International Journal of Speech Technology

Continuous Speech Recognition Using Linear Dynamic Models

--Manuscript Draft--

Manuscript Number:	IJST-D-13-00013R1
Full Title:	Continuous Speech Recognition Using Linear Dynamic Models
Article Type:	Manuscript
Keywords:	linear dynamic models; nonlinear statistical modeling; speech recognition; acoustic modeling
Corresponding Author:	Joseph Picone, Ph.D. Temple University Elkins Park, Pennsylvania UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Temple University
Corresponding Author's Secondary Institution:	
First Author:	Tao Ma, Ph.D.
First Author Secondary Information:	
Order of Authors:	Tao Ma, Ph.D.
	Sundararajan Srinivasan, Ph.D.
	Georgios Lazarou, Ph.D.
	Joseph Picone, Ph.D.
Order of Authors Secondary Information:	
Abstract:	Hidden Markov models (HMMs) with Gaussian mixture distributions rely on an assumption that speech features are temporally uncorrelated, and often assume a diagonal covariance matrix where correlations between feature vectors for adjacent frames are ignored. A Linear Dynamic Model (LDM) is a Markovian state-space model that also relies on hidden state modeling, but explicitly models the evolution of these hidden states using an autoregressive process. An LDM is capable of modeling higher order statistics and can exploit correlations of features in an efficient and parsimonious manner. In this paper, we present a hybrid LDM/HMM decoder architecture that postprocesses segmentations derived from the first pass of an HMM-based recognition. This smoothed trajectory model is complementary to existing HMM systems. An Expectation-Maximization (EM) approach for parameter estimation is presented. We demonstrate a 13% relative WER reduction on the Aurora-4 clean evaluation set, and a 13% relative WER reduction on the babble noise condition.

General Comments:

Linear dynamic models were first introduced to speech recognition applications by Digilakis et al. (1993). Unfortunately, the techniques described in those papers simply did not scale to the problems addressed here. The main contribution of this paper is to extend this work to a very challenging large-scale task – speech in noise task (Aurora 4) – that features a mismatched training condition. This paradigm is still beyond most state of the art machine learning systems.

Below we address the specific concerns of the reviewers:

Reviewer #3:

The paper "Continuous speech recognition using linear dynamic models" submitted by you to International Journal of Speech Technology with the number IJST-D-13-00013 is THE SAME PAPER AS: "Continuous Speech Recognition Using Linear Dynamic Models" of the authors T. Ma, S. Srinivasan, G. Lazarou and J. Picone submitted to be published in IEEE Signal Processing Letters on November 8, 2010 with the number SPL-06538-2010. The paper was web-published in: http://www.isip.piconepress.com/publications/unpublished/journals/2010/ieee_spl/ldm/paper_double_v11.docx As I don't have a conclusion about the revision process of IEEE SPL , lamentably I must reject it.

Perhaps there was a misunderstanding due to the blind review process. This above web site that is referenced is, in fact, the web site of the authors. An early version of this paper was submitted to the IEEE Signal Processing Letters and rejected, with no chance for resubmission. Since then, the graduate student doing the work, Tao Ma, finished the work and completed his PhD. We have made substantial revisions to the paper and resubmitted to this journal.

We have not previously published this work in a journal or conference paper. The experimental results are much more complete and overall the paper does a much better job of explaining the contributions. Due to various circumstances, the co-authors have not had time to complete the necessary revisions until recently. We see no reason why it does not qualify for publication in your journal since it is not currently under consideration in any other journal and has not been previously published.

Reviewer #2:

1. Two pass ASR system with different acoustic models in different pass has been proved and used widely to improve upon one pass ASR system. In the proposed framework LDM is used in the second pass rescoring. However, during second pass rescoring, I am not sure why the authors need to combine the LDM likelihood score with HMM likelihood score in the 2nd pass. Is it neccessary?

What if only LDM likelihood score is used for 2nd pass rescoring? Will the results be very different? If so, could the authors explain the reason?

This is an excellent observation. This problem was studied extensively in Dr. Ma's dissertation: Ma, T. (2010). *Linear dynamic model for continuous speech recognition*. Mississippi State University. This work was cited in the paper. Due to space limitations we could not explore all the details presented in the dissertation.

The continuous speech recognition experiments with various hybrid systems are described more fully in Chapter V. Table 5.2 directly addresses the concern of the reviewer. Mixing of these parameters does not dramatically change the results but does provide some benefit.

In our previous work with similar hybrid systems, we have observed that mixing scores is desirable. The basis for our hybrid system work dates back to this publication:

> Ganapathiraju, A., Hamaker, J., & Picone, J. (2004). Applications of Support Vector Machines to Speech Recognition. *IEEE Transactions on Signal Processing*, 52(8), 2348-2355.

Our understanding of this phenomenon is that it relates to the interdependency of segmentation and scoring. We refer to this at several points in the paper. Since the LDM model is not integrated inside the training loop, relying on LDM solely creates a mismatch during the search process between the HMM scores/segmentation process and the LDM model. It is very hard to decouple these in practice. Relying solely on LDM scores produces marginal improvements in performance, because the LDM model is not allowed to reach into the search pass and produce better hypotheses. Combining the scores seems to give LDM a better chance to access and rescore hypotheses that can make a difference in the WER.

Is this optimal? Of course not. We would ultimately like to integrate this approach inside the training loop, and then we believe there would be no need to decouple the scores. These experiences provided the motivation to adopt an integration of the scores.

We also explored this issue in the context of a basic phone classification task for TIDigits. The LDM model is described in Chapter IV (p. 66). In the summary, we state:

> Results for each phonetic class are presented individually in Figure 4.5. The relative differences in classification accuracy are not consistent among the phonemes. It can be seen that the classification results for fricatives and stops are high, while classification results for glides are lower (~85%). Vowels and nasals result in mediocre accuracy (89% and 93% respectively). Overall, LDMs provide a reasonably good classification performance for TIDigits.

We referenced these results in the paper, but chose not to explore them fully due to space limitations. We have modified the paper to explain this issue in more detail.

2. The experiments demonstrate LDM will be a good complementary for traditional HMM/GMM score because LDM is used in 2nd pass recoring. To directly compare LDM likelihood score with HMM/GMM likelihood score, perhaps the authors could consider putting LDM in the 1st pass recognition. My understanding of the difficulty of such experiment is that it's not straightforward to use LDM to do segmentation. What if HMM/GMM is used do segmentation, then LDM is used to do scoring on the provided segments?

Again, Dr. Ma's dissertation explores this issue in more detail (and this work is referenced). We have not been able, due to time, funding and other limitations, to fully explore integrating LDM into the first pass of the recognition system. This is not easy to do from a theoretical or implementation standpoint (the software is extremely complex). The hybrid systems approach we have taken is fairly standard and meant to give simply an indication of the promise of this method. The improvements we see are typical of other published work on such N-best rescoring approaches and demonstrate there would be benefit to integrating this into the first pass. But the effort to do this is beyond our resources at this time.

3. Several places in the manuscript mentioned that "HMMs typically assume a diagonal covariance matrix...". To be more accurate, it is HMMs/GMMs(HMMs with GMMs observation distribution) that assumes a diagonal covariance matrix for GMMs.

Excellent point. The need for brevity influenced our decision to use the term HMM to refer to both the Markov structure and the observation distributions. We have modified the manuscript appropriately to make this distinction more clear.

Note also that both diagonal and full covariance models were explored in Dr. Ma's dissertation work for a phone classification problem. Again, due to space limitations, we chose not to discuss these experiments in this paper.

4. Apprently there are some format issues with all the equations in this manuscript. Please fix those.

We apologize for this inconvenience. When we submitted the paper, we pointed this out. There are some MS Word problems encountered when we upload the paper. We have provided a pdf and Word document via email to the editor so hopefully the reviewers can access a clean copy of the paper. We have also tried to upload a different version to see if we can resolve these problems.

Summary Comments:

We would like to thank the reviewers for their thoughtful and insightful feedback. We have done our best to address each of their concerns. We hope our responses have adequately addressed your concerns and we can proceed with publication.

We want to emphasize that this is the first time this particular model has been shown to provide a statistically significant improvement on a challenging task. Recognition results on the Aurora 4 Corpus are well calibrated. These improvements are significant compared to the types of improvements typically published in human language technology papers.

The work constitutes the dissertation of a graduate student and represents four years of work on these problems, and is extremely comprehensive in nature. Again, this work has not been previously published and has been struggling through the review process for a long time. We selected ISTJ, a relatively new journal for us, because of its timely review process. The IEEE journal that we previously submitted to took an excessive amount of time. We hope we can bring this to a speedy conclusion.

CONTINUOUS SPEECH RECOGNITION USING LINEAR DYNAMIC MODELS[1]

Tao Ma[2], Sundararajan Srinivasan[3], Georgios Lazarou[4] and Joseph Picone[5]

Abstract— Hidden Markov models (HMMs) with Gaussian mixture distributions rely on an assumption that speech features are temporally uncorrelated, and often assume a diagonal covariance matrix where correlations between feature vectors for adjacent frames are ignored. A Linear Dynamic Model (LDM) is a Markovian state-space model that also relies on hidden state modeling, but explicitly models the evolution of these hidden states using an autoregressive process. An LDM is capable of modeling higher order statistics and can exploit correlations of features in an efficient and parsimonious manner. In this paper, we present a hybrid LDM/HMM decoder architecture that postprocesses segmentations derived from the first pass of an HMM-based recognition. This smoothed trajectory model is complementary to existing HMM systems. An Expectation-Maximization (EM) approach for parameter estimation is presented. We demonstrate a 13% relative WER reduction on the Aurora-4 clean evaluation set, and a 13% relative WER reduction on the babble noise condition.

***Keywords*—** Linear dynamic models, nonlinear statistical modeling, speech recognition, acoustic modeling

Manuscript submitted February 1, 2013. Revised May 20, 2013.

1. This material is based upon work supported by the National Science Foundation under Grant No. IIS-0414450. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
2. T. Ma is with Siri at Apple Inc, 2 Infinite Loop, mailstop 302-4APP, Cupertino, California 95014, USA (phone: 408-643-5909; email: tma@apple.com).
3. S. Srinivasan is with Nuance Communications Inc., 1198 East Arques Avenue Sunnyvale, CA 94085, USA (phone: 408-992-6243; email: sundararajan.srinivasan@gmail.com).
4. G. Lazarou is with The New York City Transit Authority, 30-74 38th Street, Apt 1A, Astoria, New York, New York, USA 11103 (phone: (662) 617-2064; email: glaz@ieee.org).
5. J. Picone is with the Department of Electrical and Computer Engineering at Temple University, 1947 North 12th Street, Philadelphia, Pennsylvania 19027 USA (phone: 215-204-4841; fax: 215-204-5960; email: joseph.picone@isip.piconepress.com).

I. INTRODUCTION

Over the past several decades, Hidden Markov Models (HMMs) that use Gaussian mixture models to model state observation distributions have been the most popular approach for acoustic modeling in automatic speech recognition (ASR) applications. We will refer to these as HMM/GMM. An HMM/GMM can be regarded as a finite state machine in which the states of the system evolve in accordance with an inherent deterministic mechanism and the emission probabilities map hidden states to observations. HMM modeling techniques have relied on a standard assumption that speech features are temporally uncorrelated. Recent theoretical and experimental studies (Frankel, 2007; Frankel & King, 2003; Digalakis et al., 1993) suggest that exploiting frame-to-frame correlations in a speech signal further improves the performance of ASR systems. This is typically accomplished by developing an acoustic model that includes higher order statistics or parameter trajectories (Liang, 2003).

Linear Dynamic Models (LDMs) have generated significant interest in recent years (Frankel, 2007; Tsontzos et al., 2007) due to their ability to model higher order statistics. An LDM describes a linear dynamic system as underlying states and observables using a measurement equation to link the internal states to the observables. An autoregressive model is used to capture the time evolution of states. An LDM models every word or phoneme segment as a non-separable unit that incorporates the dynamic evolution of the hidden states. Digalakis et al. (1993) first applied LDMs to the speech recognition problem by developing a maximum likelihood approach based on an Expectation-Maximization (EM) parameter estimation algorithm. In subsequent work by Frankel and King (2003), LDMs were applied to an acoustic modeling problem to characterize articulatory dynamics. Promising results have been demonstrated on a limited recognition task based on TIMIT (Garofolo et al., 1993). More recently, LDMs have been applied to noisy speech recognition problems using Aurora-2 (Wollmer et al., 2011), but not in a manner conducive to large vocabulary continuous speech recognition (LVCSR).

In this paper, we present a hybrid framework that integrates LDM into an LVCSR system, and demonstrate a significant improvement on a difficult evaluation task: the Aurora-4 Corpus (Parihar et al., 2004). This task includes clean and noisy speech data as well as conditions simulating mismatched training conditions. We show that the proposed hybrid recognizer provides 13% relative WER reduction on the Aurora-4 clean evaluation set, and a 13% relative WER reduction on a babble noise condition. In Section II, we present a brief review of LDM. In Section III, we describe a hybrid HMM/LDM recognizer architecture that effectively integrates these two technologies. Continuous speech recognition results on the Aurora-4 corpus are presented in Section IV. The paper concludes with a discussion of ongoing research on directly integrating LDM into HMMs system at the frame-level of speech signals.

II. LINEAR DYNAMIC MODELS

An LDM is an example of a Markovian state-space model, and in some sense, can be regarded as analogous to an HMM/GMM since LDMs use hidden-state modeling. In an LDM, systems are described as underlying states and observables combined by a measurement equation (Digalakis et al., 1993). Every observable has a corresponding hidden internal state as illustrated in Figure 1. The LDM formulation is based on a state-space model:

$$x_{t+1} = Fx_t + \omega_t \tag{1}$$

$$y_t = Hx_t + \nu_t \ , \tag{2}$$

where x_t is a q-dimensional internal state vector, y_t is a p-dimensional observation vector, F is the state evolution matrix and H is the observation transformation matrix. The variables ω_t and ν_t are assumed to be uncorrelated white Gaussian noise with covariance matrices Q and R, respectively.

The sequence of observations, y_t, and underlying states, x_t, are finite dimensional and are assumed to follow multivariate Gaussian distributions for every time t. The first equation can be viewed as an autoregressive state process that describes how states evolve from one time frame to the next. The second equation maps the output observations to the internal states. The system's hidden states, x_t, are the deterministic characteristic of an LDM that are also affected by random Gaussian noise. The state and noise variables can be combined into one single Gaussian random variable (Frankel & King, 2007).

Based on Figure 1, conditional density functions for the states and output can be written as follows:

$$P(y_t \mid x_t) = \exp\{-(1/2)[y_t - Hx_t]^T R^{-1}[y_t - Hx_t]\}(2\pi)^{-p/2} \mid R \mid^{-1/2} \tag{3}$$

$$P(x_t \mid x_{t-1}) = \exp\{-(1/2)[x_t - Fx_{t-1}]^T Q^{-1}[x_t - Fx_{t-1}]\}(2\pi)^{-k/2} \mid Q \mid^{-1/2} . \quad (4)$$

According to the Markovian assumption, the joint probability density function of the states and observations becomes:

$$P(\{x\},\{y\}) = P(x_1)\prod_{t=2}^{T} P(x_t \mid x_{t-1}) \prod_{t=1}^{T} P(y_t \mid x_t) . \quad (5)$$

We need to estimate the hidden state evolution given y_t and the model parameters. This can be accomplished using a Kalman filter combined with a Rauch Tung Striebel (RTS) smoother (Frankel & King, 2007). The Kalman filter provides an estimate of the state distribution at time *t* given the previous observations. The RTS smoother gives a corresponding estimate of the underlying state conditions over the entire observation sequence. For the smoothing part, a fixed interval RTS smoother is used to compute the required statistics once all data has been observed.

The RTS smoother adds a backward pass that follows the standard Kalman filter forward recursion. In addition, in both the forward and the backward pass, we need some additional recursions for the computation of the cross-covariance. The corresponding RTS equations are:

$$\hat{x}_{t-1/N} = \hat{x}_{t-1/t-1} + A_t(\hat{x}_{t/N} - \hat{x}_{t/t-1}) \quad (6)$$

$$\sum\nolimits_{t-1/N} = \sum\nolimits_{t-1/t-1} + A_t(\sum\nolimits_{t/N} - \sum\nolimits_{t/t-1})A_t^T \quad (7)$$

$$A_t = \sum\nolimits_{t-1/t-1} F^T \sum\nolimits_{t/t-1}^{-1} \quad (8)$$

$$\sum\nolimits_{t,t-1/N} = \sum\nolimits_{t,t-1/t} + (\sum\nolimits_{t/N} - \sum\nolimits_{t/t})\sum\nolimits_{t/t}^{-1}\sum\nolimits_{t,t-1/t} \quad . \quad (9)$$

A synthetic LDM model with two-dimensional states and one-dimensional observations was created to demonstrate the contribution of RTS smoothing. In Figure 2, we show the state predictions of this LDM model using a traditional Kalman filter. In Figure 3, the performance of the Kalman filter with RTS smoothing is shown. In both figures, the true state evolutions for our synthetic LDM model are compared to a scatter plot of the noisy observations of the LDM model and the RTS smoothed data. RTS smoothing produces significantly better prediction for the system's internal states.

The Expectation-Maximization (EM) algorithm (Digalakis et al., 1993) is used to find maximum likelihood estimates of parameters for a specific word or phone, where the model depends on unobserved latent variables. The relevant equations are:

$$E\left[x_t / y, \theta^{(i)}\right] = \hat{x}_{t/N} \tag{10}$$

$$E\left[x_t x_t^T / y, \theta^{(i)} g\right] = \sum{}_{t/N} + \hat{x}_{t/N} \hat{x}_{t/N}^T \tag{11}$$

$$E\left[x_t x_{t-1}^T / y, \theta^{(i)}\right] = \sum{}_{t,t-1/N} + \hat{x}_{t/N} \hat{x}_{t-1/N}^T \quad . \tag{12}$$

The E-step algorithm consists of computing the conditional expectations of the complete-data sufficient statistics for standard ML parameter estimation. Therefore, the E-step involves computing the expectations conditioned on observations and model parameters. The RTS smoother described previously can be used to compute the complete-data estimates of the state statistics. EM for LDM then consists of evaluating the ML parameter estimates by replacing x_t and $x_t x_t^T$ with their expectations.

The EM algorithm converges quickly and is stable for our synthetic LDM model of two-dimensional states and one-dimensional observations. After initilizing this LDM model with an identity state transition matrix and random observation matrix, the first iteration of ML parameter estimation was applied to update the model parameters. Log-likelihood scores of observation vectors were calculated and saved in order to perform further analysis.

EM training was applied for 30 iterations. After the training recursion, intermediate log-likelihood scores of observation vectors for all iterations of LDM were plotted as a function of the number of iterations. This plot is referred to as the EM evolution curve. We explored 1-, 4-, 6-, and 10-dimensions for each state in the LDM approach, and applied EM training for each specified dimension. In Figure 4, the EM evolution curve is shown as a function of the state dimension. The training procedure converges quickly, requiring no more than 10 iterations.

III. A HYBRID HMM/LDM ARCHITECTURE

One significant drawback of LDMs is that, they are inherently static classifiers — they are not capable of implicitly modeling the temporal evolution of a speech signal. Static classifiers are not designed to find the optimal start and stop times for a phone hypothesis. HMMs, on the other hand, are very good at optimizing segmentations while performing classification. Based on our previous work integrating a Support Vector Machine into a speech recognition system (Ganapathiraju et al., 2004), we employed a similar two-pass hybrid HMM/LDM recognizer. This system, shown in Figure 5, leverages the temporal modeling and *N*-best list generation capabilities of the traditional HMM architecture in a first-pass analysis, and uses a second pass to re-rank candidate sentence hypotheses with a phone-based LDM model. A more thorough analysis of alternate strategies for integrating LDMs into an HMM framework was explored in Ma (2010). The hybrid system *N*-best rescoring approach was found to be the most promising.

Since the hybrid architecture postprocesses *N*-best lists, high performance *N*-best list generation is critical to achieving good performance. In our research, a word graph is generated and converted to an *N*-best list using a stack-based word graph to *N*-best list converter. Word lattices or word graphs are a condensed representation of the search space. Word graphs are an intermediate representation commonly used in a multi-pass speech recognition system. Typically, a word graph contains word labels, start and stop times, a language model score and an acoustic score. To convert the word graph to an *N*-best list, a stack is initialized with the start node of the graph. A recursive procedure is then used to grow partial paths according to the word graph and to re-rank the stack to find the best partial path. During this procedure, beam pruning is applied to maintain the *K*-best partial paths in the stack. Upon completion, the *N*-best partial paths ($N < K$) are traced to produce the final *N*-best sentence hypotheses.

Once this list is produced, along with the corresponding segmentations for the acoustic units, LDM classifiers are used in a second pass to estimate the likelihood scores. In this work, a transformation-based score combination scheme is applied for simplicity. The LDM likelihood scores are first normalized (transformed) to match the range of the HMM scores, and then a weighted combination of these two scores is used:

$$Likelihood = HMM\_Score + LDM\_Scale * LDM\_Score \tag{13}$$

Ma (2010) explored methods of combining these two scores and determined that a weighted sum of the two scores provided a small gain in performance over using only the LDM score in the rescoring process. Choice of the normalization scheme and combination weight is data-dependent and requires empirical evaluation. Alternate approaches such as classifier-based score fusion and density-based score fusion could be used, but our experience was that the overall results are not sensitive to the type of score fusion used.

IV. AURORA-4 EXPERIMENTS

In order to evaluate the hybrid HMM/LDM recognizer, the Aurora-4 Corpus (Parihar et al., 2004) was chosen because it contains mismatched training and evaluation conditions, which is a fundamental problem addressed in this work. The Aurora-4 Corpus consists of the original WSJ0 data with digitally-added noise and is divided into two training sets and 14 evaluation sets. Training Set 1 (TS1) and Training Set 2 (TS2) include the complete WSJ0 training set known as SI-84. TS1 consists of the original WSJ recordings, while TS2 contains various digitally-added noise conditions. The 14 evaluation sets are derived from data defined by the November 1992 NIST evaluation set. Each evaluation set consists of a different microphone or noise combination. In this work, we use only TS1 dataset for training and use the 14 evaluation sets for performance analysis.

Traditional 39 dimensional MFCC acoustic features (12 cepstral coefficients, absolute energy, and first and second order derivatives) were computed from each of the signal frames within the phoneme segments. Before extraction, each feature dimension was normalized to the range [-1,1] to improve the convergence behvaior of our LDM training. A total of 40 phonemes are used for acoustic modeling, so there are 40 LDM classifiers in the hybrid decoder.

The scale factor for combining HMM and LDM scores is shown in Table 1. We see that performance varies slightly with changes in LDM_Scale. It is not surprising that the variation is small, since the HMM-based segmentations play an important role in the overall combination of the two scores. Segmentation is an important part of the recognition process, and generally if segmentation is correct, recognition performance is high. The *N*-best rescoring process is intimately dependent on the HMM process. A better alternative would be to embed LDM in the first pass of the recognition system, but that is the subject of future research.

The evaluation results for the clean dataset and six noisy evaluation sets are presented in Table 2. The results for the hybrid HMM/LDM decoder for the condition labeled "Clean," which represents matched training and testing in a noise-free environment, are encouraging. The hybrid HMM/LDM system achieves an 11.6% WER which represents a 12.8% relative WER reduction compared to a comparably configured HMM baseline. The hybrid decoder also achieves 13.2% relative WER reduction for the babble noise evaluation dataset, and smaller improvements for a majority of the other conditions, which represent mismatched training and evaluation conditions. The overall results are promising given that the segmentations have not been optimized for the LDM system and confirms LDM's capability to model speech dynamics in a manner that is complementary to a traditional HMM.

V. SUMMARY

In this paper, we proposed a hybrid framework to integrate LDMs within the framework of an HMM for large vocabulary continuous speech recognition tasks. The theoretical foundation of the linear dynamic model is discussed and an EM-based training paradigm is introduced. The hybrid decoder architecture is an off-line processing mechanism and is bootstrapped using a baseline HMM system. Several issues related to applying an LDM in a hybrid system have been addressed: modifications to the HMM system; implementation of the *N*-best list generation; and development of an *N*-best rescoring paradigm using HMM and LDM score fusion. Results on the Aurora-4 Corpus are encouraging.

In this work, the LDM postprocesses segmentations derived from the first pass of an HMM-based recognition. It is well known that segmentation plays a major role in high performance speech recognition systems. Future work will be focused on closely integrating the LDM into the core search loop of a speech recognizer, providing acoustic scores at the frame level that can be directly integrated into the Viterbi search, alleviating the need to do *N*-best rescoring. This would allow a deeper analysis of the utterance and improve performance beyond that achievable with *N*-best rescoring and fixed segmentations.

VI. REFERENCES

Digalakis, V., Rohlicek, J., & Ostendorf, M. (1993). ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. IEEE Transactions on Speech and Audio Processing, 1(4), 431–442.

Frankel, J., & King, S. (2007). Speech Recognition Using Linear Dynamic Models. *IEEE Transactions on Speech and Audio Processing*, 15(1), 246–256.

Frankel, J. (2003). Linear Dynamic Models for Automatic Speech Recognition. University of Edinburgh. Retrieved from http://homepages.inf.ed.ac.uk/joe/pubs/2003/Frankel_thesis2003.pdf.

Ganapathiraju, A., Hamaker, J., & Picone, J. (2004). Applications of Support Vector Machines to Speech Recognition. *IEEE Transactions on Signal Processing*, 52(8), 2348–2355.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus. *The Linguistic Data Consortium Catalog*. Philadelphia, Pennsylvania, USA: The Linguistic Data Consortium. ISBN:1-58563-019-5.

Liang, F. (2003). An Effective Bayesian Neural Network Classifier with a Comparison Study to Support Vector Machine. *Neural Computation*, 15(8), 1959–1989.

Ma, T. (2010). *Linear dynamic model for continuous speech recognition*. Mississippi State University.

Parihar, N., Picone, J., Pearce, D., & Hirsch, H.-G. (2004). Performance Analysis of the Aurora Large Vocabulary Baseline System. Proceedings of the European Signal Processing Conference (pp. 553–556). Vienna, Austria.

Tsontzos, G., Diakoloukas, V., Koniaris, C., & Digalakis, V. (2007). Estimation of General Identifiable Linear Dynamic Models with an Application in Speech Recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 4, pp. IV–453–IV–456).

Wöllmer, M., Klebert, N., & Schuller, B. (2011). Switching Linear Dynamic Models for Recognition of Emotionally Colored and Noisy Speech. *Sprachkommunikation 2010* (ITG-FB 225) (pp. 1–4). Bochum, Deutschland: Springer.

VII. LIST OF FIGURES

VIII. FIGURES

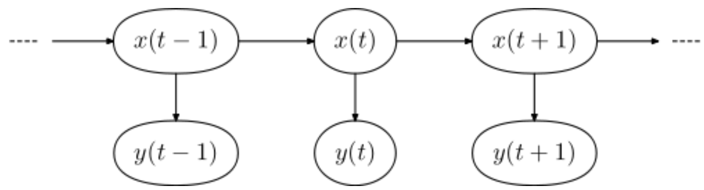

Figure 1. The internal states and observations are shown for an LDM. Every observable has a corresponding hidden internal state.

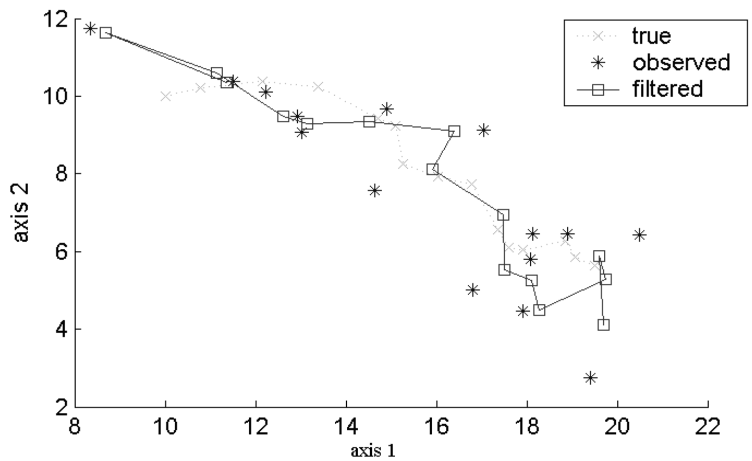

Figure 2. State predictions for an LDM model using a Kalman filter are shown.

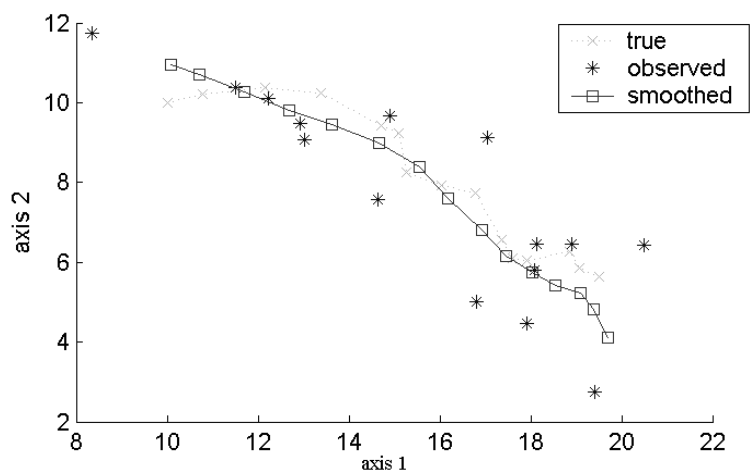

Figure 3. A Kalman filter with RTS smoothing produces smoother state trajectories.

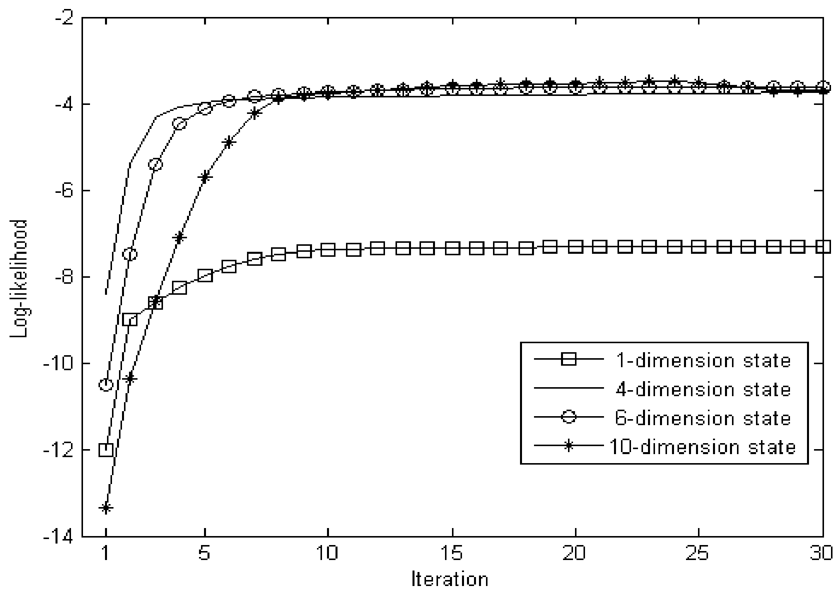

Figure 4. The EM evolution as a function of iteration is shown for a variety of state dimensions. EM training procedure converges quickly, requiring no more than 10 iterations.

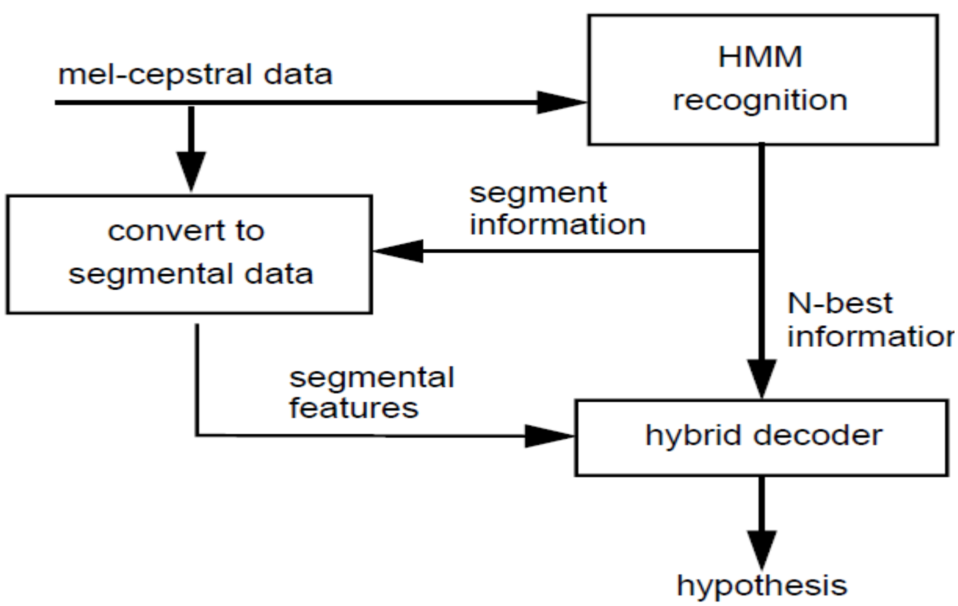

Figure 5. A hybrid HMM/LDM architecture is shown in which LDM is used to postprocess phone hypotheses using HMM segmentations.

IX. LIST OF TABLES

X. TABLES

Normalization Factor: LDM_SCALE	Hybrid Decoder WER
0.100	12.3%
0.050	12.1%
0.010	**11.8%**
0.005	11.9%
0.001	11.9%

Table 1. Experimental tuning of the scale factor combining LDM and HMM scores resulted in small improvements in performance.

Condition	HMM Baseline	Hybrid LDM	Relative Reduction
Clean	13.3	11.6	**12.8%**
Airport	53.0	50.3	5.09%
Babble	55.9	48.5	**13.2%**
Car	57.3	59.8	-4.4%
Restaurant	53.4	50.6	5.2%
Street	61.5	59.4	3.4%
Train	66.1	63.4	4.1%

Table 2. Experimental results for the hybrid HMM/LDM system are compared to a conventional HMM system. Substantial improvements were obtained on the clean and babble noise conditions.

I. LIST OF FIGURES

II. FIGURES

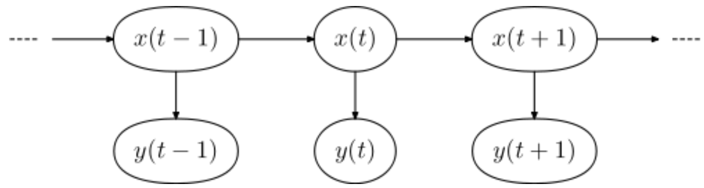

Figure 1. The internal states and observations are shown for an LDM. Every observable has a corresponding hidden internal state.

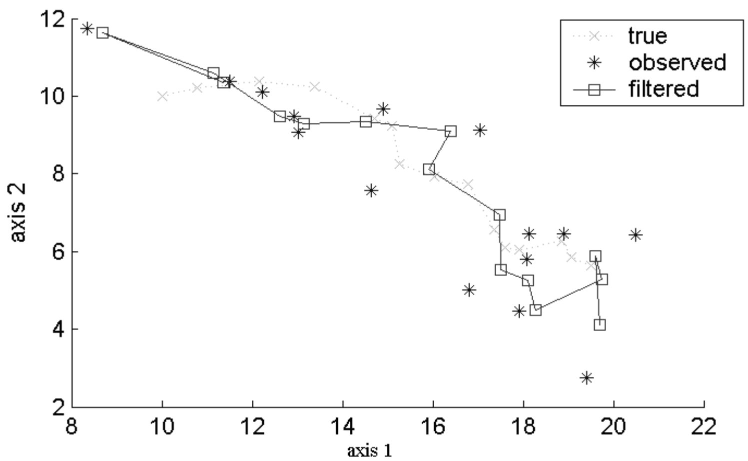

Figure 2. State predictions for an LDM model using a Kalman filter are shown.

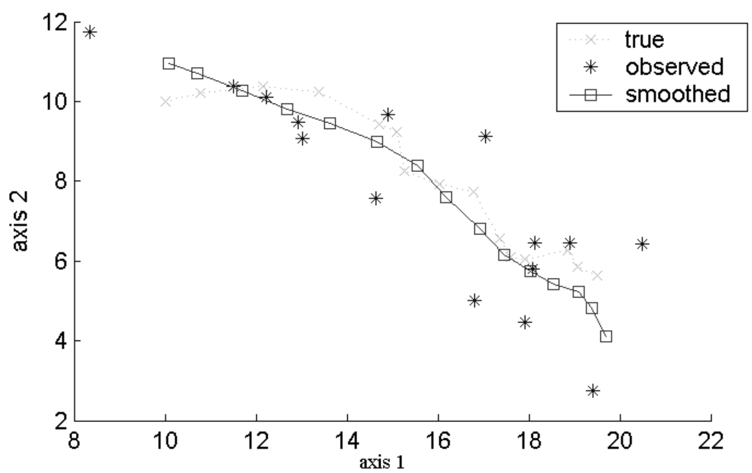

Figure 3. A Kalman filter with RTS smoothing produces smoother state trajectories.

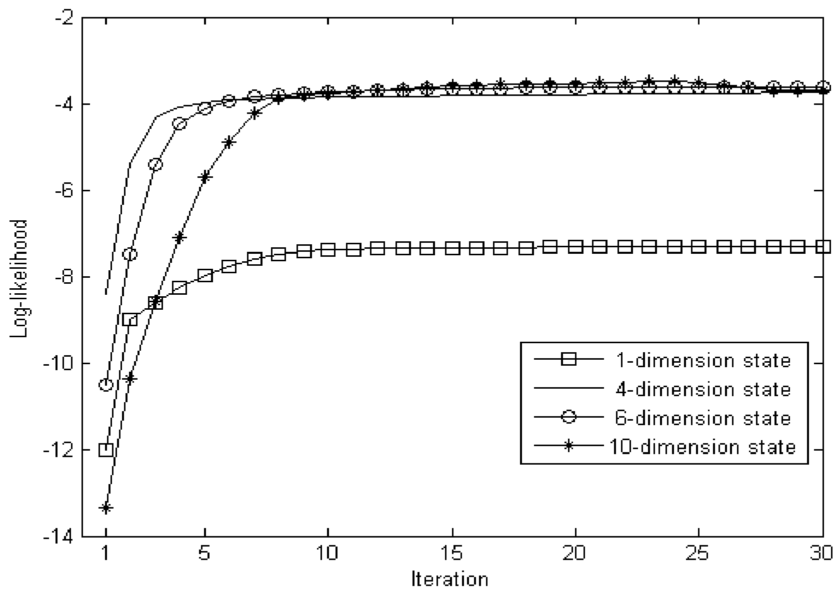

Figure 4. The EM evolution as a function of iteration is shown for a variety of state dimensions. EM training procedure converges quickly, requiring no more than 10 iterations.

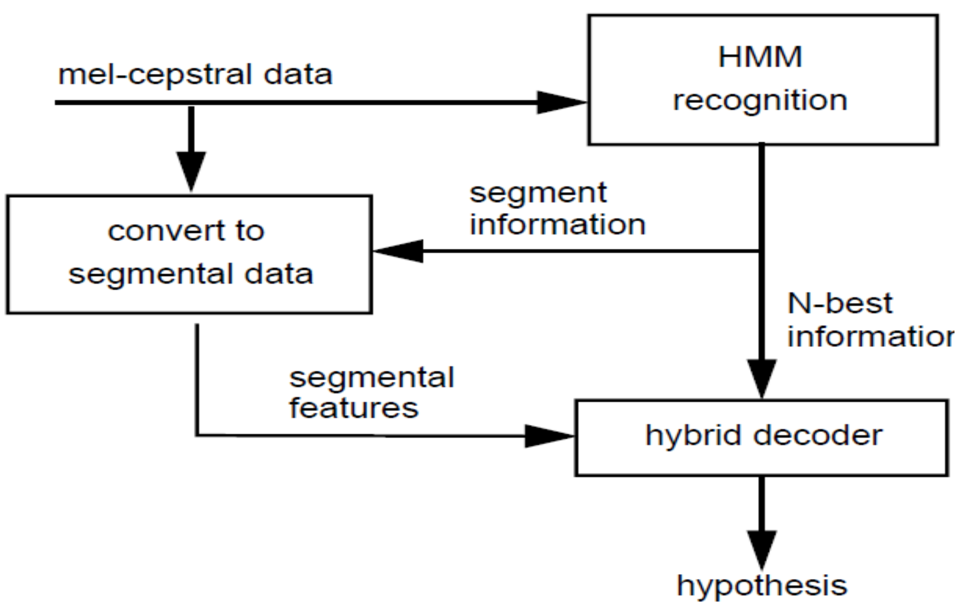

Figure 5. A hybrid HMM/LDM architecture is shown in which LDM is used to postprocess phone hypotheses using HMM segmentations.

I. LIST OF TABLES

Table 1. Experimental tuning of the scale factor combining LDM and HMM scores resulted in small improvements in performance.

Table 2. Experimental results for the hybrid HMM/LDM system are compared to a conventional HMM system. Substantial improvements were obtained on the clean and babble noise conditions.

II. TABLES

Normalization Factor: LDM_SCALE	**Hybrid Decoder WER**
0.100	12.3%
0.050	12.1%
0.010	**11.8%**
0.005	11.9%
0.001	11.9%

Table 1. Experimental tuning of the scale factor combining LDM and HMM scores resulted in small improvements in performance.

Condition	**HMM Baseline**	**Hybrid LDM**	**Relative Reduction**
Clean	13.3	11.6	**12.8%**
Airport	53.0	50.3	5.09%
Babble	55.9	48.5	**13.2%**
Car	57.3	59.8	-4.4%
Restaurant	53.4	50.6	5.2%
Street	61.5	59.4	3.4%
Train	66.1	63.4	4.1%

Table 2. Experimental results for the hybrid HMM/LDM system are compared to a conventional HMM system. Substantial improvements were obtained on the clean and babble noise conditions.