Bayesian Sensing Hidden Markov Models for Speech Recognition

George Saon[†] and Jen-Tzung Chien[‡]

[†]IBM T.J. Watson Research Center

[‡]National Cheng-Kung University, Taiwan

Introduction

- Modern ASR systems (still) use HMMs with state-dependent Gaussian mixture models for the acoustic feature vectors

- What has changed over the years is the estimation, transformation, adaptation of the Gaussian parameters

- Allocation of Gaussians to states based on heuristics (e.g. fifth root of the number of frames aligned to a state)

- Models can be easily overtrained especially with discriminative training

Shared representations

Reduce the number of parameters by sharing common structures

- Tied Gaussian mixture models: shared means and covariances, state-dependent mixture coefficients
- Subspace precision and mean (SPAM) models [Axelrod'02]: subspace constraint on precision matrices
- Subspace GMMs [Povey'10]: shared covariances, subspace constraint on component means

Parsimonious representations

Find good approximations to rich representations that use few parameters

- Diagonal covariance GMMs: $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)$
- Semi-tied covariance transforms [Gales'98]: $\Sigma = A\Lambda A^T$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$
- Extended maximum likelihood linear transforms [Olsen'02]: $\Sigma = A\Lambda A^T$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $D \leq K \leq D(D+1)/2$
- Factor-analyzed HMMs [Gopinath'98]: $\Sigma = \Lambda + \Phi\Phi^T$, Λ diagonal, $\Phi \in \mathbb{R}^{D\times K}$ is a "tall" factor loading matrix
- SPAM models: $\Sigma^{-1} = \sum_{i=1}^{n} \lambda_i B_i$, $B_i \in \mathbb{R}^{D\times D}$ are basis matrices

Bayesian estimation

- Rely on priors to prevent overfitting
- Regularized models perform better on noisy or mismatched test data
- Provides distribution estimates or “error bars” of latent variables instead of point estimates which can be unreliable
- Applications in speaker/noise adaptation: MAP, MAPLR, FMAPLR
- Little traction in acoustic modeling

Outline

- Model specification
- Some properties
- Parameter estimation
- Large scale ASR experiments

Model specification

Feature vectors $\mathbf{x}_t \in \mathbb{R}^D$ are generated from a state-dependent additive model

$$\mathbf{x}_t = \Phi_i \mathbf{w}_t + \boldsymbol{\epsilon}_t \tag{1}$$

where $\Phi_i = [\boldsymbol{\phi}_{i1}, \ldots, \boldsymbol{\phi}_{iN}]$, $\boldsymbol{\phi}_{ij} \in \mathbb{R}^D$, is the basis (or dictionary) for state i and $\mathbf{w}_t = [w_{t1}, \ldots, w_{tN}]^T$ is a time-dependent weight vector. Assumptions:

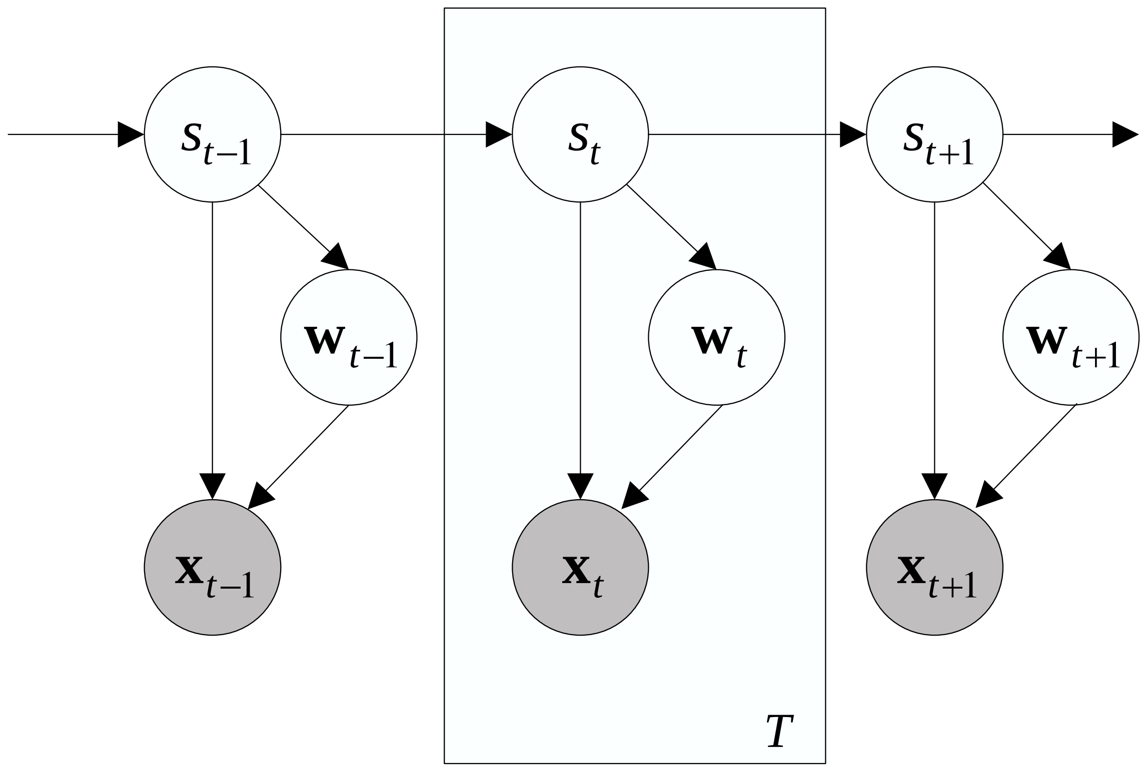
- $\boldsymbol{\epsilon}_t | s_t = i \sim \mathcal{N}(\mathbf{0}, R_i^{-1})$, i.e.

$$p(\mathbf{x}_t | \mathbf{w}_t, s_t = i) \propto |R_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \Phi_i \mathbf{w}_t)^T R_i (\mathbf{x}_t - \Phi_i \mathbf{w}_t)\right] \tag{2}$$

- $\mathbf{w}_t | s_t = i \sim \mathcal{N}(\mathbf{0}, A_i^{-1})$, i.e.

$$p(\mathbf{w}_t | s_t = i) \propto |A_i|^{1/2} \exp\left[-\frac{1}{2}\mathbf{w}_t^T A_i \mathbf{w}_t\right] \tag{3}$$

Graphical model for Bayesian sensing HMMs



Marginal state likelihood

$$p(\mathbf{x}_t|s_t = i) = \int_{\mathbb{R}^N} p(\mathbf{x}_t|\mathbf{w}_t, s_t = i)p(\mathbf{w}_t|s_t = i)\mathrm{d}\mathbf{w}_t \propto$$

$$\int_{\mathbb{R}^N} |R_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \Phi_i\mathbf{w}_t)^T R_i(\mathbf{x}_t - \Phi_i\mathbf{w}_t)\right] |A_i|^{1/2} \exp\left[-\frac{1}{2}\mathbf{w}_t^T A_i\mathbf{w}_t\right] \mathrm{d}\mathbf{w}_t$$

$$\propto |R_i|^{1/2}|A_i|^{1/2}|\Sigma_i|^{1/2} \exp\left[-\frac{1}{2}\mathbf{x}_t^T(R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i)\mathbf{x}_\mathrm{t}\right]$$

$$= |R_i|^{1/2}|A_i|^{1/2}|\Sigma_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t^T R_i\mathbf{x}_t - \mathbf{m}_{ti}^T\Sigma_i^{-1}\mathbf{m}_{ti})\right] \tag{4}$$

$\Sigma_i \triangleq (\Phi_i^T R_i\Phi_i + A_i)^{-1}$, $\mathbf{m}_{ti} \triangleq \Sigma_i\Phi_i^T R_i\mathbf{x}_t$ are the *covariance matrix* and the *mean vector* of the posterior distribution $p(\mathbf{w}_t|\mathbf{x}_t, s_t = i)$.

Gaussians with factor analyzed covariances

- Woodbury matrix inversion lemma

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{5}$$

where A, U, C and V denote matrices of compatible dimensions.

- Setting $A = R_i$, $U = R_i\Phi_i$, $V = \Phi_i^T R_i$ and $C = -\Sigma_i$, we get

$$\begin{aligned} S_i &\triangleq (R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i)^{-1} \\ &= R_i^{-1} - R_i^{-1}R_i\Phi_i((-\Sigma_i)^{-1} + \Phi_i^T R_i R_i^{-1} R_i \Phi_i)^{-1}\Phi_i^T R_i R_i^{-1} \\ &= R_i^{-1} + \Phi_i A_i^{-1} \Phi_i^T \end{aligned} \tag{6}$$

- $\Phi_i A_i^{-1/2}$ is a $D \times N$ factor loading matrix

Determinant equality

For (4) to be a Gaussian likelihood, the following has to hold

$$|R_i - R_i \Phi_i \Sigma_i \Phi_i^T R_i| = |R_i||A_i||\Sigma_i| \tag{7}$$

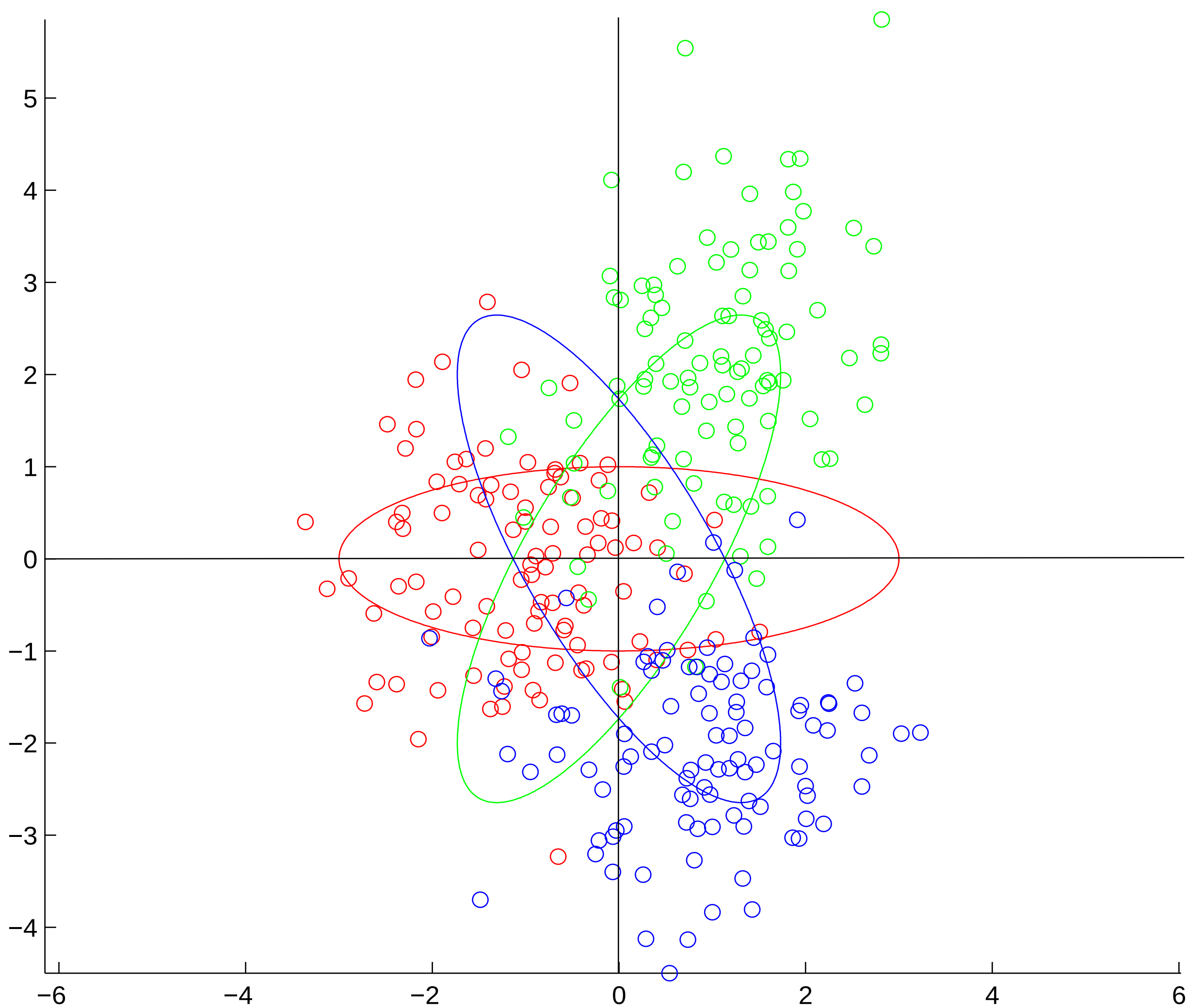
This can be shown by applying the determinant identity for a partitioned matrix

$$\begin{vmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{vmatrix} = |B_{22}||B_{11} - B_{12}B_{22}^{-1}B_{21}| = |B_{11}||B_{22} - B_{21}B_{11}^{-1}B_{12}| \tag{8}$$

to the extended matrix of size $(D+N) \times (D+N)$

$$\begin{bmatrix} (R_i)_{D\times D} & (R_i\Phi_i)_{D\times N} \\ (\Phi_i^T R_i)_{N\times D} & \Sigma_i^{-1} = (\Phi_i^T R_i \Phi_i + A_i)_{N\times N} \end{bmatrix}$$

Intuition behind the model



ML type II parameter estimation

- EM auxiliary function

$$Q(\lambda|\lambda^{(k)}) = \sum_S p(S|X, \lambda^{(k)}) \log p(X, S|\lambda) = \sum_i \sum_t \gamma_t(i) \log p(\mathbf{x}_t | s_t = i) + \mathcal{C}$$

where $\gamma_t(i) = p(s_t = i|X, \lambda^{(k)})$ is the posterior probability of being in state i at time t given observation sequence $X = \{\mathbf{x}_t\}$ and current parameters $\lambda^{(k)} = \{\Phi_i^{(k)}, A_i^{(k)}, R_i^{(k)}\}$.

- M step

$$\lambda^{(k+1)} = \operatorname*{argmax}_{\lambda} Q(\lambda|\lambda^{(k)})$$

Parameter updates

- $A_i^{(k+1)} = \left[\Sigma_i + \frac{\sum_t \gamma_t(i) \mathbf{m}_{ti} \mathbf{m}_{ti}^T}{\sum_t \gamma_t(i)} \right]^{-1}$

- $\Phi_i^{(k+1)} = \left[\sum_t \gamma_t(i) \mathbf{x}_t \mathbf{m}_{ti}^T \right] \left[\sum_t \gamma_t(i) (\Sigma_i + \mathbf{m}_{ti} \mathbf{m}_{ti}^T) \right]^{-1}$

- $R_i^{(k+1)} = \left[\Phi_i \Sigma_i \Phi_i^T + \frac{\sum_t \gamma_t(i) (\mathbf{x}_t - \Phi_i \mathbf{m}_{ti})(\mathbf{x}_t - \Phi_i \mathbf{m}_{ti})^T}{\sum_t \gamma_t(i)} \right]^{-1}$

Discriminative training

- MMI objective function (X observation sequence, W^r reference word sequence)

$$\begin{aligned}\mathcal{F}(\lambda) &= \log \frac{p(X, W^r|\lambda)}{p(X|\lambda)Pr(W^r)} \\ &= \log p(X|W^r, \lambda) - \log \sum_W p(X|W, \lambda)Pr(W) \\ &= \mathcal{F}^{num}(\lambda) - \mathcal{F}^{den}(\lambda)\end{aligned} \tag{9}$$

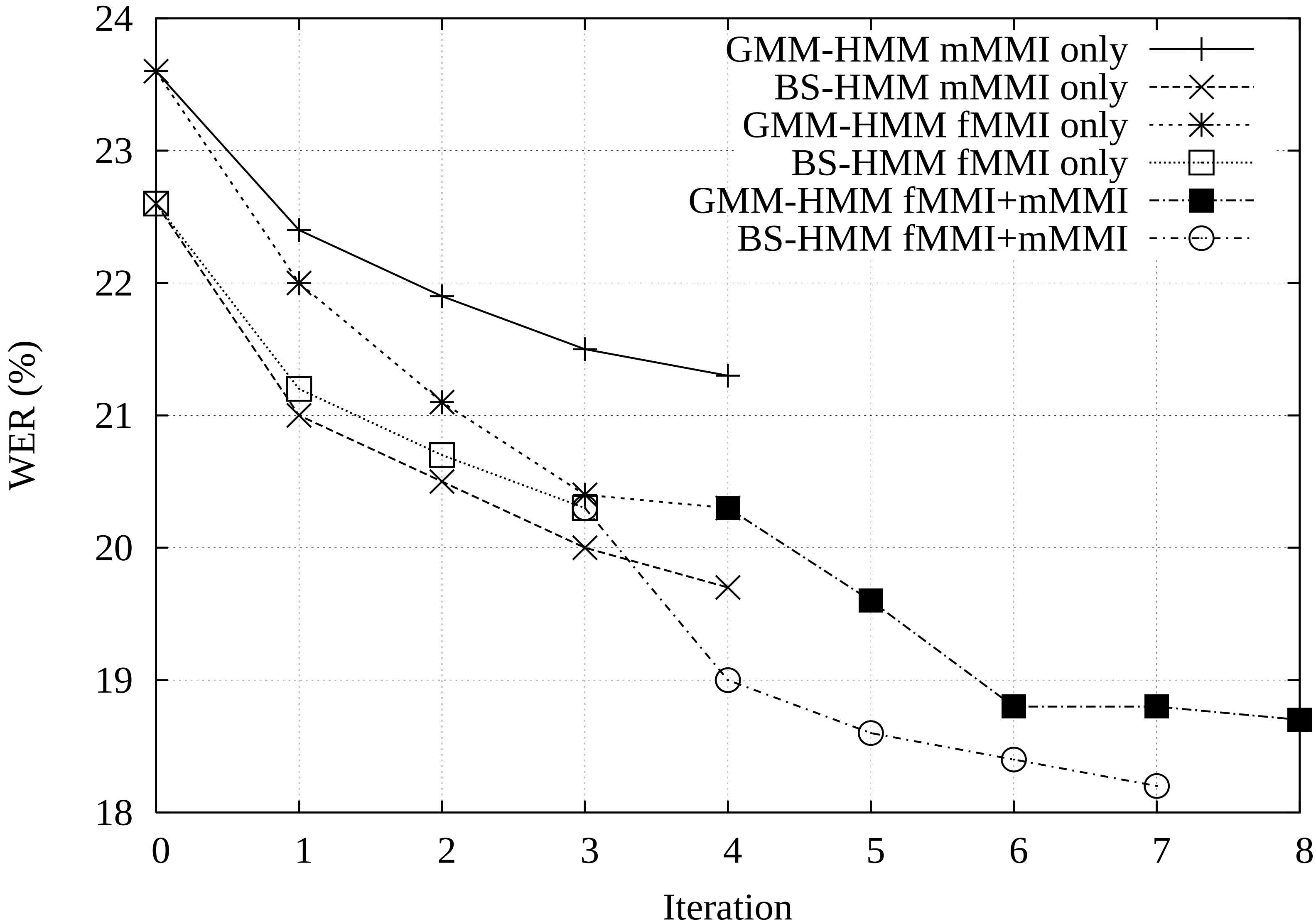
- Auxiliary function

$$Q(\lambda|\lambda^{(k)}) = Q^{num}(\lambda|\lambda^{(k)}) - Q^{den}(\lambda|\lambda^{(k)}) + Q^{sm}(\lambda|\lambda^{(k)}) \tag{10}$$

$$Q^{sm}(\lambda|\lambda^{(k)}) = \sum_i D_i \int_{\mathbb{R}^D} p(\mathbf{x}|\lambda_i^{(k)}) \log p(\mathbf{x}|\lambda_i) d\mathbf{x} \tag{11}$$

where D_i is a state-dependent smoothing constant

Example of feature and model space DT results



Automatic relevance determination

- Consider $A_i = \text{diag}(\alpha_{i1}, \ldots, \alpha_{iN})$
- If $\alpha_{ij} \to \infty$ then $w_{tj} \to 0$ because of the dimension-specific prior $\mathcal{N}(0, \alpha_{ij}^{-1})$ implying an irrelevant basis ϕ_{ij} for the Bayesian representation
- This is known as automatic relevance determination (ARD) [Tipping'01]
- Effect of α_{ij} on the factor analyzed covariance S_i from (6)

$$S_i = R_i^{-1} + \sum_{j=1}^{N} \frac{1}{\alpha_{ij}} \boldsymbol{\phi}_{ij} \boldsymbol{\phi}_{ij}^T \qquad (12)$$

- Model compression by discarding the ϕ_{ij}'s corresponding to large α_{ij}'s

Improvements for ASR I: mixture models

- Parameter initialization:
  - Train a GMM for each state and cluster the resulting means using k-means
  - Bases Φ_{ij} initialized to the partitioned means
  - Precisions R_{ij} and A_{ij} assumed diagonal and initialized to identity
- Word error rates for an English broadcast news system trained on 50 hours:

mix/state	1	2	4	8	16
A_{ij}, R_{ij} training	29.8%	27.1%	25.7%	25.2%	24.8%
$A_{ij}, R_{ij}, \Phi_{ij}$ training	29.4%	26.8%	25.3%	24.4%	24.4%

Improvements for ASR II: non-zero means

Word error rates for an Arabic broadcast news system trained on 1800 hours:

Means	Adaptation	DEV07	DEV08	DEV09
zero	none	14.3%	16.7%	19.7%
non-zero	none	14.2%	16.4%	19.6%
non-zero	MLLR	13.6%	16.0%	18.9%

Experimental setup

- 1800 hours of Arabic broadcast news training data
- VTL-warped PLP cepstra with LDA and STC
- Speaker adaptation with VTLN, FMLLR and multiple MLLR
- Feature and model space discriminative training with boosted MMI [Povey'08]
- Acoustic models have 5000 states and
  - 800K Gaussians for the baseline
  - 417K Gaussians for the BSHMMs (initialized from 2.8M Gaussians)
- Recognition vocabulary: 795K words
- Language model: 4-gram with 884M n-grams

ML type II training results

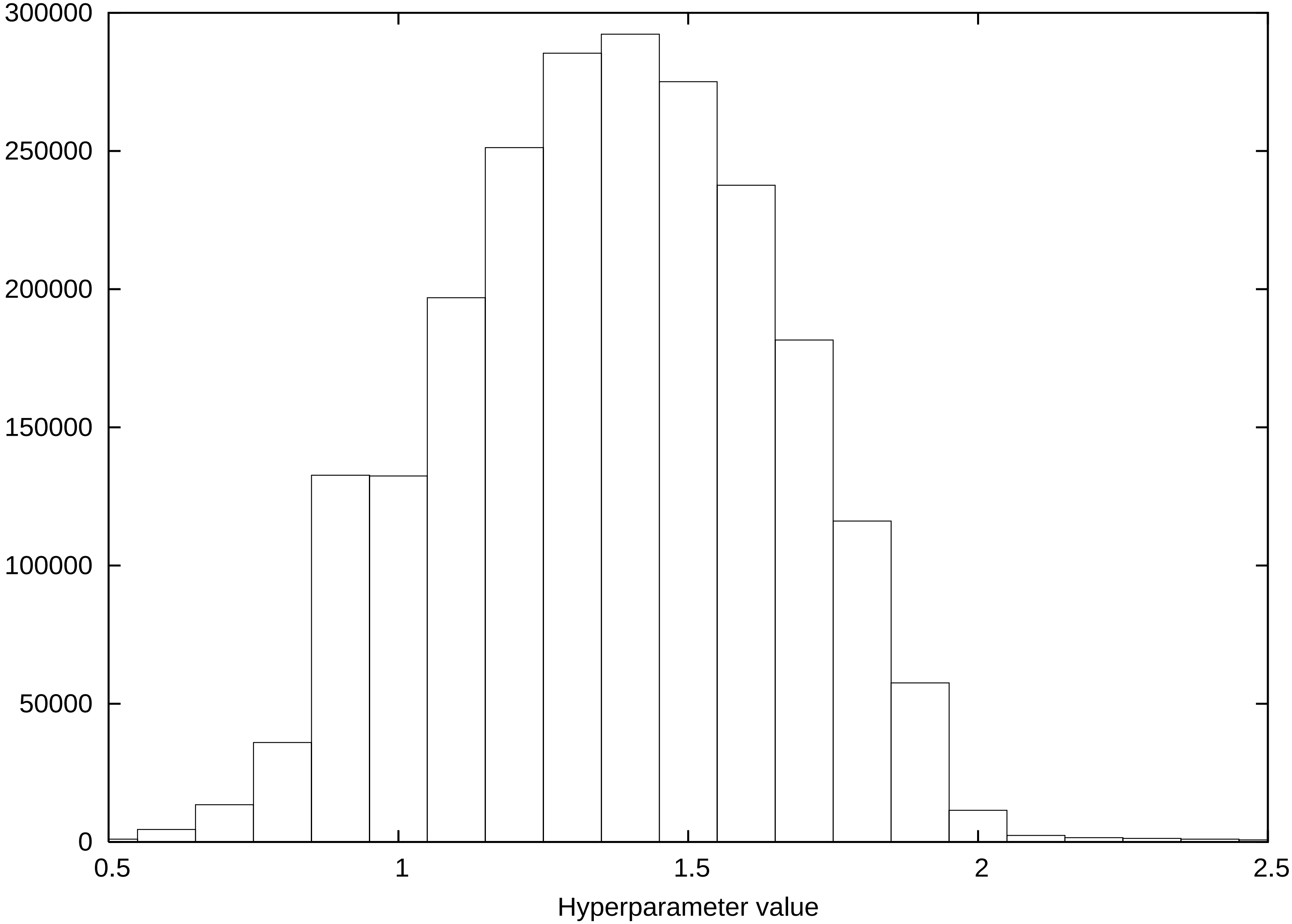
- Training regime: 5 iterations with fixed state alignments followed by one Viterbi iteration
- Word error rates:

System	DEV07	DEV08	DEV09
baseline 800K	13.8%	16.4%	19.6%
baseline 2.8M	14.1%	16.2%	19.3%
BSHMM 417K	13.6%	16.0%	18.9%

- Number of free parameters:

System	Nb. parameters
baseline 800K	64.8M
baseline 2.8M	226.8M
BSHMM 417K	148.5M

Histogram of sensing weight precisions



Model compression using ARD

- Acoustic models built with discriminative feature-space transforms [Povey'05]
- Discard 50% of basis vectors corresponding to the largest precision values after training
- Results before and after discriminative training of the parameters:

Model	Training	DEV07	DEV08	DEV09
original	ML type II	12.0%	13.9%	17.4%
compressed	ML type II	12.4%	14.2%	17.6%
original	boosted MMI	10.7%	11.9%	15.0%
compressed	boosted MMI	10.4%	11.7%	14.8%

GALE 2011 evaluation results

- All models are cross-adapted on the output of a system using SGMMs
- Evaluation testset EVAL-P5 previously unseen
- Word error rates:

System	DEV09	EVAL-P4	**EVAL-P5**
baseline 800K	13.1%	10.0%	**9.4%**
compressed BSHMM	12.8%	9.7%	**9.1%**
system combination	12.6%	9.6%	**9.0%**

Conclusion

- Gaussians with factor analyzed covariance matrices
- Bayesian smoothing (prevents overtraining)
- Model compression using ARD
- Outperformed state-of-the-art models during the last GALE evaluation
- More details:
  - G. Saon and J.-T. Chien. ”Bayesian Sensing Hidden Markov Models for Speech Recognition”, ICASSP 2011.
  - G. Saon and J.-T. Chien. ”Discriminative Training for Bayesian Sensing Hidden Markov Models”, ICASSP 2011.
  - G. Saon and J.-T. Chien. ”Some Properties of Bayesian Sensing Hidden Markov Models”, submitted to ASRU 2011.