Structured Discriminative Models for Speech Recognition

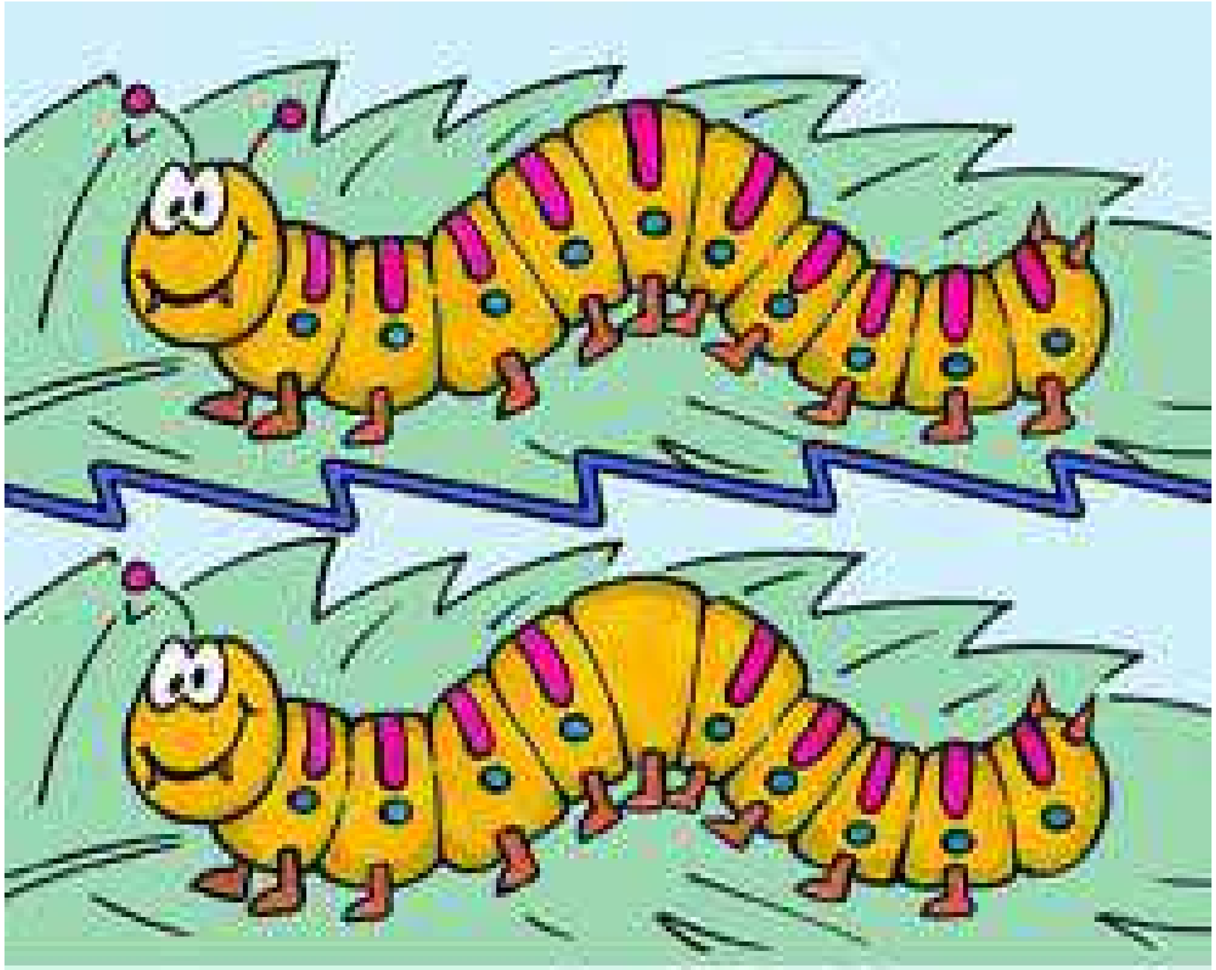
Mark Gales
with Martin Layton, Anton Ragni, Austin Zhang, Rogier van Dalen

December 2012

Cambridge University Engineering Department

ISCSLP 2012

"Spot the Difference"



Overview

- Acoustic Models for Speech Recognition
  - generative models and speech production
  - discriminative models and features
- Training Criteria
  - large-margin-based training
- Combining Generative and Discriminative Models
  - generative score-spaces and log-linear models
  - efficient feature extraction
- Initial Evaluation on Noise Robust Speech Recognition
  - AURORA-2 and AURORA-4 experimental results
- Deep discriminative models
  - integration with hybrid framework?

Generative Models

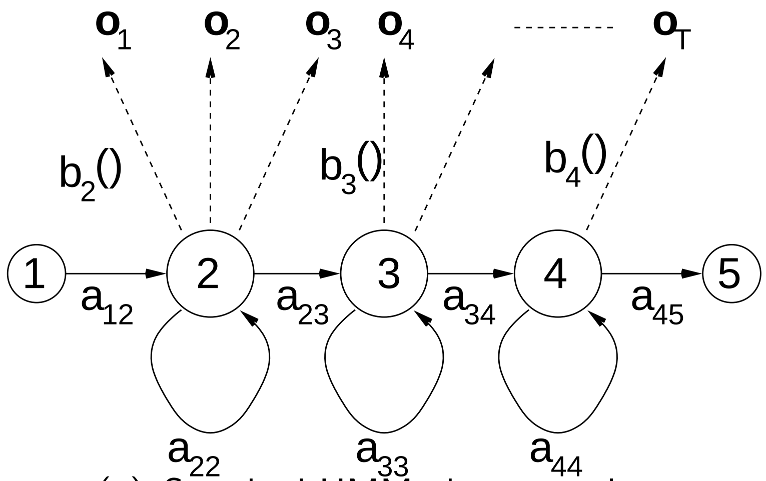
Generative Models

- Need to extract joint distribution between two sequences $p(\mathbf{w}, \mathbf{O})$
  - word sequence $\mathbf{w}$ - can't usually model at sentence level - language model
  - observation sequence $\mathbf{O}$ - usually extracted every 10ms - acoustic model
- Standard generative models - $P(\mathbf{w})p(\mathbf{O}|\mathbf{w};\boldsymbol{\lambda})$ - $\boldsymbol{\lambda}$ model parameters:

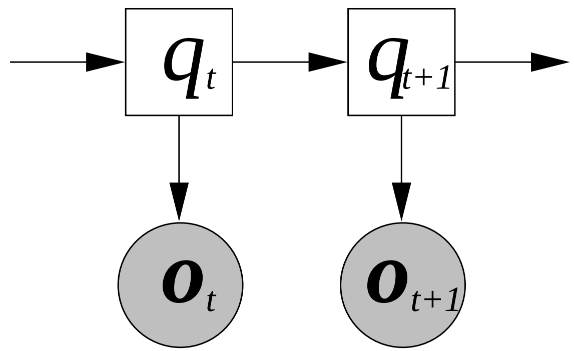
$$p(\mathbf{O}|\mathbf{w};\boldsymbol{\lambda}) = p(\boldsymbol{o}_1|\mathbf{w};\boldsymbol{\lambda})p(\boldsymbol{o}_2|\boldsymbol{o}_1,\mathbf{w};\boldsymbol{\lambda})\ldots p(\boldsymbol{o}_T|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_{T-1},\mathbf{w};\boldsymbol{\lambda})$$

  - impractical to directly model in this form
- Two possible forms of conditional independence used:
  - observed variables
  - latent (unobserved) variables
- Standard sequence model for this: Hidden Markov Model

Hidden Markov Model - A Dynamic Bayesian Network



(a) Standard HMM phone topology



(b) HMM Dynamic Bayesian Network

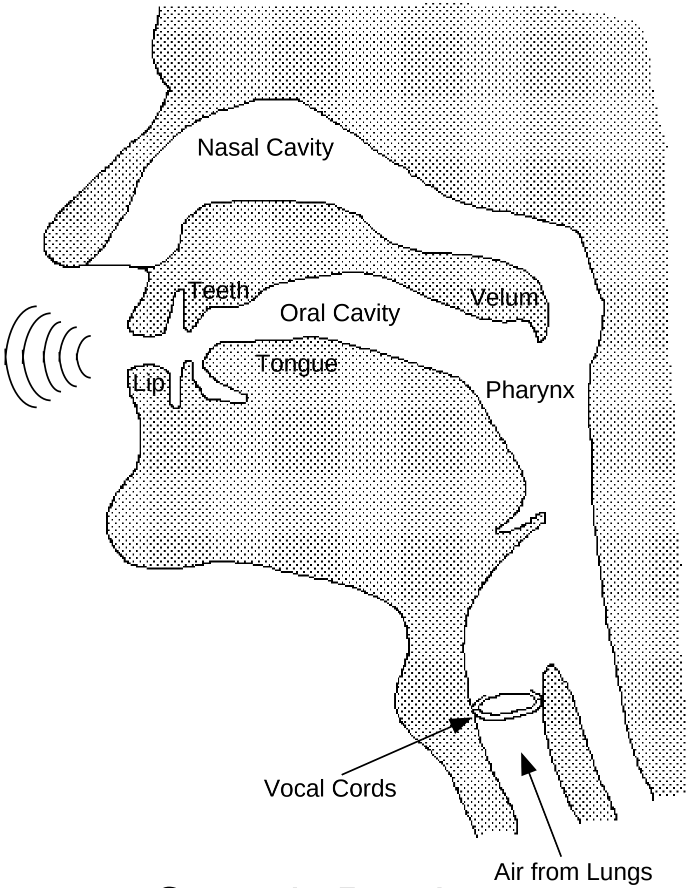
- Notation for DBNs [1]:

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

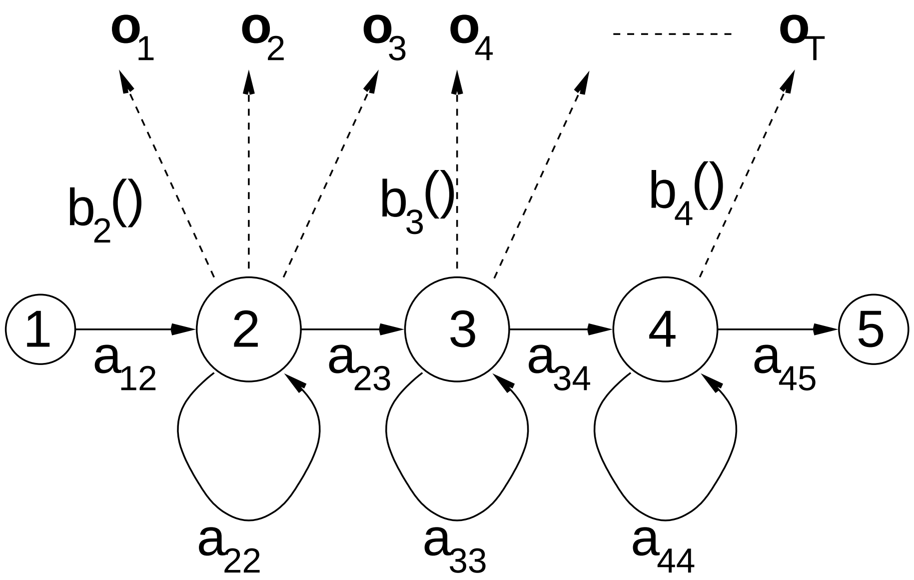
- Observations conditionally independent of other observations given state.
- States conditionally independent of other states given previous states.

$$p(\mathbf{O};\boldsymbol{\lambda}) = \sum_{\mathbf{q}} \prod_{t=1}^{T} P(q_t|q_{t-1})p(\boldsymbol{o}_t|q_t;\boldsymbol{\lambda})$$

Speech Production (1)



(c) Speech Production



(d) HMM Generative Model

- Not modelling the human production process!

Speech Production (2)

Human Production

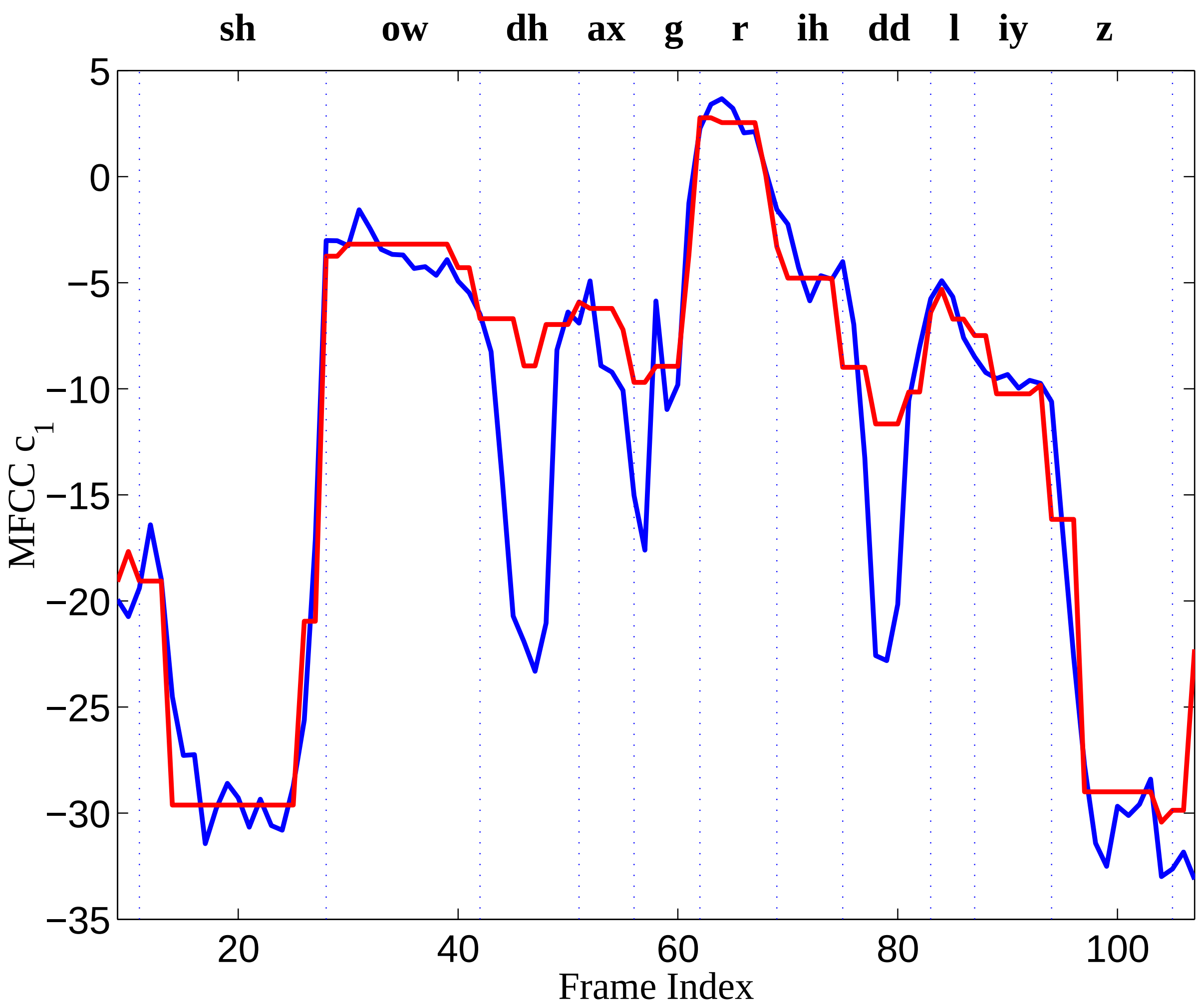
- Acoustic tube:
  - articulators move:
    alter the shape of the vocal tract;
    enable/disable nasal cavity;
  - co-articulation effect.
- Excitation source:
  - vocal cords vibrate producing quasi-periodic sounds (voiced sounds);
  - turbulence caused by forcing air through a constriction in the vocal tract (fricative sounds).
- Speech:
  - sound pressure wave.

HMM Production

- State evolution process
  - discrete state transition after each "observation";
  - probability of entering a state only dependent on the previous state.
- Observation process
  - associated with each state is a probability distribution;
  - observations are assumed independent given the current state.
- Speech representation
  - feature vector every 10ms.

HMM Trajectory Modelling

Frames from phrase:
`SHOW THE GRIDLEY'S ...`



Legend

- True
- HMM

Discriminative Models

Discriminative Models

- Classification requires class posteriors $P(\mathbf{w}|\mathbf{O})$
  - generative model classification use Bayes' rule:

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

- Discriminative model - directly model posterior [2] e.g. Log-Linear Model

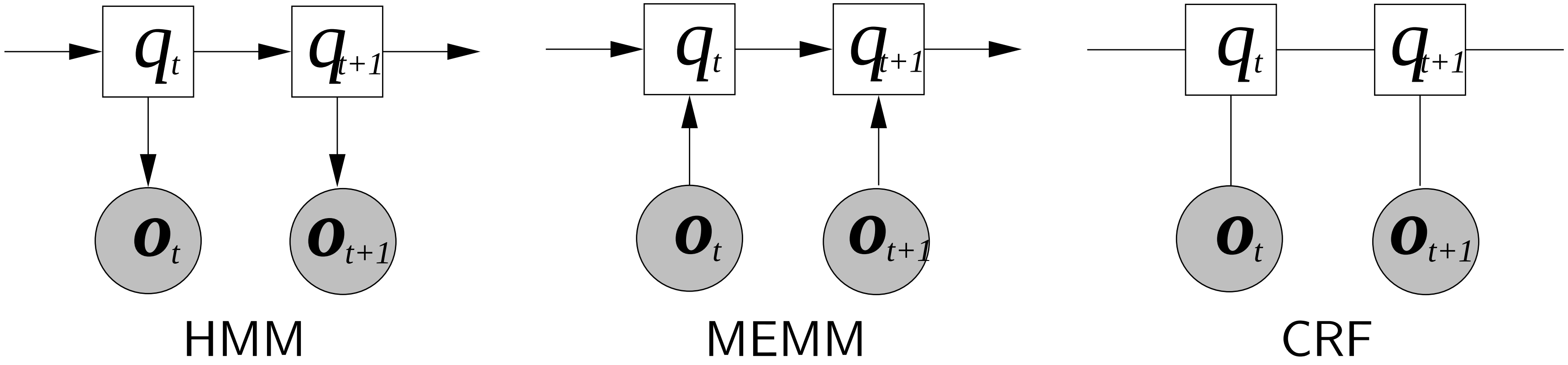
$$P(\mathbf{w}|\mathbf{O};\boldsymbol{\alpha}) = \frac{1}{Z}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O},\mathbf{w})\right)$$

  - normalisation term Z (simpler to compute than generative model)

$$Z = \sum_{\tilde{\mathbf{w}}}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O},\tilde{\mathbf{w}})\right)$$

- Able to use very rich set of features $\boldsymbol{\phi}(\mathbf{O},\mathbf{w})$

Example Standard Sequence Models



- Compute the posteriors of the state-sequence $\mathbf{q}$
  - maximum entropy Markov model [3]

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

  - conditional random field (simplified linear form only) [4]

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

Frame-Level Features

- Discriminative models performance highly dependent on the features
  - basic features - second-order statistics (almost) a discriminative HMM
  - simplest approach extend frame features (for each state $\mathtt{s}_i$) [5]

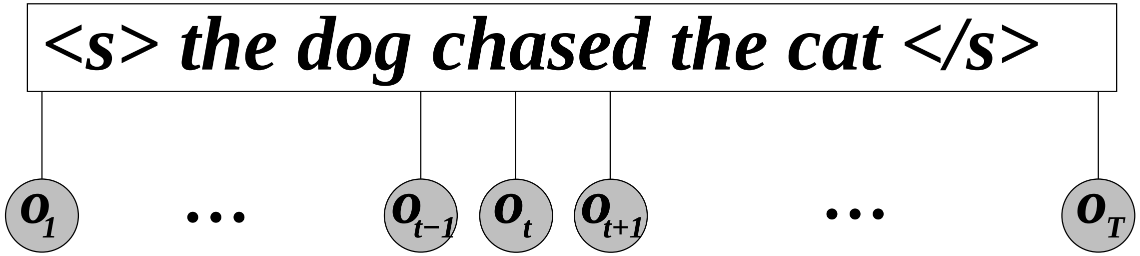
$$\boldsymbol{\phi}(q_t, q_{t-1}, \boldsymbol{o}_t) = \begin{bmatrix} \delta(q_t, \mathtt{s}_i) \\ \delta(q_t, \mathtt{s}_i)\delta(q_{t-1}, \mathtt{s}_j) \\ \delta(q_t, \mathtt{s}_i)\boldsymbol{o}_t \\ \delta(q_t, \mathtt{s}_i)\boldsymbol{o}_t \otimes \boldsymbol{o}_t \\ \delta(q_t, \mathtt{s}_i)\boldsymbol{o}_t \otimes \boldsymbol{o}_t \otimes \boldsymbol{o}_t \end{bmatrix}$$

  - features have same conditional independence assumption as HMM
- Yields a model very similar to discriminatively trained HMM!

How to extend range of features?

  - also care about word sequences $\mathbf{w}$ not state sequences $\mathbf{q}$

Flat Direct Models

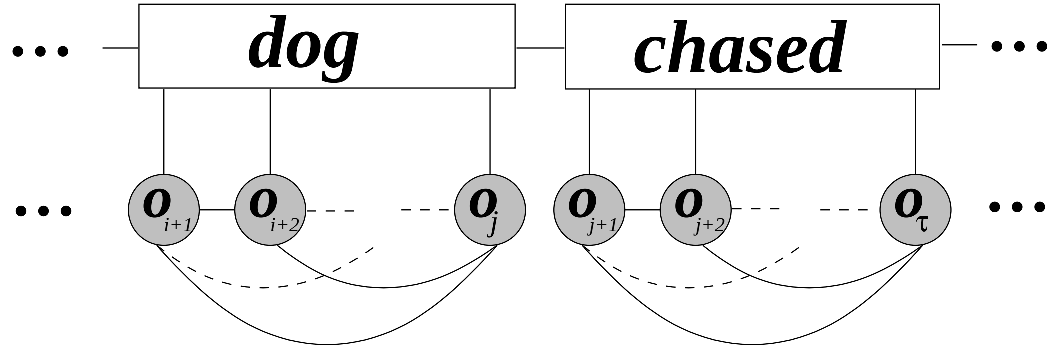


- Remove conditional independence assumptions [6]

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

- Simple model, but lack of structure may cause problems
  - extracted feature-space becomes vast (number of possible sentences)
  - associated parameter vector is vast
  - (possibly) large number of unseen examples

Structured Discriminative Models



- Introduce structure into observation sequence [7] - segmentation $\boldsymbol{a}$
  - comprises: segmentation identity $a^{\mathtt{i}}$, set of observations $\mathbf{O}_{\{a\}}$

$$P(\mathbf{w}|\mathbf{O}) = \frac{1}{Z}\sum_{\boldsymbol{a}} \exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\left[\sum_{\tau=1}^{|\boldsymbol{a}|}\boldsymbol{\phi}(\mathbf{O}_{\{a_\tau\}}, a^{\mathtt{i}}_\tau)\right]\right)$$

  - segmentation may be at word, (context-dependent) phone, etc etc
- What form should $\boldsymbol{\phi}(\mathbf{O}_{\{a_\tau\}}, a^{\mathtt{i}}_\tau)$ have?
  - must be able to handle variable length $\mathbf{O}_{\{a_\tau\}}$

"1-Best" Segmentation

- Not necessary to marginalise over all segmentations
  - could just select the best single segmentation

$$\{\hat{\mathbf{w}}, \hat{\boldsymbol{a}}\} = \operatorname*{argmax}_{\mathbf{w},\boldsymbol{a}} P(\mathbf{w}, \boldsymbol{a}|\mathbf{O}) = \operatorname*{argmax}_{\mathbf{w},\boldsymbol{a}} \left\{ \exp\left(\boldsymbol{\alpha}^{\mathsf{T}} \left[\sum_{\tau=1}^{|\boldsymbol{a}|} \boldsymbol{\phi}(\mathbf{O}_{\{a_\tau\}}, a_\tau^{\mathtt{i}}) \right] \right) \right\}$$

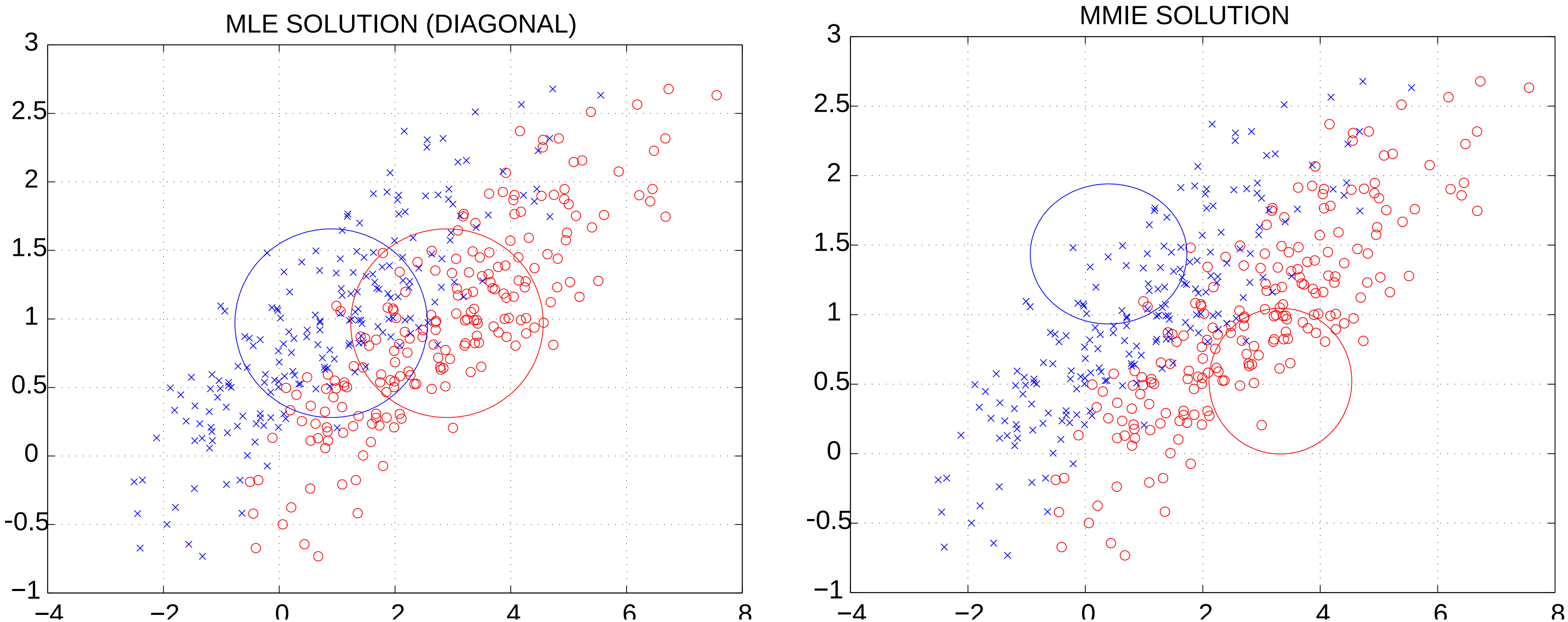
  - need to search over all possible segmentations and word sequences

- Rather than using optimal segmentation - just use a good one
  - one candidate: HMM segmentation $\hat{\boldsymbol{a}}_{\mathtt{hmm}}$
  - not optimal for model, but efficient ...

Training Criteria

Simple MMIE Example

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



- Discriminative criteria a function of posteriors $P(\mathbf{w}|\mathbf{O};\boldsymbol{\lambda})$
  - use to train the discriminative model parameters $\boldsymbol{\alpha}$

Discriminative Training Criteria

- Apply discriminative criteria to train discriminative model parameters$\boldsymbol{\alpha}$
  - Conditional Maximum Likelihood (CML) [22, 23]: maximise

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

  - Minimum Classification Error (MCE) [24]: minimise

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

  - Minimum Bayes' Risk (MBR) [25, 26]: minimise

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

MBR Loss Functions for ASR

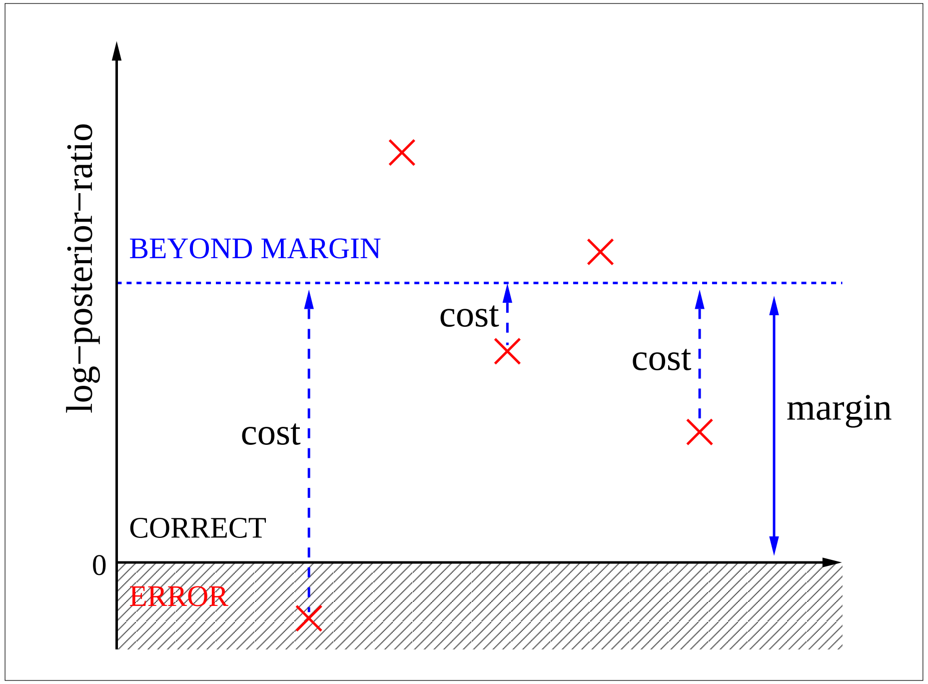
- Sentence (1/0 loss):

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

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

- 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 "errors" observed
  - this is known as Minimum Phone Error (MPE) training [27, 28].
- 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

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

to be beyond margin

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

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

use hinge-loss $[f(x)]_{+}$. Many variants possible [29, 30, 31, 32]

Relationship to (Structured) SVM

- Commonly add a Gaussian prior for regularisation

$$\mathcal{F}(\boldsymbol{\alpha}) = -\log\left(\mathcal{N}(\boldsymbol{\alpha}; \boldsymbol{\mu}_{\alpha}; \boldsymbol{\Sigma}_{\alpha})\right) + \mathcal{F}_{\texttt{lm}}(\boldsymbol{\alpha})$$

- Make the posteriors a log-linear model ($\boldsymbol{\alpha}$) with generative score-space ($\boldsymbol{\lambda}$) [33]
  - restrict parameters of the prior: $\mathcal{N}(\boldsymbol{\alpha}; \boldsymbol{\mu}_{\alpha}; \boldsymbol{\Sigma}_{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{0}, C\mathbf{I})$
  - single (best) segmentation only considered

$$\mathcal{F}(\boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{\alpha}||^2 + \frac{C}{R}\sum_{r=1}^{R}\left[\max_{\mathbf{w}\neq\mathbf{w}_{\texttt{ref}}^{(r)}}\left\{\mathcal{L}(\mathbf{w}, \mathbf{w}_{\texttt{ref}}^{(r)}) - \log\left(\frac{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O}^{(r)}, \mathbf{w}_{\texttt{ref}}^{(r)})}{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O}^{(r)}, \mathbf{w})}\right)\right\}\right]_{+}$$

- Standard result - it's a structured SVM [34, 33] - link with log-linear model
  - able to use more informative priors, for example, non-zero mean [33]

Combining Generative & Discriminative Models

Possible Features (Summary)

Feature type	Example Representation	Example papers
Gaussian sufficient statistics	$\delta(a_i^{\text{i}}, v_j)$ $\delta(a_i^{\text{i}}, v_j)\boldsymbol{o}_t$ $\delta(a_i^{\text{i}}, v_j)\text{diag}(\boldsymbol{o}_t\boldsymbol{o}_t^{\mathsf{T}})$	[4, 8, 5]
Local discriminant functions, e.g. MLP posteriors, closest Gaussians, or HMMs	$\delta(a^{\text{i}}, v_j)P(\mathbf{v}\|\boldsymbol{o}_t)$	[3, 9, 10, 11]
Segment-level score spaces	$\delta(a_i^{\text{i}}, v_1)\boldsymbol{\phi}(\mathbf{O}_{\{a_i\}})$	[12, 13, 14, 15]
Segment-level model features	$\delta(a_i^{\text{i}}, v_j)\boldsymbol{\phi}(\mathbf{O}_{\{a_i\}}, \mathbf{v},)$	[16, 11]
Suprasegmental features, e.g. word-level features	$\sum_{\tau=1}^{L}\delta\left(w_\tau, \text{dog}\right)$	[17, 16, 18, 19, 20, 21]

Interesting option: **segment-level features based on generative models**

Segment-Level Features

- Sequence data (e.g. speech) has inherent variability in the number of samples:

The	cat	sat	on	the	mat	1200 frames

$$\mathbf{O}_1 = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{1200}\}$$

The	cat	sat	on	the	mat	900 frames

$$\mathbf{O}_2 = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{900}\}$$

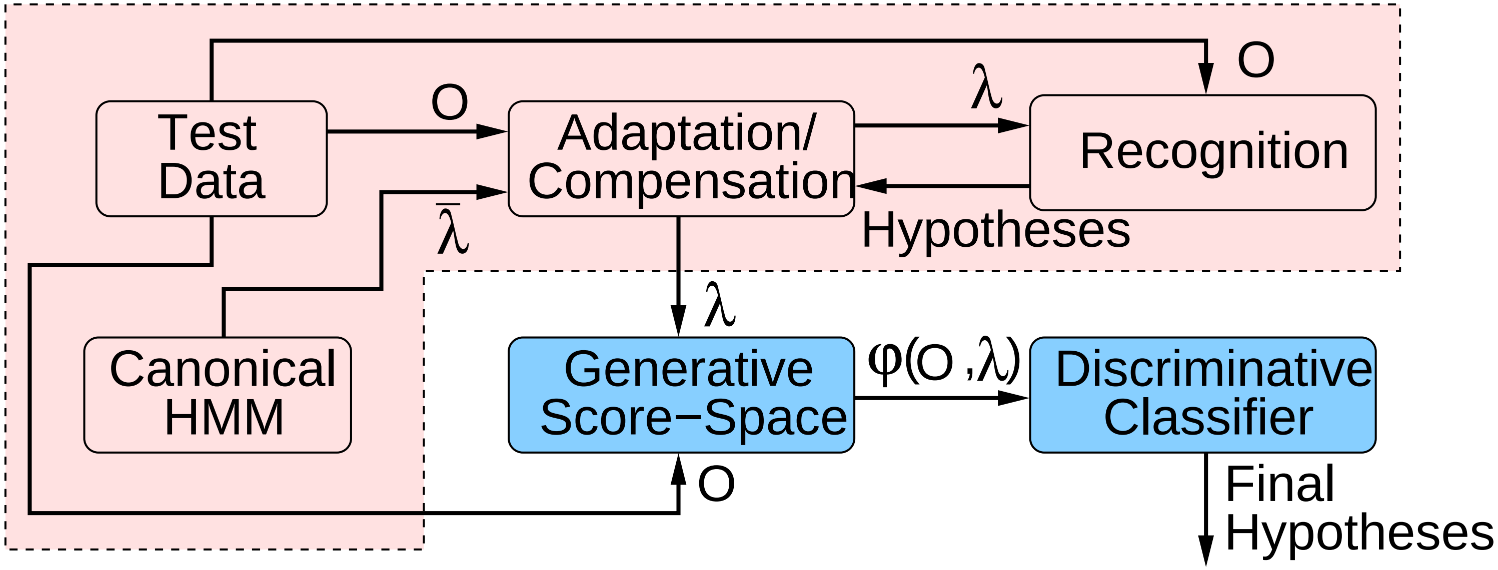
- Standard approach used in sequence generative models - log-likelihood

$$\phi(\mathbf{O}_{\{a_\tau\}}, a_\tau^{\mathtt{i}}; \boldsymbol{\lambda}) = \log\left(p(\mathbf{O}_{\{a_\tau\}}; \boldsymbol{\lambda}^{(a_\tau^{\mathtt{i}})})\right)$$

  – $\boldsymbol{\lambda}$ are the model parameters
  – standard HMM-based speech recognition has this form

- Discriminative models can make use of far richer features ...

Combining Discriminative and Generative Models



- Use generative model to extract features [12, 35] (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/structured support vector machines

Derivative Score-Spaces

- What other features can be extracted using generative models?
  - what about using score-spaces from Fisher kernels (and extensions)?

$$\phi(\mathbf{O}_{\{a_\tau\}}, a^{\mathtt{i}}_\tau; \boldsymbol{\lambda}) = \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \log\left(p(\mathbf{O}_{\{a_\tau\}}; \boldsymbol{\lambda}^{(a^{\mathtt{i}}_\tau)})\right)$$

  - does this help with the dependencies?

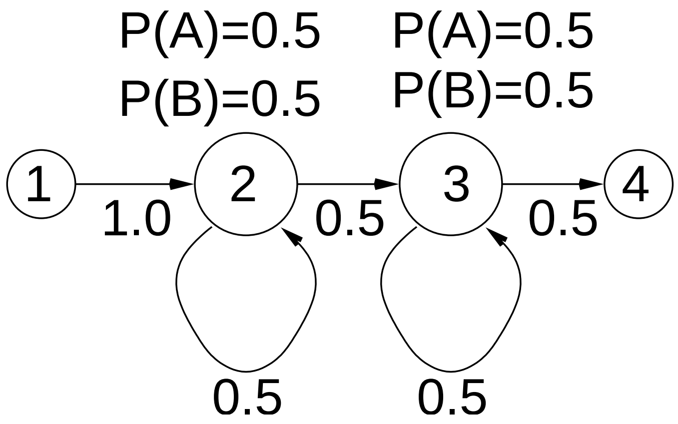
- For an HMM the mean derivative elements become

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}^{(jm)}} \log(p(\mathbf{O}_{\{a_\tau\}}, a^{\mathtt{i}}_\tau; \boldsymbol{\lambda})) = \sum_{t\in\{a_\tau\}} P(\mathbf{q}_t = \{\theta_j, m\}|\mathbf{O}; \boldsymbol{\lambda})\boldsymbol{\Sigma}^{(jm)\text{-}1}(\boldsymbol{o}_t - \boldsymbol{\mu}^{(jm)})$$

  - state/component posterior a function of complete sequence $\mathbf{O}$
  - introduces longer term dependencies
  - different conditional-independence assumptions than generative model

Score-Space Dependencies

- Consider a simple 2-class, 2-symbol $\{\mathtt{A}, \mathtt{B}\}$ problem:
  - Class ω_1: AAAA, BBBB
  - Class ω_2: AABB, BBAA



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
$\nabla_{2A}\nabla^{\mathsf{T}}_{2A}$	-3.83	0.17	-3.28	-0.61
$\nabla_{2A}\nabla^{\mathsf{T}}_{3A}$	-0.17	-0.17	-0.06	-0.06

- ML-trained HMMs are the same for both classes
- First derivative classes separable, but not linearly separable
  - also true of second derivative within a state
- Second derivative across state linearly separable

Score-Spaces for ASR

- Forms of score-space used in the experiments:

$$\boldsymbol{\phi}^{\mathsf{a}}_0(\mathbf{O};\boldsymbol{\lambda}) = \begin{bmatrix} \log\left(p(\mathbf{O};\boldsymbol{\lambda}^{(1)})\right) \\ \vdots \\ \log\left(p(\mathbf{O};\boldsymbol{\lambda}^{(K)})\right) \end{bmatrix}; \quad \boldsymbol{\phi}^{\mathsf{b}}_{1\mu}(\mathbf{O};\boldsymbol{\lambda}) = \begin{bmatrix} \log\left(p(\mathbf{O};\boldsymbol{\lambda}^{(i)})\right) \\ \boldsymbol{\nabla}_{\boldsymbol{\mu}^{(i)}}\log\left(p(\mathbf{O};\boldsymbol{\lambda}^{(i)})\right) \end{bmatrix}$$

  - appended log-likelihood: $\boldsymbol{\phi}^{\mathsf{a}}_0(\mathbf{O};\boldsymbol{\lambda})$
  - derivative (means only for class ω_i): $\boldsymbol{\phi}^{\mathsf{b}}_{1\mu}(\mathbf{O};\boldsymbol{\lambda})$
  - log-likelihood (for class ω_i): $\boldsymbol{\phi}^{\mathsf{b}}_0(\mathbf{O};\boldsymbol{\lambda}) = \left[\log\left(p(\mathbf{O};\boldsymbol{\lambda}^{(i)})\right)\right]$

- In common with most discriminative models Joint Feature Spaces,

$$\boldsymbol{\phi}(\mathbf{O},\boldsymbol{a};\boldsymbol{\lambda}) = \begin{bmatrix} \sum_{\tau=1}^{|\boldsymbol{a}|}\delta(a^{\mathtt{i}}_{\tau}, w^{(1)})\boldsymbol{\phi}(\mathbf{O}_{\{a_\tau\}};\boldsymbol{\lambda}) \\ \vdots \\ \sum_{\tau=1}^{|\boldsymbol{a}|}\delta(a^{\mathtt{i}}_{\tau}, w^{(P)})\boldsymbol{\phi}(\mathbf{O}_{\{a_\tau\}};\boldsymbol{\lambda}) \end{bmatrix}$$

for $\boldsymbol{\alpha}$-tied yielding "units" $\{w^{(1)},\ldots,w^{(P)}\}$, underlying score-space $\boldsymbol{\phi}(\mathbf{O};\boldsymbol{\lambda})$.

Handling Speaker/Noise Differences

- A standard problem with discriminative approaches is adaptation/robustness
  - not a problem with generative kernels/score-spaces
  - adapt generative models using model-based adaptation
- Standard approaches for speaker/environment adaptation
  - (Constrained) Maximum Likelihood Linear Regression [36]

$$\boldsymbol{x}_t = \mathbf{A}\boldsymbol{o}_t + \mathbf{b}; \quad \boldsymbol{\mu}^{(m)} = \mathbf{A}\boldsymbol{\mu}_{\mathrm{x}}^{(m)} + \mathbf{b}$$

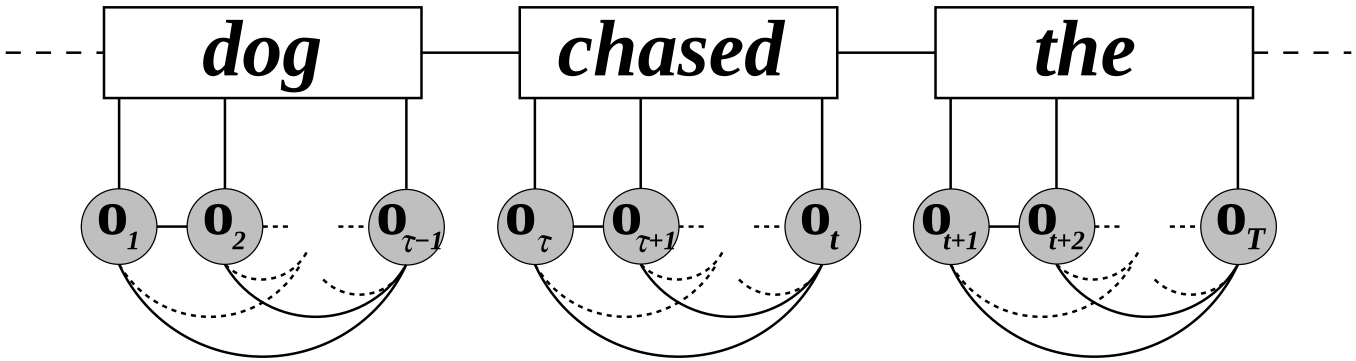
  - Vector Taylor Series Compensation [37] (used in this work)

$$\boldsymbol{\mu}^{(m)} = \mathbf{C}\log\left(\exp(\mathbf{C}^{-1}(\boldsymbol{\mu}_{\mathrm{x}}^{(m)} + \boldsymbol{\mu}_{\mathrm{h}}^{(m)})) + \exp(\mathbf{C}^{-1}\boldsymbol{\mu}_{\mathrm{n}}^{(m)})\right)$$

- Discriminative model parameters speaker/noise independent.

Efficient Feature Extraction

Structured Discriminative Models

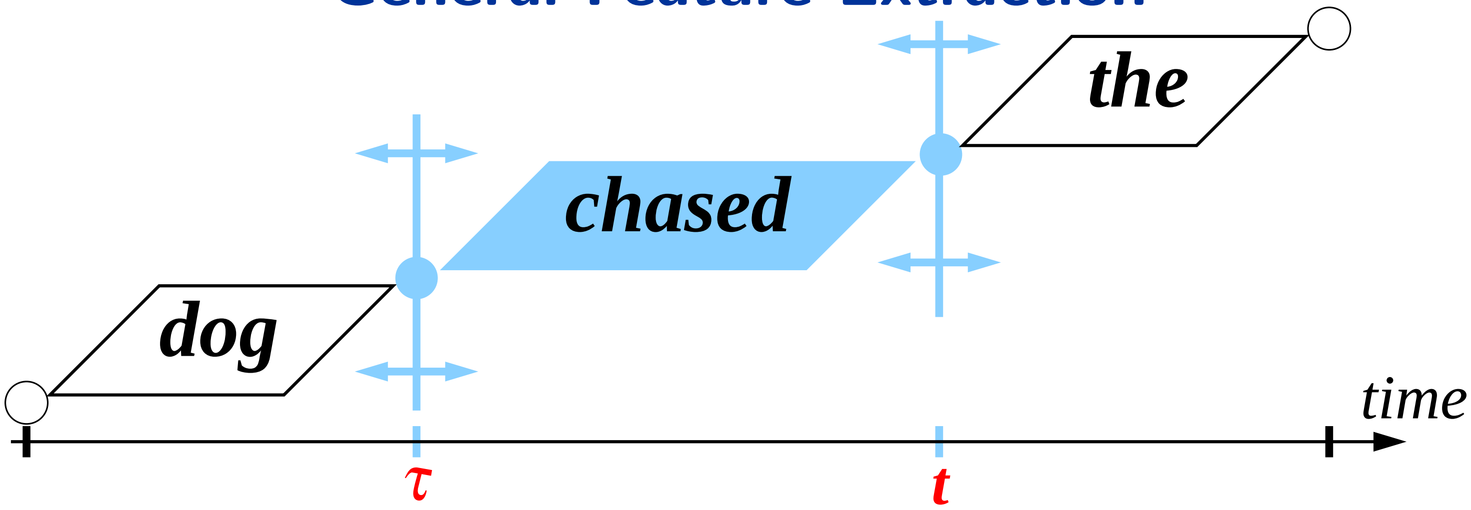


- Consider specifying speech segments as words [38, 16, 39]

$$P(\mathbf{w}_{1:L}|\mathbf{O}_{1:T};\boldsymbol{\alpha}) = \frac{1}{Z}\sum_{\boldsymbol{a}} \exp\left(\boldsymbol{\alpha}^{\mathsf{T}} \sum_{\tau=1}^{|\boldsymbol{a}|} \boxed{\boldsymbol{\phi}\left(\mathbf{O}_{\{a_\tau\}}, a_\tau^{\mathrm{i}}\right)}\right)$$

  - alignment unknown marginalised over in training (or 1-best taken)
- Features extracted from variable length observation sequence $\mathbf{O}_{\{a_\tau\}}$
  - unknown start/end times and segment identity

General Feature Extraction

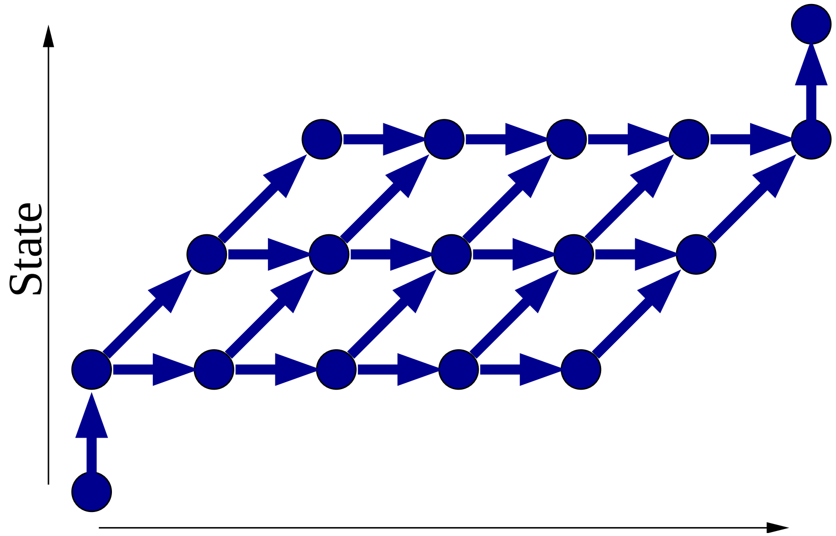


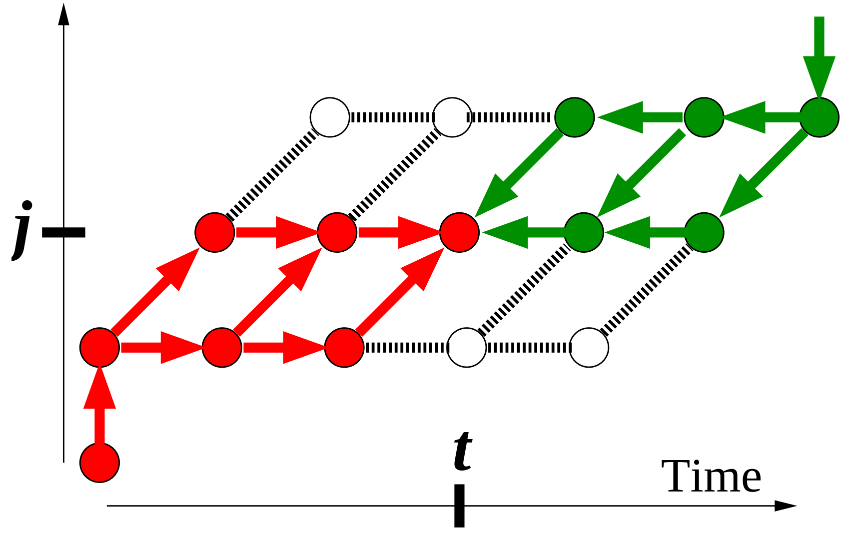
- General features depend on all elements of the observation sequence
  - consider $\phi\left(\mathbf{O}_{\tau:t}, w_l\right)$ for all possible start/end times – T^2 feature evaluations
  - general complexity $\mathcal{O}(T^3)$ – assuming each evaluation $\mathcal{O}(T)$

Computationally expensive!

- BUT extracting features based on HMMs
  - derivative features based on posteriors for each segment ...

Standard HMM Algorithms



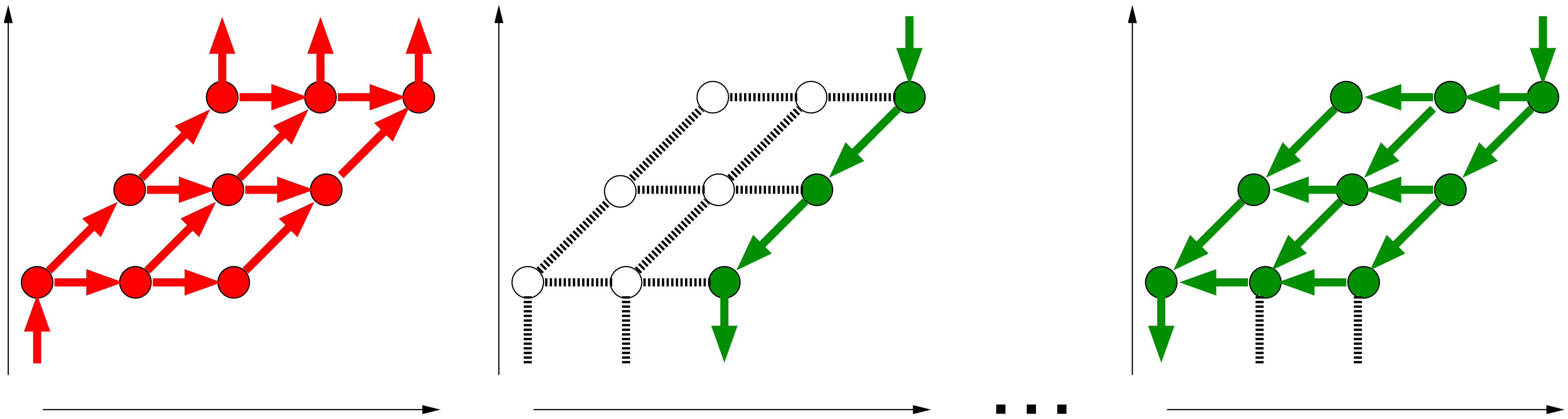


- Efficient training and inference
  - based on forward-backward/Viterbi algorithms

$$\gamma_t^{(j)} = P(q_t^{(j)}|\mathbf{O}_{1:T};\boldsymbol{\lambda}) = \frac{1}{\boxed{p(\mathbf{O}_{1:T};\boldsymbol{\lambda})}} \cdot \boxed{p(\mathbf{O}_{1:t}, q_t^{(j)};\boldsymbol{\lambda})} \cdot \boxed{p(\mathbf{O}_{t+1:T}|q_t^{(j)};\boldsymbol{\lambda})}$$

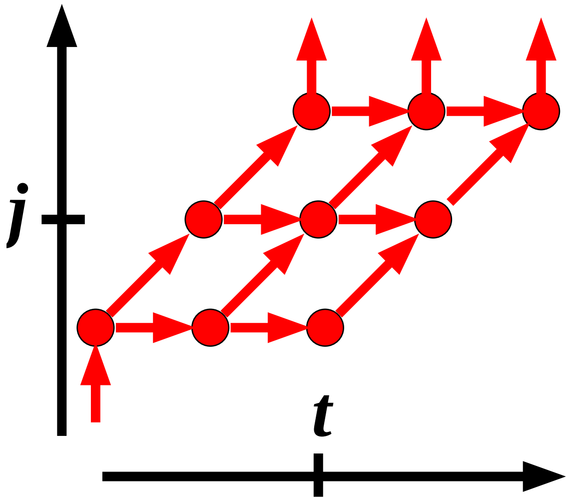
  - time/memory requirement $\mathcal{O}(T) + \mathcal{O}(T)$

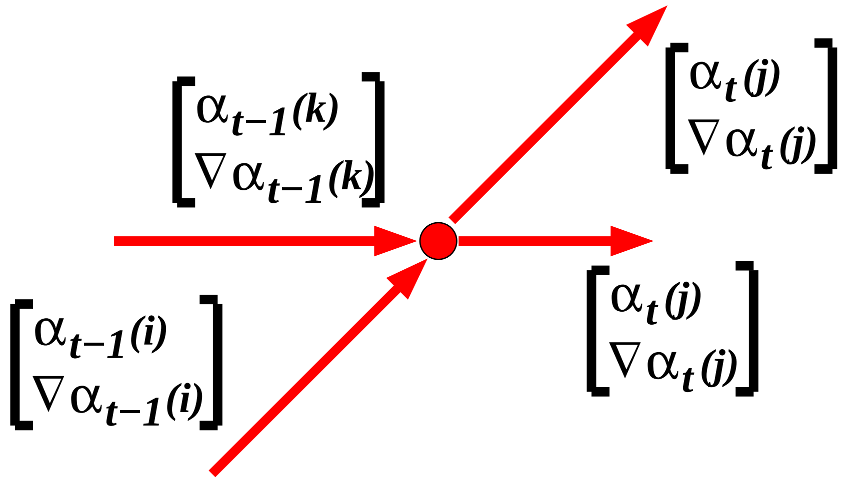
Forward/Backward Caching



- Cache all state-level forward probabilities – $\mathcal{O}(T)$ forward passes
- For each of the possible $\mathcal{O}(T)$ start-times
  - compute backward probabilities – $\mathcal{O}(T)$ possible backward passes
  - intersect of forward/backward yields required posterior
- BUT need to accumulate statistics for each start/end time – total $\mathcal{O}(T^3)$

Expectation Semiring





- Efficient calculation using expectation semirings [40, 15]
  - extend statistics propagated/combined in forward pass
  - scalar summation extended to vector summation
- Expectation semirings allows to accumulate statistics in one pass
  - derivative features can be computed for any node in the trellis - $\mathcal{O}(T^2)$

Evaluation Tasks

Preliminary Evaluation Tasks

- Select challenging task - noise robust speech recognition
  - combine with model-based noise compensation (VTS/VAT/DVAT)
  - artificial tasks reported - same results seen on in-car Toshiba data
- AURORA-2 small vocabulary digit string recognition task
  - whole-word models, 16 emitting-states with 3 components per state
  - clean training data for HMM training - HTK parametrisation SNR
  - Set B and Set C unseen noise conditions even for multi-style data
  - Noise estimated in a ML-fashion for each utterance
- AURORA-4 medium vocabulary speech recognition
  - training data from WSJ0 SI84 to train clean acoustic models
  - state-clustered states, cross-word triphones ($\approx$3K states $\approx$50k components)
  - 5-15dB SNR range of noises added
  - Noise estimated in a ML-fashion for each utterance

AURORA-2 - Derivative Score-Spaces - MWE Criterion

HMM	SDM	$\hat{a}$	Test set A	Test set B	Test set C	Avg
VTS	–	–	9.8	9.1	9.5	9.5
	$\phi^{\text{b}}_{1\mu}$	$\hat{a}_{\text{hmm}}$	7.0	6.6	7.6	**7.0**
		$\hat{a}$	6.8	6.4	7.3	**6.7**
VAT	–	–	8.9	8.3	8.8	8.6
	$\phi^{\text{b}}_{1\mu}$	$\hat{a}_{\text{hmm}}$	6.6	6.5	7.0	**6.6**
		$\hat{a}$	6.2	6.1	6.8	**6.3**
DVAT	–	–	6.7	6.6	7.0	6.7
	$\phi^{\text{b}}_{1\mu}$	$\hat{a}_{\text{hmm}}$	6.1	6.2	6.7	**6.3**
		$\hat{a}$	6.1	6.1	6.6	**6.2**

- Derivative score-spaces ($\phi^{\text{b}}_{1\mu}$) consistent gains over all baseline HMM systems
  – derivative score-space larger (1873 dimensions for each base score-space)
  – adds approximately 50% more parameters to the system

AURORA-4 - Derivative Score-Space - MPE Criterion

System	Test set A	B	C	D	Avg
VTS	7.1	15.3	12.1	23.1	**17.9**
VAT	8.6	13.8	12.0	20.1	**16.0**
DVAT	7.2	12.8	11.5	19.7	**15.3**
VAT+ϕ_0^{b}	7.7	13.1	11.0	19.5	**15.3**
VAT+$\phi_{1\mu}^{\mathrm{b}}$	7.4	12.6	10.7	19.0	**14.8**

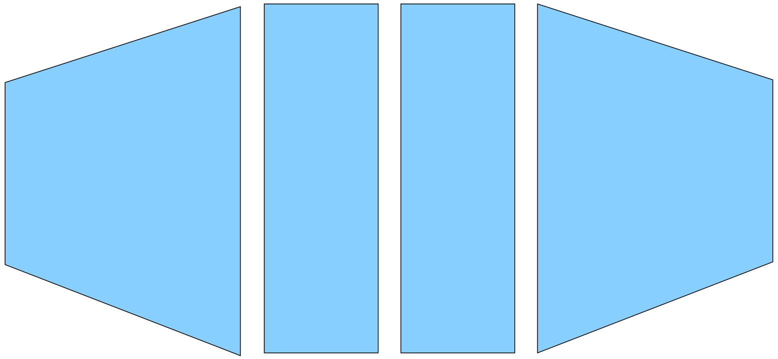
- Contrast of DVAT system with log-linear system (4020 classes)
  - single dimension space (ϕ_0^{b}) with VAT system yields DVAT performance
- Gains from derivative score-space disappointing (limited training data)
  - need to look at DVAT+$\phi_{1\mu}^{\mathrm{b}}$ (need to try on more data)

“Deep” Discriminative Models?

Very Brief History (see Tutorial)

- Form of "generative" model has been fixed =
  - standard HMMs with GMMs and MFCC/PLP features
- Vast interest in Deep Neural Networks [41]
  - resurrect hybrid systems from 1990s ...
- BUT changes to configuration and training yielded large gains
  - MLP targets the distinct states from decision tree clustering;
  - increase the number of hidden layers;
  - improved initialisation (layer by layer training, RBM initialisation).

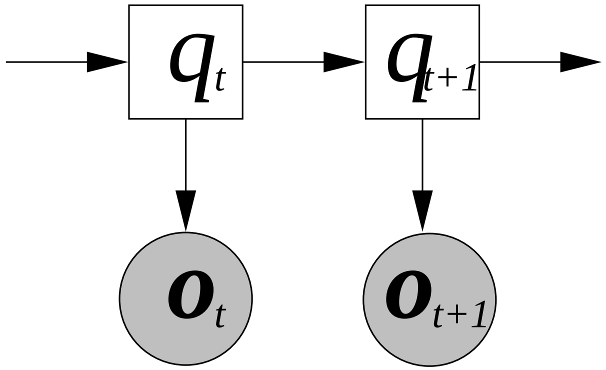
MLP Likelihoods (Hybrid Configuration)



- Replace the GMMs as state output distribution by MLP output
  - simple mapping to yield likelihoods: $p(\mathbf{o}_t|\mathtt{s};\boldsymbol{\lambda}) \propto \frac{P(\mathtt{s}|\mathbf{o}_t;\boldsymbol{\lambda})}{P(\mathtt{s})}$
- Viewed as a specific discriminative model

$$\boldsymbol{\phi}(\mathbf{O},\mathbf{a};\boldsymbol{\lambda}) = \left[\begin{array}{c} \sum_{\tau=1}^{|\mathbf{a}|}\delta(a_\tau^{\mathtt{i}}, w^{(1)})\sum_{t\in\{a_\tau\}}\log(p(\mathbf{o}_t|w^{(1)};\boldsymbol{\lambda})) \\ \vdots \\ \sum_{\tau=1}^{|\mathbf{a}|}\delta(a_\tau^{\mathtt{i}}, w^{(P)})\sum_{t\in\{a_\tau\}}\log(p(\mathbf{o}_t|w^{(P)};\boldsymbol{\lambda})) \end{array}\right]; \quad \boldsymbol{\alpha} = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \end{array}\right]$$

Hybrid Architecture



- BUT same HMM conditional independence assumptions (or features)

What about using more interesting models?

  - approaches for training MLP in sequential fashion already investigated [42]
- Simplest alternative is to train the discriminative model parameters $\boldsymbol{\alpha}$
  - equivalent of class-specific acoustic-to-language model weighting
- Can use all the log-likelihoods, similar to [43, 44]

Hybrid Segment Features?

- How to get interesting $\boldsymbol{\phi}(\mathbf{O}, \mathbf{a}; \boldsymbol{\lambda})$?
  - derive segment-level features
  - number of MLP parameters vast ...
- Alternative - use the MLP output as the parameters (like discrete HMM)
  - take derivatives with respect to parameters $(\lambda_{ti} = p(\mathbf{o}_t|\mathbf{s}_i; \boldsymbol{\lambda}))$ yields

$$\boldsymbol{\nabla}_{\lambda^{(i)}} \log(p(\mathbf{O}_{\{a_\tau\}}; \boldsymbol{\lambda})) = \sum_{t \in \{a_\tau\}} \left(\frac{\gamma_t^{(i)}}{\lambda_{ti}} - K \right)$$

  - introduces dependencies for complete segment
  - large feature-space again (number of targets)
  - could apply L_1 regularisation to achieve sparseness

Interesting combination of two research directions

Conclusions

- Structured Discriminative Models for Speech Recognition
  - flexible framework for including wide-range of features
  - structures allows direct application to speech recognition
  - range of discriminative training criteria - links with structured SVMs
- Combination of generative and discriminative models
  - use generative models to derive features for discriminative model
  - robustness and adaptation achieved by adapting underlying acoustic model
  - structured approach to adding dependencies to the features
  - efficient approaches to obtain features/optimal segmentation
- Deep Discriminative Models
  - research direction for integrating hybrid systems into framework

Interesting classifier options - without throwing away HMMs

Acknowledgements

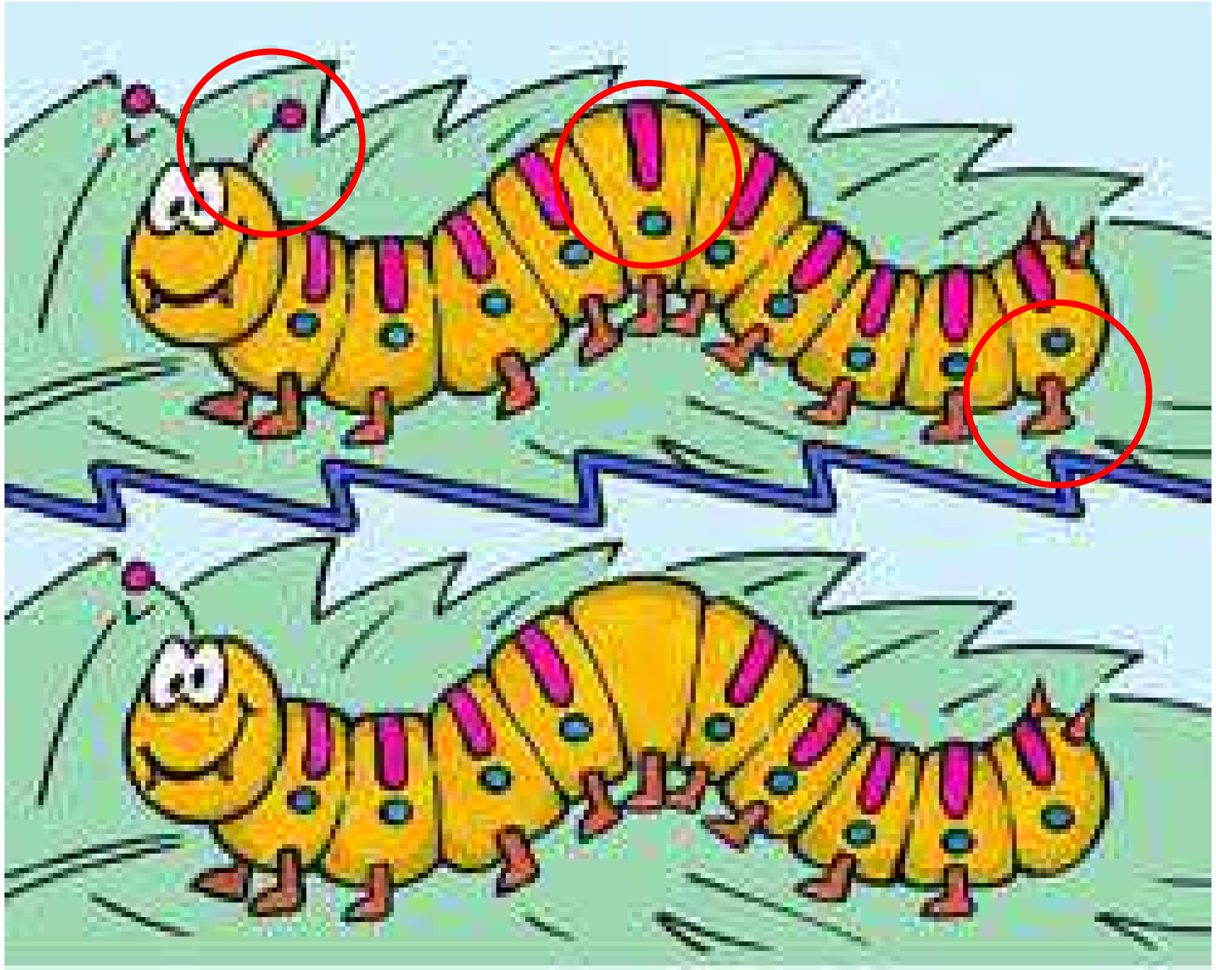
- This work has been funded from the following sources:

TOSHIBA Leading Innovation

EPSRC Pioneering research and skills

  - Toshiba Research Europe Ltd, Cambridge Research Lab
  - EPSRC - Generative Kernels and Score-Spaces for Classification of Speech

“Spot the Difference”



References

[1] J.A. Bilmes, "Graphical models and automatic speech recognition," in *Mathmatical Foundations of Speech and Language Processing*, 2003.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Verlag, 2006.

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

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

[5] S. Wiesler, M. Nußbaum-Thom, G. Heigold, R. Schlüter, and H. Ney, "Investigations on features for log-linear acoustic models in continuous speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009, pp. 52–57.

[6] P. Nguyen, G. Heigold, and G. Zweig, "Speech recognition with flat direct models," *IEEE Journal of Selected Topics in Signal Procesing*, vol. 4, pp. 994–1006, 2010.

[7] M. J. F. Gales, S. Watanabe, and E. Fosler-Lussier, "Structure discrimnative models for speech recognition," *IEEE Signal Procesing Magazine*, 2012.

[8] 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.

[9] J. Morris and E. Fosler-Lussier, "Conditional random fields for integrating local discriminative classifiers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 617–628, 2008.

[10] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 2, pp. 354–365, 2009.

[11] G. Zweig et al, "Speech recognition with segmental conditional random fields: A summary of the JHU CLSP Summer workshop," in *Proceedings of ICASSP*, 2011.

[12] T. Jaakkola and D. Haussler, "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.

[13] M. Layton, *Augmented Statistical Models for Classifying Sequence Data*, Ph.D. thesis, Cambridge University, 2006.

[14] S. Wiesler, A. Richards, Y. Kubo, R. Schlüter, and H. Ney, "Feature selection for log-linear acoustic models," in *Proc. ICASSP'11*, 2011, pp. 5324–5327.

[15] R. C. van Dalen, A. Ragni, and M. J. F. Gales, "Efficient decoding with continuous rational kernels using the expectation semiring," Tech. Rep. CUED/F-INFENG/TR.674, 2012.

[16] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *ASRU*, 2009, pp. 152–157.

[17] S. F. Chen and R. Rosenfeld, "Efficient sampling and feature selection in whole sentence maximum enropy language models," in *Proc. ICASSP'99*, 1999, pp. 549–552.

[18] S. Watanabe, T. Hori, and A. Nakamura, "Large vocabulary continuous speech recognition using WFST-based linear classifier for structured data," in *Proc. Interspeech'10*, 2010, pp. 346–349.

[19] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proc. ACL'04*, 2004.

[20] M. Lehr and I. Shafran, "Learning a discriminative weighted finite-state transducer for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1360–1367, 2011.

[21] T. Oba, T. Hori, A. Nakamura, and A. Ito, "Round-robin duel discriminative language models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1244–1255, 2012.

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.

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

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

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

[31] 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.

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

[33] S.-X. Zhang and M. J. F. Gales, "Structured svms for automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[34] I Tsochantaridis, T Joachims, T Hofmann, and Y Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, 2005.

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

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

[37] A Acero, L Deng, T Kristjansson, and J Zhang, "HMM Adaptation using Vector Taylor Series for Noisy Speech Recognition," in *Proc. ICSLP*, Beijing, China, 2000.

[38] M. Layton, *Augmented statistical models for classifying sequence data*, Ph.D. thesis, Cambridge University, 2006.

[39] A. Ragni and M. J. F. Gales, "Structured discriminative models for noise robust continuous speech recognition," in *ICASSP*, 2011, pp. 4788–4791.

[40] J. Eisner, "Parameter estimation for probabilistic finite-state transducers," in *Proc. ACL*, 2002.

[41] G. Hinton, L. Deng, D Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modelling in speech recognition," *IEEE Signal Procesing Magazine*, 2012.

[42] B. Kingsbury, T.N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed hessian-free optimization," in *Proc. Interspeech*, 2012.

[43] T Do and T Artieres, "Neural conditional random fields," in *Proceedings of the 13th International Conference on Artificial Intell igence and Statistics (AI-STATS)*, 2010.

[44] L Vinel, T Do, and T Artieres, "Joint optimization of hidden conditional random fields and non linear feature extraction," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE*, 2011.