# Improved Model Selection for the ASR-Driven Binary Mask

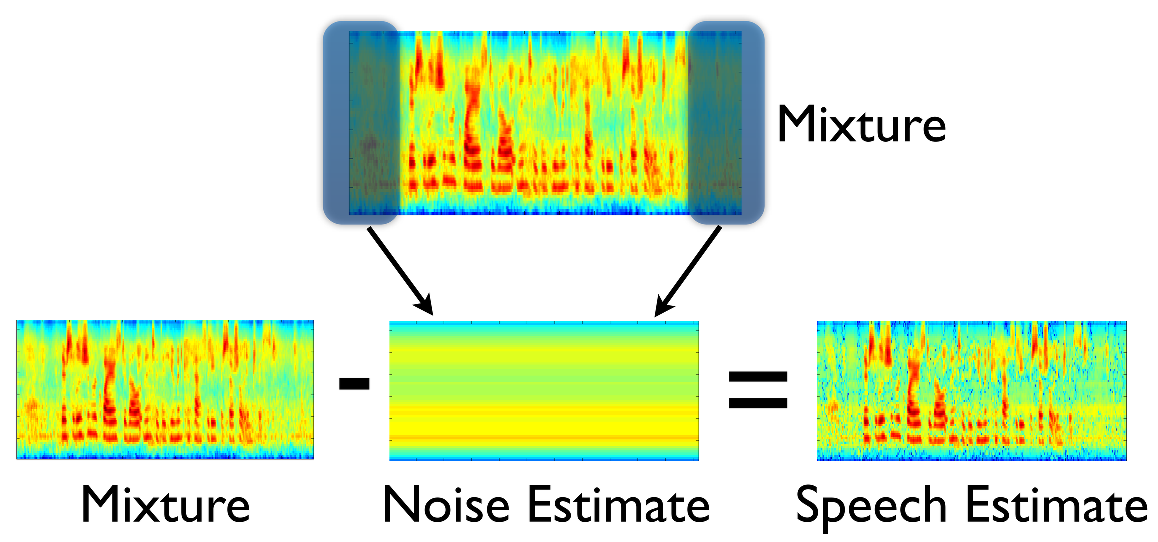
William Hartmann and Eric Fosler-Lussier (hartmann.59@osu.edu; fosler@cse.ohio-state.edu)
Department of Computer Science and Engineering, The Ohio State University

## Introduction

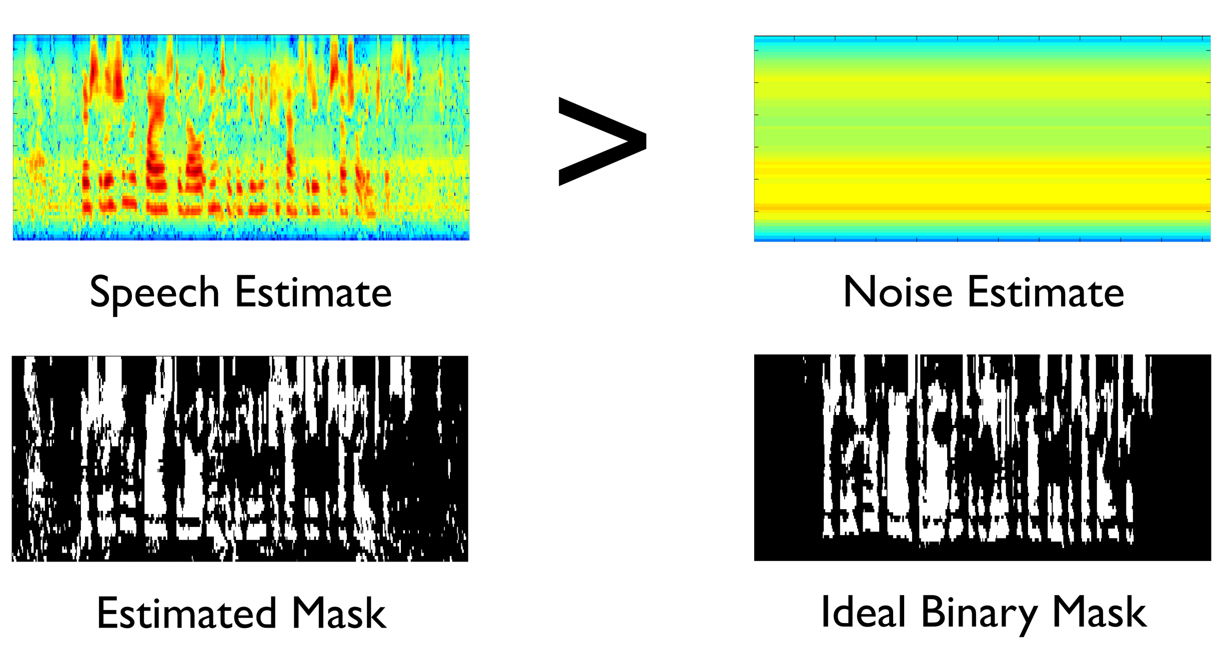
- We use a binary mask-based approach for robust automatic speech recognition.
- Our ASR-Driven mask places the focus on the underlying linguistic content of the signal.
- We propose a linear sequence model based estimation technique.
- Our method outperforms frame-independent estimation methods on the Aurora4 dataset.

## Traditional Approaches

- Traditional approaches first estimate the noise signal from the mixture.
- An estimate of the speech is obtained by subtracting away the noise estimate.




- The two estimates are directly compared to produce the estimated binary mask.
- The goal, the Ideal Binary Mask, is defined by comparing the true speech and noise signals.

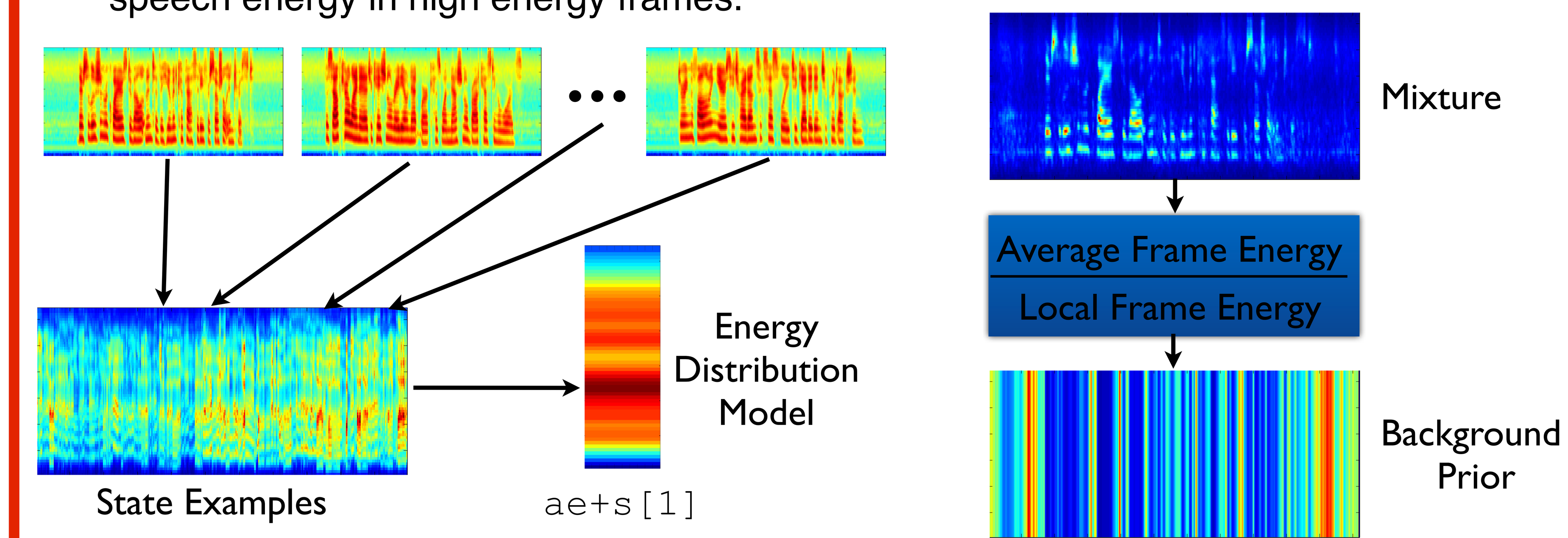



## References

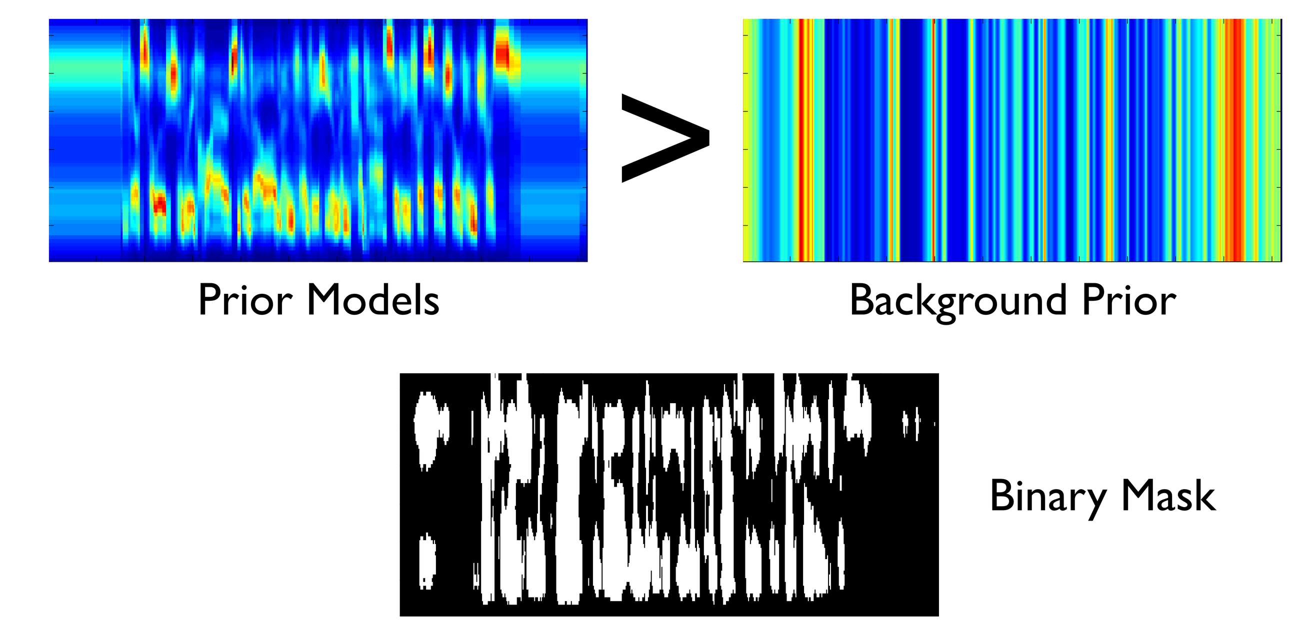
- W. Kim and J. H. L. Hansen, "A novel mask estimation method employing posterior-based representative mean estimate for missing-feature speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1434–1443, July 2011.
- W. Hartmann and E. Fosler-Lussier, "ASR-driven binary mask estimation using spectral priors," in Proceedings of IEEE ICASSP, 2012.
- M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," in Proceedings of EMNLP, 2002.

## The ASR-Driven Binary Mask

- Training sentences are force-aligned to produce state level transcripts.
- An energy prior is learned for each possible state label (triphone states in our setup).
- The background prior assumes little speech energy in low energy frames and high speech energy in high energy frames.

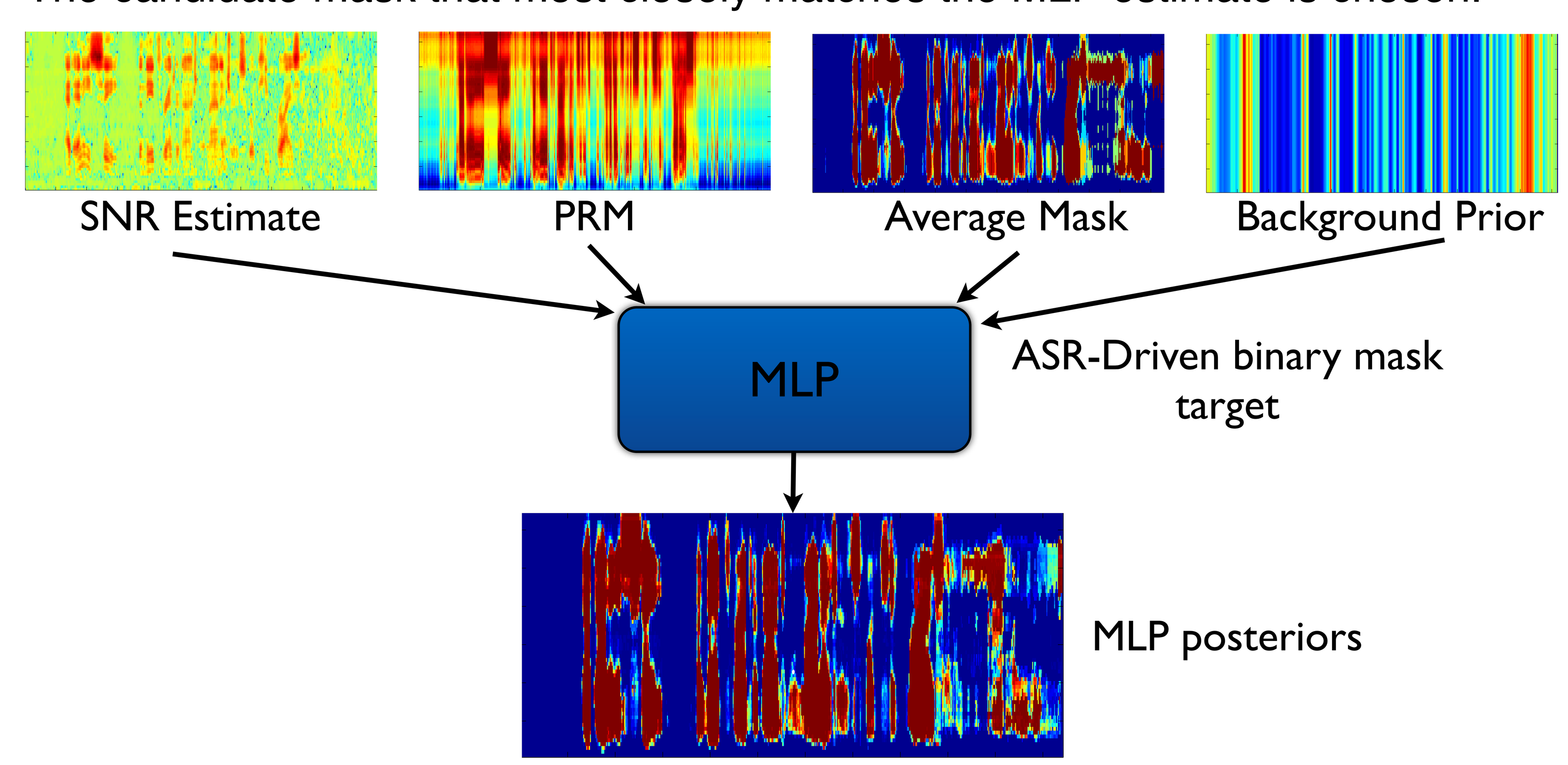



- The oracle ASR-Driven mask is generated by comparing the prior energy models to the background prior.
- The energy priors are selected by force-aligning the transcription to the speech signal.



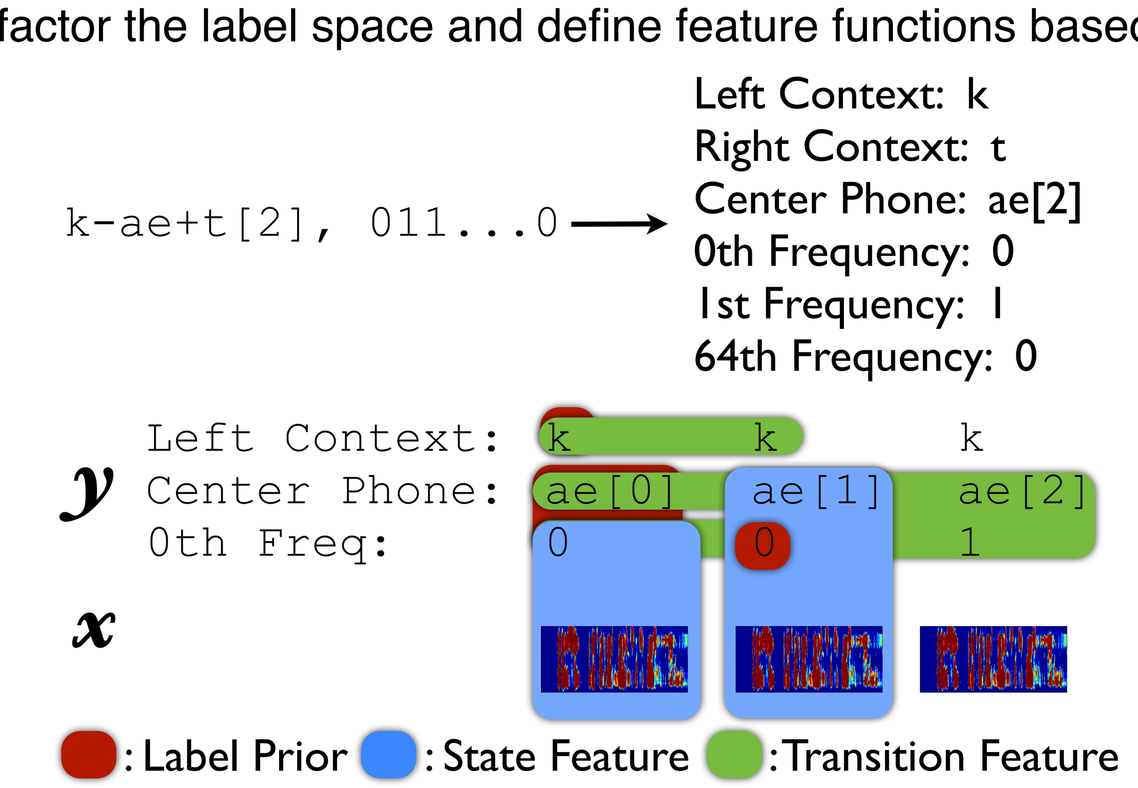

## Frame-Based Mask Estimation

- Candidate masks are generated for each frame from a lattice generated from a baseline recognizer.
- A multilayer perceptron (MLP) is trained for each frequency channel.
- The candidate mask that most closely matches the MLP estimate is chosen.



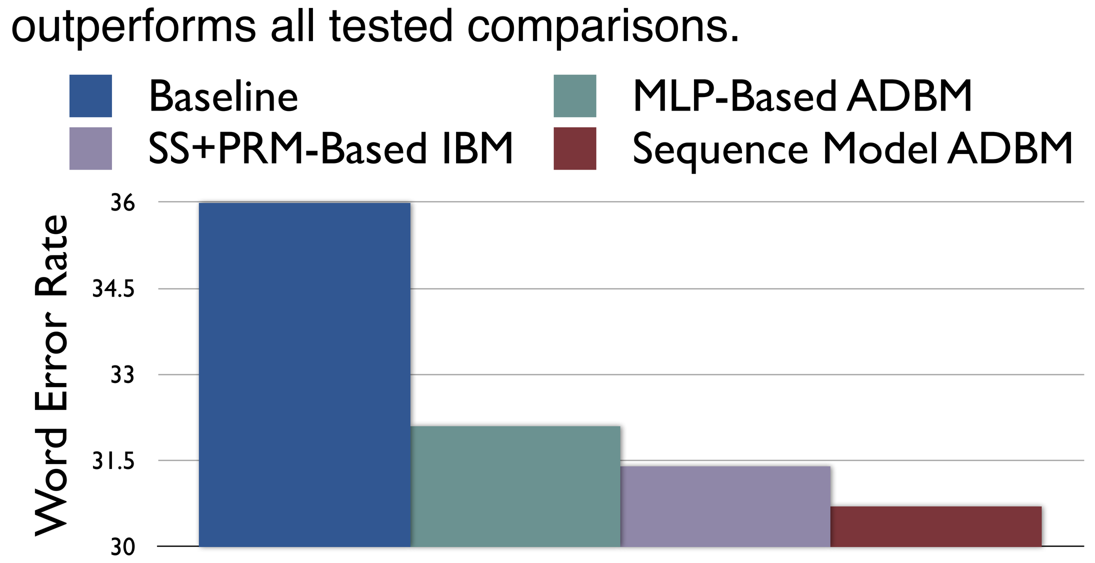

## Sequence-Based Mask Estimation

- We use a linear chain sequence model defined as $\arg\max_{y} \sum_{i} \sum_{k} \alpha_k f_k(y_i, y_{i+1}, x)$
- The model is trained using the structured perceptron.
- Our label space is the cross product of triphone states and the number of possible frame masks (20,000 x 2^64 possible labels).
- Training and decoding with this number of labels is unfeasible.
- We factor the label space and define feature functions based on properties of the labels.




## Results and Conclusions

- The proposed sequence model based approach (Sequence Model ADBM) outperforms all tested comparisons.




Average Performance on Aurora4

- We have proposed a sequence based estimation method that significantly outperforms frame based estimation methods.
- The baseline ASR system is used to provide hypotheses for mask estimation.
- By factoring the label space, we are able to overcome the difficulties associated with our large label space.
- Our approach should scale to alternative, context-dependent mask estimation methods.

We gratefully acknowledge William Hartmann's support under NSF CAREER IIS-0643901. The opinions and conclusions expressed in this work are those of the authors and not of any funding agency.