [5])



1. initialise:

- ▶ $\mathcal{M}_{ml} := \mathcal{M}$ SI model
- ▶ $\mathbf{W}_{ml}^{(s)} := (\mathbf{I} \ \mathbf{0})$ identity transform

2. estimate transforms for each speaker

$$\mathbf{W}_{ml}^{(s)} = \arg \max_{\mathbf{W}} \left\{ \log p(\mathbf{O}^{(s)} | \mathcal{H}^{(s)}; \mathbf{W}, \mathcal{M}_{ml}) \right\}$$

3. update model parameters

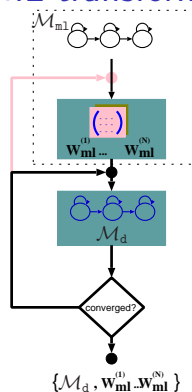
$$\mathcal{M}_{ml} = \arg \max_{\mathcal{M}} \left\{ \sum_{s=1}^S \log p(\mathbf{O}^{(s)} | \mathcal{H}^{(s)}; \mathbf{W}_{ml}^{(s)}, \mathcal{M}) \right\}$$

4. go to step 2, until converged

- ▶ state-of-the-art systems use discriminative criteria
 - ▶ how to use in adaptive framework to obtain discriminative SAT (DSAT)?



ML-transforms Based Discriminative SAT



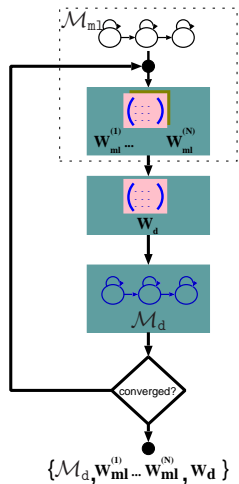
- ▶ commonly used standard DSAT (Ljolje [14])
 - ▶ only discriminative update of models
 - ▶ use ML-transforms (kept constant)
- ▶ training
 1. **initialisation:**
ML-SAT models and ML-transforms
 2. **update models, until converged**
(minimum phone error, MPE, criterion)

$$\mathcal{M}_d = \arg \min_{\mathcal{M}} \left\{ \sum_{s, \mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}_{ml}^{(s)}, \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}$$

- ▶ as discriminative training of models gives a gain, would like to use discriminative linear transforms (DLTs)
 - ▶ DLTs not used in adaptive training commonly
 - biased towards hypothesis, sensitive to supervision hypothesis errors
 - problems with unsupervised adaptation



DMT-based Discriminative SAT



1. initialisation:

- ▶ ML-SAT models and ML-transforms

2. estimate ML-transforms

$$\mathbf{W}_{ml}^{(s)} = \arg \max_{\mathbf{W}} \left\{ \log p(\mathbf{O}^{(s)} | \mathcal{H}^{(s)}; \mathbf{W}, \mathbf{W}_d, \mathcal{M}_d) \right\}$$

3. estimate DMT

$$\mathbf{W}_d = \arg \min_{\mathbf{W}} \left\{ \sum_{s, \mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}_{ml}^{(s)}, \mathbf{W}, \mathcal{M}_d) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}$$

4. update models

$$\mathcal{M}_d = \arg \min_{\mathcal{M}} \left\{ \sum_{s, \mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}_{ml}^{(s)}, \mathbf{W}_d, \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}$$

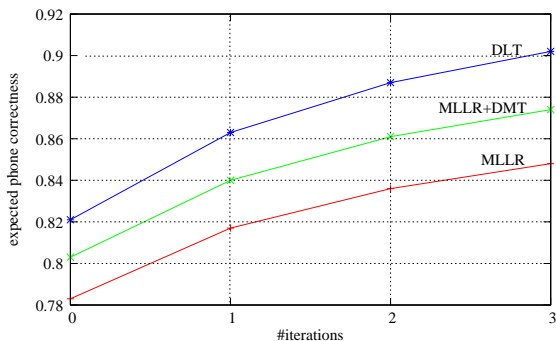
5. repeat step 2, 3 and 4 until converged

Experiments on English CTS Task

- ▶ English conversational telephone speech (CTS) task:
 - ▶ Dataset:
 - ▶ training dataset: about 290 hours data (5446 speakers)
 - ▶ test dataset (eva103): 6 hours data (144 speakers)
 - ▶ Front-end:
 - ▶ speech features: 12 PLP+ C0 and their 1st, 2nd, 3rd Δ s
 - ▶ CMN, CVN, HLDA and VTLN used
 - ▶ System:
 - ▶ state-clustered triphone HMMs with 6k distinct states
 - ▶ 16 Gaussian components per state (average)
 - ▶ MPE trained SI and SAT models, four iterations
 - ▶ 2 base classes for MLLR, 1000 for DMT
 - ▶ 58k multi-pron dictionary, trigram-LM



MPE Criteria for DSAT Iterations



- ▶ criteria obtained after applying transforms
- ▶ consistent increase with training iterations in all cases
- ▶ use of DMT shows gain in criteria compared to MLLR
- ▶ DLT-based DSAT has the highest criteria gain
 - ▶ likely to do better with training data



Comparison of WER for DSAT Schemes

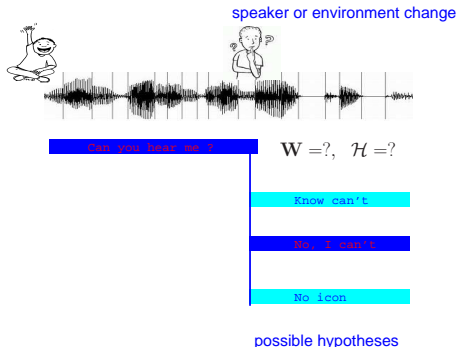
Training Scheme	Transform		WER%
	Training	Testing	
SI (hyp)	—	—	29.2
SI	—	MLLR	27.0
DSAT	MLLR	MLLR	26.4
	DLT	DLT	28.1
	MLLR+DMT	MLLR+DMT	25.3

- ▶ DLT-based DSAT best for supervised, worst for unsupervised
 - ▶ sensitive to supervision hypothesis errors
- ▶ MLLR-based DSAT gives a gain of 0.6% over MLLR adapted SI
- ▶ DMT-based DSAT gives a gain of 1.1% over MLLR-based DSAT



Instantaneous or Rapid Adaptation

- ▶ estimate transforms as soon as data becomes available

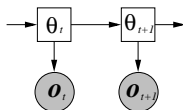


- ▶ no sufficient data for robust estimates of transforms
 - solution: Bayesian approach

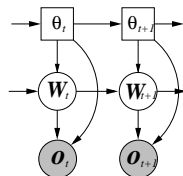


Bayesian Approach to Adaptive System

- ▶ speech state θ and transform \mathbf{W} both considered random variables
 - ▶ adaptation combined with speech state – instantaneous adaptation
 - transform constant for each acoustic condition $\mathbf{W}_t = \mathbf{W}_{t+1}$



A standard HMM DBN



An adaptive HMM DBN

- ▶ inference in adaptive systems:

$$\hat{\mathcal{H}} = \arg \max_{\mathcal{H}} \{ p(\mathbf{O}|\mathcal{H}) P(\mathcal{H}) \}$$

$$p(\mathbf{O}|\mathcal{H}) = \underbrace{\int_{\mathbf{W}} p(\mathbf{O}|\mathcal{H}, \mathbf{W}) p(\mathbf{W}|\phi) d\mathbf{W}}_{\text{intractable}}$$



a prior over transforms also imposed

- ▶ adaptation and inference – integral process though intractable
 - ▶ Viterbi decoding not possible
 - ▶ some forms of approximations required – Monte-Carlo or deterministic

Lower Bound Approximation for Bayesian Inference

- ▶ lower-bound approximation of $p(\mathbf{O}|\mathcal{H})$
 - ▶ joint variational distribution θ and transform \mathbf{W} introduced
 - ▶ Jensen's inequality applied

$$\begin{aligned} \log p(\mathbf{O}|\mathcal{H}) &= \log \int_{\mathbf{W}} p(\mathbf{O}|\mathcal{H}, \mathbf{W}) p(\mathbf{W}) d\mathbf{W} \\ &\geq \int_{\mathbf{W}} q(\theta, \mathbf{W}) \log \frac{p(\mathbf{O}, \theta|\mathbf{W}, \mathcal{H}) p(\mathbf{W})}{q(\theta, \mathbf{W})} d\mathbf{W} \end{aligned}$$

- ▶ $q(\theta, \mathbf{W}) = P(\theta|\mathbf{O}, \mathcal{H}, \mathbf{W}) p(\mathbf{W}|\mathbf{O}, \mathcal{H})$ for equality (coupled)
- ▶ Variational Bayes (VB) approximation (Yu and Gales [9])
 - ▶ θ and \mathbf{W} conditionally independent

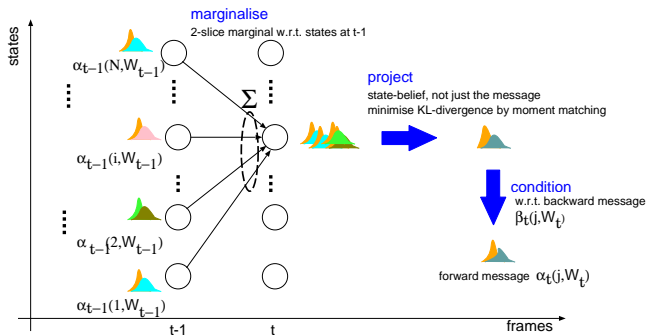
$$q(\theta, \mathbf{W}) = P(\theta|\mathbf{O}, \mathcal{H}) p(\mathbf{W}|\mathbf{O}, \mathcal{H})$$

- ▶ decoupling of θ and \mathbf{W} makes integral tractable – VB-EM
- ▶ limitations
 - ▶ rather than likelihood, gives lower-bound to likelihood – may be loose



Expectation Propagation (EP) based Inference

- ▶ loose lower bounds – bad approximation for ranking and searching
 - ▶ exact inference – infeasible due to exponential rise in mixtures
 - ▶ use EP for approximate Bayesian inference



- ▶ EP based Adaptive Inference
 - ▶ makes inference tractable by projecting mixtures to single Gaussian
 - ▶ forward-backward like formulation for iterative refinement



EP-based Forward-Backward Algorithm

1. initialise $\alpha_t(\theta_t, \mathbf{W}_t)$ and $\beta_t(\theta_t, \mathbf{W}_t)$
2. for $t=1 \dots T$, update $\alpha_t(\theta_t = j, \mathbf{W}_t)$

$$\alpha_t(j, \mathbf{W}_t) = \frac{\text{proj} \left(\sum_i \int_{\mathbf{W}_{t-1}} \frac{1}{k_t} \alpha_{t-1}(i, \mathbf{W}_{t-1}) a_{ij} \delta(\mathbf{W}_t - \mathbf{W}_{t-1}) p(\mathbf{o}_t | j, \mathbf{W}_t) \beta_t(j, \mathbf{W}_t) \right)}{\beta_t(j, \mathbf{W}_t)}$$

3. for $t=T$ to 1, update $\beta_{t-1}(\theta_{t-1} = i, \mathbf{W}_{t-1})$

$$\beta_{t-1}(i, \mathbf{W}_{t-1}) = \frac{\text{proj} \left(\sum_j \int_{\mathbf{W}_t} \frac{1}{k_t} \alpha_{t-1}(i, \mathbf{W}_{t-1}) a_{ij} \delta(\mathbf{W}_t - \mathbf{W}_{t-1}) p(\mathbf{o}_t | j, \mathbf{W}_t) \beta_t(j, \mathbf{W}_t) \right)}{\alpha_{t-1}(i, \mathbf{W}_{t-1})}$$

4. repeat steps 2 and 3, until converged.

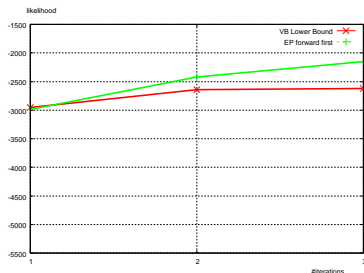
- ▶ 'proj' done by KL-divergence minimisation through moment-matching
- ▶ switching between canonical form and moment form required for distributions



EP Approximation to Marginal Likelihood

- ▶ when EP converges, likelihood computed from normalisation constants

$$p(\mathbf{O}|\mathcal{H}) \approx \prod_{t=1}^T k_t$$



- ▶ EP: not a lower-bound, not guaranteed to converge
- ▶ VB: lower-bound, convergence guaranteed
- ▶ EP computationally expensive for speech recognition
large number of matrix inversions (covariances/precision)



Maximum-a-Posteriori (MAP) Approximation

- ▶ Maximum-a-Posteriori (MAP) Approximation
 - ▶ follows from lower-bound approximation

$$\log p(\mathbf{O}|\mathcal{H}) \geq \int_{\mathbf{W}} q(\mathbf{W}) \log \frac{p(\mathbf{O}|\mathcal{H}, \mathbf{W})p(\mathbf{W})}{q(\mathbf{W})} d\mathbf{W}$$

$q(\mathbf{W}) = p(\mathbf{W}|\mathbf{O}, \mathcal{H})$ for equality

- ▶ sufficient data assumption – point estimates
 - ▶ transform posterior becomes Dirac delta function (prior still a distribution)

$$q(\mathbf{W}) = p(\mathbf{W}|\mathbf{O}, \mathcal{H}) = \delta(\mathbf{W} - \hat{\mathbf{W}})$$

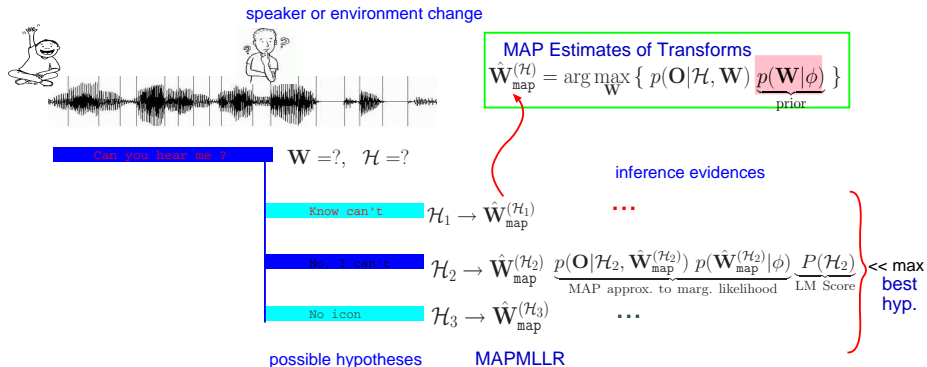
- ▶ point estimate $\hat{\mathbf{W}}$ makes integral tractable
- ▶ MAP objective function (also Chesta et al.[17])

$$\mathcal{F}_{\text{map}}(\mathbf{W}) = P(\mathbf{O}|\mathcal{H}, \mathbf{W}) \underbrace{p(\mathbf{W}|\phi)}_{\text{transform prior}}$$

- ▶ non-informative prior: MAP \Rightarrow ML estimate



N-best Based MAP Adaptive Inference



- ▶ N-best based MAP adaptive inference (Matsui & Furui [10]; Yu & Gales [9])
 - ▶ find MAP estimate $\mathbf{W}_{\text{map}}^{(\mathcal{H})}$ for each possible hypothesis
 - ▶ compute inference evidence using $\mathbf{W}_{\text{map}}^{(\mathcal{H})}$ for each hypothesis
 - ▶ rank and select the best hypothesis
- ▶ reduces hypothesis bias and can deal with insufficient amount of data



Discriminative Bayesian Adaption and Inference

- ▶ inference in discriminative adaptive system

$$\hat{\mathcal{H}} = \arg \max_{\mathcal{H}} \{P(\mathcal{H}|\mathbf{O})\} = \arg \max_{\mathcal{H}} \left\{ \int P(\mathcal{H}|\mathbf{O}, \mathbf{W}) p(\mathbf{W}|\phi) d\mathbf{W} \right\}$$

- ▶ marginalisation directly over posterior (discriminative inference evidence)
- ▶ LB cannot be found due to denominator

$$\begin{aligned} & \log \int P(\mathcal{H}|\mathbf{O}, \mathbf{W}) p(\mathbf{W}|\phi) d\mathbf{W} \\ & \geq \int \log \frac{p(\mathbf{O}|\mathcal{H}, \mathbf{W})P(\mathcal{H})}{\sum_{\tilde{\mathcal{H}}} p(\mathbf{O}|\tilde{\mathcal{H}}, \mathbf{W})P(\tilde{\mathcal{H}})} p(\mathbf{W}|\phi) d\mathbf{W} \end{aligned}$$

- ▶ use MAP approximation



Discriminative MAP Adaptive Inference

- ▶ discriminative MAP point-estimate approximation
 - ▶ MAP discriminative inference evidence

$$\begin{aligned}\hat{\mathcal{H}} &= \arg \max_{\mathcal{H}} \left\{ P(\mathcal{H}|\mathbf{O}, \mathbf{W}^{(\mathcal{H})}) p(\mathbf{W}^{(\mathcal{H})}|\phi) \right\} \\ &= \arg \max_{\mathcal{H}} \left\{ \frac{p(\mathbf{O}|\mathcal{H}, \mathbf{W}^{(\mathcal{H})})P(\mathcal{H})}{\sum_{\tilde{\mathcal{H}}} p(\mathbf{O}|\tilde{\mathcal{H}}, \mathbf{W}^{(\tilde{\mathcal{H}})})P(\tilde{\mathcal{H}})} p(\mathbf{W}^{(\mathcal{H})}|\phi) \right\}\end{aligned}$$

- ▶ denominator important, N-best list large for reasonable estimates of posteriors
- ▶ similar expressions related to MWE/MPE criteria can be derived
 - (cf. non-adaptive N-best MBR decoding in Stolke et al. [11])
- ▶ MAP discriminative transforms

$$\hat{\mathbf{W}}_{\text{dmap}}^{(\mathcal{H})} = \arg \max_{\mathbf{W}} \left\{ \frac{p(\mathbf{O}|\mathcal{H}, \mathbf{W})P(\mathcal{H})}{\sum_{\tilde{\mathcal{H}}} p(\mathbf{O}|\tilde{\mathcal{H}}, \mathbf{W})P(\tilde{\mathcal{H}})} p(\mathbf{W}|\phi) \right\}$$

- ▶ transform estimation involves discriminative criteria optimisation
 - use Extended Baum-Welch (EBW) or weak-sense auxiliary function



Estimating Discriminative MAP Transforms

- ▶ optimised by defining an auxiliary function with same gradient

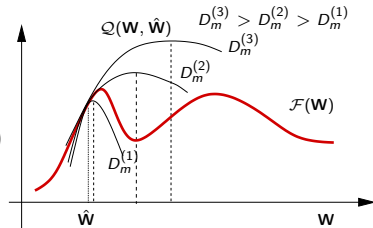
$$Q(\mathbf{W}, \hat{\mathbf{W}}) = Q^{\text{num}}(\mathbf{W}, \hat{\mathbf{W}}) - Q^{\text{den}}(\mathbf{W}, \hat{\mathbf{W}}) + Q^{\text{sm}}(\mathbf{W}, \hat{\mathbf{W}}) + Q^{\text{p}}(\mathbf{W}, \hat{\mathbf{W}})$$

- ▶ not a lower-bound due to negated denominator term
- ▶ not guaranteed to increase criteria

$$\mathcal{F}(\mathbf{W}) - \mathcal{F}(\hat{\mathbf{W}}) \not\approx Q(\mathbf{W}, \hat{\mathbf{W}}) - Q(\hat{\mathbf{W}}, \hat{\mathbf{W}})$$

- ▶ smoothing term added for stability
 - controllable by smoothing factor

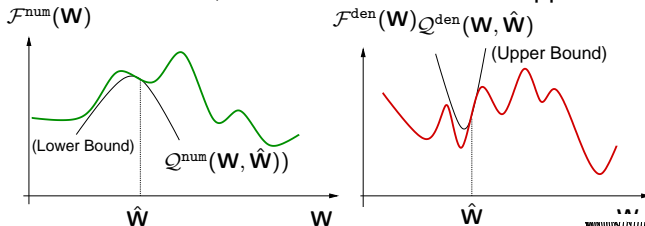
- ▶ higher smoothing factors – less but more stable update



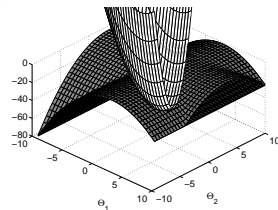
- ▶ the discriminative MAP objective function does not converge
 - ▶ due to high informative prior and its different dynamic range
 - ▶ choose higher smoothing factors, or form a lower-bound

Reverse-Jensen's Inequality for Discriminative MAP

- ▶ for an overall lower-bound, denominator should be upper-bounded



- ▶ reverse-Jensen's Inequality (Afify [12]; Jebara [13])
 - ▶ finds upper-bound to log-summation
 - ▶ applied to denominator term/mixtures
 - ▶ determines smoothing factors
 - related to first and second order statistics



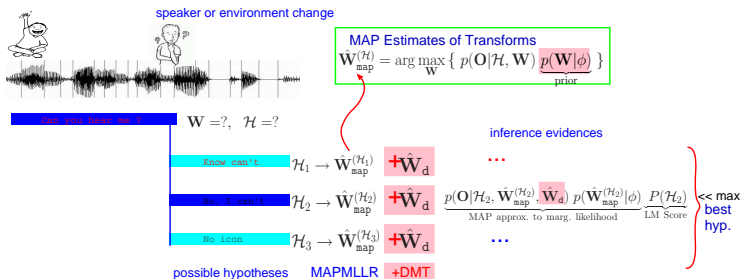
(Jebara et al.)

- ▶ gives extremely high values of smoothing factors
 - ▶ no significant update to transform parameters - unaltered rank ordering
 - ▶ possible due to loose upper-bound formed by reverse-Jensen inequality



Instantaneous Unsupervised Discriminative Adaptation

- ▶ difficult to obtain reliable discriminative MAP
 - ▶ transform ML-MAP into discriminative MAP using DMT (Bayesian + DMT)



- ▶ DMT-based Bayesian Adaptation
 - ▶ estimate MAP-MLLR for each hypothesis
 - ▶ apply DMT estimated during training
 - ▶ compute inference evidences using MAPMLLR+DMT
- ▶ DMT and VB can be also integrated

Instantaneous Utterance-Level Adaptation

- ▶ utterance-level adaptation and decoding (150-best list)

System	Transform		WER%
	Training	Testing	
SI/SI(hyp)	-	-	29.2
SI	-	MLLR	32.4
		MAPMLLR	29.0
		MAPMLLR+DMT	28.4
SAT	MLLR	MLLR	32.3
		MAPMLLR	28.8
		MAPMLLR+DMT	28.6

- ▶ significant gain compared to MAPMLLR and other systems
 - ▶ reduced gain as DMT estimated using speaker-level transforms applied to utterance level
 - ▶ SAT systems affected more by mismatch as transforms play more important role in them



Conclusion

- ▶ Adaptation and Adaptive Training
 - ▶ useful for dealing with speaker variations and non-homogeneous data
- ▶ Discriminative Adaptation and Adaptive Training
 - ▶ DMT based DSAT
 - can deal unsupervised adaptation/hypothesis-bias problem
 - significant improvement over standard systems
 - ▶ 1.1% absolute gain over standard MLLR-based DSAT
- ▶ Bayesian Adaptation and Inference
 - can deal data sparsity and hypothesis-bias problems
 - ▶ Bayesian adaptive inference
 - ML and discriminative
 - ▶ discriminative MAP transforms
 - issues
 - ▶ integration of DMT into Bayesian framework
 - significant improvement over conventional systems



References

- [1] C. K. Raut and M.J.F. Gales. Bayesian discriminative adaptation for speech recognition. In *Proc. ICASSP*, 2009.
- [2] C. K. Raut, K. Yu, and M.J.F. Gales. Adaptive training using discriminative mapping transforms. In *Proc. Interspeech*, 2008.
- [3] M. J. F. Gales, F. Diehl, C. K. Raut, M. Tomalin, P. C. Woodland, and K. Yu. Development of a phonetic system for large vocabulary arabic speech recognition. In *Proc. ASRU*, 2007.
- [4] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9:171–186, 1995.
- [5] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.
- [6] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul. A compact model for speaker adaptive training. In *Proc. ICSLP*, pages 1137–1140, 1996.
- [7] S. Matsoukas, R. Schwartz, H. Jin, and L. Nguyen. Practical implementations of speaker adaptive training. *Proc. DARPA Speech Recognition Workshop*, 1997.
- [8] K. Yu, M. J. F. Gales, and P. C. Woodland. Unsupervised discriminative adaptation using discriminative mapping transforms. In *Proc. ICASSP*, pages 4273–4276, 2008.
- [9] K. Yu and M. J. F. Gales. Bayesian adaptive inference and adaptive training. *IEEE Transactions on Audio, Speech and Language Processing*, 15(6):1932–1943, 2007.
- [10] T. Matsui and S. Furui. N-Best-based unsupervised speaker adaptation for speech recognition. *Computer Speech and Language*, 12:41–50, 1998.
- [11] A. Stolke, Y. Konig, and M. Weintraub. Explicit Word Error Minimization in N-Best List rescoring. In *Proc. Eurospeech*, 1997.
- [12] M. Afify. Extended Baum-Welch reestimation of Gaussian mixture models based on reverse Jensen inequality. In *Proc. Interspeech*, pages 1113–1116, 2005.
- [13] T. Jebara. *Discriminative, Generative and Imitative Learning*. PhD thesis, Massachusetts Institute of Technology, 2002.
- [14] A. Ljolje. The AT&T LVCSR-2001 system. In *Proc. the NIST LVCSR Workshop*, NIST, 2001.
- [15] S. Tsakalidis, V. Doumptiotis, and W. Byrne. Discriminative linear transforms for feature normalisation and speaker adaptation in HMM estimation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 13(3):367–376, 2005.
- [16] L. Wang. *Discriminative linear transforms for adaptation and adaptive training*. PhD thesis, Cambridge University, 2006.
- [17] C. Chesta, O. Siohan, and C. Lee. Maximum a posteriori linear regression for hidden Markov model adaptation. *Proc. EuroSpeech*, 1:211–214, 1999.
- [18] T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 216–223, 2002.

