Temporal Trajectory Models

Vijayaditya Peddinti

Error predictors for classifiers in an ensemble

- Given an ensemble of classifiers predict the error of these classifiers for the given test instance.
- Error prediction metrics can be used for
  - Dynamic ensemble selection
  - Dynamic weighted fusion
  - Adaptation of the classifiers in the ensemble
- For discriminative classifiers :
  - Train error prediction models
  - Classifier confidence estimators

Assume the train and test distributions are similar

Error prediction in mismatched conditions

- Use proxy measures in lieu of error rate, based on*
  - ✴ *Familiarity* similarity of classifier behavior for test data and training data
  - ✴ *Conformity* deviation of classifier behavior from gold standard data
  - ✴ *Stability* stability of classifier to minor changes in input

Conformity

- Measure performance of the classifier output on tasks that it is not trained for, but performance on which is deemed necessary for low error
- Acoustic models are trained to estimate phone posterior probabilities, however they are not trained to ensure posterior trajectory quality over time
- Measure the “quality” of the posterior trajectory

*As categorized by Emmanuel Dupoux

Design Requirement

- Design an error estimator which
  - has low data requirement
  - has low computational load
  - can be used as a cost function for unsupervised adaptation or as auxiliary function for supervised training

Design Decisions

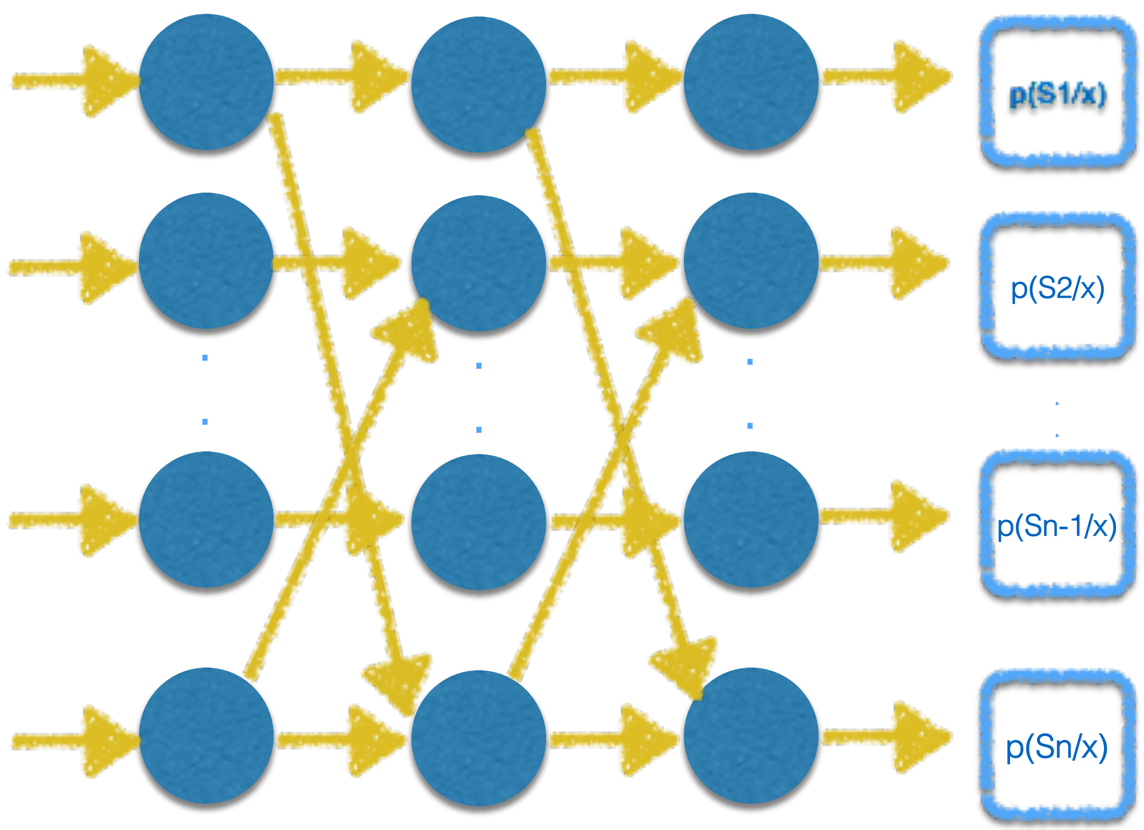
		Distance	
		Deviation of model on test instance from model on train data	Likelihood of test instance given model on train data
Complexity of Models	Simple Models	Requires estimation of only few parameters	A variety of non-speech sources could produce high likelihood instances
	Complex Models	Too many parameters to estimate, data sparsity problem	Very specific to speech Reliable measure derived with minimal data

Temporal Trajectory Models

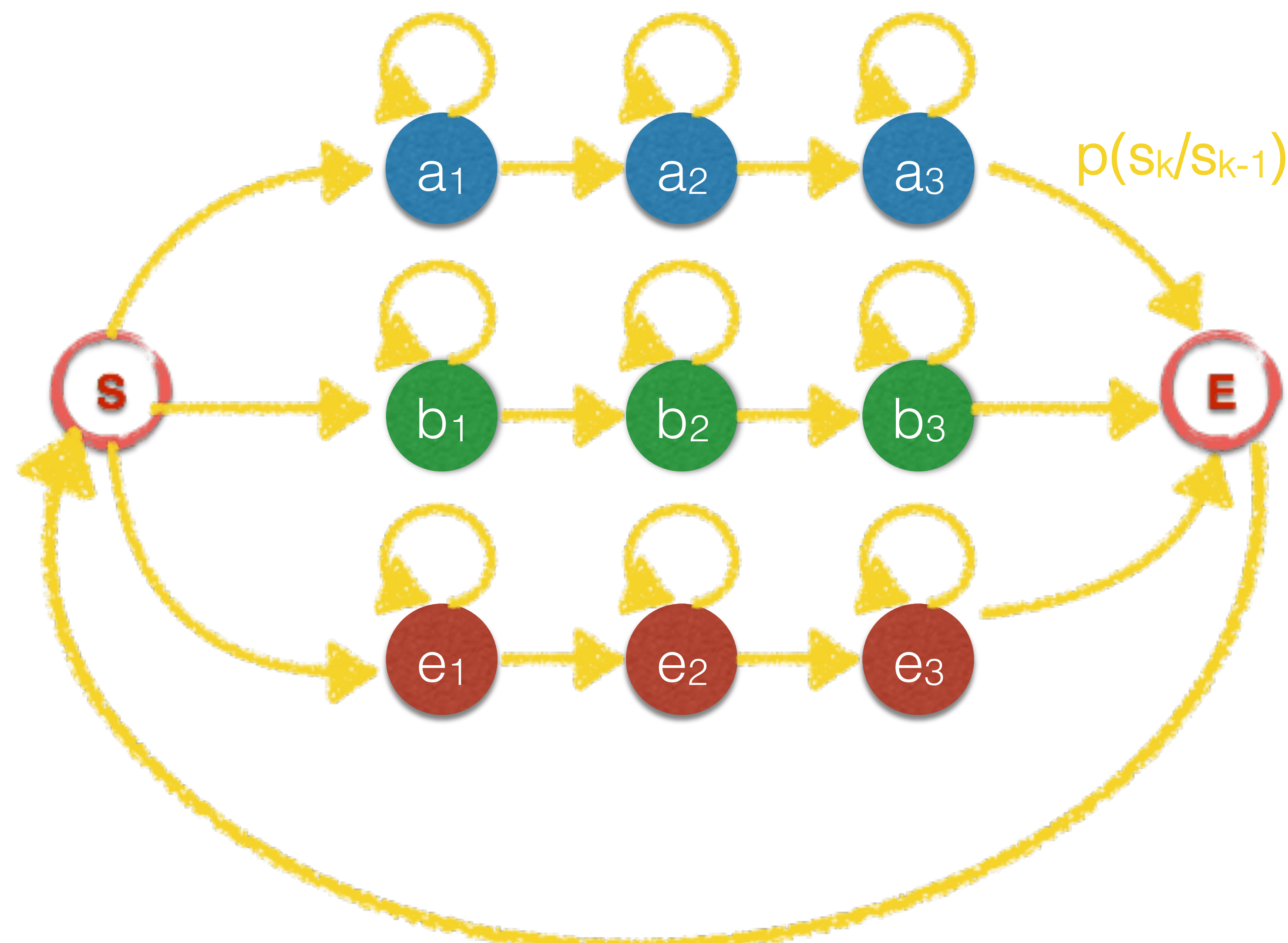
- Speech signal has very specific temporal characteristics which are distinct from noise signals
- Build temporal trajectory models using training data and score the posterior vector trajectories of the test instances
- This is similar to measuring likelihood of the utterance using HMM-DNN hybrid model where the HMM places constraints on posterior trajectory

HMM-DNN Hybrid Decoder

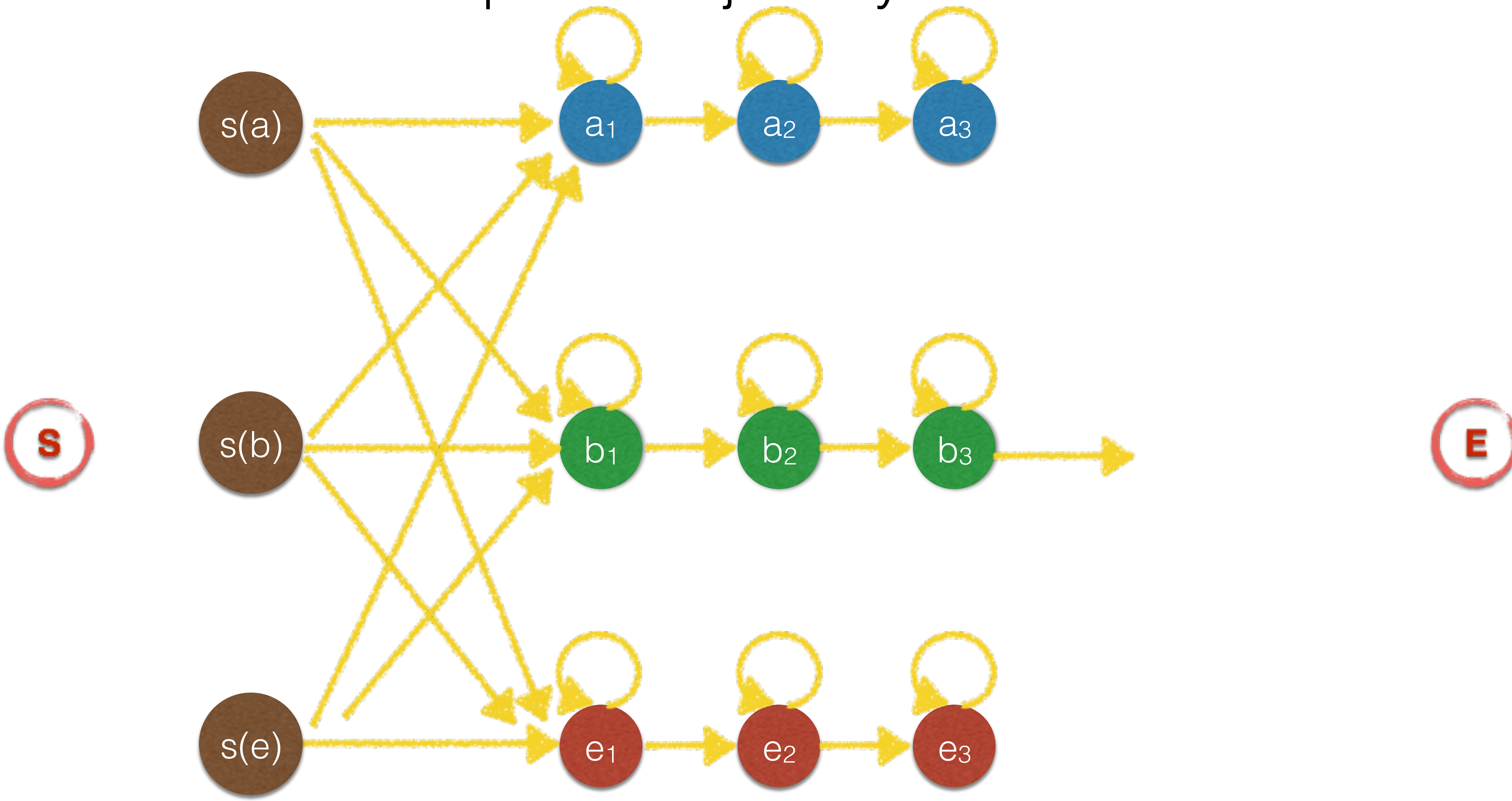
DNN - emission probabilities

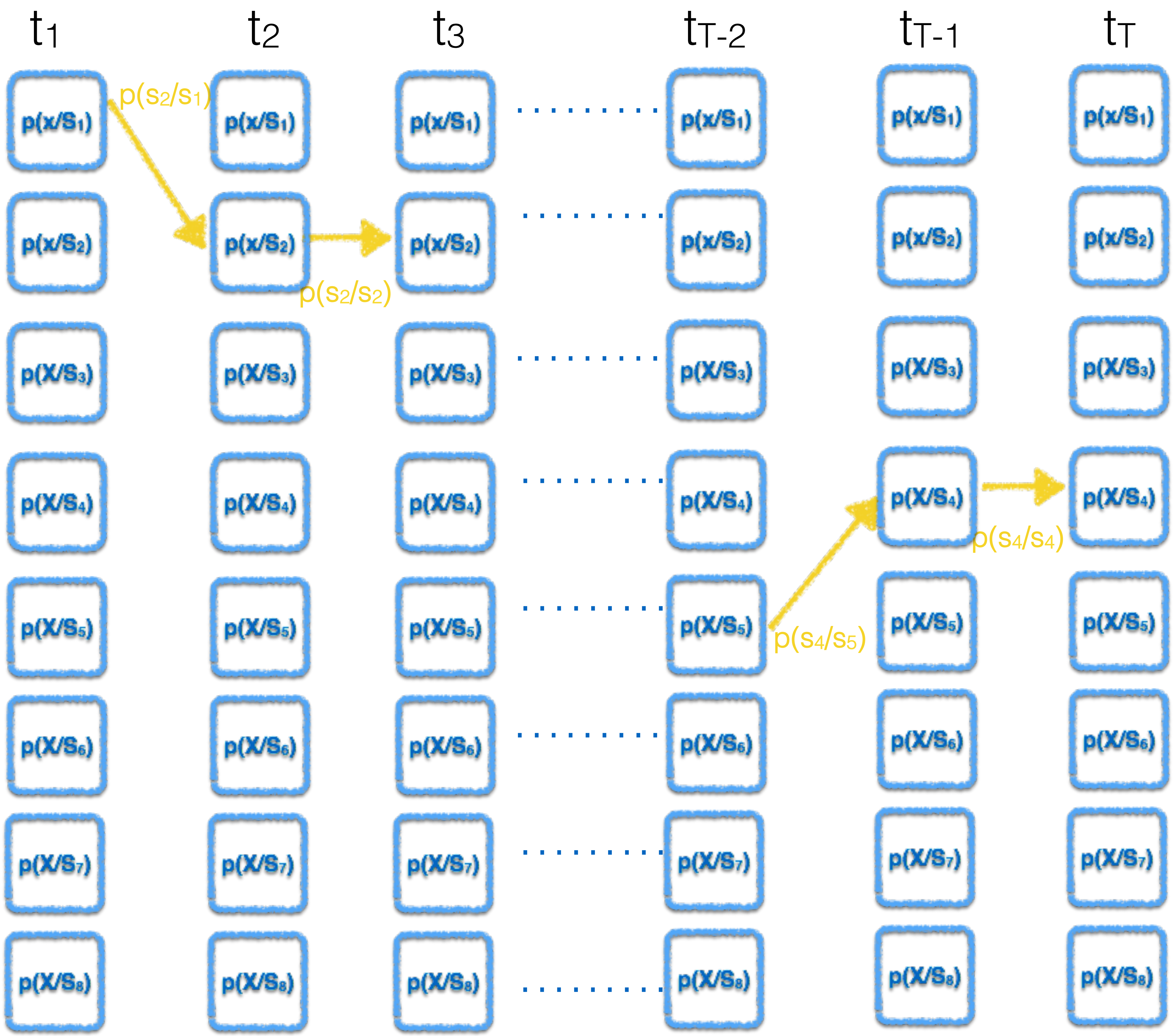


HMM transition probabilities



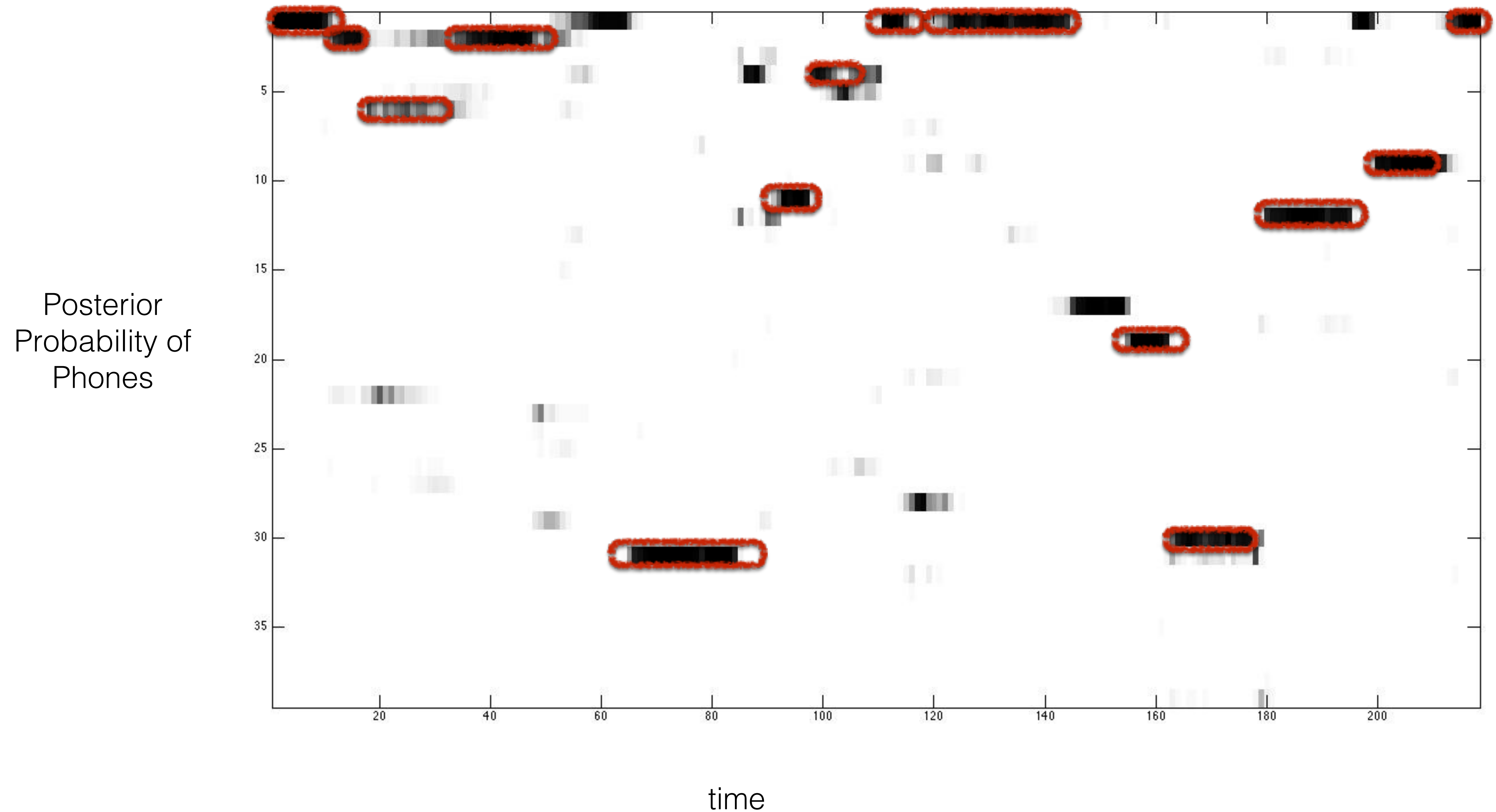
Temporal Trajectory Models



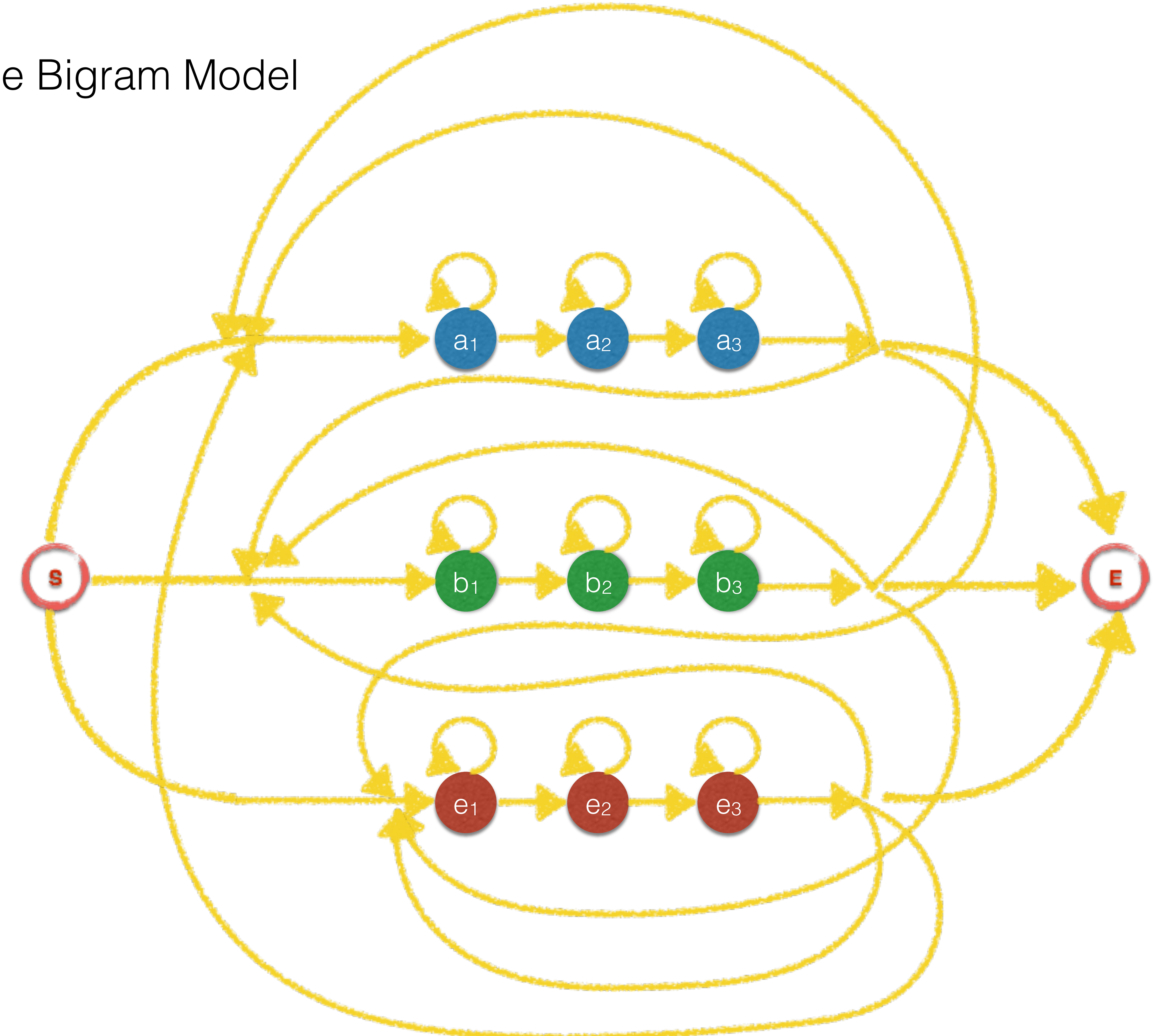


t1 t2 t3 tT-2 tT-1 tT
p(x/S1) p(x/S1) p(x/S1) p(x/S1) p(x/S1) p(x/S1)
p(x/S2) p(x/S2) p(x/S2) p(x/S2) p(x/S2) p(x/S2)
p(X/S3) p(X/S3) p(X/S3) p(X/S3) p(X/S3) p(X/S3)
p(X/S4) p(X/S4) p(X/S4) p(X/S4) p(X/S4) p(X/S4)
p(X/S5) p(X/S5) p(X/S5) p(X/S5) p(X/S5) p(X/S5)
p(X/S6) p(X/S6) p(X/S6) p(X/S6) p(X/S6) p(X/S6)
p(X/S7) p(X/S7) p(X/S7) p(X/S7) p(X/S7) p(X/S7)
p(X/S8) p(X/S8) p(X/S8) p(X/S8) p(X/S8) p(X/S8)
p(s2/s1)
p(s2/s2)
p(s4/s5)
p(s4/s4)

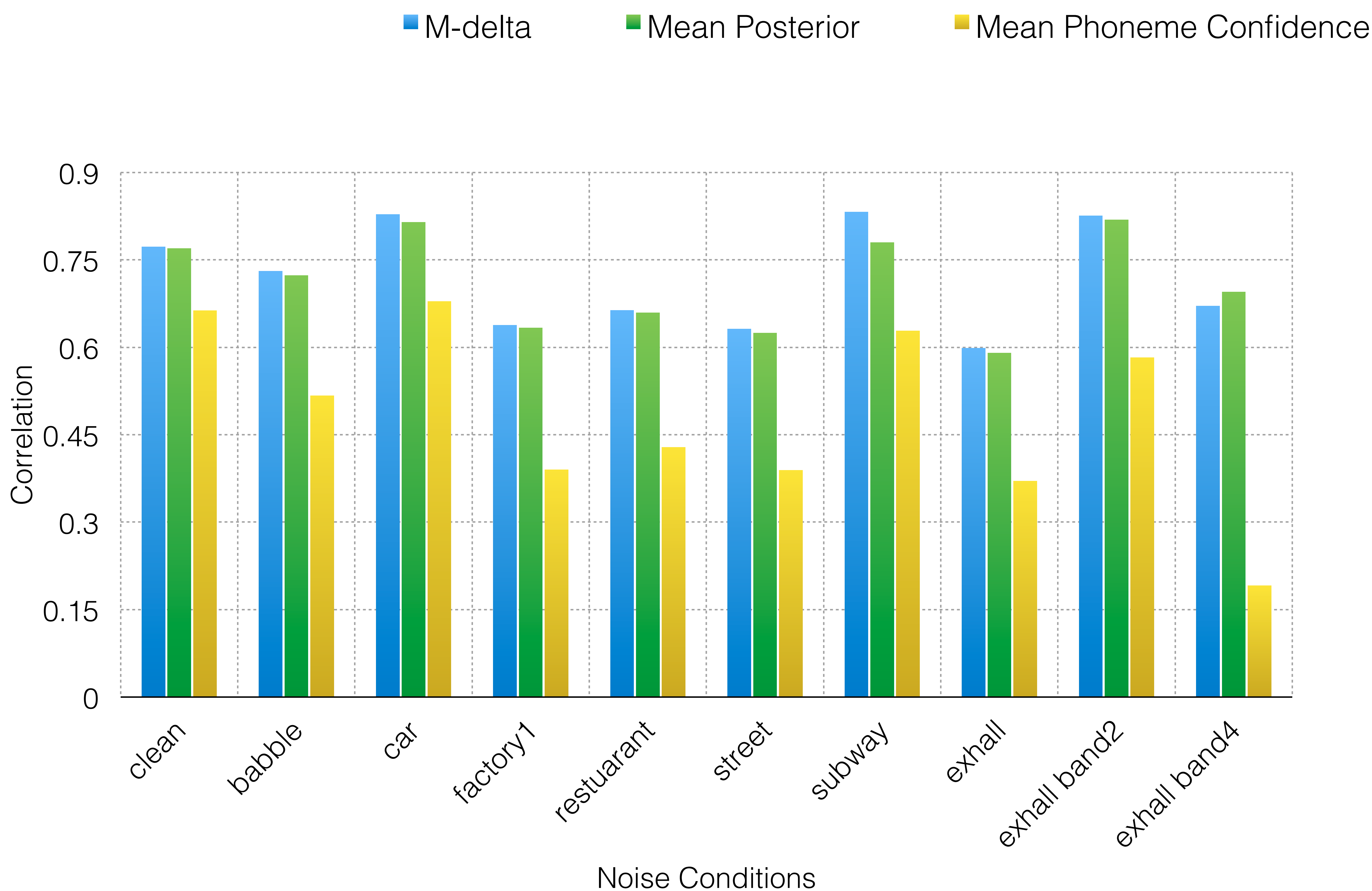
Mean Posterior along best path
Mean Phone Confusion
Mean Likelihood of segment durations



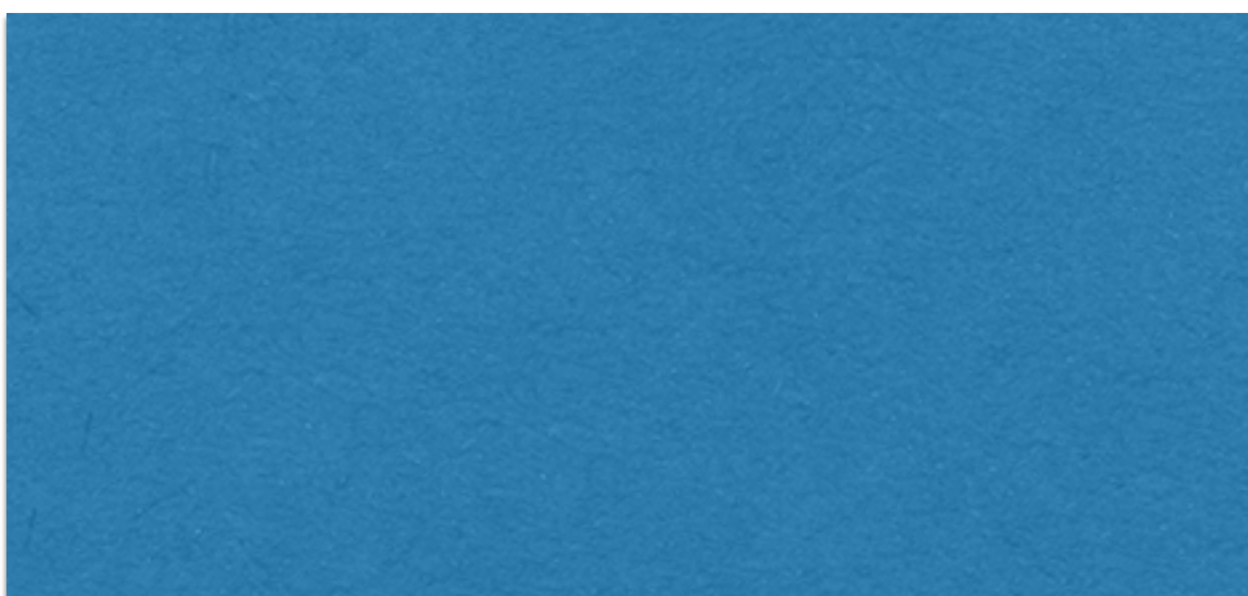
Phone Bigram Model

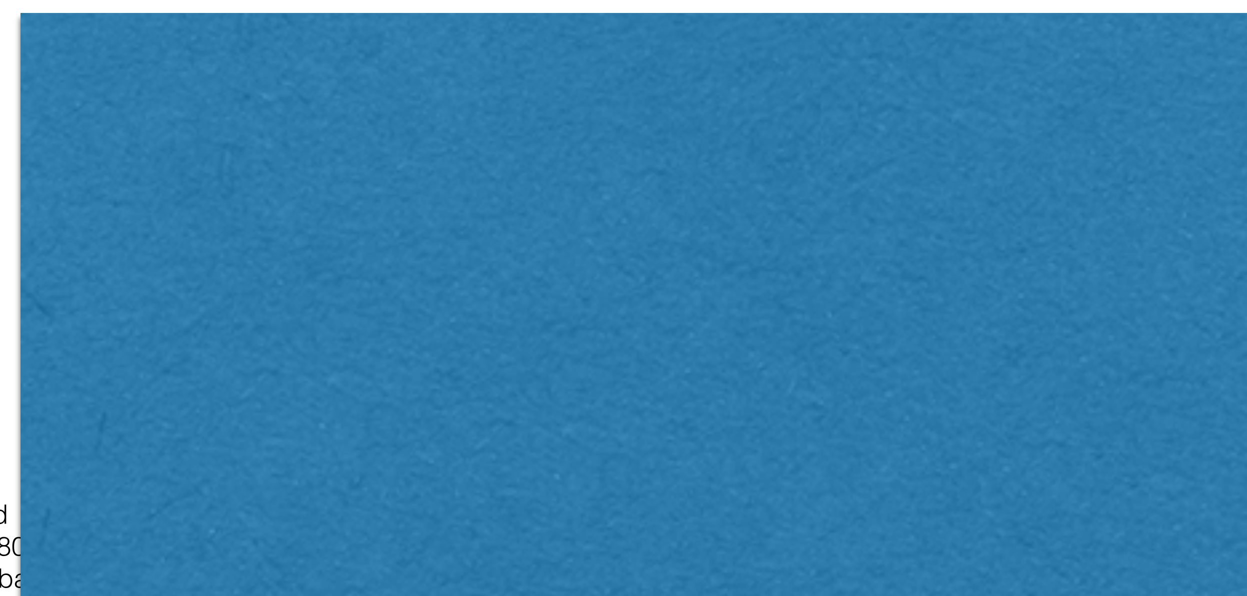


Clean training data



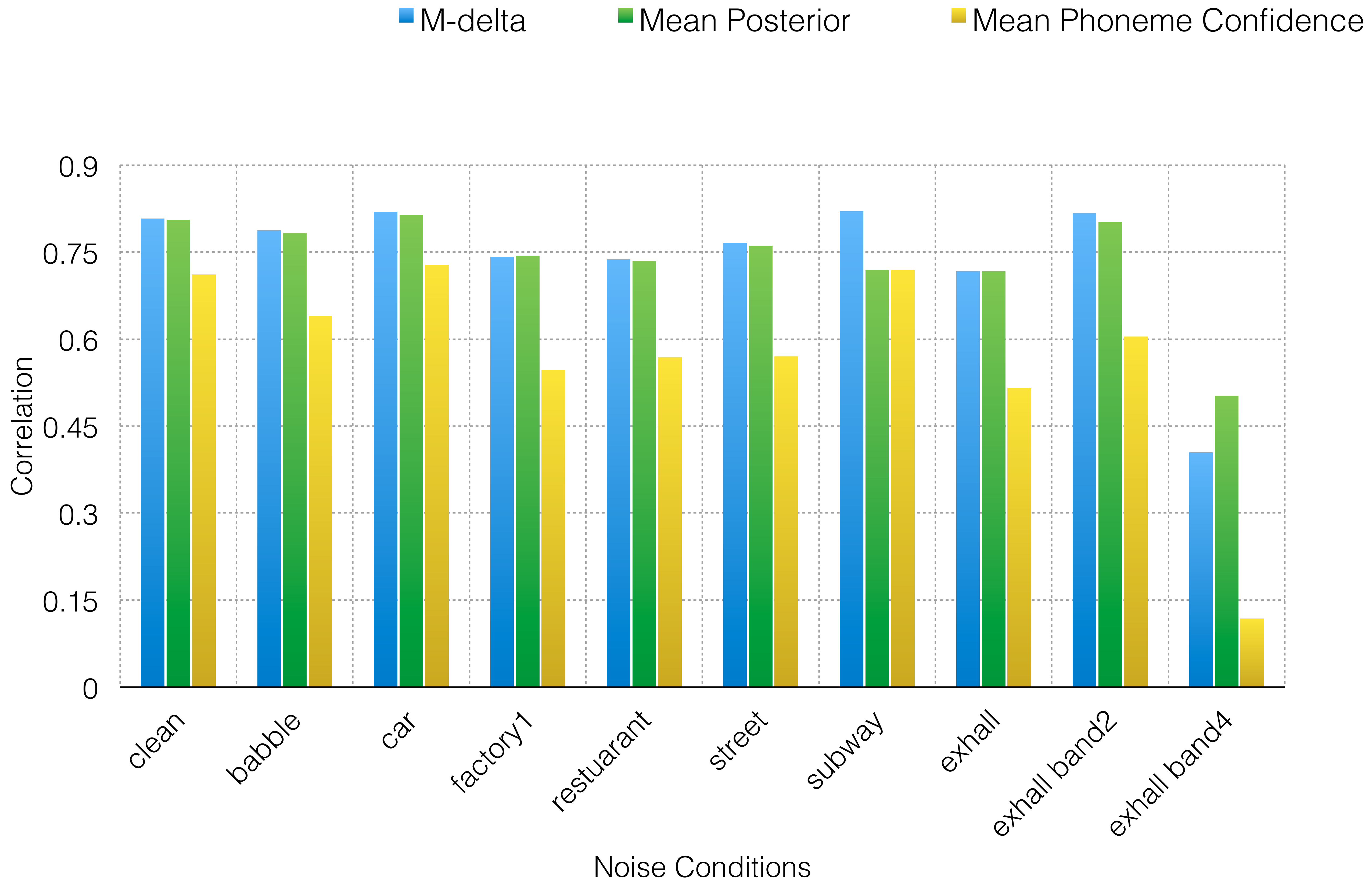
Best Path Scores



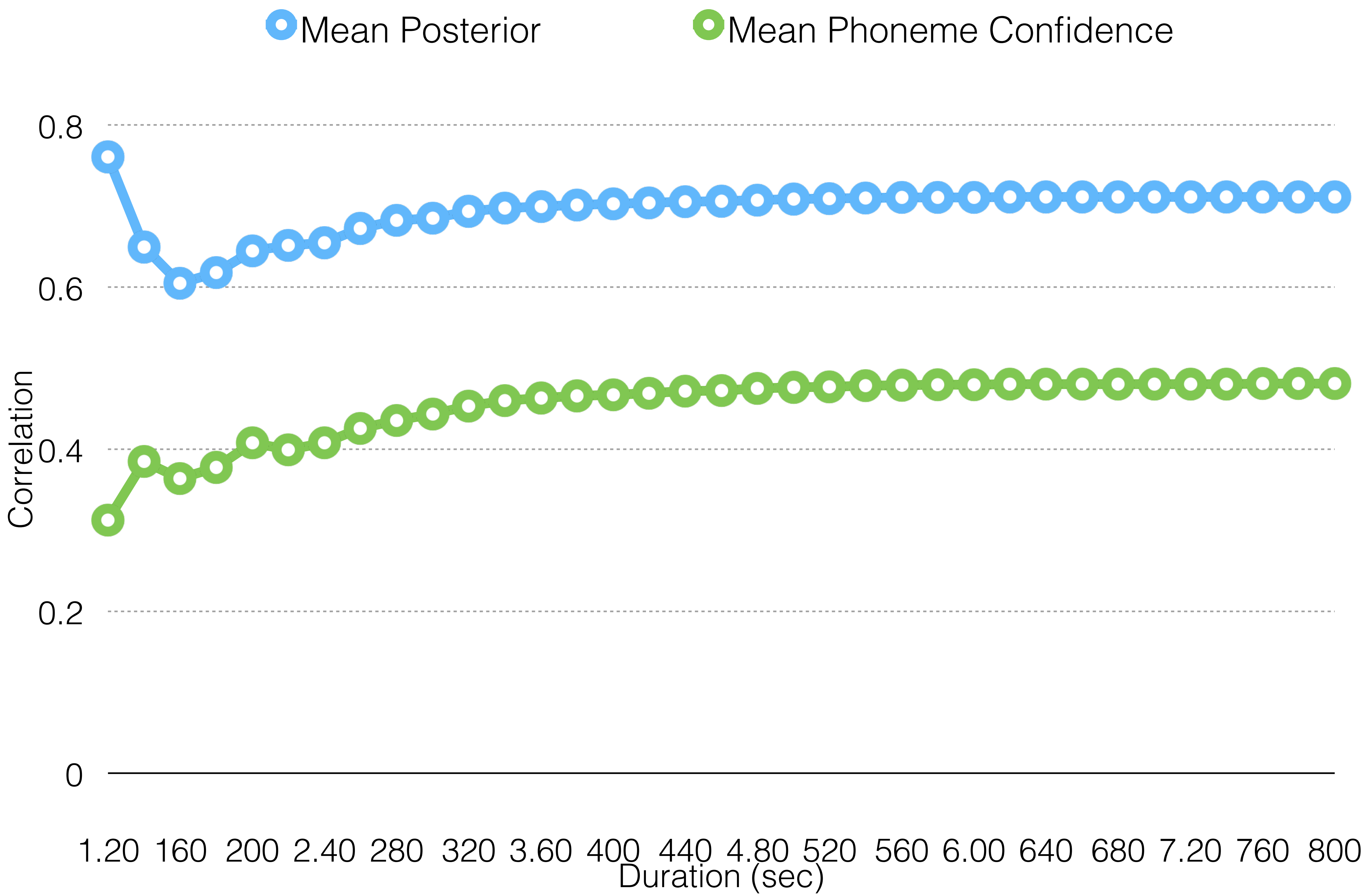


rrelated
0_SX380
exhall_ba
FMLD0_SI822

Noisy training data



Metric Vs Time



Best path scores

- Sum of posteriors along the best path
- Likelihood of the best path given a segmental duration model

Evaluation of error estimators

	Computation Cost	Data Requirement	Unsupervised training	Correlation with Phoneme Error Rate
Avg Posterior along best path	$O(T \times \|S\|^2)$ for best path computation No additional computation required for the phoneme recognition task	minimal high correlation (0.6) even for shortest utterances in TIMIT	Avg. posterior along best path can be maximized by using the best path as the target during unsupervised training This technique is widely used for unsupervised training	0.71 (clean train) 0.75 (noisy train)
Segmental Duration Likelihood	$O(K)$ where K = number of segments in best path	-	Can be used to select segments for unsupervised training	0.13 (clean train) (high variance in performance across noise conditions, bug?)
Average Phoneme Confidence from Lattice	$O(N \times T \times \|S\|^2)$ for N-best path computation Computation of N-1 additional paths for current task	reaches peak correlation in 3 secs	Can be used to select utterances for unsupervised training	0.48 (clean train) 0.57 (noisy train)

Classifier Stability (Ongoing)

- *(with Emmanuel Dupoux)*
  Stability of the classifier's posterior estimate for minor changes in the input is a cue for continuous error monitoring

Metacognition in human decision-making: confidence and error monitoring

Nick Yeung* and Christopher Summerfield

Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK

Input Jacobian

- Instantaneous input Jacobian of the classifier represents the stability of the classifier at the current input

$$f : R^N \to R^M \qquad \mathbf{J} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{pmatrix}$$

- Analytical form available for computation and shares computation with posterior estimation

- Not yet successful in showing correlation with the error

Reflections on the problem

- The current approach evaluates the classifier output on behavior its not trained to emulate however the classifier can be trained to emulate this behavior
  e.g. neural networks can be trained to produce posterior trajectories which are consistent with neighboring frames

- How to measure performance in this case ?

  Rely on long range statistics which are difficult to model (?)