COARTICULATION MODELING BY EMBEDDING A TARGET-DIRECTED HIDDEN TRAJECTORY MODEL INTO HMM – MODEL AND TRAINING

Jian-Lai Zhou, Frank Seide, and Li Deng[*]

Microsoft Research Asia, 5F Beijing Sigma Center, No. 49 Zhichun Rd., 100080 Beijing, P.R.C.
[*]Microsoft Research, One Microsoft Way, Redmond, WA 98052, U.S.A.
{jlzhou,fseide,deng}@microsoft.com

ABSTRACT

We propose and evaluate a new acoustic model that combines HMM and a special type of the hidden dynamic model (HDM) – a target-directed hidden trajectory model – into a single integrated model named HTHMM. The new model provides a computational model of coarticulation by representing the internal dynamics of human speech based on the hidden trajectory of the vocal-tract resonances. This paper focuses on the general structure of the new model and the EM training procedure. The corresponding MAP decoding algorithm and more detailed evaluation are given in [1].

Speech recognition experimental results on the Aurora2 task demonstrated that the new model, although using *only context-independent phoneme units* (no context-dependent parameters), is still slightly superior in word error rate to the corresponding cross-word triphone HMM. This provides the evidence that the coarticulatory mechanism represented by the HTHMM via the model structure matches the traditional context-dependent modeling approach based on enumeration of model parameters.

1. INTRODUCTION

Speech recognition technology has achieved significant progress with the introduction of the Hidden Markov Model (HMM). However, the current technology is often not satisfactory in real-world applications in the presence of mismatch between training and decoding. One important reason is speaking rate variation and the related coarticulation. To overcome the limitation of the current technology, we believe the it is necessary to take into account some key dynamic properties of human speech production into the mathematical representation, in particular the basic target-directed dynamic properties of human speech production.

This paper reports our recent efforts in coarticulation modeling. The novel Hidden-Trajectory HMM (HTHMM) proposed and evaluated in this paper draws heavily on previous work on the various types of Hidden Dynamic Models (HDM) [2, 3, 4] and a number of existing segment models as discussed in [5]. However, our new HTHMM drops the segmental aspects in favor of an HMM-like multi-state and frame-based mixture architecture. Similar to several previous HDM models [8, 2, 4, 9], the HTHMM presented in this paper uses a target-directed trajectory function to represent hidden dynamic variables. Continuity constraints across adjacent phones explicitly model long-span coarticulation effects.

However, the new HTHMM uses an internal noise term in the trajectory function to balance the uncertainties of both the hidden dynamic and the observable acoustics, and introduces a deterministic dynamic system, instead of a stochastic dynamic system, to characterize the hidden *vocal-tract resonance* (VTR) dynamics. Removing the frame-wise noise term used e.g. in [2] allows to eliminate the Kalman filter in training and decoding, significantly simplifying mathematical representation and recognizer implementation. Our model uses mixtures of linear mappings from the VTR-based hidden dynamic variables to the acoustic observations [2].

The main goal of our study was to compare the HTHMM's ability of modeling coarticulation, using only context-independent (CI) phoneme units, with that of a traditional context-dependent (CD) HMM. This study has so far been carried out only on a small-vocabulary task (Aurora2 TI-DIGITs) to keep training and decoding cost low and to eliminate dependence on the language model.

In contrast to all previous work on HDMs, where the lack of a direct decoding algorithm became a bottleneck to research progress, in this study we have developed a novel full MAP decoding algorithm described in detail in [1]. Specifically, this algorithm takes direct account of the continuous hidden state variables. All evaluation results presented in this paper have been obtained using this decoder rather than using N-best evalution.

This paper is organized as follows. Section 2 introduces the HTHMM model structure, and Section 3 the corresponding training algorithm. Experimental results are presented in Section 4. Finally, Section 5 concludes the paper and discusses future work.

2. THE HIDDEN-TRAJECTORY HMM (HTHMM)

The HTHMM presented in this section combines the HMM with a hidden trajectory model – a new special form of HDM in which the internal speech dynamics is modeled by a noise-free recursively-defined VTR-trajectory function of time. In contrast to the HDM in [2] which injects noise (modeling uncertainty) at each new time frame *before* predicting the hidden dynamic variable at the next time frame, our model injects no such type of noise and thereby significantly simplying the training and decoding algorithms.

2.1. Hidden Trajectory Model

The hidden trajectory model has the same general structure as the HDM. It consists of a hidden trajectory function followed by a mapping model. The former is used for representing internal speech dynamics for the cause of coarticulation, and the latter for describing the relationship between the hidden variables and the observed acoustic features. These two components are

$$z(t) \quad = \quad g(t) + w(t) \qquad (1)$$

$$o(t) \quad = \quad h^{u(t)}(z(t)) + v(t) \qquad (2)$$

with $g(t)$ being the predicted trajectory of VTR, $z(t)$ the *hidden* true trajectory, $u(t)$ the speech unit at frame t, and $h^u(z)$ a speech-unit dependent mapping function to map the hidden VTR to the acoustic measurement (such as MFCCs). In general, $h^u(z)$ is non-linear. The $w(t)$ and $v(t)$ denote i.i.d. Gaussian noise with zero mean and covariance matrices Q and R, respectively, modeling deviations of the actual values of $z(t)$ and $o(t)$ from the expected ones. Note that the noise $w(t)$ in Eq. 1 does not enter into any frame-based recursion as contrasting the model described in [2].

2.1.1. Target-directed trajectory function

In our hidden-trajectory model, the trajectory function of $g(t)$ is described by a causal, second-order discrete-time critically-damped unity-gain, low-pass filter:

$$g(t) \; = \; 2\gamma_{u(t)}g(t-1) - \gamma_{u(t)}^2 g(t-2) + (1-\gamma_{u(t)})^2 T_{u(t)} \quad (3)$$

where $T_{u(t)}$ is the *target* vector associated with the corresponding phone and $\gamma_{u(t)}$ the (scalar) system-dynamic parameter.

2.1.2. Mapping from hidden variables to observations

Here, we want to find an appropriate statistical relationship among observation o, the hidden variable z, and and speech unit u. Let's expand the non-linear function $h(z)$ by a Taylor series to obtain

$$h(z) \quad = \quad \bar{\mu} + H \cdot (z-\bar{z}) + \text{residual}(z-\bar{z}). \qquad (4)$$

Substituting $h(z)$ into Eq. (2) and combining the residual and noise terms, we obtain

$$o(t) \quad = \quad \bar{\mu} + H \cdot (z-\bar{z}) + v'(t) \qquad (5)$$

If we assume that $v'(t)$ is zero-mean Gaussian, then the parameters H, $\bar{\mu}$, and $\bar{z}$ will fully characterize the conditional PDF of $p(o|u,z)$. It is obvious that one single linear function cannot approximate the $h(z)$ well except in a small range. Hence, we employ several sets of the above parameters in forming a Gaussian mixture distribution (m shall denote the mixture component).

This mapping function bears structural similarity to the mixture linear mapping in [2]. An important difference, however, is that the model in [2] is segmental — the same mixture component is used for all frames of a speech unit — while in our model the mixture component is chosen independently in each frame. [10] also describes a similar structure for single Gaussians, but has no model for the prediction error of z.

2.2. Embedding the Hidden Trajectory Model into HMM

In the description up to now, no assumption has been made on the choice of the units u or the HTHMM's discrete states. Preliminary speech-recognition experiments showed that choosing u as phone units did not produce the desired performance. Much better performance (to be discussed in Section 4) was obtained by choosing the units to be the monophone (CI) HMM states.

It can be easily seen that when setting $H_m = 0$, the model becomes identical to a classical HMM system. This degenerated model has been used for units where the model parameters do not depend on the hidden variable z, such as silence and noise.

In the general case ($H_m \neq 0$), the HTHMM with the left-to-right HMM topology is an HMM whose Gaussian means in the output mixture distributions are adapted to the long-span context captured by the hidden-trajectory model, hence the name HTHMM.

Using the HMM topology, the continuous state g of the HTHMM becomes dependent on the HMM state sequence, and the observation becomes dependent on both u and z. Thus, we generalize the HMM by a second hidden layer, the continuous trajectory state g.

2.3. Conditional probabilities

Based on the above model components, we have the following conditional PDFs:

$$p(z(t)|m(t),U,\Theta) \quad = \quad p(z(t)|g(t),\Theta) \qquad (6)$$

$$p(z|g,\Theta) \quad = \quad \mathcal{N}(z;g;Q) \qquad (7)$$

where $U = (u(1),...,u(t),...,u(T))$ is the entire sequence of unit assignments $u(t)$, which fully determines $g(t)$, and

$$p(o|z,m,u,\Theta) \quad = \quad p(o|z,m,\Theta) \quad = \quad \mathcal{N}(o;h_m(z);R_m) \quad (8)$$

$$\begin{aligned} \omega_m(t) \quad &= \quad p(m(t)|o(t),U,\Theta) \\ &= \quad \frac{p(o(t)|m(t),\Theta)\cdot p(m(t)|U,\Theta)}{\sum_{m'=1}^{M} p(o(t)|m',\Theta)\cdot p(m'|U,\Theta)} \end{aligned} \qquad (9)$$

where

$$p(o|m,\Theta) = \int_z p(o|z,m,\Theta)\cdot p(z|m,\Theta)\; \mathrm{d}z \qquad (10)$$

3. MODEL TRAINING

One principal contribution of this study is the development of the model training algorithm, which allows automatic determination of all parameters from a given set of training data. The developed algorithm consists of two steps that are iterated for several iterations. One step is parameter estimation given speech unit boundaries, and the other step is optimization of the speech unit boundaries. These two steps are described below in detail.

3.1. Parameter Estimation Given the Speech Unit Boundaries

With the speech-unit alignment U given, the parameter estimation algorithm developed is based on the EM priciple where the hidden (missing) variables are Z (a collection of z variables) and M (the mixture assignments of each frame). Assuming observation frames to be independent of each other, and given the parameter set Θ, the joint PDF can be written as

$$\begin{aligned} p(O,Z,M|U,\Theta) \; = \; \prod_{t=1}^{T} & p(o(t)|z(t),m(t),u(t),\Theta) \qquad (11) \\ & \cdot\, p(z(t)|m(t),U,\Theta)\cdot p(m(t)|u(t),\Theta) \end{aligned}$$

3.1.1. *E Step*

The $\mathcal{Q}$-function in the E-step of the EM algorithm is computed below as the conditional expectation over the missing data:

$$\mathcal{Q}(\Theta, \hat{\Theta}) = \int\limits_{Z} \sum_{M} \log p(O, Z, M|U, \hat{\Theta}) \cdot p(Z, M|O, U, \Theta)\, \mathrm{d}Z$$

where Θ is the model parameters associated with the immediately previous iteration of the EM algorithm, and $\hat{\Theta}$ is the updated paramter set. The posterior can be decomposed as

$$p(Z, M|O, U, \Theta) \quad = \quad p(Z|M, O, U, \Theta) \cdot p(M|O, U, \Theta) \quad (12)$$

The $\mathcal{Q}$-function can be decomposed into two terms:

$$\mathcal{Q}(\Theta, \hat{\Theta}) \quad = \quad \mathcal{Q}_z + \mathcal{Q}_p \quad (13)$$

The two terms are:

$$\begin{aligned} \mathcal{Q}_z \quad &= \quad -\frac{1}{2} \sum_{t=1}^{T} \sum_{m=1}^{M} \omega_m(t) \Big[log|Q| + \log|R_m| \\ &\quad + \mathrm{E}\Big\{ e1_{t,m}^T \cdot (R_m^u)^{-1} \cdot e1_{t,m} \Big\} \\ &\quad + \mathrm{E}\Big\{ e2_{t,m}^T \cdot (Q^u)^{-1} \cdot e2_{t,m} \Big\} \Big] + \text{const} \quad (14) \\ \mathcal{Q}_p \quad &= \quad \sum_{t=1}^{T} \sum_{m=1}^{M} \log p(m(t)|U, \Theta) \cdot \omega_m(t) \quad (15) \end{aligned}$$

with $e1_{t,m} = o(t) - [H_m(z(t) - \bar{z}_m) + \mu_m]$ and $e2_{t,m} = z(t) - g(t)$. For notational simplicity, we now use $\mathrm{E}\{\cdot\}$ to represent conditional expectation $\mathrm{E}_{[\cdot|O^n, m, \Theta]}\{\cdot\}$ henceforth.

3.1.2. *M Step*

The M-step reestimation for all parameters except $\hat{\gamma}_u$ is derived via setting the partial derivatives of the $\mathcal{Q}$-function to zero, giving:

$$\begin{aligned} P(m|\hat{\Theta}) \quad &= \quad \frac{1}{T} \cdot \sum_{t=1}^{T} \omega_m(t) \quad (16) \\ \hat{H}_m \quad &= \quad \Big\{ \sum_{t=1}^{T} \Big[\omega_m(t) \cdot \Big(\hat{v}(t) \cdot \mathrm{E}\Big\{ z_t \Big\}^T \Big) \Big] \Big\} \\ &\quad \cdot \Big\{ \sum_{t=1}^{T} \Big[\omega_m(t) \cdot \mathrm{E}\Big\{ z_t \cdot z_t^T \Big\} \Big] \Big\}^{-1} \quad (17) \\ \hat{\hat{\mu}}_m \quad &= \quad \sum_{t=1}^{T} \omega_m(t) \cdot o(t) \;\; / \; \sum_{t=1}^{T} \omega_m(t) \quad (18) \\ \hat{\hat{z}}_m \quad &= \quad \sum_{t=1}^{T} \omega_m(t) \cdot \mathrm{E}\Big\{ z(t) \Big\} \;\; / \; \sum_{t=1}^{T} \omega_m(t) \quad (19) \\ \hat{R}_m \quad &= \quad \sum_{t=1}^{T} \Big[\omega_m(t) \cdot \mathrm{E}\Big\{ \hat{v}(t) \cdot \hat{v}(t)^T \Big\} \Big] \;\; / \; \sum_{t=1}^{T} \omega_m^n \quad (20) \end{aligned}$$

where

$$\begin{aligned} \hat{v}(t) \quad &= \quad o(t) - \hat{\hat{\mu}}_m - \hat{H}_m \cdot \hat{\hat{z}}_m \\ \mathrm{E}\Big\{ z_t \Big\} \quad &= \quad \Big[H_m^T R_m^{-1} H_m + Q^{-1} \Big]^{-1} \\ &\quad \cdot \Big[H_m^T R_m^{-1} \hat{v}(t) + Q^{-1} g(t) \Big] \\ \mathrm{E}\Big\{ z_t (z_t)^T \Big\} \quad &= \quad \Big\{ H_m^T R_m^{-1} H_m + Q^{-1} \Big\}^{-1} + \mathrm{E}\Big\{ z_t \Big\} \cdot \mathrm{E}\Big\{ z_t \Big\}^T \end{aligned}$$

The estimate of the speech units' target vectors T_u requires an equation system since all T_u are coupled through the continuity constraint. We define:

$$\begin{aligned} \hat{\mathbf{T}} \; &= \; \Big(\hat{T}_1, \hat{T}_2, ... \Big) \quad (21) \\ \hat{T}_u \; &= \; \hat{\mathbf{T}} \cdot e_u \quad (22) \\ e_u \; &= \; (0, 0, ..., 1, ..., 0, 0)^T \quad \text{with 1 at position } u \quad (23) \\ \hat{g}(t) \; &= \; \hat{\mathbf{T}} \cdot \hat{b}(t) \quad (24) \\ \hat{b}(t) \; &= \; 2\gamma_{u(t)} \hat{b}(t-1) - \gamma_{u(t)}^2 \hat{b}(t-2) + (1 - \gamma_{u(t)})^2 e_{u(t)} \quad (25) \end{aligned}$$

With these, we can formulate the target estimates:

$$\begin{aligned} \hat{\mathbf{T}} \quad &= \quad \Big[\sum_{t=1}^{T} \sum_{m=1}^{M} \omega_m(t) \cdot \mathrm{E}\left\{ z(t) \right\} \cdot (\hat{b}(t))_u \Big] \\ &\quad \cdot \Big[\sum_{t=1}^{T} \sum_{m=1}^{M} \omega_m(t) \cdot \hat{b}(t) \cdot (\hat{b}(t))_u \Big]^{-1} \quad (26) \end{aligned}$$

Because there is no closed-form solution for $\hat{\gamma}_u$, we use a gradient-based method (which we actually implement numerically):

$$\hat{\gamma}_u^{(r+1)} \quad = \quad \hat{\gamma}_u^{(r)} + \varepsilon \cdot \frac{\partial \mathcal{Q}_z}{\partial \hat{\gamma}_u} \bigg|_{\hat{\gamma}_u^{(r)}} \quad (27)$$

3.2. State Boundary Optimization

Most of the previous work on HDM did not optimize the speech unit boundaries because no effective decoders for finding the true optimal path were available. Running our novel HTHMM decoder [1] in forced-alignment mode, we re-segmented each training utterence into the corresponding discrete states. The optimization is similar to the Viterbi training in HMM: the decoder segments the state boundaries, and given such information reestimation as described above is carried out.

4. EXPERIMENTAL RESULTS

Since this paper focuses on the model description and training algorithms, we only include a few key results of our study. More results can be found in [1].

The HTHMM has been evaluated on a small-vocabulary task of continuous-digit recognition (the clean portion of Aurora2/TI-DIGITs). The small-vocabulary task allows us to keep dependence on the language model low. The constructed speech recognition system is gender independent, with MFCC features of 39 dimensions. The standard HTK feature configuration from the Aurora2 distribution was used. In order to compare the recognition performance among different models, we show results for baseline

Table 1. Performance comparison among baseline and the new HTHMM systems.

Id	System	WER	
	#mix comp:	16	32
baseline: traditional HMM			
B1	context-independent HMM	0.69%	0.61%
B2	cross-word triphone HMM	0.49%	0.40%
proposed HTHMM system			
M1	HTHMM (context-independent)	0.43%	0.37%

HMM using monophones and cross-word triphones. All systems used a subset of 20 phonemes with three left-to-right states.

For both HMM and HTHMM systems, we show word error rates (WER) for models with 16 and 32 mixture components per state. Table 1 provides a summary of the experiment results demonstrating the comparative WERs among the baseline HMM and the new HTHMM system.

Compared with the baseline monophone HMM system (experiment B1) with WERs of 0.69% and 0.61% for 16 and 32 mixture components, respectively, the triphone HMM system produces 29% and 34% fewer errors, respectively (0.49% and 0.40%, B2). The HTHMM system, however, achieves slightly better WERs than the HMM triphone system (0.43% and 0.37%). The improvement has been achieved by using only *context-independent units*, and context dependence was modeled structurally through the mixture-mean correction term provided by the HTHMM.

Table 2 shows the effect of Viterbi iterations. Our study is the first to iteratively optimize the state boundaries for training of a HDM-type model. The table shows the training set's log likelihood after one EM iteration and after EM convergence, after 0, 1, and 2 forced alignments. After the very first EM iteration (T0, iteration 1), we have a model that has been estimated using mixture occupation counts computed with HMM parameters, using a HMM-based state segmentation. One can see that after convergence of EM (fixed segmentation), a dramatic log-likelihood improvement (from -59.11 to -55.53) is achieved — the mixture assignment of the HMM is far from optimal for the HTHMM. A further significant likelihood reduction to -55.27 is achieved after the first forced alignment (T1), and another in the same order of magnitude to -55.05 after convergence. A small WER reduction is also observed. The second forced alignment (T2) only leads to very small increase of likelihood and no WER improvement, so in all our experiments we only use a single alignment step.

5. CONCLUSIONS AND FUTURE WORK

In this paper, the preliminary work on a new hidden trajectory model, whose discrete states have been designed using the conventional HMM topology, and its application to speech recognition is reported. With the CI HTHMM (a compact model with no CD parameters), we obtained performance improvement over the CD HMM (cross-word triphone system). This provides evidence that the coarticulatory mechanism represented by the HTHMM via the model structure matches the traditional context-dependent modeling approach based on enumeration of model parameters. However, because HTHMM cannot model all properties of coarticulation, we believe some degree of CD modeling may still be needed.

Table 2. Performance comparison for different numbers of Viterbi iterations (8 mixtures).

Id	Number of forced	log LL after EM		WER
	alignments	iteration 1	converged	[%]
T0	0	-59.11	-55.53	0.54
T1	1	-55.27	-55.05	0.52
T2	2	-54.99	-54.90	0.53

We are currently working on large vocabulary tasks and hope to demonstrate that equally excellent performance of the new HTHMM approach can be established despite the weaker phonotactic constraints and more confusable acoustic space.

6. ACKNOWLEDGEMENTS

The authors wish to thank Drs. A. Acero, E. Chang, A. Gunawardana, and X.D. Huang and other EARS-project researchers at MSR for valuable discussions and encouragement of this work.

7. REFERENCES

[1] F. Seide, J.-L. Zhou, and L. Deng. Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM — MAP decoding and evaluation. *Submitted to Proc. ICASSP'03*, 2003, Hongkong.

[2] J. Ma, L. Deng. Target-directed mixture linear dynamic models for spontaneous speech recognition. *IEEE Trans. Speech and Audio Processing*, submitted 1999, to appear 2002.

[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. 6, Dec 2000, pp. 3036-3048.

[4] J. Bridle et al. The WS98 final report on the dynamic model. `http://www.clsp.jhu.edu/ws98/projects/dynamic/presentations/finalhtml/index.html`, Johns Hopkins Univ., 1998.

[5] M. Ostendorf, V. Digalakis, and J. Rohlicek. 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.

[6] L. Deng and M. Aksmanovic. Speaker-independent phonetic classification using Hidden Markov Models with state-conditioned mixtures of trend functions. *IEEE Trans. Speech Audio Proc.*, Vol. 5, No. 4, July 1997, pp. 319-324.

[7] W. Holmes and M. Russell. Probabilistic-trajectory segmental HMMs. *Computer Speech and Language*, Vol. 13, 1999, pp. 3-37.

[8] R. Bakis. Coarticulation modeling with continuous-state HMMs. *Proc. IEEE Workshop Automatic Speech Recognition*, pp. 20-21, Harriman, New York, 1991.

[9] Y. Gao, R. Bakis, J. Huang, B. Zhang, Multistage coarticulation model combining articulatory, formant and cepstral features. *Proceedings of the ICSLP*, Vol. 1, 2000, pp. 25-28.

[10] F.-L. Chen et al. The structure and its implementation of Hidden Dynamic HMM for Mandarin speech recognition. *Proc. ICSLP*, pp. 713-716, Denver, 2002.