Instantaneous and Discriminative Adaptation for Automatic Speech Recognition

Mark Gales with Kai Yu and CK Raut

August 2008

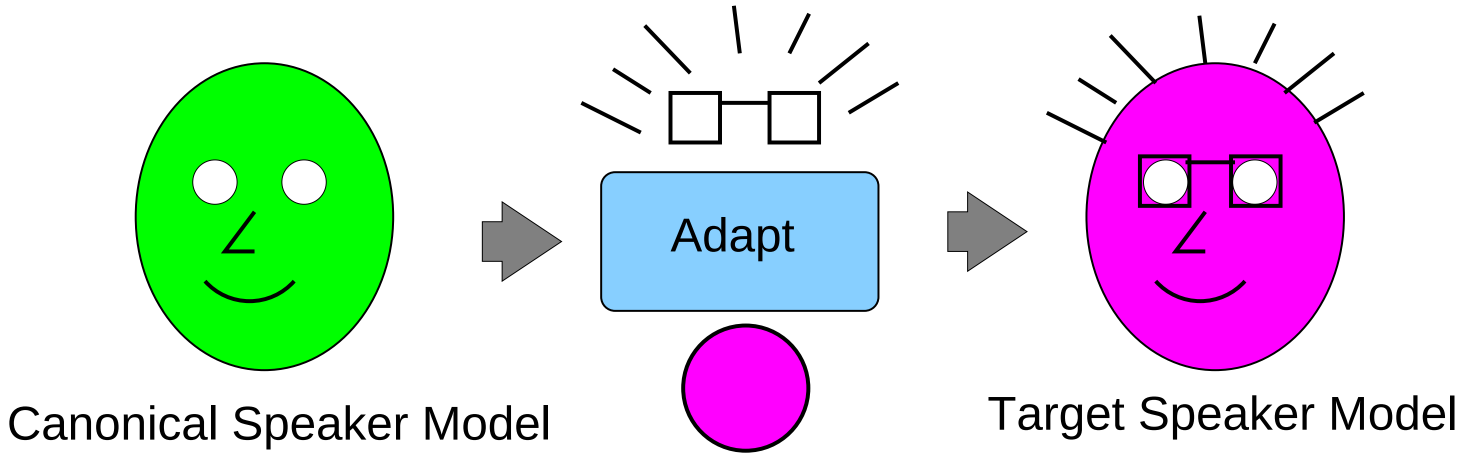
Cambridge University Engineering Department

Outline

- Adaptive Training
  - linear transform-based adaptation
  - ML and MAP estimation
  - adaptive training
- Instantaneous Adaptation
  - Bayesian adaptive training and inference
  - variational Bayes approximation
- Discriminative Mapping Transforms
  - discriminative transforms
  - discriminative adaptive training
- Current adaptive training research
  - combining for instantaneous discriminative adaptation

General Adaptation Process

- Aim: Modify a "canonical" model to represent a target speaker
  - transformation should require minimal data from the target speaker
  - adapted model should accurately represent target speaker



- Need to determine
  - nature (and complexity) of the speaker transform
  - how to train the "canonical" model that is adapted

Form of the Adaptation Transform

- There are a number of standard forms in the literature
  - Gender-dependent, MAP, EigenVoices, CAT ...
- Dominant form for LVCSR are ML-based linear transformations
  - MLLR adaptation of the means

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

  - MLLR adaptation of the covariance matrices

$$\boldsymbol{\Sigma}^{(s)} = \mathbf{H}^{(s)}\boldsymbol{\Sigma}\mathbf{H}^{(s)\mathsf{T}}$$

  - Constrained MLLR adaptation

$$\boldsymbol{\mu}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\mu} + \mathbf{b}^{(s)}; \quad \boldsymbol{\Sigma}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\Sigma}\mathbf{A}^{(s)\mathsf{T}}$$

ML and MAP Linear Transforms

- Transforms often estimated using ML (with hypothesis $\mathcal{H}$)

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

  - where $\mathbf{W}_{\tt ml}^{(s)} = \begin{bmatrix} \mathbf{A}_{\tt ml}^{(s)} & \mathbf{b}_{\tt ml}^{(s)} \end{bmatrix}$
  - however not robust to limited training data

- Including transform prior, $p(\mathbf{W})$, to get MAP estimate

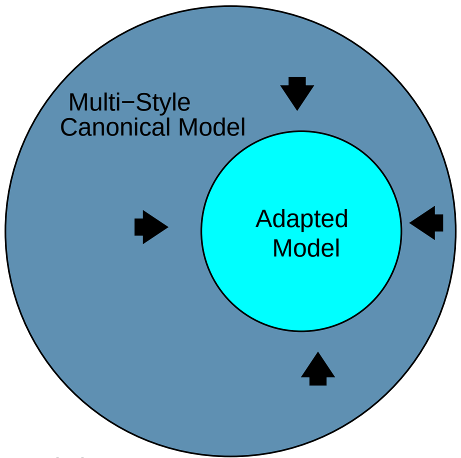
$$\mathbf{W}_{\tt map}^{(s)} = \arg\max_{\mathbf{W}} \left\{ p(\mathbf{O}^{(s)}|\mathcal{H};\mathbf{W})p(\mathbf{W}) \right\}$$

  - for MLLR Gaussian is a Gaussian prior for the auxiliary function
  - CMLLR prior more challenging ...

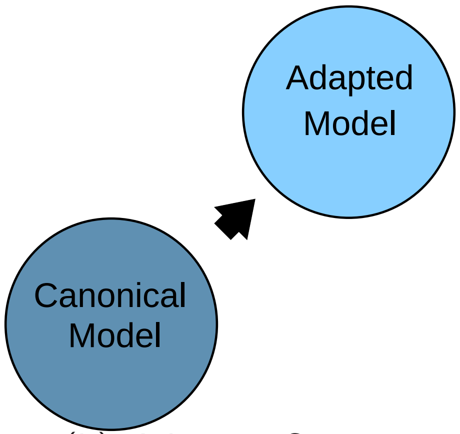
- Both approaches rely on expectation-maximisation (EM)

Training a "Good" Canonical Model

- Standard "multi-style" canonical model
  - treats all the data as a single "homogeneous" block
  - model represents acoustic realisation of phones/words (desired)
  - and acoustic environment, speaker, speaking style variations (unwanted)



(a) Multi-Style System

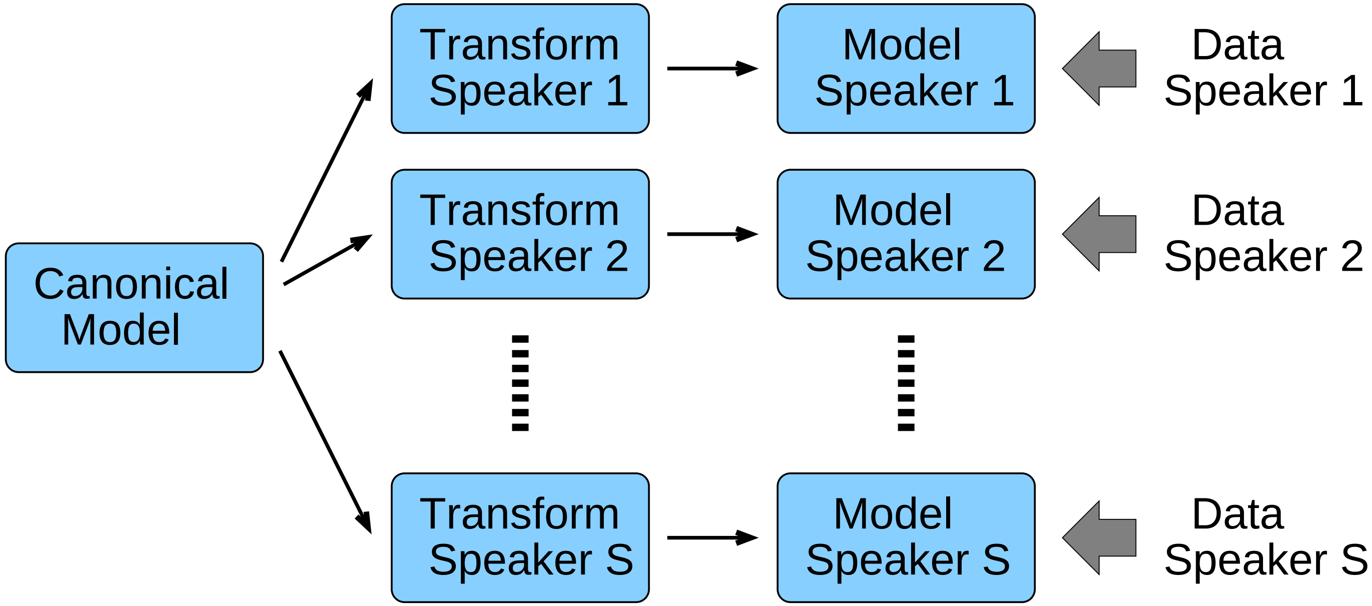


(b) Adaptive System

Two different forms of canonical model:

- Multi-Style: adaptation converts a general system to a specific condition;
- Adaptive: adaptation converts a "neutral" system to a specific condition

Adaptive Training



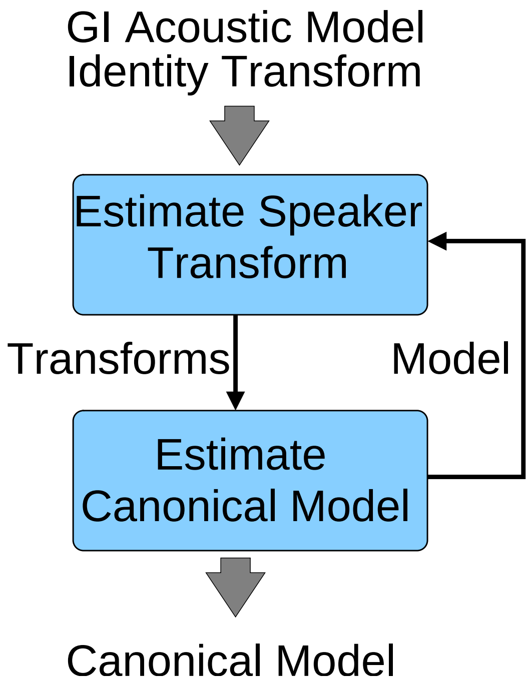
- In adaptive training the training corpus is split into "homogeneous" blocks
  - use adaptation transforms to represent unwanted acoustic factors
  - canonical model only represents desired variability
- All forms of linear transform can be used for adaptive training
  - CMLLR adaptive training highly efficient

CMLLR Adaptive Training

- The CMLLR likelihood may be expressed as:

$$\mathcal{N}(\mathbf{o}; \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathsf{T}}) = \frac{1}{|\mathbf{A}|}\mathcal{N}(\mathbf{A}^{-1}\mathbf{o} - \mathbf{A}^{-1}\mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

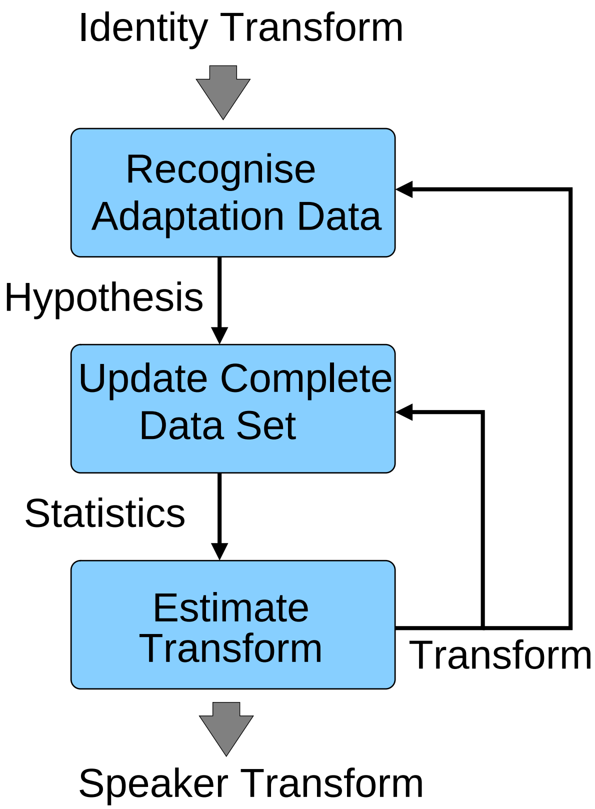
same as feature normalisation - simply train model in transformed space



- Interleave Model and transform estimation
- Adaptive canonical model not suited for unadapted initial decode
  - GI model used for initial hypothesis
- MLLR less efficient, but reasonable
  - MLLR is used in this work

Unsupervised Linear Transformation Estimation

- Estimation of all the transforms is based on EM:
  - requires the transcription/hypothesis of the adaptation data
  - iterative process using "current" transform to estimate new transform

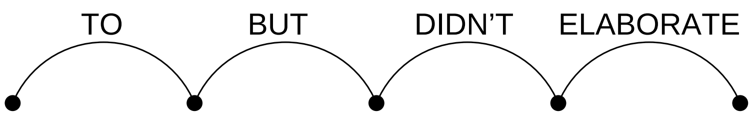


- Two iterative loops for estimation:
  1. estimate hypothesis given transform
  2. update complete-dataset given transform and hypothesis

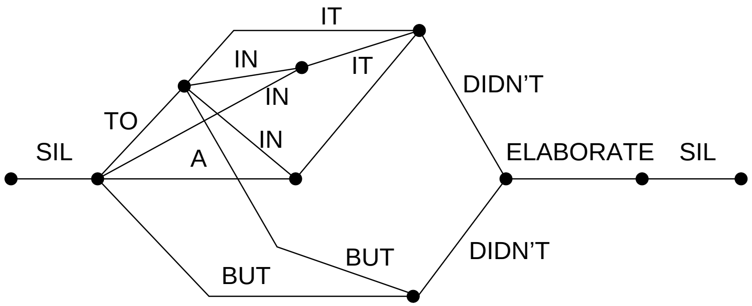
  referred to as Iterative MLLR
- For supervised training hypothesis is known
- Can also vary complexity of transform with iteration

Lattice-Based MLLR

- For unsupervised adaptation hypothesis will be error-full
- Rather than using the 1-best transcription and iterative MLLR
  - generate a lattice when recognising the adaptation data
  - accumulate statistics over the lattice (Lattice-MLLR)



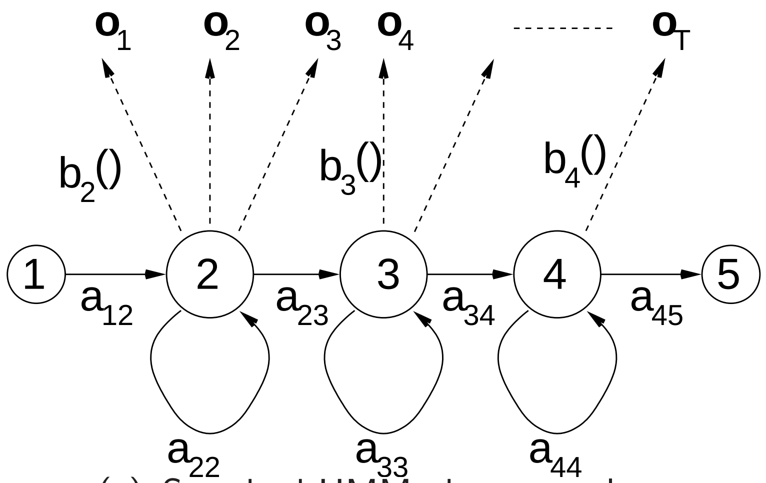
1-best transcription



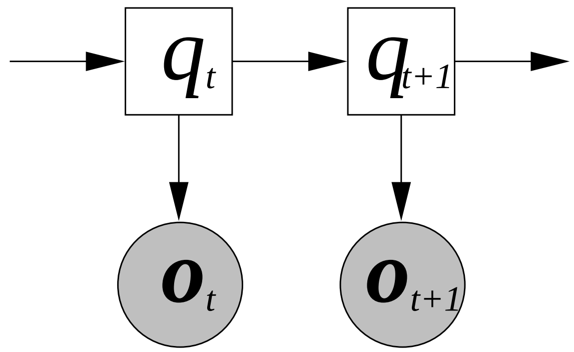
Word lattice

- The accumulation of statistics is closely related to obtaining denominator statistics for discriminative training
- No need to re-recognise the data
  - iterate over the transform estimation using the same lattice

Hidden Markov Model - A Dynamic Bayesian Network



(c) Standard HMM phone topology



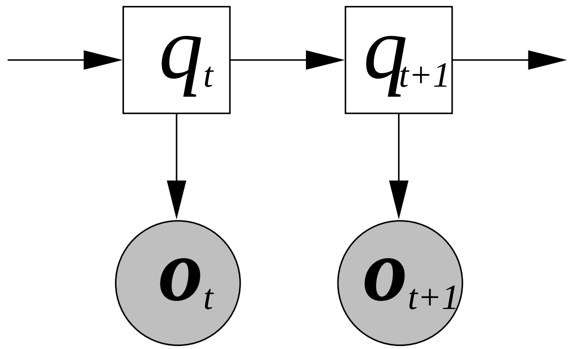
(d) HMM Dynamic Bayesian Network

- Notation for DBNs:

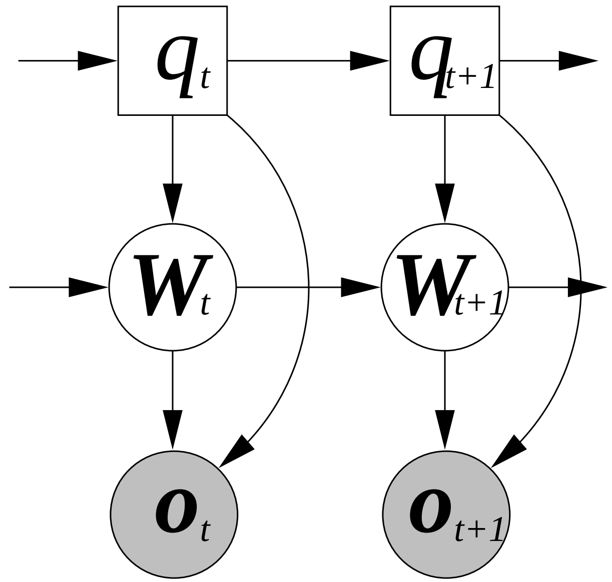
circles - continuous variables	shaded - observed variables
squares - discrete variables	non-shaded - unobserved variables

- Observations conditionally independent of other observations given state.
- States conditionally independent of other states given previous states.
- Poor model of the speech process - piecewise constant state-space.

Adaptive Training From Bayesian Perspective



(e) Standard HMM



(f) Adaptive HMM

- Observation additionally dependent on transform $\mathbf{W}_t$
  - transform same for each homogeneous block ($\mathbf{W}_t = \mathbf{W}_{t+1}$)
  - adaptation integrated into acoustic model - instantaneous adaptation
- Need to known the prior transform distribution $p(\mathbf{W})$ (as in MAP scheme)

Inference with Adaptive HMMs

- Acoustic score - marginal likelihood of the whole sequence, $\mathbf{O} = \mathbf{o}_1, \ldots, \mathbf{o}_T$
  - still depends on the hypothesis $\mathcal{H}$
  - point-estimate canonical parameters (standard complexity control schemes)

$$
\begin{aligned}
p(\mathbf{O}|\mathcal{H}) &= \int_{\mathbf{W}} p(\mathbf{O}|\mathcal{H}, \mathbf{W}) p(\mathbf{W})\, d\mathbf{W} \\
&= \int_{\mathbf{W}} \sum_{\mathbf{q} \in \mathbf{Q}^{(\mathcal{H})}} P(\mathbf{q}) \prod_{t=1}^{T} \mathcal{N}(\mathbf{o}_t; \mathbf{A}\boldsymbol{\mu}^{(q_t)} + \mathbf{b}, \boldsymbol{\Sigma}^{(q_t)}) p(\mathbf{W})\, d\mathbf{W}
\end{aligned}
$$

- Latent variables makes exact inference impractical
  - need to sum over all possible state-sequences explicitly
  - Viterbi decoding not possible to find bets hypothesis
- Need schemes to handle both these problems

Lower Bound Approximation

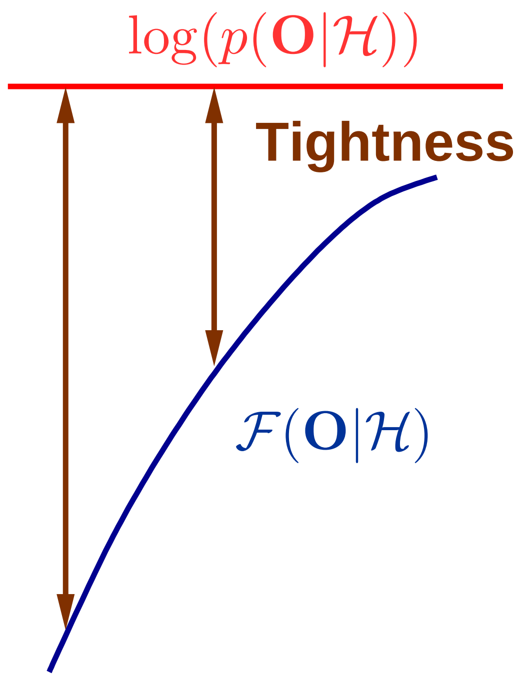
- Lower bound to log marginal likelihood using Jensen's inequality
  - introduce variational distribution $f(\mathbf{q}, \mathbf{W}|\mathcal{H})$, then [1]

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

- Equality in the above when: $f(\mathbf{q}, \mathbf{W}|\mathcal{H}) = P(\mathbf{q}, \mathbf{W}|\mathbf{O}, \mathcal{H})$
  - unfortunately this is impractical
  - need approximation that is as close as possible

Tightness of Lower Bound

- Tightness of lower bound will affect inference
  - want the bound to be as tight as possible
  - write $\log(p(\boldsymbol{O}|\mathcal{H})) \geq \mathcal{F}(\boldsymbol{O}|\mathcal{H})$ where $f(\mathbf{q}, \mathbf{W}|\mathcal{H})$ determines $\mathcal{F}(\boldsymbol{O}|\mathcal{H})$



- EM-like algorithm possible
  - iterative approach
  - more iterations - tighter bounds
- Forms of lower bound
  - point estimate - loose
  - variational Bayes - tighter bound

Point Estimate Lower Bound

- Variation distribution can be approximated by a point -estimate
  - has the form of a Dirac-delta function $\delta(\mathbf{W} - \hat{\mathbf{W}})$

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

- Basically assume that the transform posterior is a point estimate

$$P(\mathbf{W}|\mathbf{O}, \mathcal{H}) \approx \delta(\mathbf{W} - \hat{\mathbf{W}})$$

  - two forms of point estimate possible: MAP, or ML estimates
  - issues of robust transform estimation
- Theoretical motivation for ML/MAP linear transforms
  - bound is very loose (infinitely large)

Variational Bayes Lower Bound

- Useful to modify variational approximation to yield tighter bound
  - need to have a distribution over the transform distribution
- Assume that the state and transform distributions are conditionally independent

$$f(\mathbf{q},\mathbf{W}|\mathcal{H}) = f(\mathbf{q}|\mathcal{H})f(\mathbf{W}|\mathcal{H})$$

  - decoupling of $\mathbf{q}$ and $\mathbf{W}$ posteriors makes integral tractable
  - more robust than point transform estimate as distribution used
- Variational distribution $f(\mathbf{W}|\mathcal{H})$ used to calculate $\mathcal{F}(\mathbf{O}|\mathcal{H})$

$$\begin{aligned}\mathcal{F}(\mathbf{O}|\mathcal{H}) &= \log\left(\sum_{\mathbf{q}\in\mathbf{Q}^{(\mathcal{H})}} P(\mathbf{q})\prod_{t=1}^{T}\tilde{p}(\mathbf{o}_t|q_t)\right) - \mathtt{KL}(f(\mathbf{W}|\mathcal{H})||p(\mathbf{W}))\\ \tilde{p}(\mathbf{o}_t|q_t) &= \exp\left(\int_{\mathbf{W}} \log(p(\mathbf{o}_t|\mathbf{W},q_t)f(\mathbf{W}|\mathcal{H})d\mathbf{W}\right)\end{aligned}$$

Bayesian Inference Approximations

- So far assumed that hypothesis is given
  - in practice inference used to determine hypothesis
  - likelihood-based inference

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

  - lower-bound inference - "practical" approximation

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} \left\{ \mathcal{F}(\mathbf{O}|\mathcal{H}) + \log(P(\mathcal{H})) \right\}$$

- As using lower-bound approximation $\log(p(\mathbf{O}|\mathcal{H})) \geq \mathcal{F}(\mathbf{O}|\mathcal{H})$
  - assumes that lower-bound ranking is the same as the likelihood
  - strong motivation for making bound as tight as possible

N-Best Supervision

- Variational approximation is a function of the hypothesis (for VB)

$$f(\mathbf{q}, \mathbf{W}|\mathcal{H}) = f(\mathbf{q}|\mathcal{H})f(\mathbf{W}|\mathcal{H})$$

- 1-Best supervision - standard adaptation, variational approximation based on

$$f(\mathbf{q}, \mathbf{W}|\mathcal{H}^{(n)}) \approx f(\mathbf{q}, \mathbf{W}|\mathcal{H}^{(1)}) = f(\mathbf{q}|\mathcal{H}^{(1)})f(\mathbf{W}|\mathcal{H}^{(1)})$$

  - same variational approximation used for all hypotheses, $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(N)}$
  - biases the output to the supervision (standard problem)

- N-Best supervision - use different variational approximation for each hypothesis
  - variational approximation to determine $\mathcal{F}(\mathbf{O}|\mathcal{H}^{(n)})$ is

$$f(\mathbf{q}, \mathbf{W}|\mathcal{H}^{(n)}) = f(\mathbf{q}|\mathcal{H}^{(n)})f(\mathbf{W}|\mathcal{H}^{(n)})$$

  - tighter-bound than 1-best supervision
  - removes bias to 1-best supervision

N-Best Implementation

- Practical implementation based on N-best list
  1. Generate N-best list using baseline models: $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(N)}$
  2. Foreach of the N-hypotheses, $\mathcal{H}^{(n)}$:
     (a) compute variational approximation to yield $f(\mathbf{W}|\mathcal{H}^{(n)})$
     (b) compute $\mathcal{F}(\mathbf{O}|\mathcal{H}^{(n)})$
  3. Rank hypotheses using $\mathcal{F}(\mathbf{O}|\mathcal{H}^{(n)}) + \log(P(\mathcal{H}^{(n)}))$
- Simple example based on N-best list: `bat`, `fat`, `mat`

Exact Evidence	Exact	Supervision	
		1-Best	N-Best
$p(\mathbf{O}\|\texttt{bat})P(\texttt{bat})$	0.88	0.66	0.80
$p(\mathbf{O}\|\texttt{fat})P(\texttt{fat})$	0.84	0.78	0.78
$p(\mathbf{O}\|\texttt{mat})P(\texttt{mat})$	0.80	0.68	0.74

  – 1-best supervision is `fat` (same as 1-best supervision output)
  – N-best supervision output is `bat` (correct answer!!!)

Experiments on Conversational Telephone Speech Task

- Switchboard (English): conversational telephone speech task
  - Training dataset: about 290hr, 5446spkr; Test dataset: 6hr, 144spkr
  - Front-end: PLP+Energy+1^{st},2^{nd},3^{rd} derivatives, HLDA and VTLN used
  - 16 Gaussian components per state systems; state clustered triphones
  - 150-Best list rescoring in Bayesian inference (utterance-level) experiments
- Acoustic models configurations investigated
  - ML and MPE speaker independent (SI) system - baseline models
  - MLLR based speaker adaptive training (SAT) - ML and MPE version
  - transform prior distribution - single Gaussian distribution
  - MPE-SAT only discriminatively update the canonical model
- Performance investigated at an two-level
  - utterance level for instantaneous adaptation
  - side/speaker level for unsupervised adaptation

Utterance Level Bayesian Adaptation - ML

Bayesian Approx	ML Train	
	SI	SAT
—	32.8	—
ML	35.5	35.2
MAP	32.2	31.8
VB	31.8	31.5

- All experiments use N-best supervision
  - ML adaptation much worse than SI - insufficient adaptation data
  - MAP yields robust estimates - performance gains over ML
  - VB yields additional gains over MAP
- SAT performance better than SI performance
  - gains from adaptive HMM 1.3% absolute over SI baseline
  - integrated adaptation seems to be useful (though implementation an issue)

Lower Bound Tightness - N-Best Supervision

- Investigate gains of using N-best rather than 1-best supervision
  - investigated using ML-SAT models

Bayesian Approx.	Supervision	
	N-Best	1-Best
MAP	31.8	32.0
VB	31.5	32.0

- N-Best supervision significantly better than 1-Best supervision
- VB approximation more sensitive to use of N-best supervision
  - expected as VB approximation more powerful than point estimate
  - bias due to 1-best supervision has an impact

Utterance Level Bayesian Adaptation - MPE

Bayesian Approx	MPE Train	
	SI	SAT
—	29.2	—
ML	32.4	32.3
MAP	29.0	28.8
VB	28.8	28.6

- Similar trends for lower bound approximation as ML case
  - VB $>$ MAP $>$ SI $>$ ML
  - gains compared to ML acoustic models reduced (for VB 0.6% vs 1.3%)
- Reason for reduced gain compared to ML systems
  - prior distribution estimated on ML transforms
  - prior applied in a non-discriminative fashion

Discriminative Linear Transforms

- Linear transforms can be trained using discriminative criteria
  - estimation using minimum phone error (MPE) training

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

- For unsupervised adaptation discriminative linear transforms (DLTs) not used
  - estimation highly sensitive to errors in supervision hypothesis
  - more costly to estimate transform than ML training
- Not used for discriminative SAT, standard procedure
  1. perform standard ML-training (ML-SI)
  2. perform ML SAT training updating models and transforms (ML-SAT)
  3. estimate MPE-models given the ML-transforms (MPE-SAT)

Discriminative Mapping Functions

- Would like to get aspects of discriminative transform without the problems:
  - train all speaker-specific parameters in using ML training
  - train speaker-independent parameters in using MPE training
- Applying this to linear transforms yields (as one option) [2]

$$\begin{aligned}\boldsymbol{\mu}^{(s)} &= \mathbf{A}_{\mathtt{d}}\left(\mathbf{A}_{\mathtt{ml}}^{(s)}\boldsymbol{\mu} + \mathbf{b}_{\mathtt{ml}}^{(s)}\right) + \mathbf{b}_{\mathtt{d}} \\ &= \mathbf{A}_{\mathtt{d}}\boldsymbol{\mu}_{\mathtt{ml}}^{(s)} + \mathbf{b}_{\mathtt{d}}\end{aligned}$$

  - $\mathbf{W}_{\mathtt{ml}}^{(s)} = \begin{bmatrix}\mathbf{A}_{\mathtt{ml}}^{(s)} & \mathbf{b}_{\mathtt{ml}}^{(s)}\end{bmatrix}$ - speaker-specific ML transform
  - $\mathbf{W}_{\mathtt{d}} = \begin{bmatrix}\mathbf{A}_{\mathtt{d}} & \mathbf{b}_{\mathtt{d}}\end{bmatrix}$ - speaker-independent MPE transform
- Yields a composite discriminative-like transform

$$\mathbf{A}_{\mathtt{d}}^{(s)} = \mathbf{A}_{\mathtt{d}}\mathbf{A}_{\mathtt{ml}}^{(s)}; \quad \mathbf{b}_{\mathtt{d}}^{(s)} = \mathbf{A}_{\mathtt{d}}\mathbf{b}_{\mathtt{ml}}^{(s)} + \mathbf{b}_{\mathtt{d}}$$

Training DMTs

- This form of DMT results in the following estimation criterion

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

  - posterior $P(\mathcal{H}|\mathbf{O}^{(s)}; \mathbf{W}, \mathbf{W}_{\mathtt{ml}}^{(s)})$ based on speaker ML-adapted models
  - supervised training of discriminative transform
- Standard DLT update formulae can be used
- Quantity of training data vast compared to available speaker-specific data
  - use large number of base-classes
  - in these experiments 1000 base-classes used
- Can also be used for discriminative adaptive training [3]

DMT Speaker Level Adaptation - ML

- Use ML-trained models but side (speaker) level adaptation

Adaptation	ML Train	
	SI	SAT
—	32.6	—
MLLR	30.2	29.3
MLLR+DMT	27.9	27.5

- Large gains from MLLR+DMT over standard MLLR
  - 2.3% absolute reduction for SI models
- Gains using SAT models slightly less
  - 1.8% absolute reduction in error rate

DMT Speaker Level Adaptation - MPE

- Use SI-MPE models - again side (speaker) level adaptation

Adaptation	Supervision		
	1-Best	Lattice	Reference
—	29.2	—	—
MLLR	27.0	26.7	24.3
MLLR+DMT	26.2	25.9	23.4
DLT	26.8	26.6	21.7

- DMTs show consistent significant gains over standard MLLR adaptation
  - lattice-based MLLR shows gains over 1-best
- DLTs show sight gains over MLLR using both 1-best and lattices
  - performance biased to reference (or hypothesis)

DMT for Discriminative Adaptive Training

- Three versions of Discriminative SAT (DSAT) evaluated
  - transforms: MLLR (standard), DLT and MLLR+DMT
  - MPE use to train canonical model

Scheme	Training	Testing	WER
SI	—	—	29.2
	—	MLLR	27.0
	—	MLLR+DMT	26.2
DSAT	MLLR	MLLR	26.4
	DLT	DLT	28.1
	MLLR+DMT	MLLR+DMT	25.3

- DMTs useful for discriminative adaptive training
  - problems with using DLTs for unsupervised adaptation

Discriminative Instantaneous Adaptation

- Interesting to try discriminative versions of instantaneous adaptation
- Using MAP in combination with, for example, MPE difficult
  - "weak"-sense and "strong"-sense auxiliary functions don't combine well
  - implementation of DLT-MAP awkward ...
- DMTs can be directly applied to the Bayesian inference framework
  - currently only applied to the MAP Bayesian approximation
  - no theoretical issue with the VB approximation
- DMTs from speaker level adaptation used
  - known mis-match with the utterance level MAP transforms

DMT Utterance Level Bayesian Adaptation

Bayesian Approx	MPE Train	
	SI	SAT
—	29.2	—
ML	32.4	32.3
MAP	29.0	28.8
MAP+DMT	28.4	28.6

- For the SI models DMTs show gains over MAP approximation
  - gains slightly smaller than for speaker-level 0.6% vs 0.8%
- SAT gains disappointing (0.2% compared to 0.8%)
  - SAT expected to be more sensitive to transform errors
  - DMT estimated on a speaker-level

Summary

- Described two approaches and their combination
  - Bayesian adaptive training/inference for instantaneous adaptation
  - discriminative mapping transforms for robust "discriminative" transforms
- Instantaneous adaptation and interesting direction
  - current approximations impractical (N-best list rescoring)
  - examining alternative approximations (Gibbs sampling EP etc)
- DMTs show gains over standard ML and discriminative transforms
  - easy to train and implement
  - currently looking to work with CMLLR (mainly implementation)
- Combination dependent on sorting out both!
- Still disappointing gains from adaptive training
  - need to look at combinations of transforms (acoustic factorisation [4])

References

[1] K. Yu and M. Gales, “Bayesian adaptive inference and adaptive training,” *IEEE Transactions Speech and Audio Processing*, vol. 15, no. 6, pp. 1932–1943, August 2007.

[2] K. Yu, M. Gales, and P. Woodland, “Unsupervised discriminative adaptation using discriminative mapping transforms,” in *Proc. ICASSP*, 2008.

[3] C. Raut, K. Yu, and M. Gales, “Adaptive training using discriminative mapping transforms,” in *Proc. InterSPeech*, 2008.

[4] M. Gales, “Acoustic factorisation,” in *Proc. ASRU*, 2001.