

Instantaneous Unsupervised Adaptation using Discriminative Mapping Transforms

C. K. Raut, Kai Yu and Mark Gales

16th July 2008



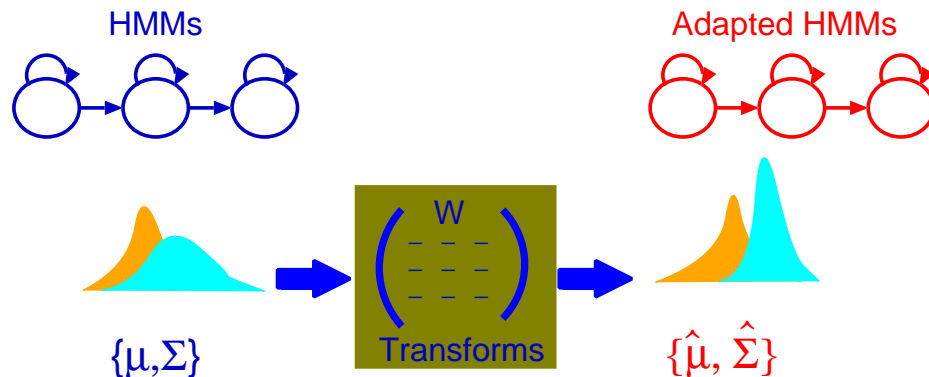
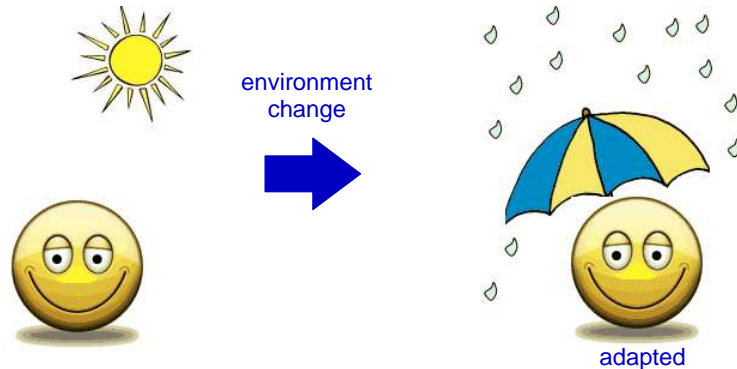
Cambridge University Engineering Department

Overview

- Speaker or Environmental Adaptation
- Unsupervised Adaptation
 - problems of unsupervised discriminative adaptation
 - solution by discriminative mapping transforms (DMTs)
- Instantaneous Adaptation
 - problems of instantaneous adaptation
 - solution by Bayesian approach
- Instantaneous Unsupervised Discriminative Adaptation
 - combine Bayesian approach with DMTs
 - experimental evaluation



Speaker or Environmental Adaptation



- Speech recognition systems work under different environment or speakers
- Speaker or environment change requires model adaptation
 - Linear transforms are widely used to adapt HMM parameters
- Mean transform:

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

$W = [A \ b]$: affine transform

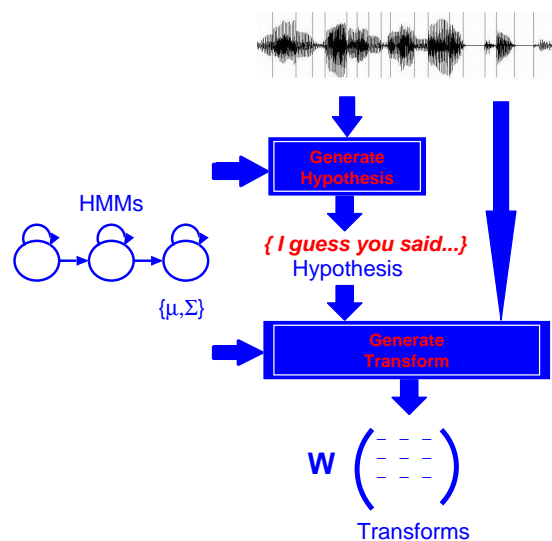
Unsupervised Adaptation

- Transform estimation requires supervision for given speech observation
- No supervision hypothesis available for unsupervised adaptation



Unsupervised Adaptation

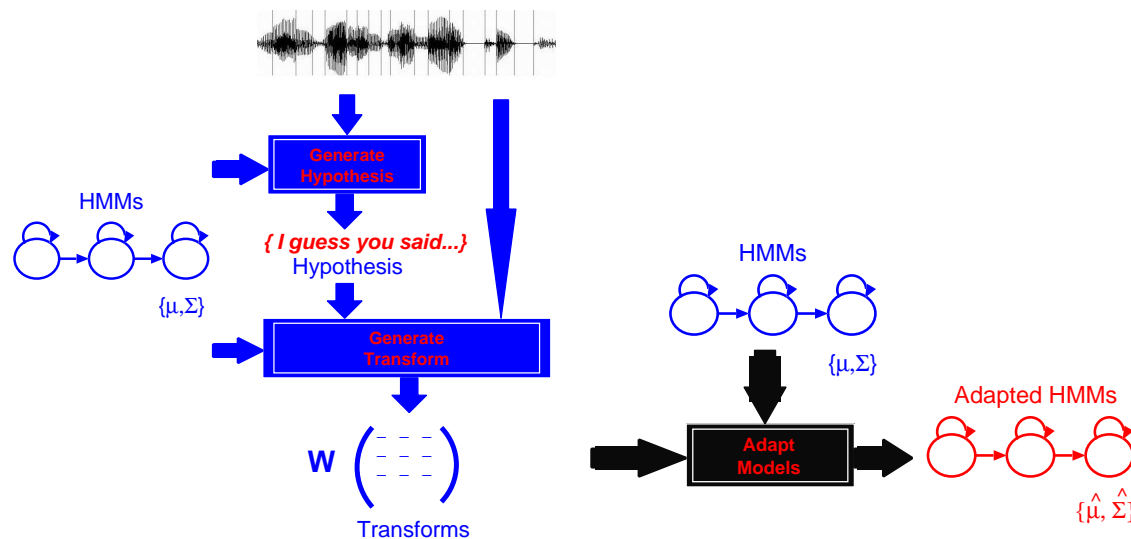
- Transform estimation requires supervision for given speech observation
- No supervision hypothesis available for unsupervised adaptation



(1) Generate 1-best hypothesis and estimate transforms

Unsupervised Adaptation

- Transform estimation requires supervision for given speech observation
- No supervision hypothesis available for unsupervised adaptation

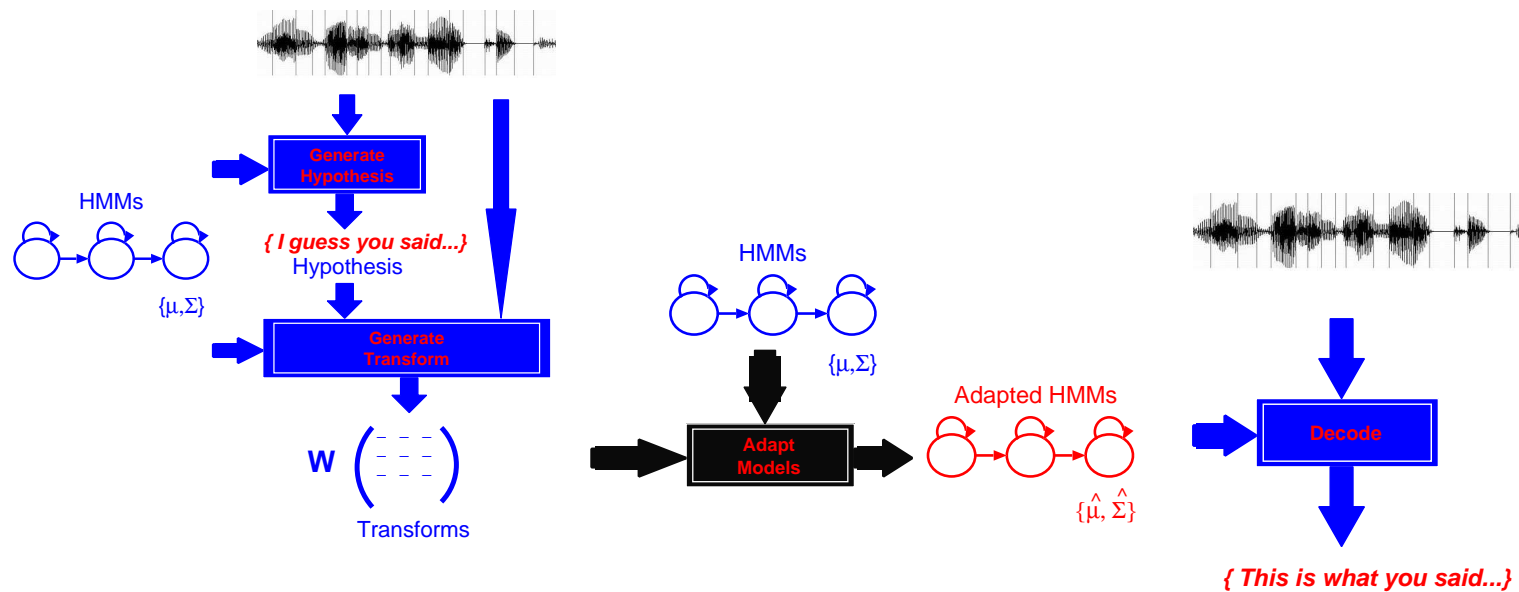


- (1) Generate 1-best hypothesis and estimate transforms
- (2) Adapt HMMs



Unsupervised Adaptation

- Transform estimation requires supervision for given speech observation
- No supervision hypothesis available for unsupervised adaptation



(1) Generate 1-best hypothesis and estimate transforms

(2) Adapt HMMs

(3) Decode with adapted models



Maximum Likelihood and Discriminative Transforms

- Maximum-likelihood Linear Regression (MLLR)

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

- Discriminative Linear Transforms (DLT)
 - Use discriminative criteria such as Minimum Phone Error (MPE)

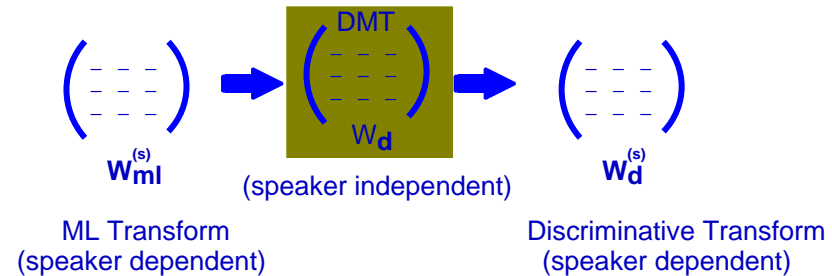
$$\mathbf{W}_{\text{d}}^{(s)} = \arg \min_{\mathbf{W}} \left\{ \underbrace{\sum_{\mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}, \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)})}_{\text{expected phone error}} \right\}$$

- Discriminative transforms give significant gain in supervised adaptation mode, but unsupervised adaptation with them is a problem
- **Problem:** Discriminative transforms are very sensitive to errors in supervision
 - **Solution:** Discriminative Mapping Transforms (DMT)



Discriminative Mapping Transforms (DMT)

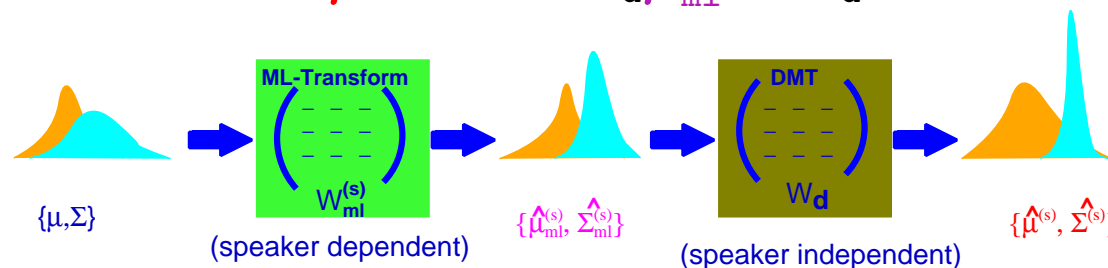
- speaker-independent discriminative transform
- maps speaker-dependent ML-transforms to discriminative transforms
- a simplified form of DMT used



– mean is adapted using MLLR $W_{ml}^{(s)} = [A_{ml}^{(s)} \quad b_{ml}^{(s)}]$, followed by DMT $W_d = [A_d \quad b_d]$

$$\hat{\mu}_{ml}^{(s)} = A_{ml}^{(s)} \mu + b_{ml}^{(s)}$$

$$\hat{\mu}^{(s)} = A_d \hat{\mu}_{ml}^{(s)} + b_d$$



Discriminative Mapping Transforms (contd.)

- DMT estimation: similar to DLT but uses training data from all speakers

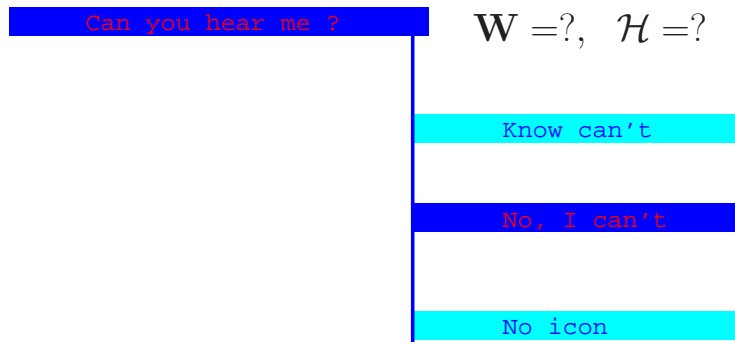
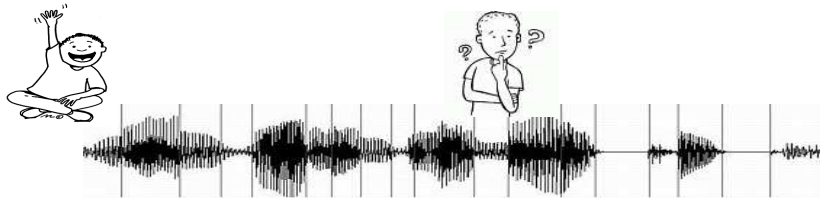
$$\mathbf{W}_d = \arg \min_{\mathbf{W}} \left\{ \sum_s \sum_{\mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}, \mathbf{W}_{ml}^{(s)}, \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}$$

- Estimated during training
- No need to re-estimate during testing
- Not sensitive to errors in supervision hypothesis
- Possible to estimate large number of transforms (as they are estimated from large training dataset)
- ICASSP'08 Paper (Kai et al.): “Unsupervised discriminative adaptation using discriminative mapping transforms”



Instantaneous Unsupervised Adaptation

speaker or environment change



possible hypotheses

- The best hypothesis is searched as

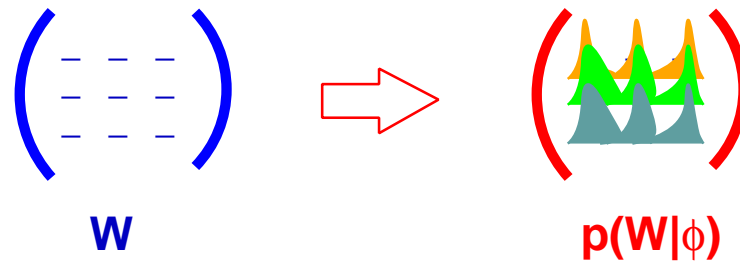
$$\hat{\mathcal{H}} = \operatorname{argmax}_{\mathcal{H}} \underbrace{\left\{ \overbrace{p(\mathbf{O}|\mathcal{H})}^{\text{acoustic score}} \overbrace{P(\mathcal{H})}^{\text{LM score}} \right\}}_{\text{inference evidence}}$$

- In standard adaptation,
 - transform \mathbf{W} estimated using best hypothesis
 - same transform \mathbf{W} used for all possible hypotheses to compute likelihood
- Need to estimate transforms as soon as data becomes available
- Problem:** no sufficient data for robust estimates of transforms
 - **Solution:** Bayesian approach



Bayesian Approach to Instantaneous Adaptation

- Impose a prior over transforms

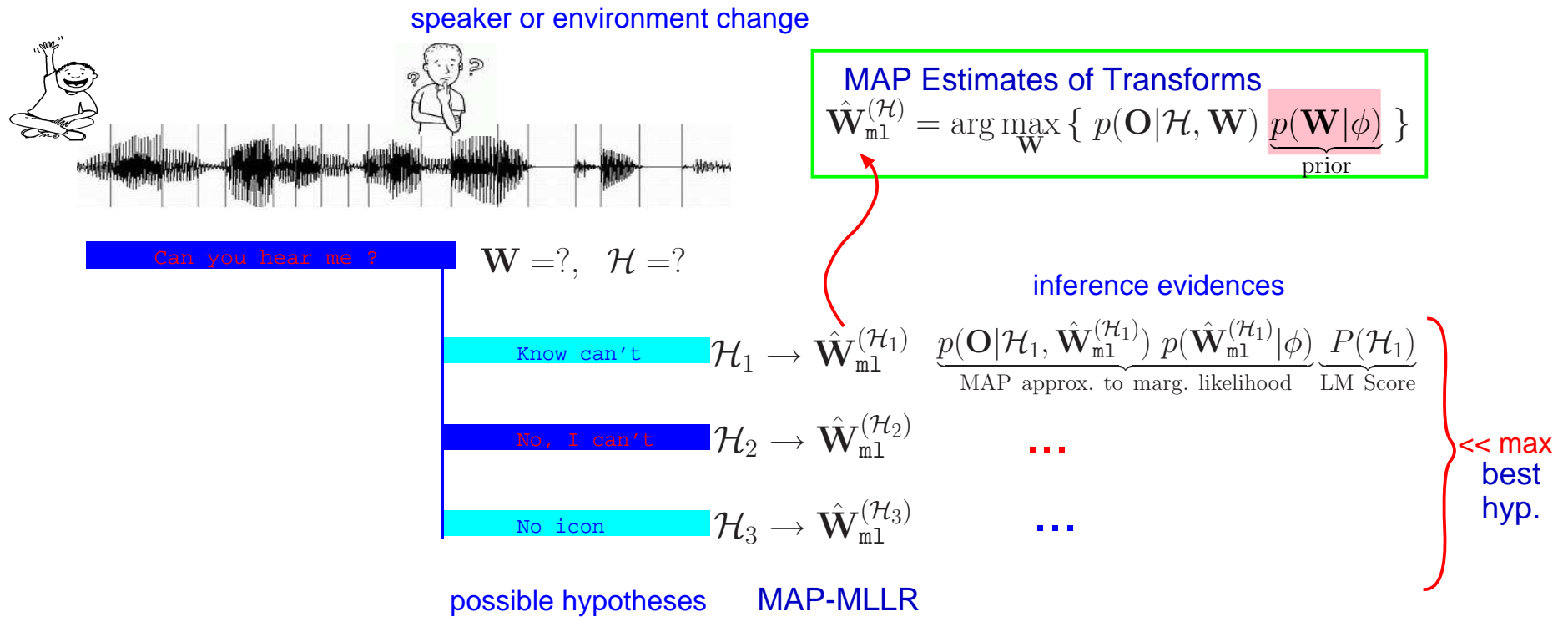


- The acoustic score is now marginal likelihood, given as

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

- Intractable integral for marginal likelihood
 - Maximum-a-Posteriori (MAP) approximation
 - * uses transform prior but gives point estimates

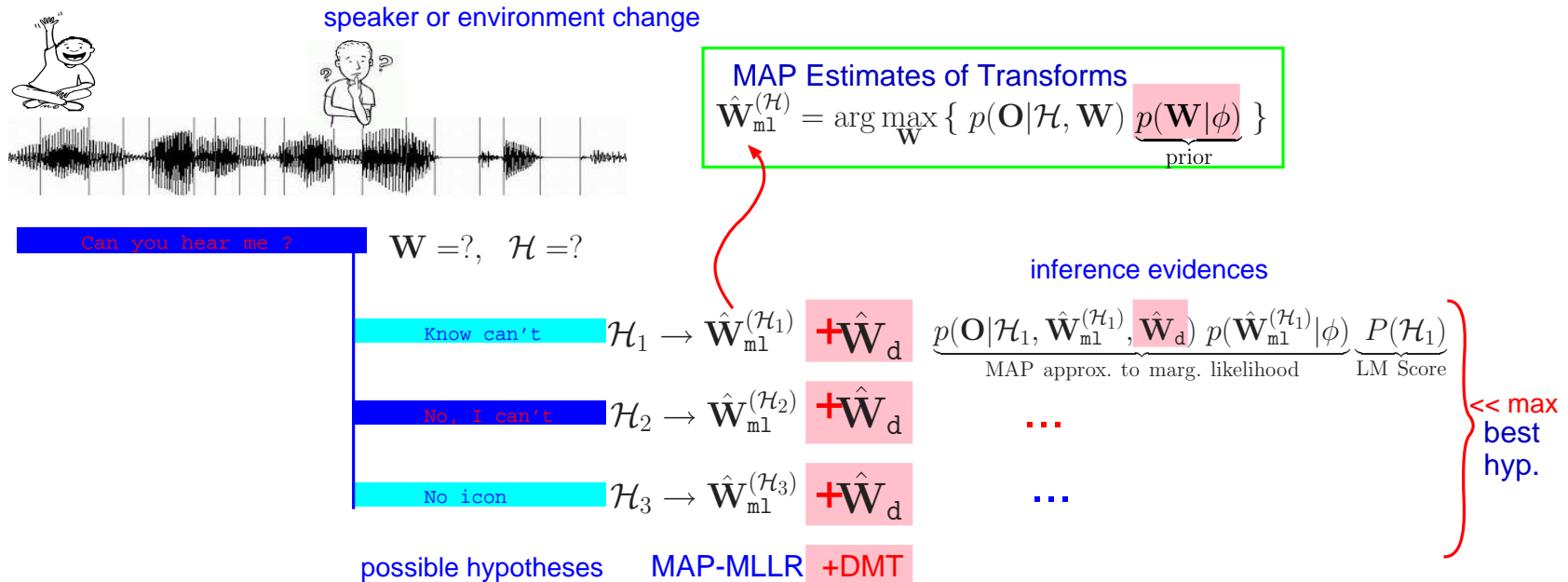
Bayesian Approach to Instantaneous Adaptation (contd.)



- MAP estimates of transforms for each possible hypothesis
- Inference evidence computed using transform for each hypothesis
- Best inference evidence \Rightarrow Best hypothesis

Instantaneous Unsupervised Discriminative Adaptation

- Combine instantaneous and discriminative adaptation (Bayesian + DMT)



- MAP-MLLR estimated for each hypothesis
- DMT estimated during training used
- inference evidence computed by applying MAPMLLR+DMT
- Issues:
 - DMT estimated using MLLR, but applied to MAPMLLR
 - DMT estimated for speaker-level MLLR, but applied to utterance-level

Experiments on English CTS Task

- English conversational telephone speech (CTS) task:
 - Dataset:
 - * Training dataset: about 290 hour data (5446 speakers)
 - * Test dataset (eva103): 6 hour data (144 speakers)
 - Front-end:
 - * Speech features: 12 PLP+ C0 and their 1st, 2nd, 3rd derivatives
 - * 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
 - Decoding:
 - * single Gaussian prior for transforms
 - * 2 baseclass for MLLR, 1000 for DMT
 - * 58k multi-pron dictionary, trigram-LM
 - * 150-best list rescoring in inference



Speaker-Level Adaptation

System	Transform		WER%
	Training	Testing	
SI/SI(hyp)	-	-	29.2
SI	-	MLLR	27.0
		MLLR+DMT	26.2
SAT	MLLR	MLLR	26.4
		MLLR+DMT	25.6

- DMT gives gains of **0.8% absolute** on both systems



Instantaneous Utterance-Level Adaptation

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

- DMT gives gains of **0.6% absolute on MPE-SI** system and **0.2% absolute on MPE-SAT** compared of MAPMLLR
 - reduced gain as DMT applied over MAP-MLLR (mismatched), not MLLR
 - SAT systems affected more by mismatch as transforms play more important role in them



Conclusion

- Adaptation to Speaker or Environment
 - maximum-likelihood linear transforms
 - discriminative linear transforms
- Unsupervised Adaptation
 - no correct supervision available
 - discriminative transforms are sensitive to errors in supervision
 - * Discriminative Mapping Transforms (DMTs)
- Instantaneous Adaptation
 - no sufficient data for robust transform estimation
 - * Bayesian approach to adaptation
- Instantaneous Unsupervised Discriminative Adaptation
 - DMTs over Bayesian estimates of ML-transforms
 - N-best list based rescoring
 - found to improve performance on a CTS task

