AN EKF-BASED ALGORITHM FOR LEARNING STATISTICAL HIDDEN DYNAMIC MODEL PARAMETERS FOR PHONETIC RECOGNITION

Roberto Togneri

The University of Western Australia, Australia

Li Deng

University of Waterloo, Canada
currently at Microsoft Research, Redmond, WA, USA

ABSTRACT

This paper presents a new parameter estimation algorithm based on the Extended Kalman Filter (EKF) for the recently proposed statistical coarticulatory Hidden Dynamic Model (HDM). We show how the EKF parameter estimation algorithm unifies and simplifies the estimation of both the state and parameter vectors. Experiments based on N-best rescoring demonstrate superior performance of the (context-independent) HDM over a triphone baseline HMM in the TIMIT phonetic recognition task. We also show that the HDM is capable of generating speech vectors close to those from the corresponding real data.

1. INTRODUCTION

Hidden Dynamic models (HDMs) [1, 8, 9, 2, 3, 4] attempt to model the intrinsic dynamics in the human speech production system in an effort to address some of the known weaknesses of the current hidden Markov modelling (HMM) paradigm when applied to acoustic modelling for unconstrained, spontaneous speech recognition. Such weaknesses include the HMM's inability to adequately model coarticulation and phonological variation without resorting to the use of very large numbers of context-dependent models of an enumerative type. There is an increasing demand for speech recognisers to cope with larger vocabularies, less constrained task grammars, large populations of speakers and backgrounds, and different speaking styles. With the current HMM paradigm this can only be achieved by using copious amounts of training data and sophisticated clustering algorithms to reliably estimate the many parameters. The very large number of model parameters makes it difficult for a speech recognizer to adapt to a new speaker, a new speaker style, and a new environment. This is a direct consequence of the blind, data-driven approach of the current HMM approach when applied to acoustic modelling.

The HDM described in this paper adopts a more structured model of the underlying human speech production dynamics by describing the acoustic features as the observations measured from a state-space model description of the speech production process. While describing the articulatory dynamics of speech production is in itself an unsolved problem, a simpler alternative is to describe the known spectral manifestations of the process. In this paper the statistical coarticulatory modelling of the vocal-tract-resonances (VTRs) proposed by Deng [2, 1, 3] is further investigated. The principal advantage of such a model lies in the compact structure for representing long-term contextual dependence in the observable speech acoustics. This is based on the lower-dimensional, less variable, VTR feature space compared to the higher-dimensional, highly variable MFCC feature space. The compact structure also results in fewer parameters that need to be estimated and less training data needed to estimate the parameters reliably.

In this paper we present a new parameter estimation algorithm based on the full use of the extended Kalman filter (EKF) for both state and parameter estimation as an alternative to the use of the EM algorithm, where the EKF was used only for state estimation in the E-step. [2]. Experiments on the TIMIT phone recognition are performed to evaluate the performance of the HDM based on rescoring of the N-best lists generated by a baseline HMM. Investigation of the new learning algorithm demonstrates the convergence of the model parameters to the corresponding realistic acoustic observations.

2. MODEL FORMULATION

The hidden dynamic model presented in this paper is based on the statistical coarticulatory model described in [2, 1]. The system model consists of a target-directed hidden dynamic state process coupled with a non-linear observation process.

The hidden dynamic "state" equation is used to describe the vocal-tract resonance (VTR) dynamics according to:

$$z(k+1) = \Phi^j z(k) + (I - \Phi^j)T^j + w(k) \qquad (1)$$

where $z(k)$ is the three-dimensional state vector and T^j and Φ^j are the phone target and diagonal "time-constant" system matrix parameters associated with the phone regime

j. The process noise, $w(k)$, is represented in this study by an i.i.d, zero-mean, Gaussian process with covariance matrix Q. A feature of this model is its ability to switch state-space parameters when crossing over to new phone dynamic regimes and continuity of the hidden state variable $z(k)$ across phone regimes. The latter provides a long-span continuity across phone regimes and structurally models the inherent context dependencies and coarticulatory manifestations between adjacent phone regimes.

The observation equation is used to describe the mapping between the hidden state dynamic to the observable acoustic features. The most general form of the observation equation is a static, nonlinear mapping as follows:

$$O(k) = h^r(z(k)) + v(k) \tag{2}$$

where the acoustic observation $O(k)$ is the set of Mel cepstral coefficients (MFCCs) at frame k, and $v(k)$ is modelled by an i.i.d, zero-mean, Gaussian process with covariance matrix R and represents the additive observation noise which captures the residual errors in mapping from $z(k)$ to $O(k)$. The multivariate nonlinear mapping, $h^r(z(k))$, is implemented by a multi-layer perceptron (MLP) for each distinct manner of articulation r.

A three-layer feedforward multi-layer perceptron was implemented for the nonlinear function $h^r(z(k))$ with linear activation function on the output layer and the antisymmetric hyperbolic tangent function:

$$g(x) = 1.7159 \tanh((2/3)x)$$

on the hidden layer.

3. PARAMETER ESTIMATION BY EKF

The parameter estimation method for the hidden dynamic model can be based on a generalised EM algorithm [2]. However due to the nonlinear equations in the M-step and the crude approximation for estimation of the MLP weights in this paper we propose to use the EKF algorithm for joint state and parameter estimation. This is achieved by using the appropriate augmented form of the state equation defined as:

$$\theta(k) = \begin{pmatrix} z(k) \\ \tilde{\Phi}^j(k) \\ T^j(k) \end{pmatrix} \tag{3}$$

where $\tilde{\Phi}^j(k)$ is the $m^2 \times 1$ time-constant vector and $m = 3$ is the dimension of the state vector. $\tilde{\Phi}^j(k)$ is related to the time-constant matrix $\Phi^j(k)$ as follows:

$$\tilde{\Phi}^j(k) = \begin{pmatrix} \Phi_1^{j'}(k) \\ \Phi_2^{j'}(k) \\ \vdots \\ \Phi_m^{j'}(k) \end{pmatrix}$$

where $\Phi_i^j(k)$ is row i of $\Phi^j(k)$. The new state equation becomes

$$\theta(k+1) = f(\theta(k)) + w(k), \tag{4}$$

which is now non-linear in the state variable, $\theta(k)$, and can be decomposed as follows:

$$\begin{pmatrix} z(k+1) \\ \tilde{\Phi}^j(k+1) \\ T^j(k+1) \end{pmatrix} = \begin{pmatrix} \Phi^j(k)z(k) + (I - \Phi^j(k))T(k) \\ \tilde{\Phi}^j(k) \\ T^j(k) \end{pmatrix} + \begin{pmatrix} w_z(k) \\ w_\Phi(k) \\ w_T(k) \end{pmatrix}$$

The measurement equation becomes

$$O(k) = h^r(\theta(k)) + v(k), \tag{5}$$

where it is noted that the nonlinear mapping function $h^r(.)$ is strictly dependent only on $z(k)$.

The standard EKF algorithm recursion [7, 5] is used to yield joint state and parameter estimates at each time-step. This use of the EKF obviates the need for an additional EM algorithm step for parameter estimation, and conveniently estimates both the system dynamic and MLP weight parameters.

The expression for the $(2m + m^2)$ x $(2m + m^2)$ state equation Jacobian matrix $F_\theta[\hat{\theta}(k|k)] = \frac{\partial f(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}(k+1|k)}$ used in the EKF recursion can be shown to be:

$$F_\theta[\hat{\theta}(k|k)] = \begin{bmatrix} \hat{\Phi}^j(k|k) & \frac{\partial f}{\partial \Phi}(k|k) & I_m - \hat{\Phi}^j(k|k) \\ 0 & I_{m^2} & 0 \\ 0 & 0 & I_m \end{bmatrix},$$

where $\hat{\Phi}^j$ is the current estimate of the $m \times m$ time-constant matrix and

$$\frac{\partial f}{\partial \Phi}(k|k) = \begin{bmatrix} [z(k|k) - T^j(k|k)]' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & [z(k|k) - T^j(k|k)]' \end{bmatrix}$$

is the $m \times m^2$ partial derivative submatrix expression for $\frac{\partial f}{\partial \Phi}$.

The $(n) \times (2m + m^2)$ observation equation Jacobian matrix, $H_\theta[\hat{\theta}(k+1|k)]$ used in the EKF recursion, is only dependent on $z(k+1|k)$ and is expressed as:

$$H_\theta[\hat{\theta}(k+1|k)] = \begin{bmatrix} H_z[\hat{z}(k+1|k)] & 0 & 0 \end{bmatrix}$$

where n is the dimension of the acoustic observation vector and the elements of the $n \times m$ Jacobian submatrix, $H_z[\hat{z}(k+1|k)]$, at row j and column i are defined as:

$$H_z^{ji}[\hat{z}(k+1|k)] = \left[\frac{\partial O_j(k+1)}{\partial z_i(k+1)}\right]$$
$$= \left[\sum_{h=1}^{J} W_{2j}(h) g'(W_{1h}^{'} z(k+1)) W_{1h}(i)\right]$$

where W_{lj} is the MLP weight vector of node j in layer l and $g'(x)$ is the derivative of the activation function.

Use of the EKF for joint state and parameter estimation requires initial values for both the augmented state vector, $\theta(0|0) = (z(0|0), \tilde{\Phi}(0|0), T(0|0))^{'}$, state error covariance matrix, $P(0|0)$ and specification of the noise covariances $Q(k)$ and $R(k)$. Selection of the covariance parameters is crucial in guaranteeing convergence of the EKF recursion.

Evaluation of the HDM is based on a rescoring task which requires the model to output a score (i.e. likelihood) of a given utterance given the segmented phone transcription. The log-likelihood scoring function used in this work is identical to that reported in [3].

4. EXPERIMENTS

Due to the complexity of direct search and of lattice rescoring with the HDM, the evaluation of the new HDM learning algorithm was carried out performing N-best rescoring on time aligned transcriptions produced by a baseline HMM system [8]. Evaluation of the HDM in this paper was based on the phone recognition task using the TIMIT corpus. Due to the extremely large computational requirements of the current implementation of the estimation algorithm, only the speaker dr8 subset was used for training the HDM phone models and performing the evaluations. To produce time-aligned transcriptions a baseline context-dependent phone HMM system was trained on the complete TIMIT training data and tested on the dr8 test data subset.

The acoustic features used were 13 dimensional static MFCC vectors for the HDM models and 39 dimensional static, delta and delta-delta MFCC vectors for the HMM models. The HDM hidden dynamic was a 3-dimensional VTR state vector requiring a 3-input, 13-output MLP nonlinear mapping function, $h^r(z(k))$ to map the VTR dynamic to the observable MFCC observation vectors.

Two implementations of the HDM were evaluated. The HDMm implementation used one 3-layer, 12 hidden node, MLP per phone model for the nonlinear mapping in the observation process. The HDMc implementation used only three broad class (Silence, Voiced, Unvoiced) 3-layer, 16 hidden node MLPs. For both implementations 5 iterations of the EKF parameter estimation were used, the noise covariance $Q(k)$ associated with the parameters was set to zero and arbitrary values for state error covariance $P(k)$ were chosen to drive the EKF state and parameter estimation updates. Identical initialisation of the state and parameter vectors was used for all phone models.

The HMM was used to generate the 100-best and 5-best time-aligned transcription for the dr8 test data utterances and the corresponding reference transcription. The HMM, HDMm and HDMc rescored the 100-best, 100-best+ref, 5-best and 5-best+ref transcriptions and the top score was used to evaluated the WER and sentence error rate (SER) performance of the system. The bounds on performance were provided by the Chance and Oracle systems. A random transcription was chosen for Chance (lower bound) and the best transcription was chosen for Oracle (upper bound). In addition to recognition performance the number of parameters to be estimated for each system was also derived. The results of the evaluations obtained so far are presented in Table 1. While the WERs are roughly the same for the HMM and HDM when the recognizer is not exposed to the reference transcription, upon the exposure the WER drops significantly for the HDM but not for the HMM. Further, such error rate reduction is achieved with the use of much fewer HDM model parameters than HMM. This is a highly desirable property since it would make any adaptive learning algorithm (to be developed) much more effective.

System	Oracle	Chance	HMM	HDMm	HDMc
100-best	18.4 *94.6*	29.1 *100.0*	28.7 *100.0*	28.9 *100.0*	29.1 *100.0*
100-best + ref	0.0 *0.0*	28.4 *99.1*	28.7 *100.0*	22.4 *80.0*	22.2 *81.8*
5-best	24.7 *99.1*	27.6 *100.0*	27.8 *100.0*	28.2 *100.0*	27.7 *100.0*
5-best + ref	0.0 *0.0*	25.8 *90.9*	27.7 *99.1*	17.7 *65.4*	13.1 *51.8*
Parameters	N/A	N/A	778245	11350	1115

Table 1. Analysis of WER (normal), SER (italics), and number of parameters on the phone recognition task using the dr8 subset from the TIMIT corpus. Evaluation was performed on rescoring the 100-best and 5-best HMM time-aligned transcriptions with and without the reference transcription.

To investigate the generative properties of the HDM, we show a typical plot in Figure 1 of the real MFCC acoustic feature vector, $O(k)$, together with the corresponding HDMm and HDMc outputs, $h(z(k))$. It is evident that the HDMs attempt to converge to the observation output as a consequence of the EKF parameter estimation being driven by minimisation of the innovation sequence. Some evidence of the target-directed nature of the underlying production process can also be seen by the positioning of the phone

segment centers where there is a change in the target and time-constant dynamics.

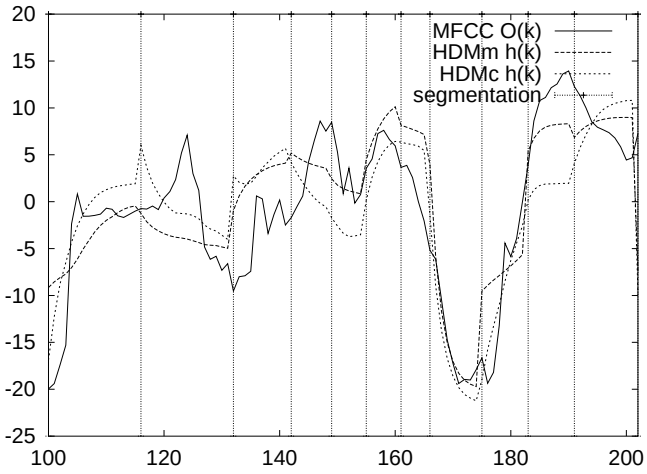

Fig. 1. Plot of the first component of the MFCC acoustic observation and HDM model output vectors from frame 100 to 202. The phone segment centers are indicated by the vertical lines.

5. SUMMARY AND CONCLUSION

The HDM approach represents an important new paradigm for acoustic modelling based on a more structured and parsimonious model of the human speech generation process. The results presented in this paper indicate the superior performance of the HDM, especially when exposed to the reference transcription. The HMM often failed to provide high scores for the reference transcription while the HDM succeeded in this in most cases. The main contribution of this paper is the novel EKF-based parameter learning algorithm which enables such success.

Further work is needed to develop a lattice scoring algorithm with optimal segmentation of the dynamic regimes to properly evaluate the performance of the HDM and to investigate alternative EKF and EM parameter estimation algorithms that incorporate estimation of the phone boundaries.

6. ACKNOWLEDGEMENTS

We are grateful to Jeff Ma for many helps in carrying out this work and for his suggestions. Main components of this work were developed while the first author was taking sabbatical leave at U. Waterloo in 1999.

7. REFERENCES

[1] L. Deng, “A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition”, Speech Communication, Vol. 24, 1998, pp. 299-323.

[2] L. Deng, J. Ma, “A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics”, Proc Eurospeech, pp. 1499-1502, September 1999.

[3] L. Deng and J. Ma. “Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics,” J. Acoust. Soc. Am., Vol. 108, No. 5, November 2000.

[4] Y. Gao, R. Bakis, J. Huang, B. Ziang, “Multistage coarticulation model combining articulatory, formant and cepstral features”, Proc. ICSLP, October 2000, pp. 25-28.

[5] L. Ljung, “Asymptotic Behavior of the Extended Kalman Filter as a Parameter Estimator for Linear Systems”, IEEE Trans. Automat. Control, 1972, pp. 693-698.

[6] J. Ma, L. Deng, “A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech”, Computer Speech and Language, Vol. 14, 2000, pp. 101-114.

[7] J. Mendel, “Lessons in Estimation Theory for Signal Processing, Communications, and Control”, Prentice-Hall, 1995

[8] J. Picone, et. al, “Initial evaluation of hidden dynamic models on conversational speech”, Proc. ICASSP’99, pp. 109-112, March 1999.

[9] H.B. Richards, and J.S. Bridle, “The HDM: A segmental hidden dynamic model of coarticulation”, Proc. ICASSP’99, pp. 357-360, March 1999.