Acoustic Modelling for Speech Recognition: Hidden Markov Models and Beyond?

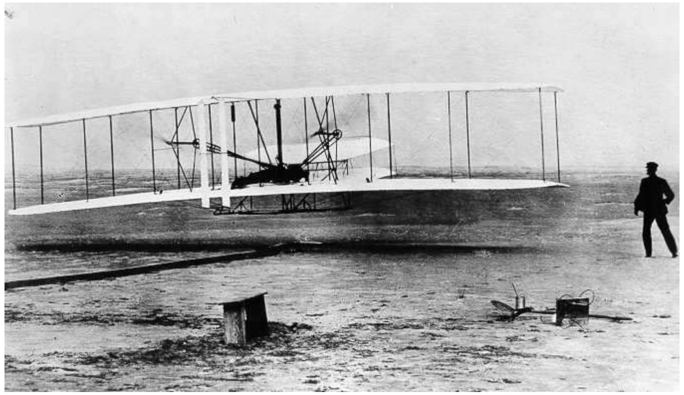
Mark Gales

December 2009

Cambridge University Engineering Department

ASRU 2009

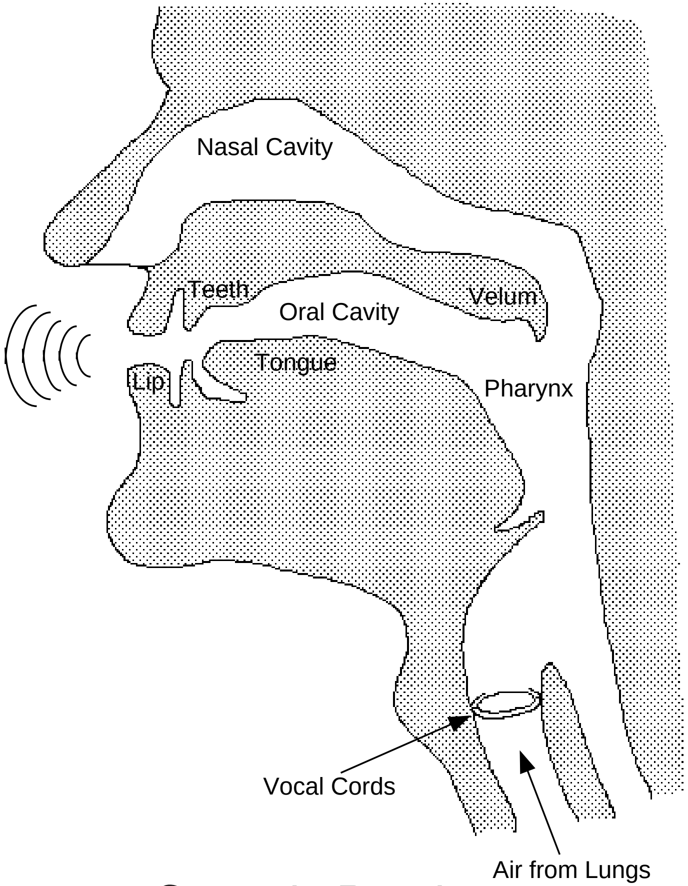
An Engineering Solution - should planes flap their wings?



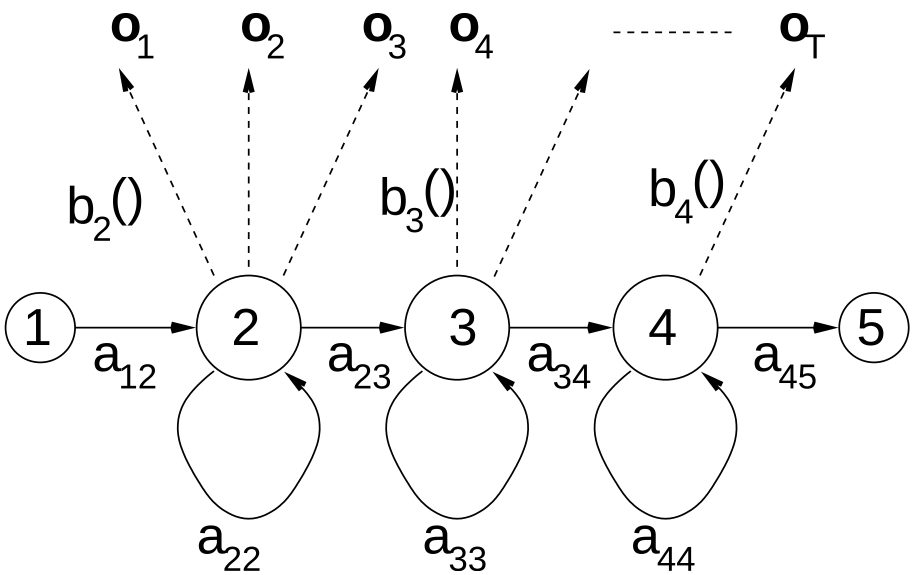
Overview

- Engineering solutions to speech recognition
  - machine learning (statistical) approaches
  - the acoustic model: hidden Markov model
- Noise Robustness
  - model-based noise and speaker adaptation
  - adaptive training
- Discriminative Criteria and (Possibly) not a HMM?
  - discriminative training criteria
  - discriminative models
  - combined generative and discriminative models

Acoustic Modelling



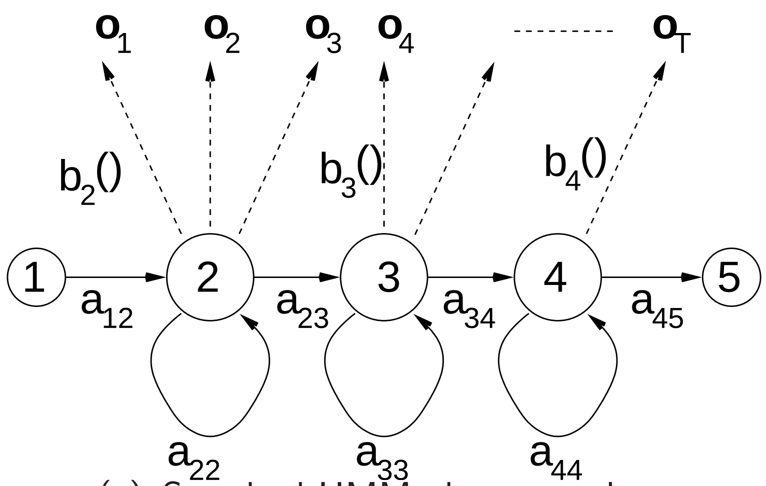
(a) Speech Production



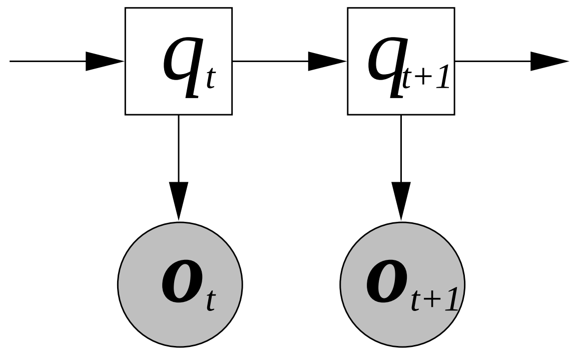
(b) HMM Generative Model

- Not modelling the human production process!

Hidden Markov Model



(c) Standard HMM phone topology



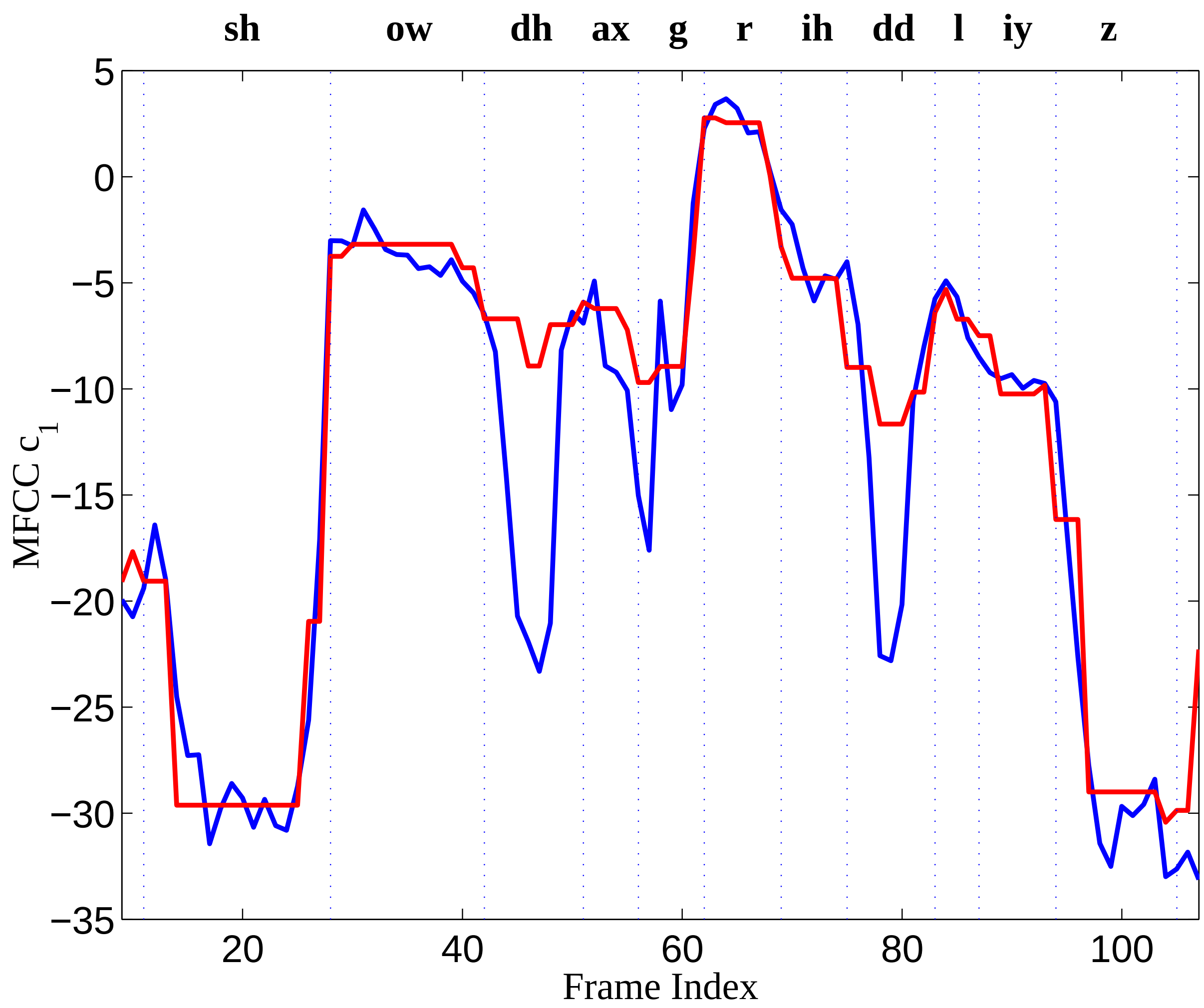
(d) HMM Dynamic Bayesian Network

- HMM generative model
  - class posteriors, $P(\mathbf{w}|\mathbf{O};\boldsymbol{\lambda})$, obtained using Bayes' rule
  - requires class priors, $P(\mathbf{w})$ - language models in ASR
- Parameters trained
  - ASR - Gaussian Mixture Models (GMMs) as state output distributions
  - efficiently implemented using Expectation-Maximisation (EM)
- Poor model of the speech process - piecewise constant state-space.

HMM Trajectory Modelling

Frames from phrase:

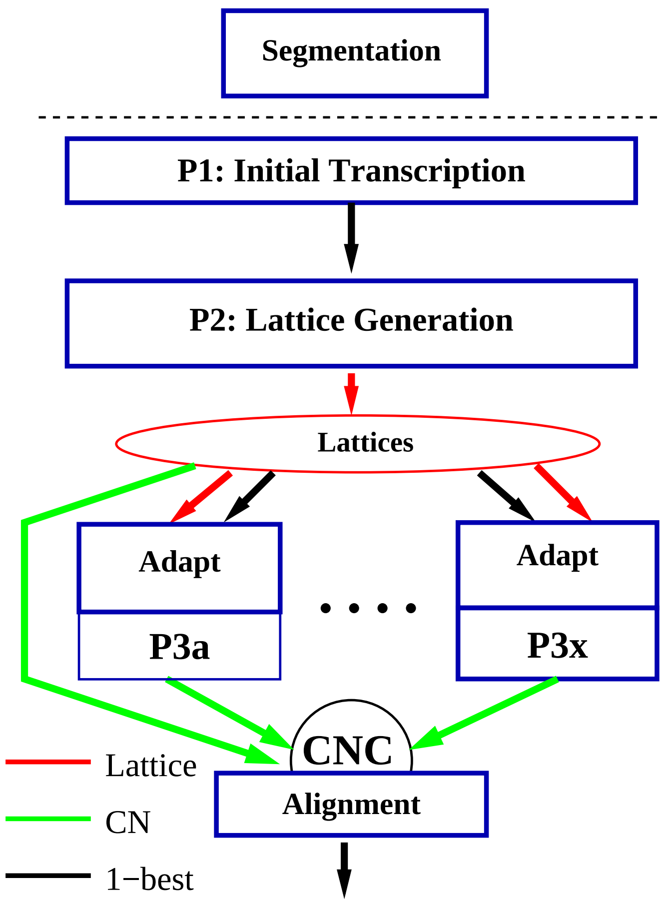
`SHOW THE GRIDLEY'S ...`



Legend

- True
- HMM

CU-HTK Multi-Pass/Combination Framework

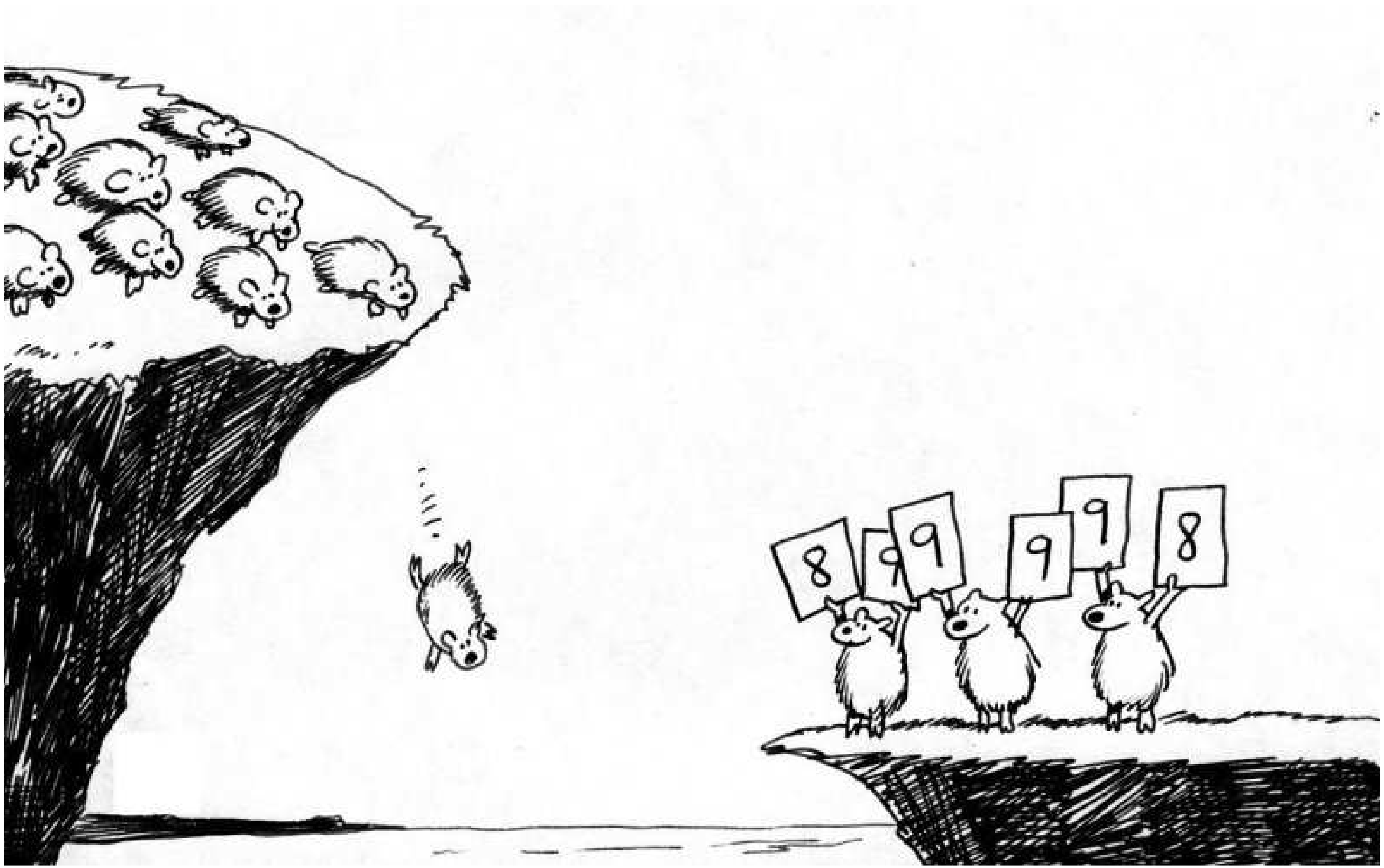


- Structure for CU-HTK systems [1]
- P1 used to generate initial hypothesis
- P1 hypothesis used for rapid adaptation
  - LSLR, diagonal variance transforms
- P2: lattices generated for rescoring
  - apply complex LMs to trigram lattices
- P3 Adaptation of “diverse” systems
  - 1-best/lattice-based CMLLR/MLLR
- CN Decoding/Combination

Large Vocabulary Speech Recognition Systems

- "Typical" LVCSR system acoustic models comprise:
  - thousands of hours acoustic training data
  - PLP/MFCC/MLP/TANDEM-based feature-vectors
  - decorrelating transforms/projections
  - decision tree state-clustered tri/quin/septa phone
  - thousands of distinct states, hundreds of thousands of Gaussian components
  - discriminative training criteria
  - speaker adaptation and adaptive training
  - combination of multiple diverse (possibly cross-site) systems
- **Why we like HMMs** - example broadcast news/conversation results

System	WER (%)		
	BN	BC	Avg
English	6.7	—	6.7
Mandarin (CER%)	2.3	12.6	7.1
Arabic	8.6	16.6	11.7

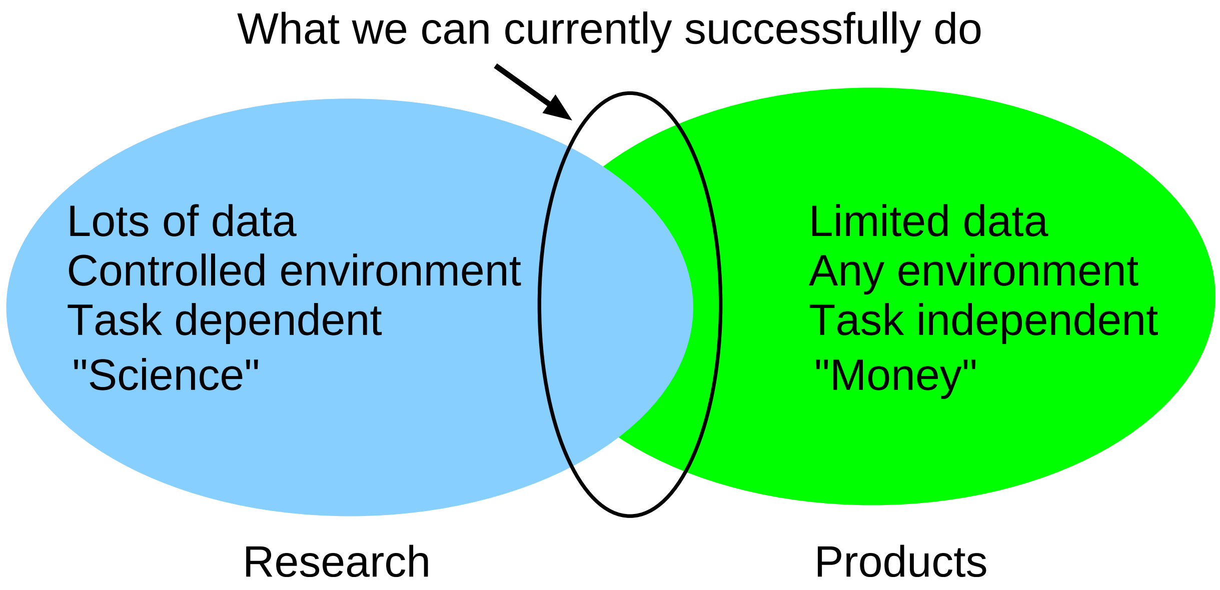


"One hundred thousand lemmings can't be wrong"

“Five hundred thousand Gaussians can’t be wrong”

"Five hundred thousand Gaussians can't be wrong"

Generalisation of our systems still poor

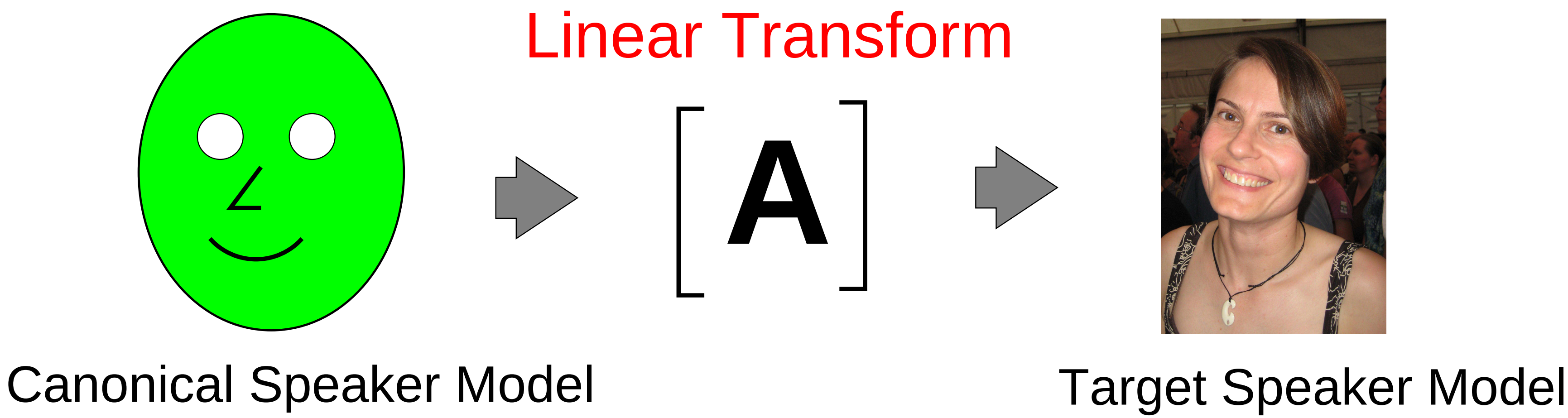


Noise Robustness

Example Application - In-Car Navigation

"Adaptive" Linear Model Compensation

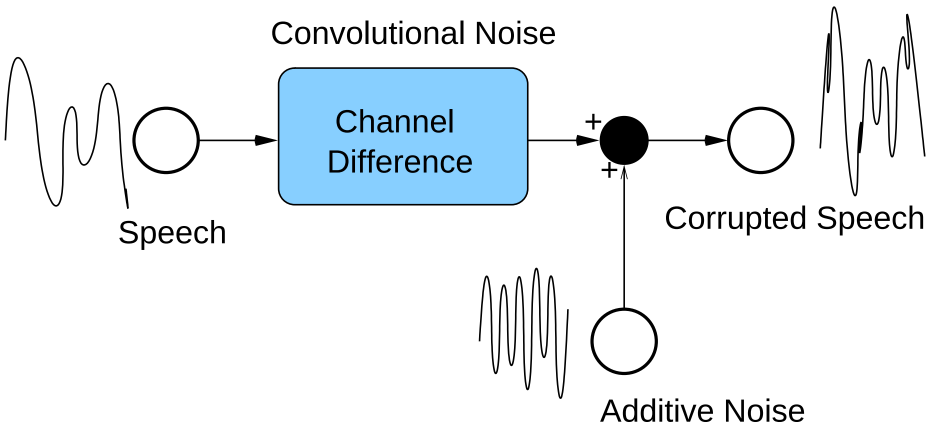
- Standard scheme for speaker/environment adaptation is linear transforms [2, 3]:
  - all speaker difference can be modelled as a linear transform



- Common form is $\boxed{\boldsymbol{\mu}^{(ms)} = \mathbf{A}\boldsymbol{\mu}^{(m)} + \mathbf{b}}$
- General approach, but large numbers of model parameters
  - a single full-transform has about 1560 parameters to train
  - the impact of noise is non-linear, so many transforms useful

“Predictive” Compensation Schemes

- Predict impact of noise of clean-speech: mismatch function



- Ignore effects of stress:
- Group noise sources

$$y(t) = x(t) * h(t) + n(t)$$

- Squared magnitude of the Fourier Transform of signal

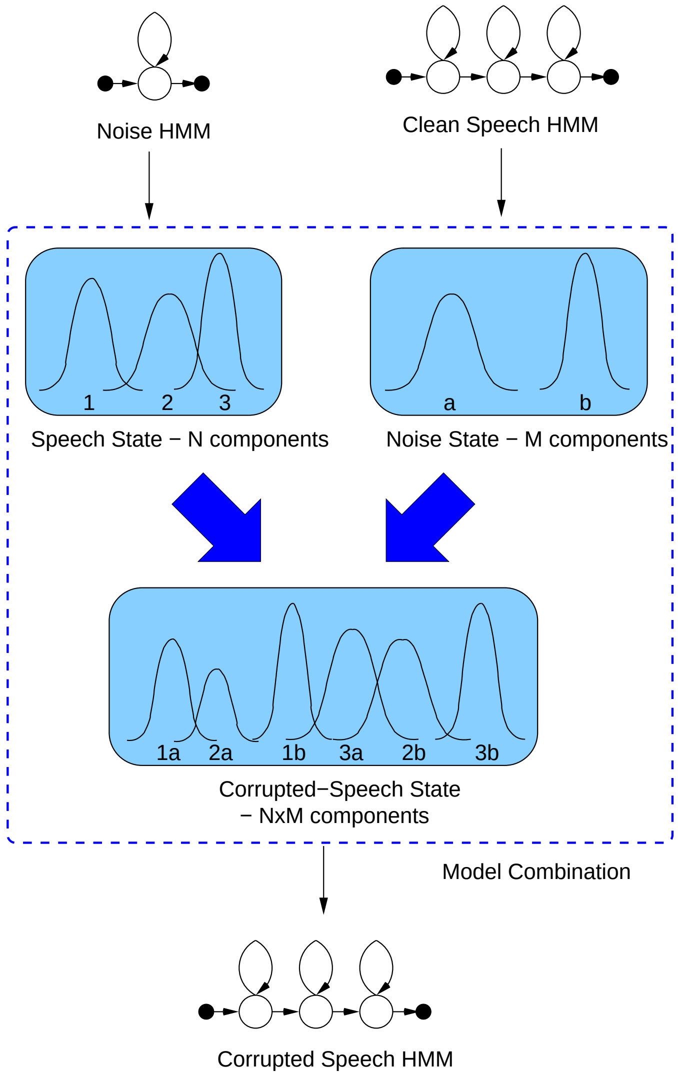
$$Y(f)Y^*(f) = |H(f)X(f)|^2 + |N(f)|^2 + 2|N(f)||H(f)X(f)|\cos(\theta)$$

θ is the angle between the vectors $N(f)$ and $H(f)X(f)$.

- Average (over Mel bins), assume speech and noise independent and $\log()$ [4]

$$\boldsymbol{y}_t = \mathbf{C}\log\left(\exp\left(\mathbf{C}^{-1}(\boldsymbol{x}_t + \boldsymbol{h})\right) + \exp\left(\mathbf{C}^{-1}\boldsymbol{n}_t\right)\right) = \boldsymbol{x}_t + \boldsymbol{h} + \mathbf{f}\left(\boldsymbol{x}_t, \boldsymbol{h}, \mathbf{n}_t\right)$$

Model-Based Predictive Compensation Procedure



- Each speech/noise pair considered
  - yields final component
- VTS approximation [5, 6]

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{y}}^{(mn)} &= \mathcal{E}\{\boldsymbol{y}_t|\mathbf{s}_m, \mathbf{s}_n\} \\ &\approx \boldsymbol{\mu}_{\mathrm{x}}^{(m)} + \boldsymbol{\mu}_{\mathrm{h}} + \mathbf{f}(\boldsymbol{\mu}_{\mathrm{x}}^{(m)}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{\mu}_{\mathrm{n}}^{(n)})\end{aligned}$$

- Also multiple-states possible
  - 3-D Viterbi decoding [7]
  - usually single component/single state
- Only need to estimate noise model
  - $\boldsymbol{\mu}_{\mathrm{n}}$, $\boldsymbol{\Sigma}_{\mathrm{n}}$ $\boldsymbol{\mu}_{\mathrm{h}}$

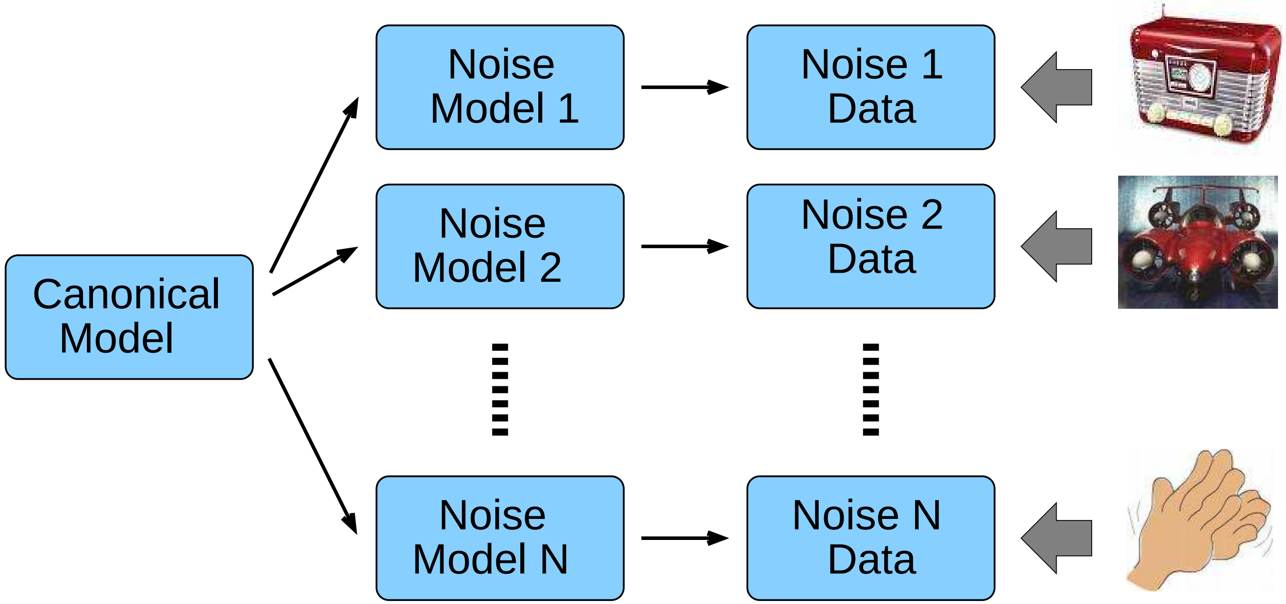
“Adaptive” vs “Predictive” Schemes

- Adaptive and predictive schemes complementary to one another

Adaptive	Predictive
general approach	applicable to noise
linear assumption - use many linear transforms	mismatch function required - may be inaccurate
transform parameters estimated - large numbers of parameters	noise model estimated - small number of parameters

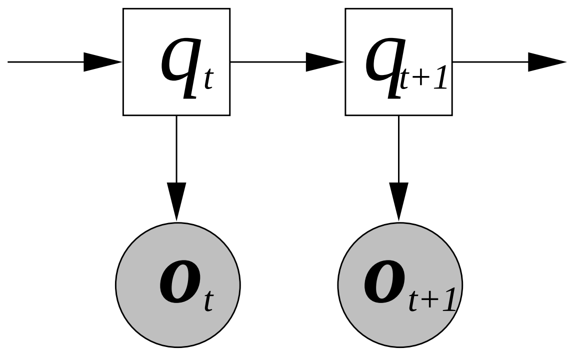
- Possible to combine both predictive and adaptive models [8]
  - would be nice to get “orthogonal” transforms acoustic factorisation
- Need to decide on form of canonical model to adapt:
  - Multi-Style: adaptation converts a general system to a specific condition;
  - Adaptive: adaptation converts “neutral” system to specific condition [9, 3]

Noise Adaptive Training

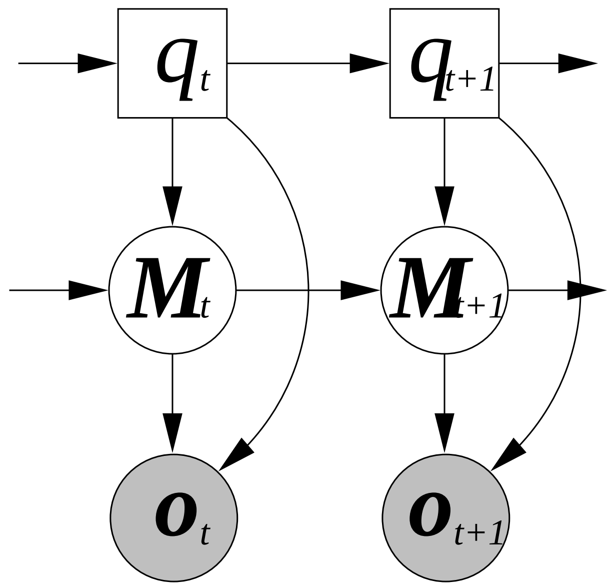


- In adaptive training the training corpus is split into "homogeneous" blocks
  - use adaptation transforms to represent unwanted acoustic noise factors
  - canonical model only represents desired variability
- Adaptive training possibly more important for noise than speakers [10, 11, 12]
  - very wide range of possible noise conditions - hard to cover with multi-style
  - contribution of low SNR training examples to canonical model de-weighted

Adaptive Training From Bayesian Perspective



(e) Standard HMM



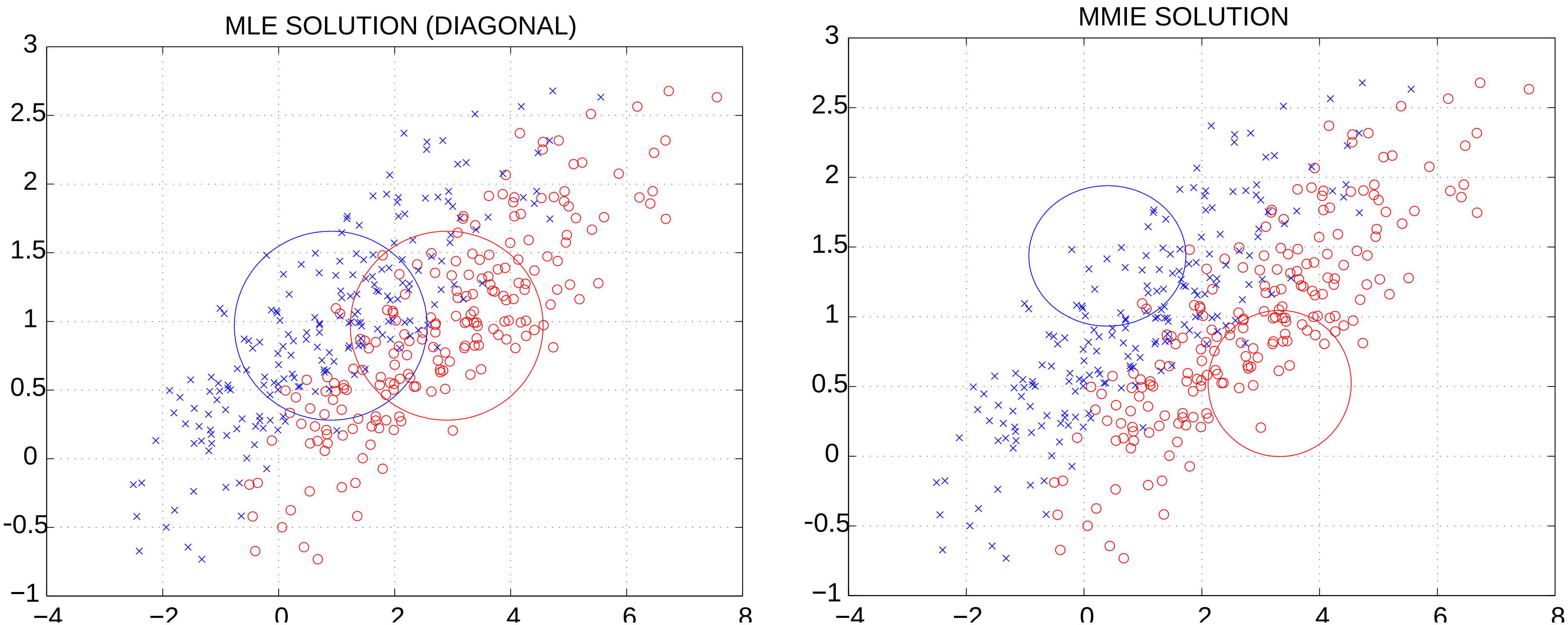
(f) Adaptive HMM

- Observation additionally dependent on noise model $\mathcal{M}_t$ [13]
  - noise model same for each homogeneous block ($\mathcal{M}_t = \mathcal{M}_{t+1}$)
  - model-compensation integrated into model (cf instantaneous adaptation)
- Need to known the prior noise model distribution
  - inference computationally will be expensive (but interesting)

Discriminative Criteria and Models (Possibly) not an HMM

Simple MMIE Example

- HMMs are not the correct model - discriminative criteria a possibility



- Discrimnative criteria a function of posteriors $P(\mathbf{w}|\mathbf{O};\boldsymbol{\lambda})$
  - NOTE: same generative model, and conditional independence assumptions

Discriminative Training Criteria

- Discriminative training criteria commonly used to train HMMs for ASR
  - Maximum Mutual Information (MMI) [14, 15]: maximise

$$\mathcal{F}_{\mathtt{mmi}}(\boldsymbol{\lambda}) = \frac{1}{R}\sum_{r=1}^{R}\log(P(\mathbf{w}_{\mathtt{ref}}^{(r)}|\mathbf{O}^{(r)};\boldsymbol{\lambda}))$$

  - Minimum Classification Error (MCE) [16]: minimise

$$\mathcal{F}_{\mathtt{mce}}(\boldsymbol{\lambda}) = \frac{1}{R}\sum_{r=1}^{R}\left(1+\left[\frac{P(\mathbf{w}_{\mathtt{ref}}^{(r)}|\mathbf{O}^{(r)};\boldsymbol{\lambda})}{\sum_{\mathbf{w}\neq\mathbf{w}_{\mathtt{ref}}^{(r)}}P(\mathbf{w}|\mathbf{O}^{(r)};\boldsymbol{\lambda})}\right]^{\varrho}\right)^{-1}$$

  - Minimum Bayes' Risk (MBR) [17, 18]: minimise

$$\mathcal{F}_{\mathtt{mbr}}(\boldsymbol{\lambda}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{\mathbf{w}}P(\mathbf{w}|\mathbf{O}^{(r)};\boldsymbol{\lambda})\mathcal{L}(\mathbf{w},\mathbf{w}_{\mathtt{ref}}^{(r)})$$

MBR Loss Functions for ASR

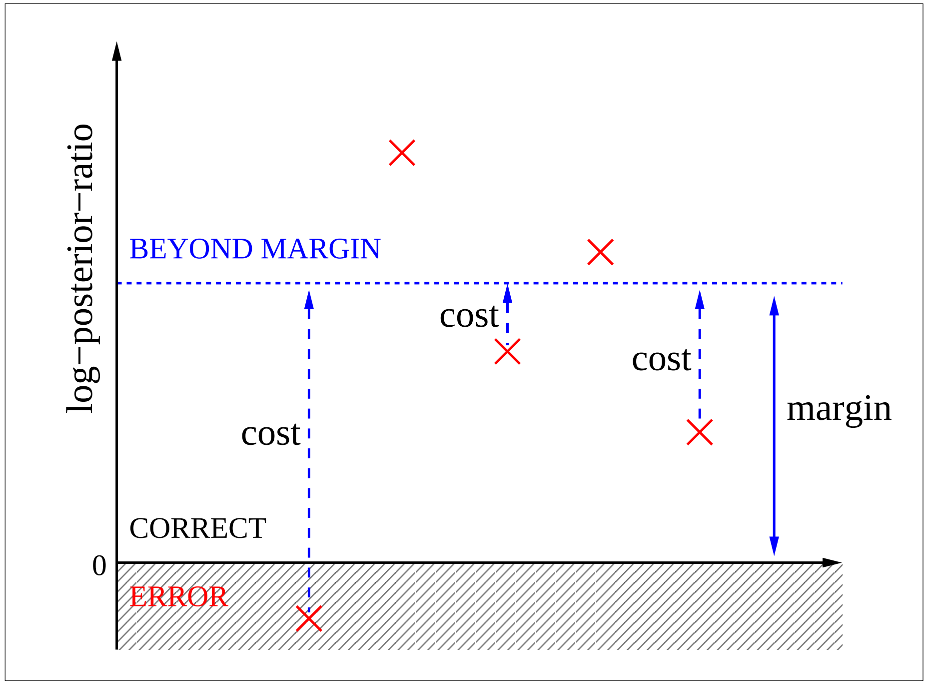
- Sentence (1/0 loss):

$$\mathcal{L}(\mathbf{w}, \mathbf{w}^{(r)}_{\texttt{ref}}) = \begin{cases} 1; & \mathbf{w} \neq \mathbf{w}^{(r)}_{\texttt{ref}} \\ 0; & \mathbf{w} = \mathbf{w}^{(r)}_{\texttt{ref}} \end{cases}$$

  When $\varrho = 1$, $\mathcal{F}_{\texttt{mce}}(\boldsymbol{\lambda}) = \mathcal{F}_{\texttt{mbr}}(\boldsymbol{\lambda})$

- Word: directly related to minimising the expected Word Error Rate (WER)
  - normally computed by minimising the Levenshtein edit distance.
- Phone: consider phone rather word loss
  - improved generalisation as more "error's" observed
  - this is known as Minimum Phone Error (MPE) training [19, 20].
- Hamming (MPFE): number of erroneous frames measured at the phone level

Large Margin Based Criteria



- Standard criterion for SVMs
  - improves generalisation
- Require log-posterior-ratio

$$\min_{\mathbf{w}\neq\mathbf{w}_{\tt ref}}\left\{\log\left(\frac{P(\mathbf{w}_{\tt ref}|\mathbf{O};\boldsymbol{\lambda})}{P(\mathbf{w}|\mathbf{O};\boldsymbol{\lambda})}\right)\right\}$$

to be beyond margin

- As sequences being used can make margin function of the "loss" - minimise

$$\mathcal{F}_{\tt lm}(\boldsymbol{\lambda})=\frac{1}{R}\sum_{r=1}^{R}\left[\max_{\mathbf{w}\neq\mathbf{w}_{\tt ref}^{(r)}}\left\{\mathcal{L}(\mathbf{w},\mathbf{w}_{\tt ref}^{(r)})-\log\left(\frac{P(\mathbf{w}_{\tt ref}^{(r)}|\mathbf{O}^{(r)};\boldsymbol{\lambda})}{P(\mathbf{w}|\mathbf{O}^{(r)};\boldsymbol{\lambda})}\right)\right\}\right]_{+}$$

use hinge-loss $[f(x)]_{+}$. Many variants possible [21, 22, 23, 24]

Generative and Discriminative Models

- HMMs are a generative model where Bayes' rule is used to get the posterior

$$P(\mathbf{w}|\mathbf{O};\boldsymbol{\lambda}) = \frac{p(\mathbf{O}|\mathbf{w};\boldsymbol{\lambda})P(\mathbf{w})}{\sum_{\tilde{\mathbf{w}}} p(\mathbf{O}|\tilde{\mathbf{w}};\boldsymbol{\lambda})P(\tilde{\mathbf{w}})}$$

- Also possible to directly model the posterior - a discriminative model
  - simple, standard, form log-linear model

$$P(\mathbf{w}|\mathbf{O};\boldsymbol{\alpha}) = \frac{1}{Z}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O}_{1:T},\mathbf{w})\right)$$

  - features from sequence: $\boldsymbol{\phi}(\mathbf{O}_{1:T},\mathbf{w})$ - determines dependencies
  - model parameters: $\boldsymbol{\alpha}$
- Can use any of the previous training criteria ...

Direct Flat Models

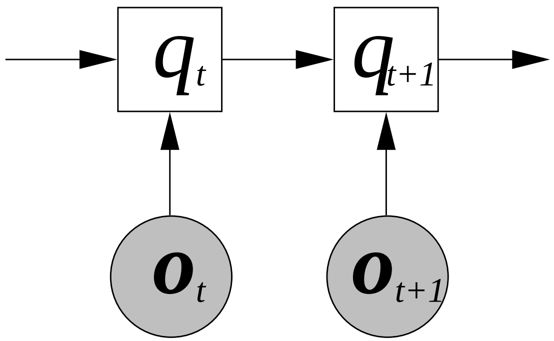
- Based on log-linear model feature set has the form [25]

$$\boldsymbol{\phi}(\mathbf{O}_{1:T}, \mathbf{w}) = \left[\begin{array}{c} \boldsymbol{\phi}_{\mathrm{l}}(\mathbf{w}) \\ \boldsymbol{\phi}_{\mathrm{a}}(\mathbf{O}_{1:T}, \mathbf{w}) \end{array} \right]$$

- Text Features $\boldsymbol{\phi}_{\mathrm{l}}(\mathbf{w})$: from the sequence $\mathbf{w}$
  - N-gram features (word or level), related to N-gram language model
- Acoustic Feature $\boldsymbol{\phi}_{\mathrm{a}}(\mathbf{O}_{1:T}, \mathbf{w})$: for hypothesis $\mathbf{v}$
  - rank feature of hypothesis $\mathbf{v}$
  - HMM posterior features $P(\mathbf{v}|\mathbf{O}_{1:T}; \boldsymbol{\lambda})$
  - DTW distance to closest template (or set of templates)
- "Spotter" features nearest neighbour DTW templates
  - utterance, or N-gram features

Maximum Entropy Markov Models

- Attempt to model the class posteriors directly - MEMMs one example
  - The DBN and associated word sequence posterior [26]



$$P(\mathbf{w}|\mathbf{O}_{1:T};\boldsymbol{\alpha}) = \sum_{\mathbf{q}} P(\mathbf{w}|\mathbf{q}) \prod_{t=1}^{T} P(q_t|\mathbf{o}_t, q_{t-1};\boldsymbol{\alpha})$$

$$P(q_t|\mathbf{o}_t, q_{t-1};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha},\mathbf{o}_t)} \exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{o}_t, q_t, q_{t-1})\right)$$

- Features extracted - transitions $\boldsymbol{\phi}(q_t, q_{t-1})$, observations $\boldsymbol{\phi}(\mathbf{o}_t, q_t)$
  - same features as standard HMMs
- Problems incorporating language model prior
  - gains over standard (ML-trained) HMM with no LM
  - does yield gains in combination with standard HMM

Hidden Conditional Random Fields

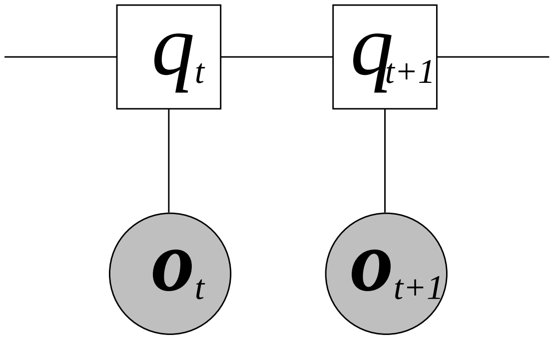
- Conditional random fields hard to directly apply to speech data
  - observation sequence length T doesn't word match label sequence L
  - introduce latent discrete sequence (similar to HMM)
- The feature dependencies in the HCRF and word sequence posterior [27]

$$P(\mathbf{w}|\mathbf{O}_{1:T};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha},\mathbf{O}_{1:T})}\sum_{\mathbf{q}}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O}_{1:T},\mathbf{w},\mathbf{q})\right)$$

$$\boldsymbol{\phi}(\mathbf{O}_{1:T},\mathbf{w},\mathbf{q}) = \left[\begin{array}{c}\boldsymbol{\phi}_{1}(\mathbf{w}) \\ \boldsymbol{\phi}_{\mathsf{a}}(\mathbf{O}_{1:T},\mathbf{w},\mathbf{q})\end{array}\right]$$

  - $\boldsymbol{\phi}_1(\mathbf{w})$ may be replaced by $\log(P(\mathbf{w}))$
  - allows LM text training data to be used

HCRF Features

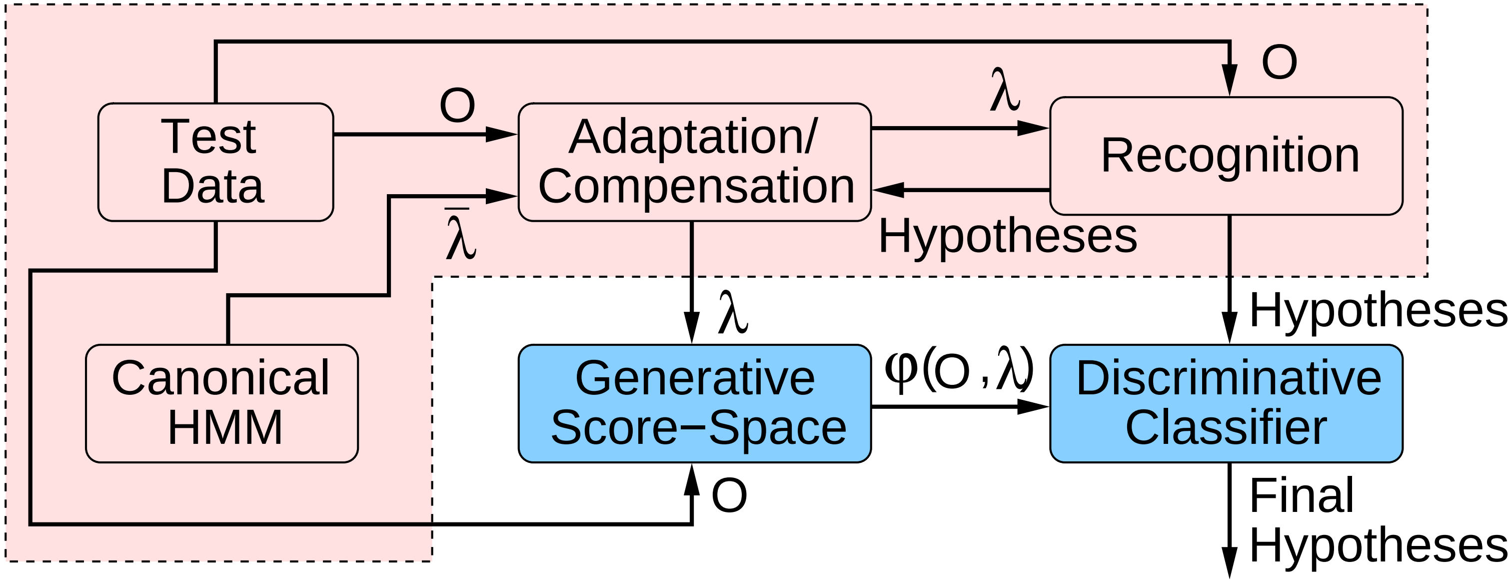


$$\phi_{\text{a}}(\mathbf{O}_{1:T}, \mathbf{w}, \mathbf{q}) = \left[\begin{array}{c} \vdots \\ \sum_{t=1}^{T} \delta(q_{t-1} - s_i)\delta(q_t - s_i) \\ \sum_{t=1}^{T} \delta(q_t - s_i) \\ \sum_{t=1}^{T} \delta(q_t - s_i)\mathbf{o}_t \\ \sum_{t=1}^{T} \delta(q_t - s_i)\text{vec}(\mathbf{o}_t\mathbf{o}_t^{\mathsf{T}}) \\ \vdots \end{array}\right]$$

- Example features used with HCRFs:
  - features the same as those associated with a generative HMM
  - state "distributions" not required to be valid individual PDFs
- Using these features closely related to discriminatively trained HMM [28]

Interest in modifying features extracted from sequence

Combined Discriminative and Generative Models



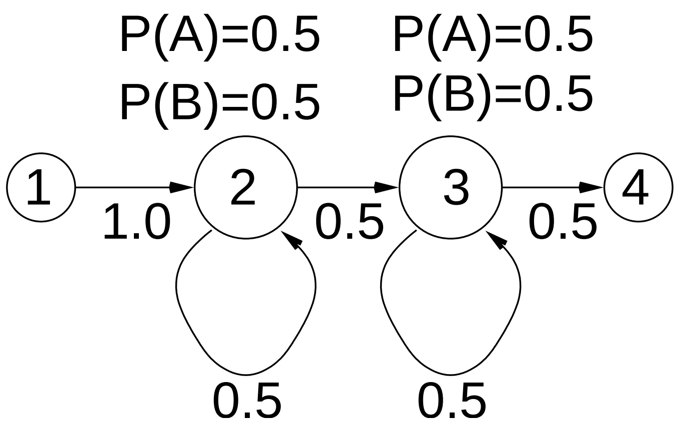
- Use generative model to extract features [29, 30] (we do like HMMs!)
  - adapt generative model - speaker/noise independent discriminative model
- Use favourite form of discriminative classifier for example
  - log-linear model/logistic regression
  - binary/multi-class support vector machines

Generative Score-Spaces (Features)

- Possible generative score-spaces:

$$\phi(\boldsymbol{O};\boldsymbol{\lambda}) = \left[\begin{array}{c} \log(P(\boldsymbol{O};\boldsymbol{\lambda}^{(1)})) \\ \vdots \\ \log(P(\boldsymbol{O};\boldsymbol{\lambda}^{(K)})) \end{array}\right]; \quad \phi(\boldsymbol{O};\boldsymbol{\lambda}) = \left[\begin{array}{c} \log\left(P(\boldsymbol{O};\boldsymbol{\lambda}^{(1)})\right) \\ \boldsymbol{\nabla}_{\lambda}\log\left(P(\boldsymbol{O};\boldsymbol{\lambda}^{(1)})\right) \\ \vdots \end{array}\right]$$

- Derivatives extend dependencies - Consider 2-class, 2-symbol $\{\mathtt{A},\mathtt{B}\}$ problem:
  - Class ω_1: AAAA, BBBB — not separable using ML HMM
  - Class ω_2: AABB, BBAA — linearly separable with second-order-features

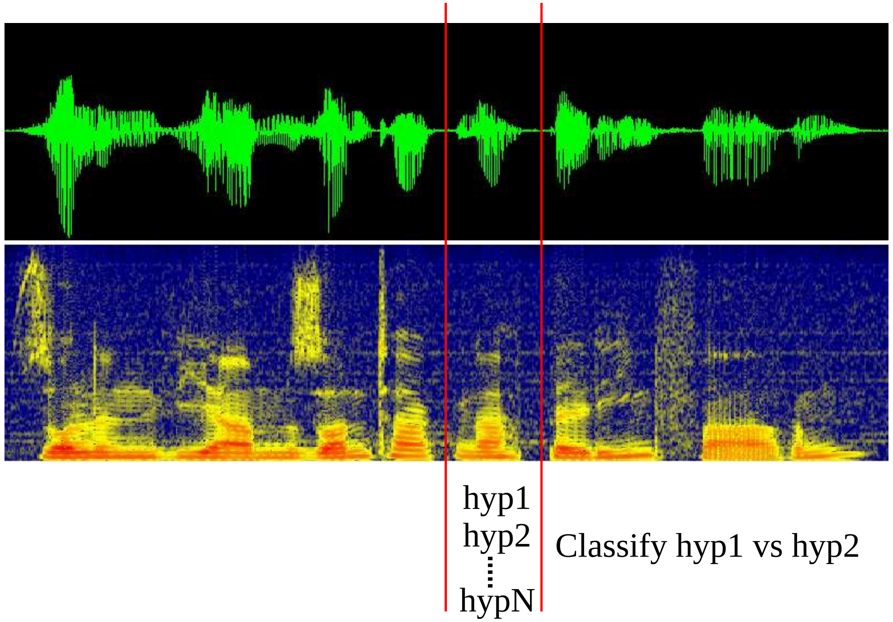


Feature	Class ω_1		Class ω_2	
	AAAA	BBBB	AABB	BBAA
Log-Lik	-1.11	-1.11	-1.11	-1.11
∇_{2A}	0.50	-0.50	0.33	-0.33
$\nabla_{2A}\nabla_{2A}^{\mathsf{T}}$	-3.83	0.17	-3.28	-0.61
$\nabla_{2A}\nabla_{3A}^{\mathsf{T}}$	-0.17	-0.17	-0.06	-0.06

Combined Generative and Discriminative Classifiers

- For continuous speech recognition number of possible word sequence $\mathbf{w}$ vast
  - makes discriminative style models problematic
  - hard to simply incorporate structure into discriminative models

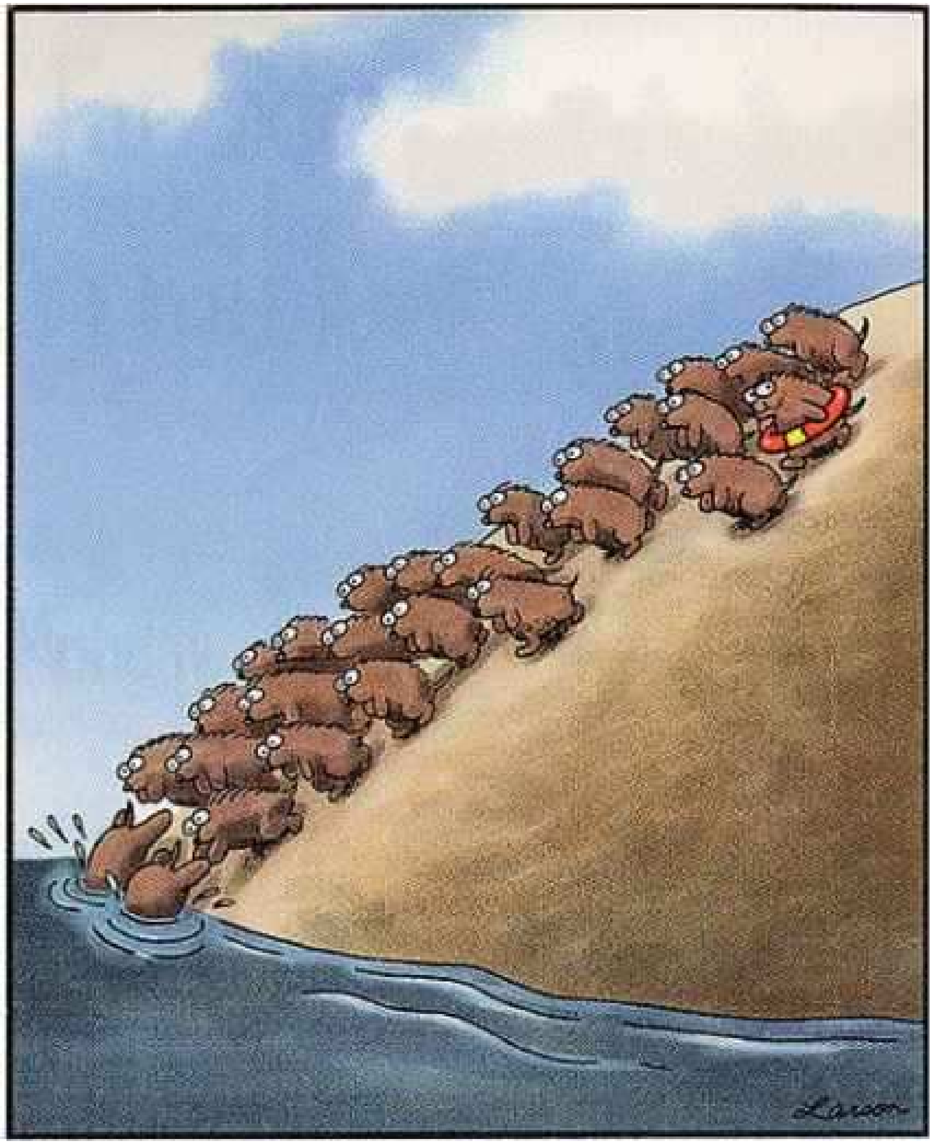
- Acoustic Code-Breaking [31]



- Use HMM-based classifier to:
  - identify possible boundaries
  - identify possible confusions
- Use classify to resolve confusions
  - can use binary classifiers
  - or limit possible alternatives

Summary

- Hidden Markov Models still the dominant form of acoustic model
  - generalisation is still a major problem
- Adaptive training handles inhomogeneous data
  - probably more important for noise than speaker
- Discriminative training yields significant performance gains over ML
  - large margin approaches currently popular and very interesting
- Discriminative models alternative to generative models
  - able to use a wide-range of features (generative scores one option)
  - hard to determine how to incorporate structure


Larson

References

[1] M.J.F. Gales, D.Y. Kim, P.C. Woodland, H.Y. Chan, D. Mrva, R. Sinha, and S.E. Tranter, "Progress in the CU-HTK Broadcast News transcription system," *IEEE Transactions Audio, Speech and Language Processing*, 2006.

[2] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[3] M J F Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[4] A Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.

[5] P. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[6] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector taylor series for noisy speech recognition," in *Proc. ICSLP*, Beijing, China, Oct. 2000.

[7] AP Varga and RK Moore, "Hidden Markov model decomposition of speech and noise," in *Proc ICASSP*, 1990, pp. 845–848.

[8] F. Flego and M. J. F. Gales, "Incremental predictive and adaptive noise compensation," in *Proc. ICASSP*, Taipei, Taiwan, 2009.

[9] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proceedings ICSLP*, 1996, pp. 1137–1140.

[10] H. Liao and M. J. F. Gales, "Adaptive Training with Joint Uncertainty Decoding for Robust Recognition of Noisy Data," in *Proc. ICASSP*, Honolulu, USA, Apr. 2007, vol. 4, pp. 389–392.

[11] Q. Huo and Y. Hu, "Irrelevant variability normalization based hmm training using vts approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1042–1045.

[12] O. Kalinli, M.L. Seltzer, and A. Acero, "Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3825–3828.

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

[14] P.S. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Information Theory*, 1991.

[15] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, pp. 25–47, 2002.

[16] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, 1992.

[17] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Proc. ICSLP*, 2000.

[18] W. Byrne, "Minimum Bayes risk estimation and decoding in large vocabulary continuous speech recognition," *IEICE Special Issue on Statistical Modelling for Speech Recognition*, 2006.

[19] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 2002.

[20] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[21] F. Sha and L.K. Saul, "Large margin gaussian mixture modelling for phonetic classification and recognition," in *ICASSP*, 2007.

[22] J. Li, M. Siniscalchi, and C-H. Lee, "Approximate test risk minimization theough soft margin training," in *ICASSP*, 2007.

[23] G Heigold, T Deselaers, R Schluter, and H Ney, "Modified MMI/MPE: A direct evaluation of the margin in speech recognition," in *Proc. ICML*, 2008.

[24] G Saon and D Povey, "Penalty function maximization for large margin HMM training," in *Proc. Interspeech*, 2008.

[25] G Heigold, G Zweig, and P Nguyen, "A flat deirect model for speech recognition," in *ICASSP*, 2009.

[26] H-K. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Transactions Audio Speech and Language Processing*, 2006.

[27] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden conditional random fields for phone classification," in *Interspeech*, 2005.

[28] G Heigold, R Schlter, and H Ney, "On the equivalence of Gaussian HMM and Gaussian HMM-like hidden conditional random fields," in *Interspeech*, 2007, pp. 1721–1724.

[29] T. Jaakkola and D. Hausser, "Exploiting generative models in disciminative classifiers," in *Advances in Neural Information Processing Systems 11*, S.A. Solla and D.A. Cohn, Eds. 1999, pp. 487–493, MIT Press.

[30] N.D. Smith and M.J.F. Gales, "Speech recognition using SVMs," in *Advances in Neural Information Processing Systems*, 2001.

[31] V. Venkataramani, S. Chakrabartty, and W. Byrne, "Support vector machines for segmental minimum Bayes risk decoding of continuous speech," in *ASRU 2003*, 2003.