ESTIMATION OF GENERAL IDENTIFIABLE LINEAR DYNAMIC MODELS WITH AN APPLICATION IN SPEECH RECOGNITION

G. Tsontzos, V. Diakoloukas, Ch. Koniaris and V. Digalakis

Dept. of Electronics & Computer Engineering
Technical University of Crete, GR-73100 Chania, Greece
{gtsntzs,vdiak,chkoniaris,vas}@telecom.tuc.gr

ABSTRACT

Although Hidden Markov Models (HMMs) provide a relatively efficient modeling framework for speech recognition, they suffer from several shortcomings which set upper bounds in the performance that can be achieved. Alternatively, linear dynamic models (LDM) can be used to model speech segments. Several implementations of LDM have been proposed in the literature. However, all had a restricted structure to satisfy identifiability constraints. In this paper, we relax all these constraints and use a general, canonical form for a linear state-space system that guarantees identifiability for arbitrary state and observation vector dimensions. For this system, we present a novel, element-wise Maximum Likelihood (ML) estimation method. Classification experiments on the AURORA2 speech database show performance gains compared to HMMs, particularly on highly noisy conditions.

Index Terms— Speech Recognition, Modeling, Identification

1. INTRODUCTION

Hidden Markov Models (HMMs) dominate in today's speech recognition engines. This is primarily attributed to their ability to efficiently model the time varying statistical characteristics of the speech signal through a set of discrete states. However, they still posses many modelling inadequacies that derive from the numerous assumptions that are made to simplify the speech recognition problem. For instance, dynamic information in HMMs is included through the time-derivatives in the observation vector under the false frame-independence assumption and the spatial correlation of the observation vector is ignored when diagonal covariance matrices are considered.

This work is motivated from our belief that these assumptions set upper limits in the progress that can be made when using HMMs in speech recognition. In an effort to improve robustness, particularly under noisy conditions, we examine new modeling schemes that can explicitly model time and

spatial correlations such as the linear dynamical models (LDM). LDMs were first proposed to be used for speech recognition in [1]. They characterize complete speech segments such as words, phonemes or sub-phoneme units with a linear state evolution process and a linear observation process. Thus, they can be seen as a variation of segment-based modeling which, in turn, can be considered as a generalization of the HMMs with a continuous state-space instead of a discrete one[2].

There are several variations of the LDMs that can be found in the literature. In [1] LDMs were used to obtain a smoothed realization of a Gauss-Markov model. In [3] and [4] several statistical modeling techniques such as factor analysis (FA) and principle component analysis (PCA) are presented as special cases of a general LDM. Other variations are also discussed in [5]. In all cases, several modeling constraints were applied in an effort to obtain good system convergence, stability and identifiability. However, these constraints alter the properties of the model, and diminish the benefits of the general system architecture.

In this paper, we introduce a generalized linear dynamic system in an identifiable canonical form. The system is a multivariate state-space linear dynamic model which follows the identifiable form that was proposed by Ljung [6]. We begin by introducing the linear system and its parametric structure. We describe our novel element-wise estimation method based on the Expectation-Maximization (EM) algorithm. Finally, we present classification results on the AURORA2 speech database.

2. THE LINEAR DYNAMIC SYSTEM

The LDM is described from the following pair of equations

$$x_{k+1} = Fx_k + w_k \qquad (1)$$

$$y_k = Hx_k + v_k \qquad (2)$$

where the state x_k at time k is a $(n \times 1)$ vector, the observation y_k is $(m \times 1)$ and w_k, v_k are uncorrelated, zero-mean Gaussian vectors with covariances

$$E\{w_k w_l^T\} = P\delta_{kl} \qquad (3)$$

$$E\{v_k v_l^T\} = R\delta_{kl} \qquad (4)$$

In the above equation δ_{kl} denotes the Kronecker delta and T denotes the transpose of a matrix. The initial state x_0 is Gaussian with known mean and covariance μ_0, Σ_0. Equation (1) describes the state dynamics, while (2) shows a prediction of the observation based on the state estimation.

The parametric structure of our multivariate state-space model has the following identifiable canonical form for the case in which x_k is a 5×1 vector and the observation vector y_k is a 3×1 vector.

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & 1 \\ \times & \times & \times & \times & \times \end{bmatrix} \tag{5}$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{6}$$

The number of rows with $\times$' s in F represents the free parameters of the matrix and equals the size of the output vector m. The ones in matrix H are equal to the number of the rows in F that are filled with free parameters, and their position is related to the location of these rows in F.

To construct the form of the state transition matrix F we follow the process described in [6]. First, we set its elements along the superdiagonal equal to one and the remaining elements are zeroed. Then, we choose arbitrarily the m row numbers r_i to be filled with free parameters, where $i = 1, ..., m$. There is only one constraint, that $r_m = n$, where m denotes the dimension of the observation and n the dimension of the state vector. In addition, we set $r_0 = 0$.

The observation matrix H is then constructed as follows. First, we define H to be $m \times n$ in size and filled with zeros. Then we set each row $i = 1, ..., m$ of the H matrix to have a one in column $c_i = r_{i-1} + 1$. For instance, for the example shown in (5) and (6) we get:
$r_1 = 2 \Rightarrow c_1 = r_{1-1} + 1 = r_0 + 1 = 1$
$r_2 = 3 \Rightarrow c_2 = r_{2-1} + 1 = r_1 + 1 = 3$
$r_3 = 5 \Rightarrow c_3 = r_{3-1} + 1 = r_2 + 1 = 4$.
Hence, the observation matrix H will have ones in columns 1, 3 and 4 for its rows 1, 2 and 3, respectively.

Ljung [6] proves that the above canonical form is identifiable if and only if it is also controllable. Furthermore, this canonical form does not impose any loss of generality in the LDM, that is, any state-space system described by equations 1 and 2 can be transformed to have the structure of equations 5 and 6.

3. ELEMENT-WISE ESTIMATION WITH EM

The matrices of the LDM presented in section 2 contain, by-construction, free parameters at very specific positions. An estimation algorithm for linear state-space systems that is based on the Expectation-Maximization (EM) algorithm was introduced in [1]. This algorithm assumed that all matrices $\theta = F, H, P, R$ are filled with free parameters. In our case, however, the free parameters of the system are located in specific position, hence the estimation must be performed in an element-wise fashion. Given the observations $\mathbf{Y} = [y_0 \ldots y_N]$ and the state vectors $\mathbf{X} = [x_0 \ldots x_N]$, the ML estimates of θ are obtained by minimizing the quantity:

$$L(\mathbf{X}, \mathbf{Y}, \theta) =$$
$$-\sum_{k=1}^{N} \left\{ \log |P| + (x_k - Fx_{k-1})^T P^{-1} (x_k - Fx_{k-1}) \right\}$$
$$-\sum_{k=0}^{N} \left\{ \log |R| + (y_k - Hx_k)^T R^{-1} (y_k - Hx_k) \right\}$$

It can be shown that the estimates of the system's parameters are given by:

$$\hat{F}_{ij} = \frac{\sum_{c=1}^{M} \left\{ (cof(\hat{P}_{ic}))(S_{cj}^{(4)}) \right\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})}$$
$$- \frac{\sum_{c=1, c\neq i}^{M} \left\{ (cof(\hat{P}_{ic}))(\hat{F}_{cj})(S_{jj}^{(3)}) \right\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})}$$
$$- \frac{\sum_{c=1}^{M} \left\{ (cof(\hat{P}_{ic})) \sum_{r=1, r\neq j}^{M} \left\{ (\hat{F}_{cr})(S_{rj}^{(3)}) \right\} \right\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})} \tag{7}$$

$$\hat{P}_{ij} = (S_{ij}^{(2)}) - \sum_{r=1}^{M}(\hat{F}_{ir})(S_{jr}^{(4)}) - \sum_{r=1}^{M}(\hat{F}_{jr})(S_{ir}^{(4)})$$
$$+ \sum_{c=1}^{M}\sum_{r=1}^{M}(\hat{F}_{ic})(\hat{F}_{jr})(S_{cr}^{(3)}) \tag{8}$$

$$\hat{R} = S^{(5)} - S^{(6)}(S^{(1)})^{-1}(S^{(6)})^T \tag{9}$$

where $cof(\hat{P}_{ic})$ is the cofactor of the element $\hat{P}_{ic}$ of the covariance $\hat{P}$. Index i denotes the $i - th$ row of a matrix, and j denotes the $j - th$ column. The sufficient statistics that in-

volved in the previous equations are given by[1]

$$S^{(1)} = \frac{1}{N+1}\sum_{k=0}^{N} x_k x_k^T \tag{10}$$

$$S^{(2)} = \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T \tag{11}$$

$$S^{(3)} = \frac{1}{N}\sum_{k=1}^{N} x_{k-1} x_{k-1}^T \tag{12}$$

$$S^{(4)} = \frac{1}{N}\sum_{k=1}^{N} x_k x_{k-1}^T \tag{13}$$

$$S^{(5)} = \frac{1}{N+1}\sum_{k=0}^{N} y_k y_k^T \tag{14}$$

$$S^{(6)} = \frac{1}{N+1}\sum_{k=0}^{N} y_k x_k^T. \tag{15}$$

The statistics shown above require the following quantities at each iteration p:

$$E_{\theta^{(p)}}\{y_k x_k^T | \mathbf{Y}\} = y_k \hat{x}_{k|N} \tag{16}$$

$$E_{\theta^{(p)}}\{y_k y_k^T | \mathbf{Y}\} = y_k y_k^T \tag{17}$$

$$E_{\theta^{(p)}}\{x_k x_{k-1}^T | \mathbf{Y}\} = \Sigma_{k,k-1|N} + \hat{x}_{k|N}\hat{x}_{k-1|N}^T \tag{18}$$

$$E_{\theta^{(p)}}\{x_k x_k^T | \mathbf{Y}\} = \Sigma_{k|N} + \hat{x}_{k|N}\hat{x}_{k|N}^T. \tag{19}$$

Equations (7) through (9) form the Maximization step of the EM algorithm. For the Expectation step of the EM algorithm we need to compute the required statistics, and we use the fixed interval smoothing form of the Kalman filter (RTS smoother) [7]. It consists of a backward pass that follows the standard Kalman filter forward recursions [8]. In addition, we compute also the cross-covariances proposed by Digalakis [1] in both the forward and the backward pass.

4. APPLICATION TO SPEECH RECOGNITION

A straightforward way to model speech units using LDMs is to train separate segment-specific models, each one corresponding to a sub-word or sub-phoneme unit. The correlation between consecutive frames within the same segment is modelled with the same set of parameters. Furthermore, the inter-segment correlation is also captured since the initial state estimate of a segment derives from the last state estimate of the previous segment. The process is also illustrated in Figure 1 for a 4 segment example.

During classification, each model segment is classified based on the log-likelihood computed by:

$$L(\mathbf{Y}, \theta) = -\sum_{k=0}^{N}\left\{ \log|\Sigma_{e_k}(\theta)| + e_k^T(\theta)\Sigma_{e_k}^{-1}(\theta)e_k^T(\theta)\right\} + C$$
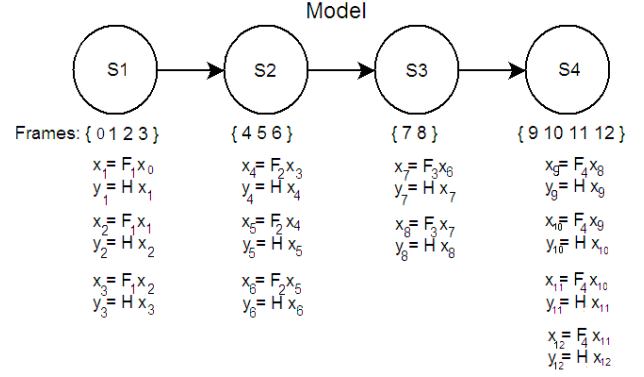

Fig. 1. Example of an LDM with 4 segments.

where $e_k^T(\theta)$, $\Sigma_{e_k}(\theta)$ is the prediction error and its covariance obtained from the Kalman filter equations and C is a constant.

5. EXPERIMENTS

We have performed a series of word-classification experiments in order to validate our LDM system for speech recognition and evaluate the estimation algorithm. In specific, we used the AURORA2 speech database[9], which is a connected digit corpus based on TIDIGITS, downsampled to 8KHz and with several types of noise artificially added at several SNRs. The front-end uses a total of 13 Mel-warped cepstral coefficients plus energy. In some experiments we also augmented the observation vector with the first (δ) and second order derivatives ($\delta\delta$).

We used 11 word-models corresponding to the words in the AURORA2 corpus (digits 1 to 9, zero and oh). Each word-model has a number of time-invariant regions (segments) ranging from 2 to 8, depending on the phonetic transcription of each word. Table 1 shows the number of regions for each word-model that we considered.

one	two	three	four	five	six
6	4	6	6	6	4

seven	eight	nine	oh	zero
8	4	6	2	6

Table 1. Number of regions for each word-model

The first issue in implementing the dynamical system is the dimensionality of the state-space. Based on the general canonical forms of the LDM that we examine, the size of the state-vector can be equal or larger than the size of the observation vector. When the state and observation vectors are at equal size, the observation matrix becomes the identity matrix and the observation vector is just a noisy version of the state vector. Even in this case, our scheme relaxes the constraints

of other approaches (i.e. in [1]).

Another important issue is the initialization of system parameters. The noise covariance matrices are initialized randomly, while the initial state-transition matrices, and the covariance of the initial state x_0 are directly estimated from the observations.

As far as the classification is concerned, at this moment we do not perform any search over all possible segmentations, but we keep the true word-boundaries produced by an HMM fixed. We do search, however, over all possible word histories given the segmentation. To speed-up the classification process we apply a suboptimum search and pruning algorithm which keeps the 11 most probable word-histories for each word in the sentence.

For our experiments, we used a clean training set consisting of 104 gender-balanced speakers and 8444 sentences. The evaluation was done on a separate test set defined as the AURORA2-A test set, with subway additive noise at several SNRs, which consisted of 1000 sentences from the training speakers. Table 2 summarizes the classification performance of the LDM for several SNR values. As can be seen, appending the derivatives in the MFCCs results in performance gains which increase as the noise level rises.

AURORA 2/	LDM	
Subway	Mfcc,energy	$+\delta+\delta\delta$
clean	97.53%	97.61%
SNR20	93.23%	95.12%
SNR15	87.91%	91.13%
SNR10	76.29%	82.69%
SNR5	54.87%	63.56%

Table 2. Word-classification performance of the LDM system

AURORA 2/	HMM	
Subway	Mfcc,energy	$+\delta +\delta\delta$
clean	97.19%	97.57%
SNR20	90.91%	95.71%
SNR15	80.09%	91.76%
SNR10	57.68%	81.93%
SNR5	36.01%	64.24%

Table 3. Performance of an HMM system

To compare the performance of our system to HMMs, we also performed a set of classification experiments using the standard HTK configuration. Each word was modelled as a 16-state continuous density HMM with a mixture of 3 Gaussian components associate in each state. The front-end configuration and the word-boundaries were the same as with the LDM. The recognition accuracy of the HMM is shown in Table 3. Without derivatives, the LDM outperforms sig-

nificantly the HMM, especially as the SNR level decreases. When derivatives are used for both models, their performance is similar.

6. CONCLUSIONS

In this paper, we presented the application of linear dynamic models with general, canonical forms for their parameters in speech recognition and we showed a novel and efficient element-wise Maximum Likelihood estimation. We evaluated our scheme with a series of classification experiments on the AURORA2 speech database. Since we have now introduced a methodology to use general identifiable forms of state-space systems in speech recognition, we plan to investigate in the future several combinations of state and observation vector dimensions.

7. REFERENCES

[1] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the em algorithm and its application to speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, Oct. 1993.

[2] M. Ostendorf, V. Digalakis, and O.A. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 360–378, 1996.

[3] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, 1999.

[4] A. Rosti and M. Gales, "Generalised linear gaussian models," Tech. Rep., Engineering, Cambridge University, 2001.

[5] J. Frankel and S. King, "Speech recognition using linear dynamic models," *IEEE Transactions on Speech and Audio Processing*, January 2007.

[6] L. Ljung, *System Identification: Theory for the User (2nd Edition)*, Prentice Hall PTR, 1998.

[7] H.E. Rauch, F. Tung, and C.T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, pp. 1445–1450, August 1965.

[8] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, Series D, J. Basic Eng.*, vol. 82, pp. 35–45, March 1960.

[9] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, 2000.