Noise Robust Speech Recognition with a Switching Linear Dynamic Model

Jasha Droppo and Alex Acero

{jdroppo|alexac}@microsoft.com

Microsoft Research

1 Overview

This paper presents a nonlinear, non-stationary, stochastic model for estimating and removing the effects of background noise on speech cepstra. The model is the union of dynamic system equations for speech and noise, and a model describing how speech and noise are mixed.

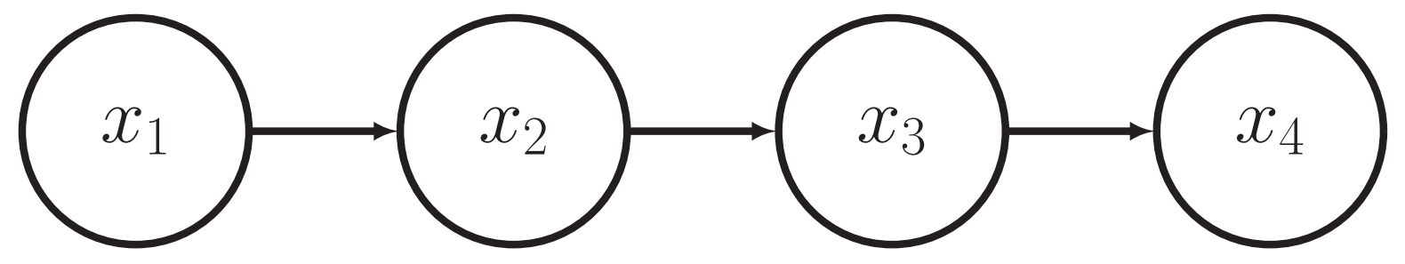
We replace the Gaussian mixture model (GMM) or hidden Markov model (HMM) for speech commonly found in standard model based feature enhancement techniques with a switching linear dynamic model (LDM). The main advantages of using a LDM are:

- Linear dynamics capture the smooth time evolution.
- Switching states capture piecewise stationarity.

This paper show how substantial word error rate improvement can be achieved with a relatively small model sizes under reasonable computational requirements.

2 Modeling Equations

2.1 Linear Dynamic Model



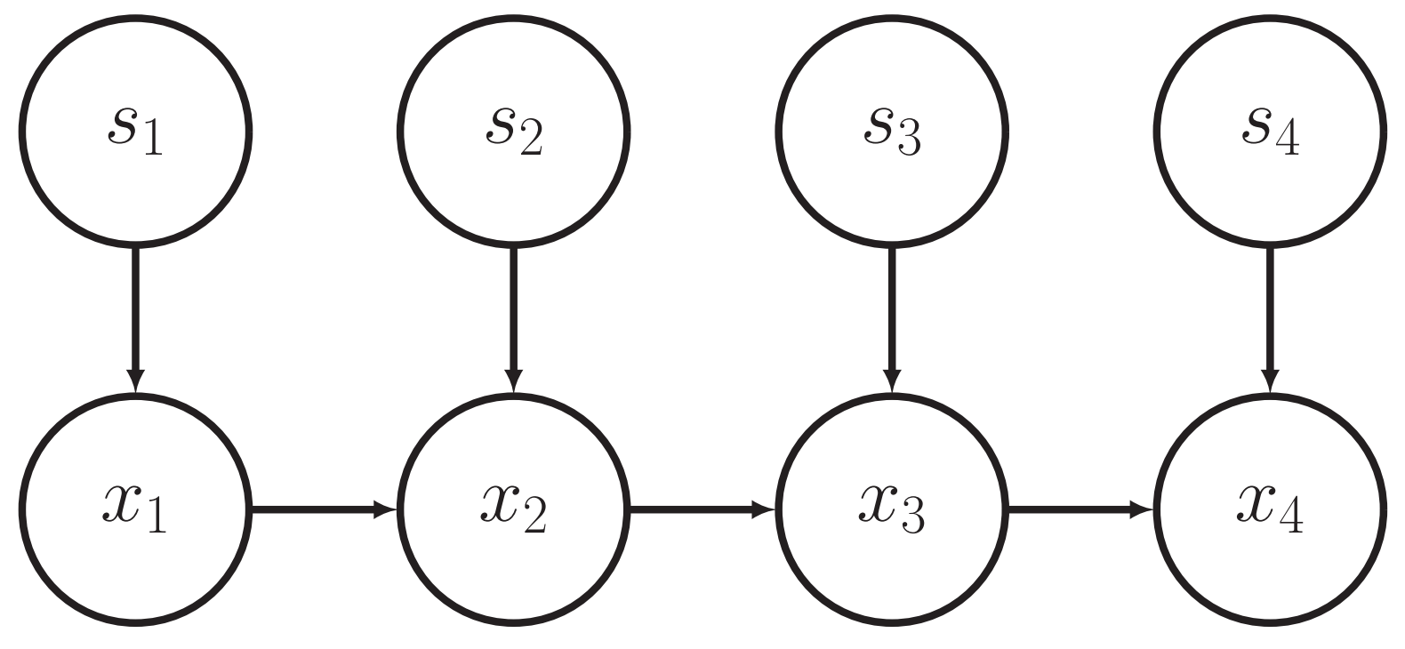
A standard LDM obeys equations,

$$p(x_t|x_{t-1}) = N(x_t; Ax_{t-1} + b, C)$$

$$p(x_1^T) = p(x_1)\prod_{t=2}^{T} p(x_t|x_{t-1})$$

Here, A and b describe how the process evolves over time, and the covariance C is induced by the zero-mean Gaussian noise source which drives the system. The LDM parameters are time-invariant, and are useful in describing signals such as colored Gaussian noise.

2.2 Switching LDM



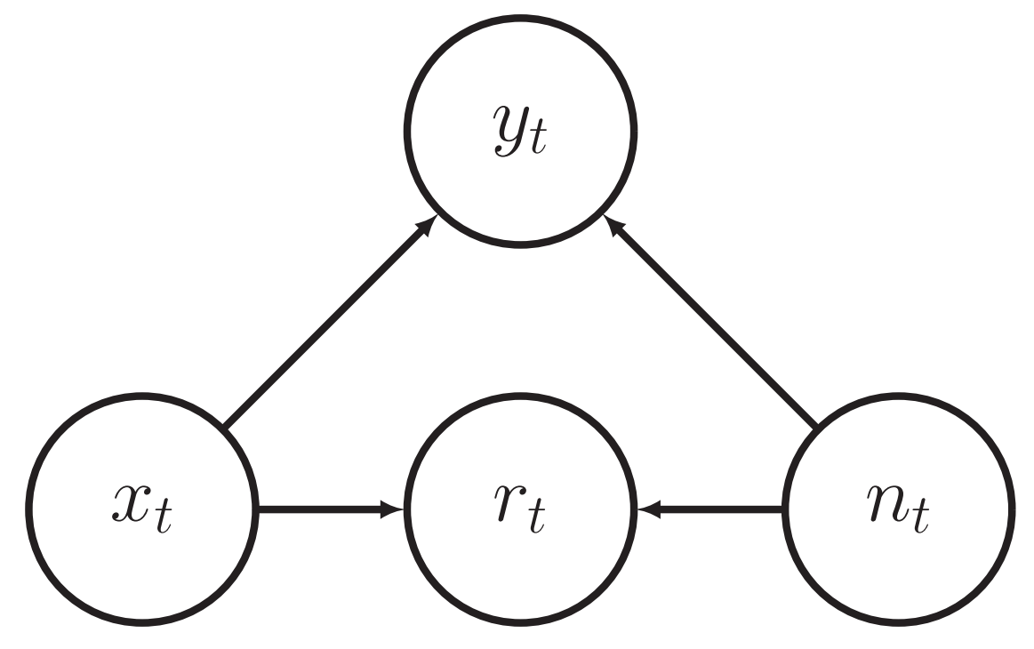
In a switching LDM, the A and b are dependent on a hidden variable at each time t.

$$p(x_t, s_t|x_{t-1}) = N(x_t; A_{s_t}x_{t-1} + b_{s_t}, C_{s_t})p(s_t)$$

$$p(x_1^T, s_1^T) = p(x_1, s_1)\prod_{t=2}^{T} p(x_t, s_t|x_{t-1})$$

Every unique state sequence s_1^T describes a non-stationary LDM. As a result, it is appropriate for describing a number of time-varying systems, including the evolution of speech and noise features over time.

2.3 Observation Model



$$\left.\begin{aligned} p(r|x,n) &= \delta(x-n-r) \\ p(y|x,n) &= \delta(\ln(e^x+e^n)-y) \\ p(r,y) &= N(y-\ln(e^r+1)+r;\mu_x,\sigma_x) \\ & \quad N(y-\ln(e^r+1);\mu_n,\sigma_n) \end{aligned}\right\} \Rightarrow p(y|x)$$

The observation model relates the noisy observation to the hidden speech and noise features. The model used in this paper is the zero variance model with SNR inference[Droppo2003]. It is similar to several related techniques including those by Moreno, Frey, and Stouten.

3 System Behavior

The system, like other model based feature enhancement systems, produces clean cepstral estimates from noisy cepstra.

But, when we replace the more traditional GMM with a switching LDM, it causes the enhancement problem to become intractable.

- Enhancement running time under a GMM is proportional to the length of the utterance.
- An exact implementation of the switching LDM is exponential in the length of the utterance.

To overcome this drawback, the standard generalized pseudo-Bayesian technique is used to provide an approximate solution of the enhancement problem.

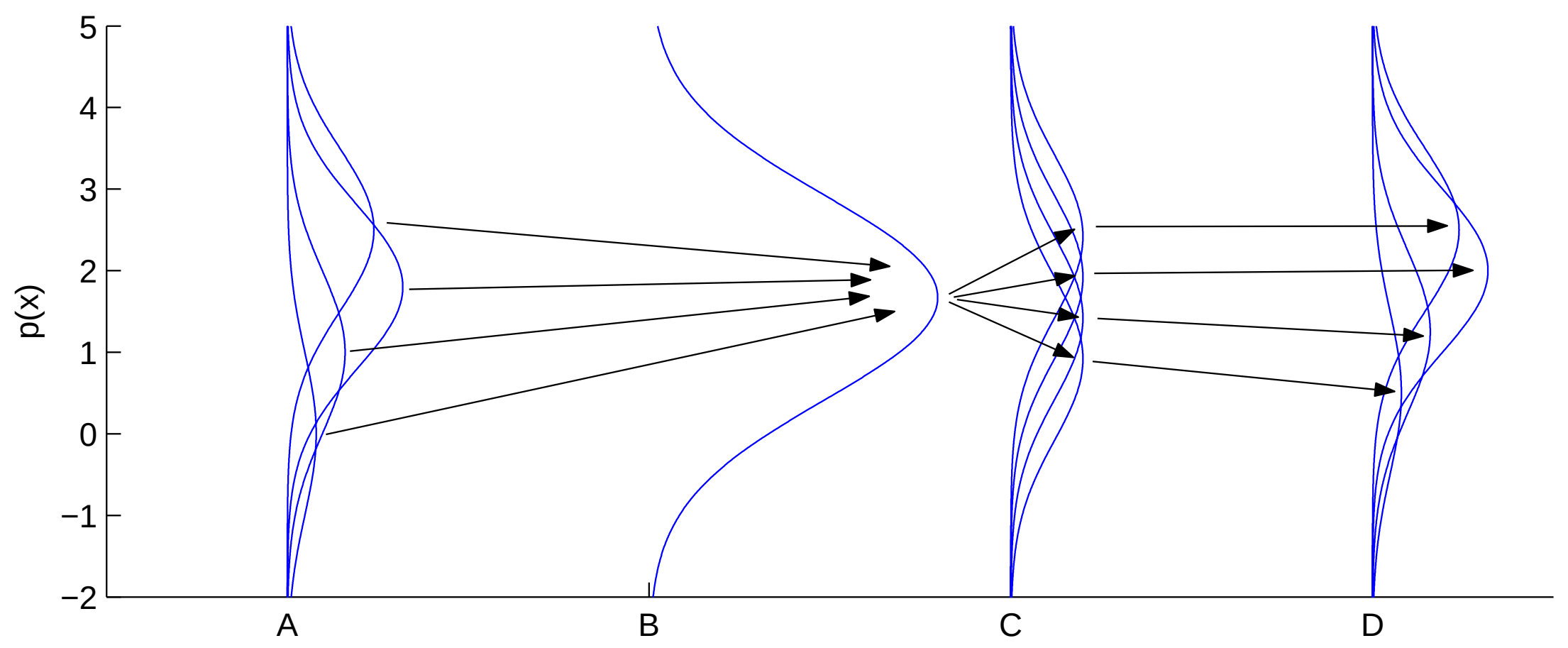


Figure 1: The GPB(1) approximation.

A Obtain the posterior for frame $t-1$, $p(x_{t-1}|s_{t-1}, y_1^{t-1})$.

B Use moment matching to approximate [A] as a single component, $p(x_{t-1}|y_1^{t-1})$.

C Combine [B] with the switching linear dynamic model to create a new multi-component prior for the current frame, $p(x_t|s_t, y_1^{t-1})$.

D Combine [C] with the observation model to produce a posterior for the current frame $p(x_t|s_t, y_1^t)$.

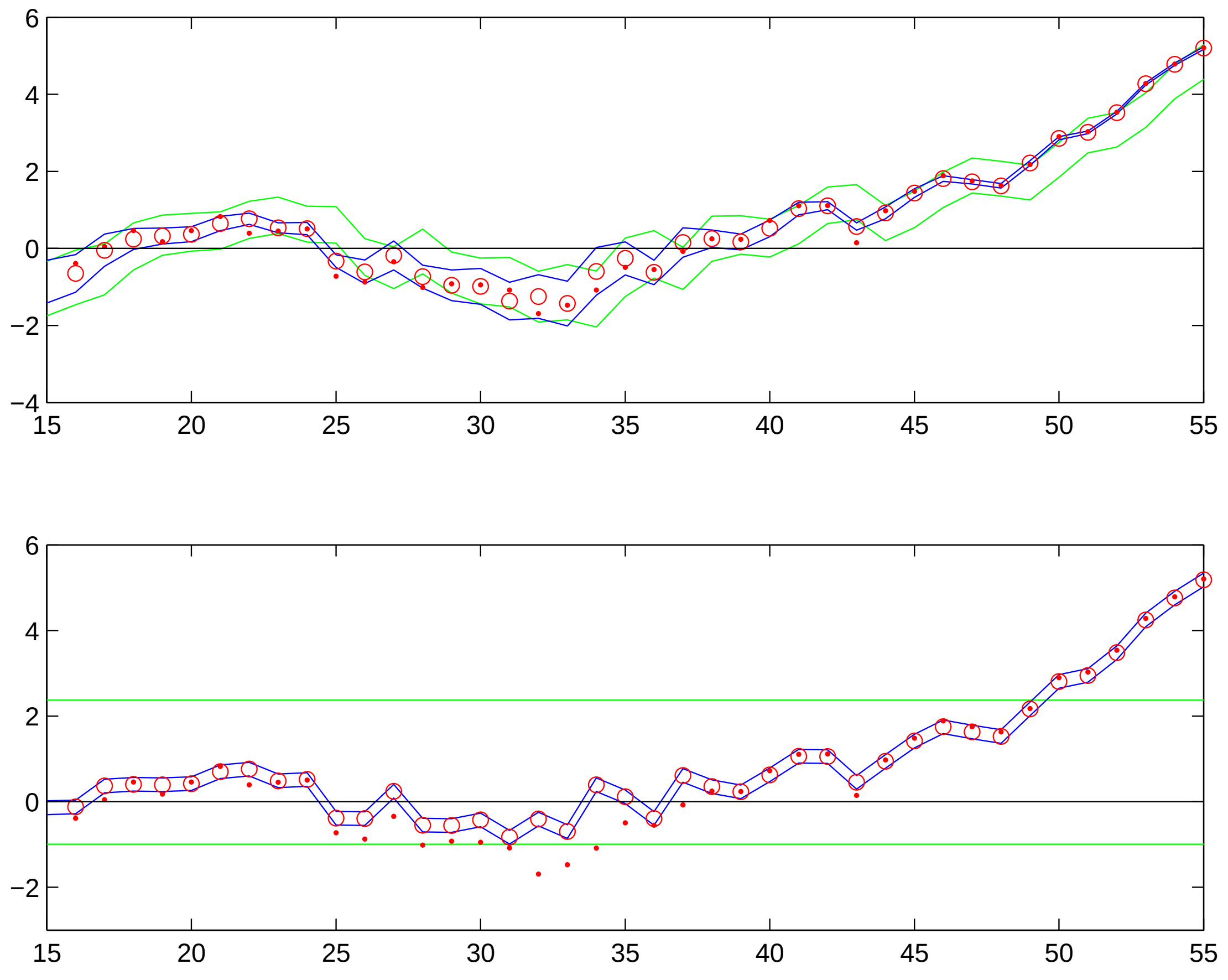


Figure 2: Estimating x from noisy input.

The prior for x, $p(x_t|y_1^{t-1})$, (solid green lines) comes either from a linear dynamic model (top) or a static Gaussian model (bottom). The posterior for x, $p(x_t|y_1^t)$ (solid blue lines) is created from the prior and the observation model. The linear dynamic model produces estimates for x (red circles) that are closer to the true values (red dots).

4 Experimental Results

Recognition accuracy is measured on the Aurora 2 task with the "complex" back-end with a clean acoustic model. Results shown are measured on enhanced data from test set A, and have been averaged over the 0 dB to 20 dB conditions. Models with up to 128 hidden components are evaluated.

Components	Subway	Babble	Car	Exhibition	Ave.
0	65.8	43.2	57.6	67.8	58.6
1	73.1	61.6	80.1	71.6	71.6
2	74.7	64.2	81.8	73.5	73.6
4	80.4	65.1	85.4	76.9	77.0
8	80.5	66.9	86.1	78.1	77.9
16	80.6	67.5	86.2	77.8	78.0
32	83.2	69.6	87.0	79.2	79.8
64	83.6	68.8	87.4	79.2	79.8
128	83.7	69.7	87.4	79.1	80.0

Table 1: Enhancement is performed in forward direction.

The average results for forward enhancement saturate at just under 80% digit accuracy, which indicates that a model with only 16 or 32 mixture components is sufficient.

Components	Subway	Babble	Car	Exhibition	Ave.
0	65.8	43.2	57.5	67.8	58.6
1	76.5	68.3	83.9	76.2	76.2
2	77.0	70.0	84.5	76.5	77.0
4	80.9	69.4	86.5	77.6	78.6
8	81.4	71.0	87.2	79.4	79.7
16	81.6	71.3	87.5	79.4	80.0
32	83.6	72.4	87.8	80.2	81.0
64	84.1	72.1	88.3	80.3	81.2
128	84.1	73.1	88.3	80.3	81.5

Table 2: Forward and backward enhancement are combined.

5 Summary

These preliminary results indicate that this model can reduce digit error rate, even with relatively small number of mixture components.

To expand upon this initial result, future work should include:

- Increasing the history length of GPB to more closely approximate the true posterior distribution.
- Modeling the linear dynamics of noise in addition to speech.
- Augmenting the switching LDM with discrete state transition probabilities.
- Exploring other approximation strategies for this system.