A DYNAMICAL SYSTEM APPROACH TO CONTINUOUS SPEECH RECOGNITION

V.Digalakis† *J.R. Rohlicek*‡ *M.Ostendorf*†

† Boston University
44 Cummington St.
Boston, MA 02215

‡ BBN Inc.
10 Moulton St.
Cambridge, MA 02138

ABSTRACT

A dynamical system model is proposed for better representing the spectral dynamics of speech for recognition. We assume that the observed feature vectors of a phone segment are the output of a stochastic linear dynamical system and consider two alternative assumptions regarding the relationship of the segment length and the evolution of the dynamics. Training is equivalent to the identification of a stochastic linear system, and we follow a nontraditional approach based on the Estimate-Maximize algorithm. We evaluate this model on a phoneme classification task using the TIMIT database.

INTRODUCTION

A new direction in speech recognition via statistical methods is to move from frame-based models, such as Hidden Markov Models (HMMs), to segment-based models that provide a better framework for modeling the dynamics of the speech production mechanism. The Stochastic Segment Model (SSM) is a joint model for a sequence of observations, allowing explicit modeling of time correlation. Originally in the SSM, a phoneme was modeled as a sequence of feature vectors that obeyed a multivariate Gaussian distribution. The variable length of an observed phoneme was handled either by modeling a fixed-length transformation of the observations [6] or by assuming the observation was a partially observed sample of a trajectory represented by a fixed-length model [7]. In the first case, the maximum likelihood estimates of the parameters can be obtained directly, but the Estimate-Maximize algorithm [2] may be required in the second case.

Unfortunately, the joint Gaussian model suffers from estimation problems, given the number of acoustic features and the analysis-frame rate that modern continuous speech recognizers use. Therefore, a more constrained assumption about the correlation structure must be made. In previous work [3], we chose to constrain the model to a time-inhomogeneous Gauss-Markov process. Under the Gauss-Markov assumption, we were able to model well the time correlation of the first few cepstral coefficients, but the performance decreased when a larger number of features were used. We attribute the performance decrease to insufficient training data and the noisy nature of the cepstral coefficients. In this work we deal with the problem of noisy observations through a time-inhomogeneous dynamical system formalism, including observation noise in our model.

Under the assumption that we model speech as a Gaussian process at the frame-rate level, a linear state-space dynamical system can be used to parameterize the density of a segment of speech. This a natural generalization of our previous Gauss-Markov approach, with the addition of modeling error in the form of observation noise.

We can make two different assumptions to address the time-variability issue:

1. *Trajectory invariance* (A1): There are underlying unobserved trajectories in state-space that basic units of speech follow. In the dynamical system formalism, this assumption translates to a fixed sequence of state transition matrices for any occurrence of a speech segment. Then, the problem of variable segment length can be solved by assuming that the observed feature vectors are not only a noisy version of the fixed underlying trajectory, but also an incomplete one with missing observations. Successive observed frames of speech have stronger correlation for longer observations, since the underlying trajectory is sampled at shorter intervals (in feature space).
2. *Correlation invariance* (A2): The underlying trajectory in phase space is not invariant under time-warping transformations. In this case, the sequence of state transition matrices for a particular observation of a phoneme depends on the phoneme length, and we have a complete (albeit noisy) observation of the state sequence. In this case, we assume that it is the correlation between successive frames that is invariant to variations in the segment length.

Under either assumption, the training problem with a known segmentation is that of maximum likelihood identification of a dynamical system. We use here an nontraditional

method based on the EM algorithm, that can be easily used under either correlation or trajectory invariance. The model is described in Section , and the identification algorithms are in Section . In Section we shall briefly describe phoneme classification and recognition algorithms for this model, and finally in Section we present phone classification results on the TIMIT database [5].

A DYNAMICAL MODEL FOR SPEECH SEGMENTS

A segment of speech is represented by an L-long sequence of q-dimensional feature vector $\mathbf{Z} = [z_1\ z_2\ \ldots\ z_L]$. The original stochastic segment model for Z had two components [7]: i) a time transformation T_L to model the variable-length observed segment in terms of a fixed-length unobserved sequence $\mathbf{Z} = \mathbf{Y}T_L$, where $\mathbf{Y} = [y_1\ y_2\ \ldots\ y_M]$, and ii) a probabilistic representation of the unobserved feature sequence Y. We assumed in the past [3] that the density of Y was that of an inhomogeneous Gauss-Markov process. We then showed how the EM algorithm can be used to estimate the parameters of the models under this assumption.

In this work, we extend the modeling of the feature sequence, to the more general Markovian representation for each different phone model α

$$\begin{aligned} x_{k+1} &= F_k(\alpha)x_k + w_k \\ y_k &= H_k(\alpha)x_k + v_k \end{aligned} \tag{1}$$

where w_k, v_k are uncorrelated Gaussian vectors with covariances

$$\begin{aligned} E\{w_k w_l^T|\alpha\} &= Q_k(\alpha)\delta_{kl} \\ E\{v_k v_l^T|\alpha\} &= R_k(\alpha)\delta_{kl} \end{aligned}$$

where δ_{kl} is the Kronecker delta. We further assume that the initial state x_0 is Gaussian with mean and covariance $\mu_0(\alpha), \Sigma_0(\alpha)$. In this work, we arbitrarily choose the dimension of the state to be equal to that of the feature vector and $H_k(\alpha) = I$, the identity matrix. The sequence Y is either fully or partially observed under the assumptions of correlation and trajectory invariance respectively. In order to reduce the number of free parameters in our model, we assume that a phone segment is locally stationary over different regions within the segment, where those regions are defined by a fixed time warping that in this work we simply choose as linear. In essence, we are tying distributions, and the way this is done under the correlation and trajectory invariance assumptions is shown in Figure 1.

The likelihood of the observed sequence Z can be obtained by the Kalman predictor, as

$$\log p(\mathbf{Z}|\alpha) = -\sum_{k=1}^{L}\Big\{\log|\Sigma_k^{(e)}(\alpha)| + e_k^T(\alpha)[\Sigma_k^{(e)}(\alpha)]^{-1}e_k(\alpha)\Big\} + \text{constant} \tag{2}$$

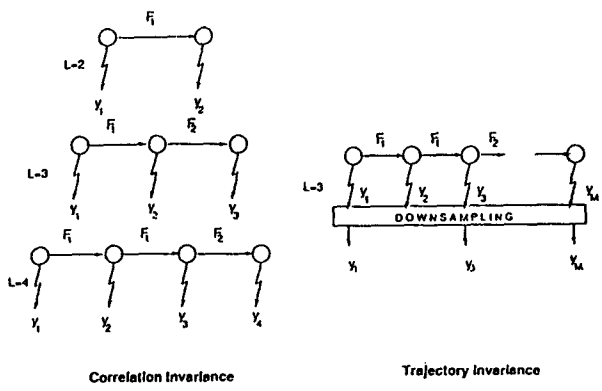

Figure 1: Distribution tying for (a) Correlation and (b) Trajectory invariance.

where $\Sigma_k^{(e)}(\alpha)$ is the prediction error variance given phone model α. In the trajectory invariance case, innovations are only computed at the points where the output of the system is observed, and the predicted state estimate for these times can be obtained by the l-step ahead prediction form of the Kalman filter, where l is the length of the last "black-out" interval - the number of missing observations y immediately before the last observed frame z.

TRAINING

The classical method to obtain maximum likelihood estimates involves the construction of a time-varying Kalman predictor and the expression of the likelihood function in terms of the prediction error as in (2) [1]. The minimization of the log-likelihood function is equivalent to a nonlinear programming problem, and iterative optimization methods have to be used that all require the first and perhaps the second derivatives of the log-likelihood function with respect to the system parameters. The solution requires the integration of adjoint equations, and the method becomes too involved under the trajectory invariance assumption, where we have missing observations.

We have developed a nontraditional iterative method for maximum likelihood identification of a stochastic dynamical system, based on the observation that the computation of the estimates would be simple if the state of the system were observable: using simple first and second order sufficient statistics of the state and observation vectors. The Estimate-Maximize algorithm provides an approach for estimating parameters for processes having unobserved components, in this case the state vectors, and therefore can be used for maximum likelihood identification of dynamical systems.

If we denote the parameter vector of phone model α by θ, then at the p-th iteration of the EM algorithm the new estimate of the parameter vector is obtained by minimizing

$$Q(\theta^{(p+1)}|\theta^{(p)}) = -E\Big\{\log p(\mathbf{X},\mathbf{Y}|\theta^{(p+1)}) \;\Big|\; \mathbf{Z},\theta^{(p)}\Big\}$$

$$= E\Big\{\sum\sum_{k=1}^{L}[(x_k - F_k x_{k-1})^T Q_k^{-1}(x_k - F_k x_{k-1}) + \log|Q_k| + (y_k - H_k x_k)^T R_k^{-1}(y_k - H_k x_k) + \log|R_k|] + \text{constant} \;\Big|\; \mathbf{Z},\theta^{(p)}\Big\} \quad (3)$$

where we have suppressed the parameterization of the system parameters on phone model α and the first summation is over all occurrences of a specific phone model in the training data.

Since the noise process is assumed to be Gaussian, the EM algorithm simply involves iteratively computing the expected first and second order sufficient statistics given the current parameter estimates. It is known from Kalman filtering theory [1] that the conditional distribution of the state $\mathbf{X}$ given the observations $\mathbf{Z}$ on an interval is Gaussian. The sufficient statistics are then

$$\begin{aligned}
E\{x_k|\mathbf{Z}\} &= \hat{x}_{k|L}\\
E\{x_k x_k^T|\mathbf{Z}\} &= \Sigma_{k|L} + \hat{x}_{k|L}\hat{x}_{k|L}^T\\
E\{x_k x_{k-1}^T|\mathbf{Z}\} &= \Sigma_{k,k-1|L} + \hat{x}_{k|L}\hat{x}_{k-1|L}^T\\
E\{y_k|\mathbf{Z}\} &= \begin{cases} z_k, & \text{if observed;}\\ H_k\hat{x}_{k|L}, & \text{if missing.}\end{cases}\\
E\{y_k y_k^T|\mathbf{Z}\} &= \begin{cases} z_k z_k^T, & \text{if obs.;}\\ R_k + H_k E\{x_k x_k^T|\mathbf{Z}\}H_k^T, & \text{if mis.}\end{cases}\\
E\{y_k x_k^T|\mathbf{Z},\theta\} &= \begin{cases} z_k\hat{x}_k^T, & \text{if observed;}\\ H_k E\{x_k x_k^T|\mathbf{Z}\}, & \text{if missing.}\end{cases}
\end{aligned}$$

where the quantities on the right, $\hat{x}_{k|L}, \Sigma_{k|L}, \Sigma_{k,k-1|L}$ are the fixed interval smoothed state estimate, its variance and the one lag cross-covariance respectively. The computation of these sufficient statistics can be done recursively. Under A2, since $\mathbf{Y} = \mathbf{Z}$, it reduces to the fixed-interval smoothing form of the Kalman filter, together with some additional recursions for the computation of the cross-covariance. These recursions consist of a forward pass through the data, followed by a backward pass and are summarized in Table 1. Under A1, the recursions take the form of a fixed interval smoother with blackouts, and can be derived similarly to the standard Kalman filter recursions.

To summarize, assuming a known segmentation and therefore a known sequence of system models, the EM algorithm involves at each iteration the computation of the sufficient statistics described previously using the recursions of Table 1 and the old estimates of the model parameters (Estimate step). The new estimates for the system parameters can then be obtained from these statistics as simple multivariate regression coefficients (Maximize step). In addition, the structure of the system matrices can be constrained in order to satisfy identifiability conditions. When the segmentation is unknown, since the estimates obtained from our known segmentation method are Maximum Likelihood ones, training can be done in an iterative fashion, as described in [6].

Forward recursions		
$\hat{x}_{k\|k}$	=	$\hat{x}_{k\|k-1} + K_k e_k$
$\hat{x}_{k+1\|k}$	=	$F_k\hat{x}_{k\|k}$
e_k	=	$y_k - H_k\hat{x}_{k\|k-1}$
K_k	=	$\Sigma_{k\|k-1}H_k^T[\Sigma_k^{(e)}]^{-1}$
$\Sigma_k^{(e)}$	=	$H_k\Sigma_{k\|k-1}H_k^T + R_k$
$\Sigma_{k\|k}$	=	$\Sigma_{k\|k-1} - K_k\Sigma_k^{(e)}K_k^T$
$\Sigma_{k,k-1\|k}$	=	$(I - K_k H_k)F_{k-1}\Sigma_{k-1\|k-1}$
$\Sigma_{k+1\|k}$	=	$F_k\Sigma_{k\|k}F_k^T + Q_k$
Backward Recursions		
$\hat{x}_{k-1\|L}$	=	$\hat{x}_{k-1\|k-1} + A_k[\hat{x}_{k\|L} - \hat{x}_{k\|k-1}]$
$\Sigma_{k-1\|L}$	=	$\Sigma_{k-1\|k-1} + A_k[\Sigma_{k\|L} - \Sigma_{k\|k-1}]A_k^T$
A_k	=	$\Sigma_{k-1\|k-1}F_{k-1}^T\Sigma_{k\|k-1}^{-1}$
$\Sigma_{k,k-1\|L}$	=	$\Sigma_{k,k-1\|k} + [\Sigma_{k\|L} - \Sigma_{k\|k}]\Sigma_{k\|k}^{-1}\Sigma_{k,k-1\|k}$

Table 1: Summary of E-step recursions

RECOGNITION

When the phonetic segmentation is known, under both assumptions A1 and A2 the model sequence can be determined from the segmentation and therefore the MAP rule can be used for phone classification, where the likelihood of the observations is obtained from the Kalman predictor (2).

For connected-phone recognition, with unknown segmentation, the MAP rule for detecting the most likely phonetic sequence involves computing the total probability of a certain sequence by summing over all possible segmentations. Because of the computational complexity of this approach, one can jointly search for the most likely phone sequence and segmentation given the observed sequence. This can be done with a Dynamic-Programming recursion. In previous work we have also introduced alternative fast algorithms for both phone classification and recognition [4] which yield performance similar to Dynamic-Programming with significant computation savings.

EXPERIMENTAL RESULTS

We have implemented a system based on our correlation invariance assumption and performed phone classifi-

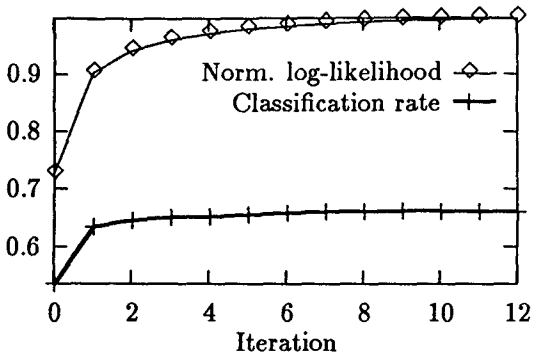

Figure 2: Classification performance of test data vs. number of iterations and log-likelihood ratio of each iteration relative to the convergent value for the training data.

cation experiments on the TIMIT database [5]. We used Mel-warped cepstra and their derivatives together with the derivative of log power.The number of different distributions (time-invariant regions) for each segment model was 5. We used 61 phonetic models, but in counting errors we folded homophones together and effectively used the reduced CMU/MIT 39 symbol set. The measurement-noise variance was common over all different phone-models and was not reestimated after the first iteration. In experiments with class-dependent measurement noise, we observed a decrease in performance, which we attribute to "over-training"; a first order Gauss-Markov structure can adequately model the training data, because of the small length of the time-invariant regions in the model. In addition, the observed feature vectors were centered around a class-dependent mean. Duration probabilities as well as a priori class probabilities where also used in these experiments. The training set that we used consist of 317 speakers (2536 sentences), and evaluation of our algorithms is done on a separate test set with 12 speakers (96 sentences).

The effectiveness of the training algorithm is shown in Figure 2, where we present the normalized log-likelihood of the training data and classification rate of the test data versus the number of iterations. We used 10 cepstra for this experiment, and the initial parameters for the models where uniform across all classes, except the class-dependent means. We can see the fast initial convergence of the EM algorithm, and that the best performance is achieved after only 4 iterations.

In Figure 3 we show the classification rates for no correlation modeling (independent frames), the Gauss-Markov model and the Dynamical system model for different numbers of input features. We also include in the same plot the classification rates when the derivatives of the cepstra are

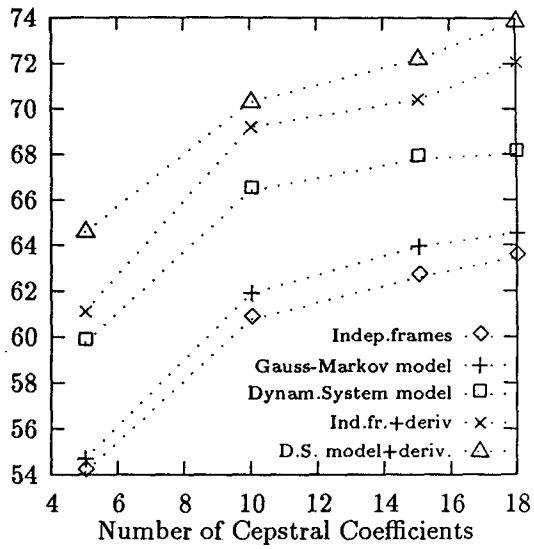

Figure 3: Classification rates for various types of Correlation modeling and numbers of cepstral coefficients

included in the feature set, so that some form of correlation modeling is included in the independent-frame model. We can see that the proposed model clearly outperforms the independent-frame model. Furthermore, we should notice the significance of incorporating observation noise in the model, by comparing the performance of the new model to the earlier, Gauss-Markov one.

CONCLUSION

In this paper, we have shown that segment model based on a stochastic linear system model which incorporates a modeling/observation noise term is effective for speech recognition. We have shown that classification performance using this model is significantly better than is obtained using either an independent-frame or a Gauss-Markov assumption on the observed frames. Finally, we have presented a novel approach to the system parameter estimation problem based on the EM algorithm.

ACKNOWLEDGEMENTS

This work was supported jointly by NSF and DARPA under NSF grant number IRI-8902124. This paper also appears in the *Proceedings of the International Conference on Acoustics, Speech and Signal Processing.*

REFERENCES

1. P.E.Caines, "Linear Stochastic Systems", John Wiley & Sons, 1988.

2. A.P.Dempster, N.M.Laird and D.B.Rubin, "Maximum Likelihood Estimation from Incomplete Data," in *Journal of the Royal Statistical Society (B)*, Vol. 39, No. 1, pp. 1–38, 1977.

3. V. Digalakis, M. Ostendorf and J. R. Rohlicek, "Improvements in the Stochastic Segment Model for Phoneme Recognition," in *Proceedings of the Second DARPA Workshop on Speech and Natural Language*, pp. 332–338, October 1989.

4. V. Digalakis, M. Ostendorf and J. R. Rohlicek, "Fast Search Algorithms for Connected Phone Recognition Using the Stochastic Segment Model," manuscript submitted to *IEEE Trans. Acoustic Speech and Signal Processing* (a shorter version appeared in *Proceedings of the Third DARPA Workshop on Speech and Natural Language*, June 1990).

5. L.F. Lamel, R. H. Kassel and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," in *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 100–109, Feb. 1986.

6. M. Ostendorf and S. Roucos, "A Stochastic Segment Model for Phoneme-based Continuous Speech Recognition," in *IEEE Trans. Acoustic Speech and Signal Processing*, Vol. ASSP-37(12), pp. 1857–1869, December 1989.

7. S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 127–130, New York, New York, April 1988.