Noise Robust Speech Recognition with a Switching Linear Dynamic Model

Jasha Droppo and Alex Acero
{jdroppo|alexac}@microsoft.com

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 shows how substantial word error rate improvement can be achieved with a relatively small model size under reasonable computational requirements.

2 Modeling Equations

2.1 Linear Dynamic Model

\[ p(x_t, n_t | x_{t-1}) = N(x_t; A_t x_{t-1} + b_t, C_t) p(n_t) \]

\[ p(x_t^2, n_t^2) = p(x_t, x_t) \sum_{t-1}^{T} p(x_t, x_{t-1}) \]

Every unique state sequence \( s_t^2 \) 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

\[ p(x_t | y_t) = \delta(x - n - r) \]
\[ p(x_t | y_t) = \delta(\ln(c^t + e^t) - y) \]
\[ p(x_t | y_t) = N(y - \ln(c^t + 1) + r; \mu_t, \sigma^2) \]

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[Dropko2003]. 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.

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 GBP 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.