Efficient Decoding Strategy for Conversational Speech Recognition Using State-Space Models for Vocal-Tract-Resonance Dynamics

Jeff Z. Ma and Li Deng

Department of Electrical and Computer Engineering
University of Waterloo, Canada
Current addresses: BBN Technologies, Cambridge MA (jma@bbn.com)
Microsoft Research, Redmond WA (deng@microsoft.com)

Abstract

In this paper, we present an efficient strategy for likelihood computation and decoding in a continuous speech recognizer using underlying state-space dynamic models for the hidden speech dynamics. The state-space models have been constructed in a special way so as to be suitable for the conversational or casual style of speech where phonetic reduction abounds. The interacting multiple model (IMM) state estimation algorithm for switching state-space models is first introduced, which uses a merging strategy derived from Bayes's rule to meet the challenge of exponential growth in the switching combination. Then one specific dynamic-programming based decoding algorithm, incorporating the merging strategy, are derived, implemented, and evaluated. It successfully overcomes the exponential growth in the original search paths by using the path-merging strategy. Evaluation experiments on conversational speech using the Switchboard corpus demonstrate that the use of the new decoding strategy is capable of reducing the recognizer's word error rate compared with the baseline recognizers, including the HMM system and the state-space dynamic model using the HMM-produced phonetic boundaries, under identical test conditions.

1. Introduction

We have in recent years developed various versions of statistical coarticulatory dynamic models for spontaneous speech recognition [2]. Rather than using a large number of unstructured Gaussian mixture components to account for the tremendous variation in the observable acoustic data of highly coarticulated spontaneous speech, the new model provides a rich structure for the partially observable (hidden) dynamics in the domain of vocal-tract-resonances (VTRs). In the design of the speech recognizer, we used the target-directed state-space model to describe the physical process of spontaneous speech production where knowledge of the VTR dynamic behavior in speech production is naturally incorporated into the model training and decoding.

One major challenge for our dynamic model applied to speech recognition is the decoding difficulty, as is for other types of stochastic segmental models (SSM) [8]. In [5] we analyzed in detail the special decoding difficulty for our dynamic models and developed a path-stack decoding algorithm therein based on a path deletion strategy. In this paper we report our new progress along this direction. Herein we have developed a new decoding algorithms, which is based on path merging strategies.

The work in this paper has been partly motivated by some earlier work on the various versions of the switching state-space model developed in control engineering, time series analysis, and in econometrics [3, 4, 9, 10]. In the earlier work such switching models were used to deal with several simplified cases for applications of multiple targets tracking [10] and prediction of recession and booming in economy [4], etc.

Moving beyond the original applications in target tracking and in econometrics, the research in this paper represents our novel contribution of applying the specially constructed switching state-space model to functional modeling of speech production and to speech recognition. The speech production process can be well fitted into the switching state-space model since each phone in a finite number of phones in a language can be associated with a largely distinct target vocal-tract shape and its related acoustic resonance structure. When a speech utterance is produced, the vocal tract shape or resonance (continuous state in the model) changes relatively smoothly from one target phone to another, where the target shapes determined by the model parameters, are modeled to be switching from one target phone to its temporally adjacent one.

A further contribution of this research is to extend the solutions to the state estimation problem to those of the state *sequence* estimation (i.e., decoding) problem. All the earlier work on the switching model focused on the state estimation problem where the source of the difficulty is exponential growth in the number of switching combinations over time. The earlier solutions to the state estimation problem have been modified and extended to the decoding problem necessary for the use of our new speech model for speech recognition.

This paper is organized as follows. In Section 2, the switching state-space models for the dynamics of VTRs are reviewed in order to set up the context in which the formulation of the decoding strategy is established as the problem of optimal state sequence estimation. The state estimation problem and one approximate solution for it are contained in Section 3. Armed with the approximate solution to the state estimation problem, a one-pass dynamic programming based decoding algorithm is developed and presented in Section 4. This algorithm aims at finding the globally optimal path under the constraints imposed on the switching process of the model. Details of the evaluation experiments using an N-best re-scoring paradigm are reported in Section 5. Finally, we draw conclusions in Section 6.

2. A switching state-space model for speech dynamics

The constrained, nonlinear, and switching state-space model (details in [2]) for the target-directed dynamics of VTRs is re-

viewed in this section, for the purpose of setting up the context where the Bayesian formulation of the decoding strategy will be established in later sections. The mathematical structure of the model is described by the following state-space system:

$$Z(k) = \Phi^{(j)} Z(k-1) + (I - \Phi^{(j)}) T^{(j)} + W^{(j)}(k) \quad (1)$$

$$O(k) = h(Z(k)) + V^{(j)}(k). \quad (2)$$

where the state variable $Z(k)$ represents vector-valued VTR, $T^{(j)}$ and $\Phi^{(j)}$ are the vector-valued VTR target and "time constant", respectively. $O(k)$ are the acoustic observation (MFCC). $h(Z)$ describes the noisy nonlinear relationship between the VTR space and the acoustic space, which is implemented by a global multi-layer perceptron (MLP) herein. $W^{(j)}(k)$ and $V^{(j)}(k)$ are zero-mean Gaussian i.i.d. noises with variances $Q^{(j)}$ and $R^{(j)}$ respectively. j is the index of the switching dynamic regime which corresponds to a set of unique model parameters. In this work each dynamic regime occupies a segment of speech roughly of the size of a surface form of a phone, so j is also the index of phone.

We have developed a new version for the above state-space dynamic model, a mixture linear dynamic model (MLDM) [6]. In the evaluation experiment section we will also be giving the results of the decoding algorithm applied to this new version of the model.

In short, the model parameter set consists of $\Theta = \{\Phi, T, R, Q\}$ plus the parameters contained in the nonlinear function $h(\cdot)$ (e.g. MLP weights). As the speech utterance traverses from left to right in time, phone-sized dynamic regimes switch from one to another, which induces the switching process among M parameter sets $\Theta^{(j)} = \{\Phi^{(j)}, T^{(j)}, R^{(j)}, Q^{(j)}\}$, where $j = 1, 2, ..., M$ and M is the total number of phones.

Given a sequence of observations $O_1^K = \{O(1), O(2), \cdots, O(K)\}$, and assuming that this entire sequence is generated by model j (i.e., no model switching occurs during the generation of the observation sequence), then the log-likelihood of the model in Eqns.(1) and (2) can be computed according to [11]

$$L(O_1^K|\Theta^{(j)}) = \sum_{k=1}^{K} \log p(O(k)|O_1^{k-1}, \Theta^{(j)})$$

$$= -\frac{1}{2} \sum_{k=1}^{K} \{|\Sigma_{\tilde{O}_k}^{(j)}| + [\tilde{O}_k^{(j)}]' [\Sigma_{\tilde{O}_k}^{(j)}]^{-1} \tilde{O}_k^{(j)}\} + cnst. \quad (3)$$

where we assume $p(O(1)) = p(O(1)|O(0))$. $\tilde{O}_k^{(j)}$ is the innovation at time k and $\Sigma_{\tilde{O}_k^{(j)}}$ its variance, they are calculated by using Kalman filter [7, 11]. The Kalman filter converts the likelihood of the entire observation sequence (which is temporally correlated) into the summation of likelihoods at local time points (uncorrelated innovations). However, the Kalman filter at the current time step depends on the previous time step. This makes the local likelihood computation depend on the entire past history.

If the model switches among M different modes (or phones) during the generation of the observation sequence (j is not fixed any more, which turns into a decoding problem), the switching combinations grow exponentially. Different switching histories bring different local scores. Therefore, no path deletion is theoretically allowed during the decoding (see details in [5]), which makes the search for the new model difficult. In the next sections, we first review how the earlier work on state estimation overcomes the exponential growth problem for the switching models. We then exploit and extend these results to solve the decoding problem which is formulated as optimal state-sequence estimation.

3. State estimation for switching models

For conventional state-space models with *fixed* parameters, state estimation (filtering) is to calculate the conditional mean and covariance of the hidden state $Z(k)$ given the observations up to time k. Generalizing from the fixed-parameter case to the case with switching parameters like the model given in Eqns.(1) and (2), the mean and covariance of $Z(k)$ will be conditioned not only on the observations up to time k but also on the evolution history of the model switching. Let's use $\hat{Z}_{k|k}$ to denote the conditional mean, and $\Sigma_{k|k}$ the conditional covariance, then

$$\hat{Z}_{k|k} = \sum_{i \in \Psi_k} P_k(i) \hat{Z}_{i,k|k} \quad (4)$$

$$\Sigma_{k|k} = \sum_{i \in \Psi_k} P_k(i) \{\Sigma_{i,k|k} + [\hat{Z}_{i,k|k} - \hat{Z}_{k|k}][\hat{Z}_{i,k|k} - \hat{Z}_{k|k}]'\} \quad (5)$$

where Ψ_k represents all possible switching evolution histories up to time k, $P_k(i)$ is the probability of the i-th switching history being true, and $\hat{Z}_{i,k|k}$ and $\Sigma_{i,k|k}$ is the state estimate given the i-th switching history being true, which is

$$\hat{Z}_{i,k|k} = E[Z(k)|i_k, i_{k-1}, \cdots, i_1, O_1^k] \quad (6)$$

$$\Sigma_{i,k|k} = Cov[Z(k)|i_k, i_{k-1}, \cdots, i_1, O_1^k] \quad (7)$$

where we use $\{i_k, i_{k-1}, \cdots, i_1\}$ ($1 \le i_k \le M$) to represent the i-th switching history.

We refer to Eqns.(6) and (7) as the elemental estimator in the overall state estimation procedure.

At each time point, any one of the M models could be chosen, and such as, there are potentially as many as M^k paths for the switching evolution up to time k. Each of these paths forms an elemental estimator based on the conventional state estimation, and hence there are a prohibitively large number M^k of the elemental estimators at time k. How to obtain the overall state estimate from the huge number of elemental estimators in a computationally efficient manner? Earlier work on state estimation has provided several efficient approximation solutions. In this paper we review one of them, the interacting multiple model approach (IMM), and show how to extend it to our decoding problem.

In the IMM approach [9], the state estimate is carried out under each current model at each time step k. That is, the switching history $\{i_k, i_{k-1}, \cdots, i_1\}$ is approximated by $\{i_k\}$. Therefore, the elemental estimator in Eqns.(6) and (7) is approximated by

$$E[Z(k)|i_k, O_1^k] \ and \ Cov[Z(k)|i_k, O_1^k] \quad (8)$$

For simplicity purposes, we will use j and i to represent i_k and i_{k-1} respectively, and use $\hat{Z}_{k|k}^{(j)}$ and $\Sigma_{k|k}^{(j)}$ to represent $E[Z(k)|j, O_1^k]$ and $Cov[Z(k)|j, O_1^k]$ respectively.

By this approximation, the state estimation in Eqns.(4) and

(5) becomes

$$\hat{Z}_{k|k} = \sum_{j=1}^{M} P(j|O_1^k)\ \hat{Z}_{k|k}^{(j)} \quad (9)$$

$$\Sigma_{k|k} = \sum_{j=1}^{M} P(j|O_1^k)\ \{\Sigma_{k|k}^{(j)} + [\hat{Z}_{k|k}^{(j)} - \hat{Z}_{k|k}][\hat{Z}_{k|k}^{(j)} - \hat{Z}_{k|k}]^{'}\} \quad (10)$$

where $P(j|O_1^k)$ is called model probability, $\hat{Z}_{k|k}^{(j)}$ and $\Sigma_{k|k}^{(j)}$ is the elemental estimator.

The elemental estimator, $\hat{Z}_{k|k}^{(j)}$ and $\Sigma_{k|k}^{(j)}$, is implemented by the Kalman filter [7, 11]. Its initial value is obtained by merging according to:

$$\bar{Z}_{k-1|k-1}^{(j)} = \sum_{i=1}^{M} P(i|j, O_1^{k-1})\hat{Z}_{k-1|k-1}^{(i)} \quad (11)$$

$$\bar{\Sigma}_{k-1|k-1}^{(j)} = \sum_{i=1}^{M} P(i|j, O_1^{k-1})\ \{\Sigma_{k-1|k-1}^{(i)} + (\hat{Z}_{k-1|k-1}^{(i)} - \bar{Z}_{k-1|k-1}^{(j)}) \cdot (\hat{Z}_{k-1|k-1}^{(i)} - \bar{Z}_{k-1|k-1}^{(j)})^{'}\} \quad (12)$$

where $P(i|j, O_1^{k-1})$ is called the mixing probability, and is computed recursively from Bayes' rule:

$$P(i|j, O_1^{k-1}) = \frac{P(j|i, O_1^{k-1})P(i|O_1^{k-1})}{\sum_{i=1}^{M} P(j|i, O_1^{k-1})P(i|O_1^{k-1})} \quad (13)$$

The model probability $P(j|O_1^{k-1})$ in Eqns.(11), (12) and (13) is updated according to the following two equations,

$$P(j|O_1^k) = \frac{\sum_{i=1}^{M} p(O(k), j, i|O_1^{k-1})}{\sum_{j=1}^{M}\sum_{i=1}^{M} p(O(k), j, i|O_1^{k-1})} \quad (14)$$

$$p(O(k), j, i|O_1^{k-1}) = p(O(k)|j, i, O_1^{k-1})P(j|i, O_1^{k-1})P(i|O_1^{k-1}) \quad (15)$$

where due to the merging $p(O(k)|j, i, O_1^{k-1}) = p(O(k)|j, O_1^{k-1})$, which is calculated from the elemental estimator (EKF) at model j, $P(j|i, O_1^{k-1})$ is transition probability from model i to j (known), and $P(i|O_1^{k-1})$ is the model probability at the previous time.

A graphical illustration of the IMM merging strategy is given in Fig.(1) (for convenience, (Z, V) is used to represent $(\hat{Z}, \Sigma)$). The elemental estimates at previous time $k-1$ are merged first to obtain initial values for the elemental estimators at the current time k. Then, the EKF is carried out to obtain the elemental estimates, $\hat{Z}_{k|k}^{(j)}$ and $\Sigma_{k|k}^{(j)}$ $(1 \le j \le M)$, at the current time. Finally, these elemental estimates are summed with weights, according to Eqn.(9) and (10), to obtain the overall state estimate. Note that the elemental estimates at the current time k is used to initialize the elemental estimators at the next time $k+1$. At the same time, the mixing probability and model probability are updated.

4. Decoding strategy using the IMM merging method

In this section, we incorporate the IMM method into the decoding strategy for our new speech model, aiming to search (time-synchronous) for the *global* optimal path through all discrete states (i.e., the entire parameter switching history). For the IMM,

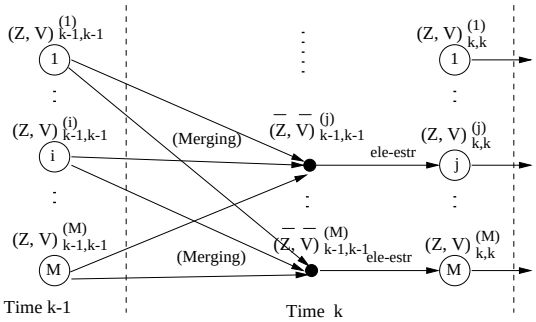

Figure 1: *The IMM merging strategy for state estimation*

at each time step the posterior probability, $P(j|O_1^k)$ (Eqn.(14)) , can be computed for each phone (j) . Therefore, at each time step, if we were to "decode" the discrete state based on the highest posterior probability ($argmax_{1 \le j \le M}\ P(j|O_1^k)$), we would have a global "maximum a posterior" path. However, the posterior probability $P(j|O_1^k)$ is conditioned only on the observations up to time k rather than on the entire sequence O_1^K. Hence, the "maximum a posteriori" path is not consistent with the usual decoding criterion of maximizing the joint likelihood of observation sequence and path: $max_{\psi_K \in \Psi_K}\ l(O_1^K, \psi_K)$ ($\propto l(\psi_K|O_1^K)$).

As analyzed in [5], the search space for our new dynamic model grows exponentially with time (like the switching evolution in the switching models), which make the search difficult. However, the IMM algorithm described in the preceding section has provided an merging strategy to effectively overcome this difficulty. We have incorporated the IMM merging strategy into the dynamic programming based decoding strategy to implement a desirable decoding rule, $max_{\psi_K \in \Psi_K}\ l(O_1^K, \psi_K)$. During the trellis search, at time k $(1 \le k \le K)$ all the paths entering node j are merged to a single one using Eqns.(11) and (12). Therefore, by merging, there are only a fixed, small number M of paths expanding into the future (as in the Viterbi search, only one path (the most likely one) at each node is kept for the future expansion). The computational burden is greatly reduced by this merging strategy. In the mean time, the merging is carried out according to the posterior probabilities $P(j|i, O_1^{k-1})$. This makes the merging efficient as well. We call this one-pass dynamic programming decoding algorithm with incorporation of the IMM merging strategy the IMM-decoding (IMM-D) algorithm.

5. Experiments on conversational speech recognition

We first tested the IMM-D algorithm on simulated data. The results show that it is generally capable of recovering the hidden switching time points that had been used to artificially generate the observation data. Given the correct implementation of the algorithm as confirmed in these simulation experiments, we apply them to real speech data and carry out speech recognition experiments.

To design evaluation experiments on conversational speech, we choose one male speaker's data (speaker ID: 1028) extracted from **"train-ws97-a"** training set of the Switchboard corpus to train our dynamic models. The data consist of 966 utterances (about half an hour long). The state-space dynamic model given in Eqns.(1) and (2) was trained on this "1/2 hour" data [1, 2],

which is denoted by "MLP version" in Table 1. The MLDM model (Section 2) with two mixture components was trained also using this "1/2 hour" data, which is denoted by "MLDM-2mix" in Table 1. Finally, for comparison purposes, one HMM system was also trained on the "1/2 hour" data [1], which is denoted by **HMM-baseline** system in Table 1 .

N-best re-scoring paradigm was used for the evaluation of the new decoding approaches. The test set is chosen to be the male side of **"test-ws97-dev-l"** Switchboard test set (a total of 23 male speakers, 24 conversation sides, 1243 utterances and 50 minutes of speech)[1]. One hundred hypotheses for each utterance in the test set have been generated by a triphone HMM system which was developed for the Workshop'97 [1]. The phone segmentations generated by the HMM system are not consistent with our new model's dynamic regimes. In our earlier work [1, 2] the computation of the acoustic likelihoods for each hypothesis transcription in the 100-best list was carried out using the dynamic regimes fixed sub-optimally from the phone boundaries provided by the HMM system. In this work the computation of the acoustic likelihoods is carried out using the dynamic regimes optimized by the decoding strategies. For decoding, each hypothesis becomes a simple lattice, where each phone is only allowed to transit to itself and to the next phone.

In order to focus on the acoustic modeling issue, we ignore the language model scores in our experiments. However, any improvement due to language models should be equally applicable to the new model.

The re-scoring word error rate results are shown in Table 1. The "fix" in Table 1 means the dynamic systems using the phone dynamic regimes sub-optimally fixed based on the HMM phone alignments, the "IMM-D" means the systems using the dynamic regimes optimized by the "IMM-D" decoding algorithm. For both "MLP version" and "MLDM-2mix" version of the dynamic model, the "IMM-D" algorithm gives word error rate (WER) reduction of 4.1% and 2.3% for the "Ref+100" case (100 best list plus reference hypothesis), and the WER reduction of 1.2% for the "100 best" case.

Compared with the HMM system under identical conditions, with the use of the "IMM-D" decoding algorithm, the dynamic models outperform the HMM baseline system with 8.4% lower absolute WER for the "Ref+100" case, and with 3.2% lower absolute WER for the "100 best" case. These results demonstrated the effectiveness of the IMM-D algorithm for conversational speech recognition using the new state-space dynamic models.

Table 1: *Evaluation Results (WERs) of IMM-D on the male side of* **"test-ws97-test"** *Switchboard test data*

Recognizers	Ref+100	100 best
MLP version: fix (1/2 hour)	55.6	58.3
MLP version: IMM-D (1/2 hour)	51.5	57.1
MLDM-2mix: fix (1/2 hour)	50.0	56.9
MLDM-2mix: IMM-D (1/2 hour)	47.7	55.7
HMM-baseline (1/2 hour)	56.1	58.9

6. Conclusion

In this paper, we report our continuing efforts in the development of a specialized dynamic model for the VTR speech dynamics. The specific contribution of this work is the establishment of a novel path merging strategy for recognizer decoding with the optimized dynamic regimes, one associated with each discrete state (or phone), in the speech model. The dynamic-programming based decoding algorithm using the IMM merging strategy (IMM-D) is developed. It has successfully overcome the formidable exponential growth in the original search space. Speech recognition experiments using a subset of the Switchboard corpus have demonstrated that use of the developed decoding strategies can effectively reduce the recognizer's word error rate over a baseline recognizer under identical test conditions.

With the efficient decoding algorithms developed and described in this paper, it becomes possible to move the evaluation of the dynamic model from N-best list re-scoring to lattice re-scoring and possibly to real-time speech recognition. The algorithms can also be incorporated into the model training process to improve the training; that is, at each EM iteration, the decoding algorithm can be used to re-align the dynamic regimes before parameter estimation takes place. With this improved model training, more significant recognizer performance improvement will be expected.

[1] See *http://www.clsp.jhu.edu/ws97/ws97_general.html*

7. References

[1] J. Bridle, L. Deng, J. Picone , etc., "An Investigation of Segmental Hidden Dynamic Models of Speech Coarticulation for Automatic Speech Recognition," Final Report for the 1998 Workshop on Language Engineering, Center for Language and Speech Processing at Johns Hopkins University, 1998, pp. 1-61.

[2] L. Deng and J. Z. Ma, "Spontaneous Speech Recognition Using a Statistical Coarticulatory Model for the Vocal-Tract-Resonance Dynamics", *J. of American Society of Acoustics*, Vol.108, No. 6, Dec 2000, pp. 3036-3048.

[3] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle", *Econometrica*, Vol. 57, No. 2, pp. 357-384, 1989.

[4] C. J. Kim, "Dynamic linear models with Markov-switching", Journal of Econometrics, Vol. 60, pp1-22, 1994.

[5] J. Z. Ma and L. Deng, "A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech", *Computer, Speech and Language*, vol. 14, pp. 101-114, 2000.

[6] J. Z. Ma and L. Deng, "Target-Directed Mixture Linear Dynamic Models for Spontaneous Speech Recognition", *submitted to the same conference, Eurospeech 2001.*

[7] J. M. Mendel, *Lessons in estimation Theory for Signal Processing, Communications and Control*, Prentice Hall, 1995.

[8] M. Ostendorf, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition" *IEEE Trans. Speech Audio Proc.*, Vol. 4, 1996, pp. 360-378.

[9] Y. Bar-Shalom and X. R. Li, *Estimation and Tracking*, Artech House, Boston, MA., 1993.

[10] R. H. Shumway and D. S. Stoffer, "Dynamic linear models with switching", *Journal of the American Statistical Association*, Vol. 86, No. 415, pp. 763-769, 1991.

[11] H. Tanizaki, *Nonlinear Filters*, Second Ed., Springer Verlag, 1996.