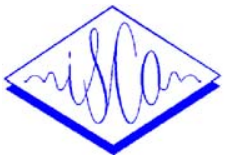

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A STATISTICAL COARTICULATORY MODEL FOR THE HIDDEN VOCAL-TRACT-RESONANCE DYNAMICS

Li Deng and Jeff Ma

Department of Electrical and Computer Engineering,
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

10.21437/Eurospeech.1999-260

ABSTRACT

A statistical coarticulatory model is presented for spontaneous speech recognition, where knowledge of the dynamic, target-directed behavior in the vocal tract resonance responsible for the production of highly coarticulated speech is incorporated into the recognizer design, training, and in likelihood computation. The principal advantage of the new speech model over the conventional HMM is the use of a compact, internal structure that parsimoniously represents long-span context dependence in the observable domain of speech acoustics without using additional, context-dependent model parameters. The new model is formulated mathematically as a constrained, nonstationary, and nonlinear dynamic system, for which a version of the generalized EM algorithm is developed and implemented for automatically learning the compact set of model parameters. Experiments for speech recognition using spontaneous speech data from SWITCHBOARD corpus are reported.

1. INTRODUCTION

We present in this paper a new dynamic approach to the challenging problem of spontaneous or conversational speech recognition. This approach is based on statistical Bayesian decision theory but is a radical departure from the current statistical modeling approaches. Rather than using a large number of unstructured Gaussian mixture components to account for tremendous variation in the observable acoustic data of highly coarticulated, spontaneous speech, our new approach provides a rich structure in the statistical model for the partially observable (hidden) dynamics in the domain of vocal-tract-resonances (VTRs).

The research reported in this paper is aimed specifically to provide a superior solution to the Switchboard conversational speech recognition task over the conventional HMM technology. We use statistical nonlinear dynamic system 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 design, training, and decoding/scoring. The statistical nature of the model design allows computation of the probability for acoustic observations of spontaneous, highly coarticulated speech in a more accurate fashion than the conventional HMM.

The VTRs are pole locations of the vocal tract configured to produce speech sounds, and have acoustic correlates of formants which are directly measurable for vowel or glide sounds, but often are hidden or perturbed for consonantal sounds due to the concurrent spectral zeros and turbulence noises. A noisy, causal and linear dynamic system is used to describe the VTR dynamics. The system matrix (encompassing the concept of time constant or the rate of articulation) is structured and constrained to ensure the asymptotic behavior in the dynamics. The output nonlinearity is multiple, switching multi-layer perceptrons (MLPs), with each MLP associated with the distinct manner of articulation of a phone. The criterion for model training (and for recognition) is maximum likelihood on observable MFCCs only (i.e., not on the partially-hidden VTRs). Some detail of the model training and recognition algorithms, based on extended Kalman filtering embedded in the EM algorithm, will be presented in the paper.

The principal advantage of the new model over the conventional HMM lies in the compact structure of the new model (in the space of hidden VTR domain) for representing long-term context dependence in the observable speech acoustics. The long-term context dependence is embedded in the model construct as a continuity constraint across the sequence of dynamic regimes (distinct for each phonological unit). Due to the compact structure in the model, the entire recognizer has approximately only 15,000 free parameters, significantly fewer than the HMM-based recognizer (3,500,000 parameters in total) designed for the identical task of recognition of spontaneous speech in the Switchboard corpora.

A series of experiments for speech recognition and model synthesis using the Switchboard corpus have been carried out. The promise of the new model is demonstrated by showing its consistently superior performance, over a state-of-the-art benchmark HMM system under controlled experimental conditions, when exposed to the reference transcription. Experiments on model synthesis and analysis shed insight into the mechanism underlying such superiority in terms of the target-directed behavior and of the long-span context-dependence property, both inherent in the designed structure of the new dynamic model of speech. Due to the space limitation, only some representative speech recognition results are reported in this paper.

2. MODEL FORMULATION

The coarticulatory speech model presented in this paper has been formulated in mathematical terms as a constrained and simplified nonlinear dynamic system. This is a special

version of the general statistical hidden dynamic model described in [4, 5]. The dynamic system model consists of two separate but related components: 1) state equation, and 2) observation equation, both of which are described in this section.

2.1. State equation

A noisy, causal, and linear first-order "state" equation is used to describe the three-dimensional (F1, F2, and F3) VTR dynamics according to

$$Z(k+1) = \Phi^j Z(k) + (I - \Phi^j)T^j + W_d(k),\ j = 1, 2, ..., J_P \quad (1)$$

where $Z(k)$ is the three-dimensional "state" vector at discrete time step k, Φ^j and T^j are the system matrix and goal (or target) vector associated with dynamic regime j which is related to the initiation of dynamic patterns in phone j, and J_P is the total number of phones in a speech utterance. Both Φ^j and T^j are a function of time k via their dependence on dynamic regime j, but the time switching points are not synchronous with the phone boundaries. The time scale for evolution of dynamic regime j is significantly larger than that for time frame k. In Eqn.(1), $W_d(k)$ is the discrete-time state noise, modeled by an i.i.d., zero-mean, Gaussian process with covariance matrix Q.

The special structure in Eqn.(1), which is linear in the state vector $Z(k)$ but nonlinear with respect to its parameters Φ^j and T^j, gives rise to two significant properties of the VTRs modeled by the state vector $Z(k)$. The first property is local smoothness; i.e., the state vector $Z(k)$ is smooth within the dynamic regime associated with each phone. The second, attractor or saturation property is related to the target-directed, temporally asymptotic behavior in $Z(k)$. This target-directed behavior of the dynamics described by Eqn.(1) can be seen by setting $k \to \infty$, which forces the system to enter the local, asymptotic regime where $Z(k+1) \approx Z(k)$. With the assumption of mild levels of noise $W_d(k)$, Eqn. (1) then directly gives the target-directed behavior in $Z(k)$: $Z(k) \to T^j$.

An additional significant property of the state equation is the left-to-right structure in Eqn.(1) for $j = 1, 2, ..., J_P$ and the related global-smoothness characteristics. That is, the local smoothness in state vector $Z(k)$ is extended across each pair of adjacent dynamic regimes, making $Z(k)$ continuous or smooth across an entire utterance. This continuity constraint is implemented in the current model by forcing the state vector $Z(k)$ at the end of dynamic regime j to be identical to the initial state vector for dynamic regime $j+1$. That is, the Kalman filter which implements optimal state estimation for dynamic regime $j+1$ is initialized by the $Z(k)$ value computed at the end of dynamic regime j.

2.2. Observation equation

The observation equation in the dynamic system model developed is nonlinear, noisy, and static, and is described by

$$O(k) = h^{(r)}[Z(k)] + V(k), \quad (2)$$

where the acoustic observation $O(k)$ is MFCC measurements computed from a conventional speech preprocessor, $V(k)$ is the additive observation noise modeled by an i.i.d., zero-mean, Gaussian process with covariance matrix R, intended to capture residual errors in the nonlinear mapping from $Z(k)$ to $O(k)$. The multivariate nonlinear mapping, $h^{(r)}[Z(k)]$, is implemented by multiple switching MLPs, with each MLP associated with a distinct manner (r) of articulation of a phone. A total of ten MLPs (i.e., $r = 1, 2, ..., 10$) are used in the experiments reported in this paper.

The nonlinearity is necessary because the physical mapping from VTR frequencies $(Z(k))$ to MFCCs $(O(k))$ is highly nonlinear in nature. The noise used in the model Eqn.(2) captures the effects of VTR bandwidths (i.e., formant bandwidths for vocalic sounds) and relative VTR amplitudes on the MFCC values. These effects are secondary to the VTR frequencies but they nevertheless contribute to the variability of MFCCs. Such secondary effects are quantified by the determinant of matrix R, which, in combination with the relative size of the state noise covariance matrix Q, plays important roles in determining relative amounts of state prediction and state update in the state estimation procedure.

In implementing the nonlinear function $h[Z(k)]$ (omitting index r for clarity henceforth) in Eqn.(2), we used a MLP network of three linear input units ($Z(k)$ of F1, F2, and F3), of 100 nonlinear hidden units, and of 12 linear output units ($O(k)$ of MFCC1-12). Denoting the MLP weights from input to hidden units as w_{jl}, and the MLP weights from hidden to output units as W_{ij}, we have

$$h_i(Z) = \sum_j W_{ij} \cdot g(\sum_l w_{jl} \cdot Z_l), \quad (3)$$

where, $i = 1, 2, ..., 12$ is the index of output units (i.e., component index of observation vector O_k), $j = 1, 2, ..., 100$ is the index of hidden units, and $l = 1, 2, 3$ is the index of input units. In Eqn.(3), the hidden units' activation function is the standard sigmoid function

$$g(x) = \frac{1}{1 + exp(-x)}$$

with its derivative

$$g'(x) = g(x)(1 - g(x)).$$

The Jacobian matrix for Eqn.(3), which will be needed for the Extended Kalman Filter (EKF), can be computed in an analytical form:

$$H_z(Z) \equiv \frac{d}{dZ} h(Z) = [H_{il}(Z)]$$

where

$$H_{il}(Z) = \sum_j W_{ij} g[\sum_l w_{jl} g(Z_l)][1 - g(\sum_l w_{jl} g(Z_l))] w_{jl}.$$

3. MODEL LEARNING AND SCORING

The learning or parameter estimation method for the new speech model is based on the generalized EM algorithm. To derive the generalized EM algorithm for the model, we first use the i.i.d. noise assumption for $W_d(k)$ and $V(k)$

in Eqns.(1) and (2) so as to express the log-likelihood for acoustic observation sequence $O = [O(1), O(2), \ldots, O(N)]$ and hidden sequence $Z = [Z(1), Z(2), \ldots, Z(N)]$. The model parameters Θ to be learned include those in the state equation Eqn.(1) and those in the MLP nonlinear mapping functions Eqn.(2): $\Theta = \{T, \Phi, W_{ij}, w_{jl}, i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; l = 1, 2, \ldots, L\}$.

3.1. Expectation-step

E-step of the EM algorithm involves computation of the following conditional expectation (together with a set of related sufficient statistics needed to complete evaluation of the conditional expectation):

$$\begin{aligned} & Q(Z, O, \Theta) = E\{\log L(Z, O, \Theta) | O, \Theta\} \\ & = -\frac{N}{2}\log|Q| - \frac{N}{2}\log|R| \\ & -\frac{1}{2}\sum_{k=0}^{N-1} E[e_{k1}^{'} Q^{-1} e_{k1} | O, \Theta] - \frac{1}{2}\sum_{k=1}^{N} E[e_{k2}^{'} R^{-1} e_{k2} | O, \Theta] \end{aligned}$$

where $e_{k1} = Z(k+1) - \Phi Z(k) - (I - \Phi)T$ and $e_{k2} = O(k) - h(Z(k))$.

This can be simplified by algebraic manipulation to

$$\begin{aligned} Q(Z, O, \Theta) \quad = \quad & \underbrace{-\frac{N}{2}\log\{|\frac{1}{N}\sum_{k=0}^{N-1} E[e_{k1} e_{k1}^{'} | O, \Theta]|\}}_{Q_1(Z, O, \Phi, T)} \\ & \underbrace{-\frac{N}{2}\log\{|\frac{1}{N}\sum_{k=1}^{N} E[e_{k2} e_{k2}^{'} | O, \Theta]|\}}_{Q_2(Z, O, W_{ij}, w_{jl})} + const. \end{aligned}$$

Note that the state-equation's parameters (Φ, T) are contained in Q_1 only and the MLP weight parameters (W_{ij}, w_{jl}) in the observation equation are contained in Q_2 only. These two sets of parameters can then be optimized independently in the subsequent M-step.

3.2. Maximization-step (approximation)

M-step of the EM algorithm aims to optimize the Q function with respect to model parameters $\Theta = \{T, \Phi, W_{ij}, w_{jl}\}$. For the model at hand, it seeks solutions for

$$\frac{\partial Q_1}{\partial \Phi} \propto \sum_{k=0}^{N-1} E[\frac{\partial}{\partial \Phi}\{[Z(k+1) - \Phi Z(k) - (I - \Phi)T]^2\} | O, \Theta] = 0$$

$$\frac{\partial Q_1}{\partial T} \propto \sum_{k=0}^{N-1} E[\frac{\partial}{\partial T}\{[Z(k+1) - \Phi Z(k) - (I - \Phi)T]^2\} | O, \Theta] = 0$$

$$\frac{\partial Q_2}{\partial W_{ij}} \propto \sum_{k=1}^{N} E[\frac{\partial}{\partial W_{ij}}\{[O(k) - h(Z(k))]^2\} | O, \Theta] = 0 \qquad (4)$$

$$\frac{\partial Q_2}{\partial w_{jl}} \propto \sum_{k=1}^{N} E[\frac{\partial}{\partial w_{jl}}\{[O(k) - h(Z(k))]^2\} | O, \Theta] = 0. \qquad (5)$$

The first two equations above are both third-order nonlinear algebraic equations (in Φ and T):

$$N\Phi TT' - \Phi TA' - \Phi AT' - NTT' - TA' + BT' + \Phi C - D = 0,$$
$$N\Phi'\Phi T - \Phi'\Phi A - N\Phi' T - N\Phi T + \Phi' B + \Phi A + NT - B = 0.$$

where

$$A = \sum_{k=0}^{N-1} E[Z(k) | O, \Theta], \qquad B = \sum_{k=0}^{N-1} E[Z(k+1) | O, \Theta],$$
$$C = \sum_{k=0}^{N-1} E[Z(k) Z(k)^{'} | O, \Theta], D = \sum_{k=0}^{N-1} E[Z(k+1) Z(k)^{'} | O, \Theta].$$

The coefficients A, B, C, and D above constitute sufficient statistics, which are computed by the standard technique of EKF [8].

Solutions to Eqns.(4) and (5) for finding (W_{ij}, w_{jl}) to maximize Q_2 have to rely on approximation due to the complexity in the nonlinear function $h(Z)$. The approximation involves first finding estimates of hidden variables $Z(k)$, $Z(k|k)$, via the EKF algorithm. Given such estimates, the conditional expectations are approximated to give

$$\frac{\partial Q_2}{\partial W_{ij}} \quad \propto \quad \sum_{k=1}^{N} [O(k) - h(Z(k|k))]^{'} \frac{\partial h(Z(k|k))}{\partial W_{ij}}$$

$$\frac{\partial Q_2}{\partial w_{jl}} \quad \propto \quad \sum_{k=1}^{N} [O(k) - h(Z(k|k))]^{'} \frac{\partial h(Z(k|k))}{\partial w_{jl}}.$$

As a crude approximation, if the estimated state variable, $Z(k|k)$, is treated as the input to the MLP neural network and the observation, $O(k)$, as the output of the MLP, then the gradients above will be exactly the same as those in the backpropagation algorithm. This crude approximation has been used in this work to learn the MLP parameters in the speech model.

3.3. Likelihood-scoring algorithm

Using a single Gaussian to approximate the multiple-modal distribution of the output of the nonlinear dynamic system, the log-likelihood scoring function for our speech model can be computed from the pseudo-innovation [1] sequence $\tilde{O}(k|k-1)$ according to

$$\begin{aligned} \log L(O|\Theta) = -\frac{1}{2}\sum_{k=1}^{N}\{\log|P_{\tilde{O}\tilde{O}}(k|k-1)| + \\ \tilde{O}(k|k-1)' P_{\tilde{O}\tilde{O}}^{-1}(k|k-1)\tilde{O}(k|k-1)\} + const., \end{aligned} \qquad (6)$$

where the pseudo-innovation sequence

$$\tilde{O}(k|k-1) = O(k) - h(\hat{Z}(k|k-1)), \quad k = 1, 2, ..., N$$

is computed from the EKF recursion, and $P_{\tilde{O}\tilde{O}}$ is the covariance matrix of the pseudo-innovation sequence:

$$P_{\tilde{O}\tilde{O}}(k|k-1) = H_z(\hat{Z}(k|k-1)) P(k|k-1) H_z(\hat{Z}(k|k-1))' + R,$$

computed also from the EKF recursion.

4. SPEECH RECOGNITION EXPERIMENTS

Evaluation of the new speech model and recognizer described in Sections 2 and 3 has been carried out during the 1998 NSF/DoD Workshop on Language Engineering. In the experiment, a total of 23 male speakers comprising 24 conversation sides, 1241 utterances, and 50 minutes of speech were chosen as the test set. Only one male speaker's 30 minutes of speech data were chosen as the training set. N-best rescoring paradigm was used to compare the new recognizer with the conventional HMM-based one.[1] Given fixed dynamic regimes (i.e., the boundaries precomputed by the HMM system) for each phone of each hypothesis in the N-best list, the computation of the acoustic likelihoods using the VTR-based new recognizer can be performed efficiently, according to Eqn.(6), by running the EKF algorithm for each hypothesis in the N-best list with known dynamic switching-regimes. The results obtained can be summarized as follows. First, when N=5 and these 5-best hypotheses are enriched by the reference transcripts, the new recognizer gives 32.2% word error rate (WER), significantly lower than the HMM counterpart system (WER 44.8%) under identical training and testing conditions. Second, when N is increased from 5 to 100, the WER reduction is somewhat less significant, from 56.1% (HMM system) to 50.3% (new recognizer). Third, when the phone boundaries in the reference hypotheses and the associated dynamic regimes for each phone are manually adjusted by reading spectrograms, use of the same scoring algorithm drastically reduces the WER to as low as 19%.

5. SUMMARY AND CONCLUSION

Spontaneous speech process is a combination of cognitive (linguistic or phonological) and physical (phonetic) subprocesses. The new statistical coarticulatory model presented in this paper focuses on the physical aspect of the spontaneous speech process, where a main novelty is the introduction of the VTR as the internal, structured model state (continuous-valued) for representing phonetic reduction and target undershoot in human production of spontaneous speech. The continuity constraint imposed on the VTR state across speech units as implemented in the model is physically motivated, and it enables phonetic information to flow from one unit to another with no use of additional, context-dependent model parameters. Such continuity is not valid in the acoustic domain because of the nonlinear, "quantal" nature of the distortion in the peripheral speech production process, and in order for the model to ultimately score on the acoustic domain, we explicitly represent the nonlinear distortion as a model component integrated with the VTR dynamic component. With the complex model structure formulated mathematically as constrained, nonstationary, and nonlinear dynamic system, a version of the generalized EM algorithm has been developed and implemented for automatically learning the compact set of model parameters. The key characteristics summarized above for the model make it distinct from all earlier types of dynamic models of speech used in speech recognition (e.g., [2, 3, 6, 7, 10]).

One research direction we are currently pursuing is motivated by the experimental results briefly described in Section 4 which underscore the critical role of using "true" dynamic regimes of the VTR model in speech recognition performance. An initial algorithm we have developed which is capable of joint optimization of dynamic regimes and of the regime-bound acoustic match scores has been evaluated in a preliminary experiment showing promising results [9]. Algorithms of such a type will also be extended to training, enabling automatic learning of all model parameters without use of heuristically supplied dynamic regimes in the training data.

[1] See details of this HMM system and the experimental paradigm in [11].

Acknowledgments:
The evaluation results reported in this paper were based mainly upon work supported by the National Science Foundation under Grant No. (#IIS-9732388) and carried out at the 1998 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. We thank M. Schuster, J. Picone, J. Bridle, H. Richards, S. Pike, R. Reagan, T. Kamm who contributed neural network programs, HMM benchmark results, discussions, and error-analysis software tools which made this evaluation possible. We also thank F. Jelinek, M. Ostendorf, C. Lee, and G. Doddington for support, encouragement, insightful comments and discussions of this work.

6. REFERENCES

[1] Anderson B. and Moore J. *Optimal Filtering,* Prentice-Hall, N.J., 1970, pp. 103.

[2] Bakis R. "Coarticulation modeling with continuous-state HMMs," *Proc. IEEE Workshop Automatic Speech Recognition,* Arden House, N.Y., 1991, pp. 20-21.

[3] Deng L. "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, Vol.27, 1992, pp. 65-78.

[4] Deng L. "Computational models for speech production," in *Computational Models of Speech Pattern Processing* (NATO ASI), 1997.

[5] Deng L. "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Communication,* Vol. 24, No. 4, pp. 299-323, 1998.

[6] Deng L., Aksmanovic M., Sun D., and Wu J. "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. Speech Audio Proc.*, Vol. 2, 1994, pp. 507-520.

[7] Deng L., Ramsay G., and Sun D. "Production models as a structural basis for automatic speech recognition," *Speech Communication* (special issue on speech production modeling), Vol. 22, No. 2, August 1997, pp. 93-112.

[8] Jazwinski, A. H. *Stochastic Processes and Filtering Theory,* Academic Press, New York, 1970.

[9] Ma J. and Deng L. "Optimization of dynamic regimes in a statistical hidden dynamic model for spontaneous speech recognition," *these proceedings.*

[10] Ostendorf M. "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.

[11] Picone J., S. Pike, R. Reagan, T. Kamm, J. Bridle, L. Deng, J. Ma, H. Richards, M. Schuster. "Initial evaluation of hidden dynamic models on conversational speech," *Proc. of ICASSP*, March 1999, pp 109-112.