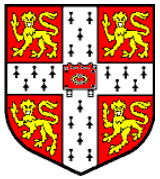
Instantaneous Adaptation: Integrated Recognition and Adaptation

Mark Gales

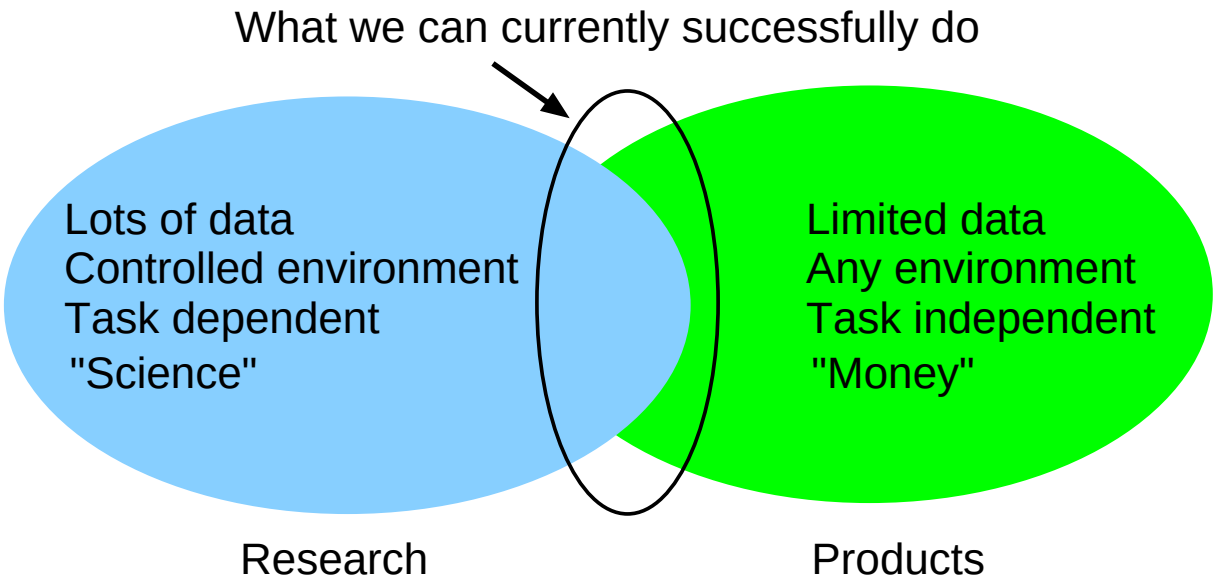
11 December 2007



Cambridge University Engineering Department

ASRU Statistical Modelling Panel 2007

Research vs Products



- Large mismatch between Research conditions and Industry desires
  - **robustness is a fundamental problem**
- Adaptation attempts to address this - can be run at many levels
  - task/speaker/environment adaptation extends the region of overlap

Ideal Acoustic Model Adaptation

- What we would like adaptation to be be:
  - **rapid**: adaptation starts with little data
    global: allow all model parameters to be adapted (not necessarily observed)
  - reliable: the adapted models are "close" to target domain/conditions
  - robust: for unsupervised cases adaptation is robust to hypothesis errors

- For some situations, such as environment mismatch function possible:

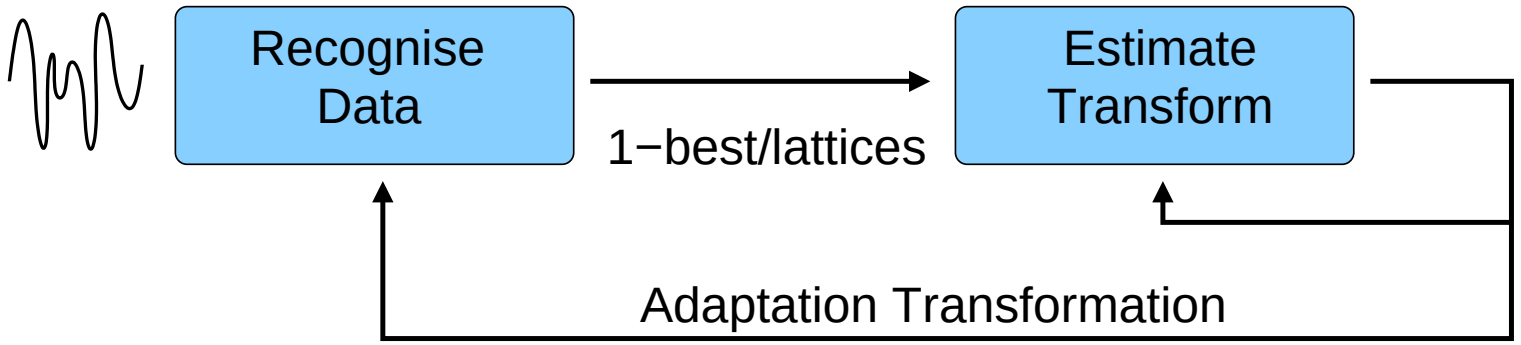
$$o_t = h * s_t + n_t$$

  but harder for many applications

  - assume linear/simple interpolation
    e.g. speaker adaptation using MLLR (Leggetter & Woodland 1995)

Unsupervised Acoustic Model Adaptation

- Standard approach to unsupervised adaptation:

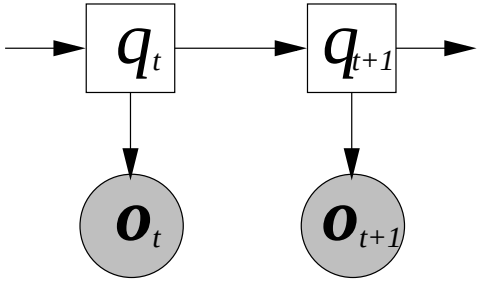


1. initial recognition of adaptation
2. estimate transform given hypotheses/lattices
3. iterate over loop as time/computation allows

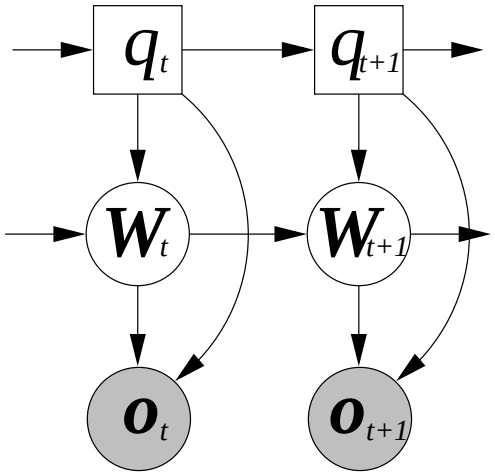
- Distinct two stage approach:
  - early decision made about hypothesis/lattices
  - issues with using adaptively trained models
- Preferable to delay all decisions as late as possible (as in rest of system)

Integrated Adaptation

- Integrate the adaptation process into the acoustic model: DBN becomes



Hidden Markov Model



Adaptive HMM

- Treat the adaptation transform as a latent variables
  - **Adaptive HMM** (Yu& Gales 2007): $p(\boldsymbol{o}_t|q_t, \boldsymbol{W}_t) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{W}_t \begin{bmatrix} \boldsymbol{\mu} \\ 1 \end{bmatrix}, \boldsymbol{\Sigma}\right)$
- Same form of DBN as Switching LDS, different interaction of latent variables.

Adaptive HMMs

- "Adaptation" starts instantly, adaptation/recognition integrated
  - simplify so that $\mathbf{W}_{t+1} = \mathbf{W}_t$ - speaker/environment blocks

$$p(\mathbf{O}_{1:T}) = \int \sum_{\mathbf{Q}} \prod_{t=1}^{T} P(q_t|q_{t-1}) p(\boldsymbol{o}_t|q_t, \mathbf{W}) p(\mathbf{W}) d\mathbf{W}$$

  - Viterbi/BW cannot be used (conditional independence assumptions broken)
- **Variational Bayes EM** (Beal 2003): lower-bound approximation relates to
  - N-best supervision schemes (Matsui & Furui 1998)
  - MAP Linear Regression (Chou 1999, Chesta et al 1999)
  - iterative MLLR (Woodland et al 1996)
  - Speaker Adaptive Training (Anastasakos et al 1996, Gales 1998)
- Adaptively trained models can be used directly

Comments

- Theoretically interesting and yields gains
  - elegant framework for adaptive training/decoding
  - distribution over transforms used - robust to limited data
  - no issues with errors in the supervision hypotheses
  - but too slow - improved/faster approximations required
- In practice many factors need to be simultaneously addressed
  - how to simultaneously address speaker/environment/task changes
  - simply using linear transforms too depressing ...
- Life's not linear
  - moving beyond linear transforms for speaker/tasks etc.