A FUNCTIONAL ARTICULATORY DYNAMIC MODEL FOR SPEECH PRODUCTION

[*]Leo J. LEE, [†]Paul FIEGUTH

University of Waterloo
[*]Dept. of Electrical & Computer Engineering
[†]Dept. of Systems Design Engineering
Waterloo, ON, N2L 3G1
CANADA

[*]Li DENG

Microsoft Research
Speech Technology Group
One Microsoft Way
Redmond WA 98052-6399
USA

ABSTRACT

This paper introduces a new statistical speech production model. The model synthesizes natural speech by modeling some key dynamic properties of vocal articulators in a linear/nonlinear state-space framework. The goal-oriented movements of the articulators (tongue tip, tongue dorsum, upper lip, lower lip, and jaw) are described in a linear dynamic state equation. The resulting articulatory trajectories, combined with the effects of the velum and larynx, are nonlinearly mapped into the acoustic feature space (MFCCs). The key challenges in this model are the development of a nonlinear parameter estimation methodology, and the incorporation of appropriate prior assumptions to assert in the articulatory dynamic structure. Such a model can also be directly applied to speech recognition to better account for coarticulation and phonetic reduction phenomena with considerably fewer parameters than HMM based approaches.

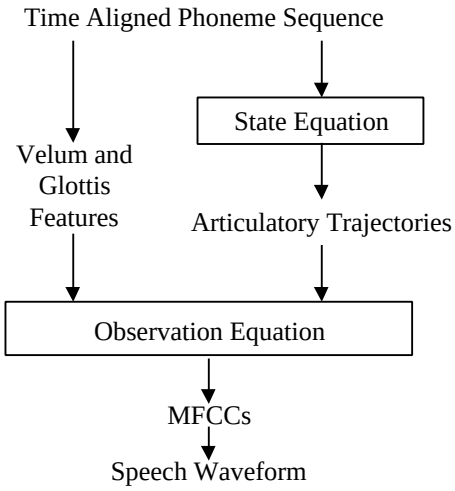

Fig. 1. A block diagram of the speech production model.

1. INTRODUCTION

Despite forty years of past studies into human speech production and an array of increasingly detailed and sophisticated proposed models [1], they have had at best only a limited impact on computer synthesized human speech and on automatic speech recognition. From a practical point of view, the proposed models are either too complicated to implement or they lack the comprehensiveness in covering all classes of sounds. On the other hand, the current *cut and paste* approach used in commercial speech synthesizers clearly fails to provide phoneme transition as naturally as human articulatory system.

The focus of this paper is to *model* the key dynamic characteristics of the articulators which are essential to natural speech. We will not describe the detailed underlying physiological mechanism which governs the movement of articulators in our model, although clearly our model must reflect the constraints in this movement. Rather, we propose a target-oriented, parameterized linear state equation where the parameters are learned automatically from real speech data, hence the name *functional* model. In terms of speech recognition, such an articulatory model represents a considerable departure from traditional HMM modeling, because we are directly representing the physical speech production process. However, our model can represent the underlying articulatory movement quite accurately with *very* low state dimension (between 4 and 10), implying that only very few parameters need to be estimated, in sharp contrast with HMM models. The simplicity of our model promises convenience in applying it to problems of statistical speech recognition.

The remainder of this paper is organized as follows: The proposed model is detailed in Section 2, followed by automatic parameter estimation strategies in Section 3. Section 4 shows preliminary results and lists possible further improvements and research directions of the model, followed by a brief discussion and conclusion in Section 5.

2. MODEL DESCRIPTION

A block diagram of the proposed speech production model is shown in Fig. 1. The underlying articulatory dynamics and the acoustic features are related by the following state-space model:

$$\mathbf{z}(k+1) = \Phi\mathbf{z}(k) + \Psi\mathrm{T} + \mathbf{w}(k), \quad (1)$$

$$\mathbf{o}(k) = h\left[\mathbf{z}(k)\right] + \mathbf{v}(k). \quad (2)$$

The state vector $\mathbf{z}(k)$ represents the positions of key articulators, listed in Fig. 2, at time k. The acoustic features $\mathbf{o}(k)$, chosen to be Mel-frequency cepstral coefficients (MFCCs) in our model, are generated through nonlinear function $h[\,]$.

The *key* parameterized elements of our model are the quantities Φ, Ψ, and T: Matrix Φ encodes the time interaction among the articulatory components; vector T is the *target* position of the articulators (in hypothetical steady-state); and matrix Ψ describes the control effect of the targets on the articulatory movement. All three quantities are phone dependent, although for implementation purposes the values Φ and Ψ may each be tied for broader classes of phones. Due to the well-known forward-anticipation property of the articulators, the boundaries for these parameters (*especially* the target T) should be in advance of the actual acoustic boundaries. The quantitative degree of anticipation is one additional parameter to be learned from the articulatory data. Finally the nonlinear observation function $h[\,]$ represents the articulatory-to-acoustic mapping. Both $\mathbf{w}(k)$ and $\mathbf{v}(k)$ are discrete-time white Gaussian noise processes, with time-invariant covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ respectively.

By definition, our state equation (1) must satisfy the assumed asymptotic target; that is, $\mathbf{z}(k) \rightarrow \mathrm{T}$ as $k \rightarrow \infty$, which places constraints on Φ and Ψ. A convenient choice, which we assume throughout this paper, is to let $\Psi = \mathbf{I} - \Phi$, in which case (1) becomes

$$\mathbf{z}(k+1) = \Phi\mathbf{z}(k) + (\mathbf{I} - \Phi)\mathrm{T} + \mathbf{w}(k), \quad (3)$$

As shown in Fig. 2, the state variable $\mathbf{z}$ is chosen to be the joint positions of jaw, upper lip, lower lip, tongue tip and tongue dorsum (each with x and y positions), i.e.,

$$\mathbf{z} = [Jx,\ Jy,\ ULx,\ ULy,\ LLx,\ LLy,\ TTx,\ TTy,\ TDx,\ TDy]^T. \quad (4)$$

The purpose of matrix Φ is to represent our assumptions regarding the interrelationships between $\mathbf{z}(k)$ and $\mathbf{z}(k+1)$. In particular, we can identify approximate conditional independences among articulators; for example the movement of the upper lip, related to that of the lower lip, is largely independent of the jaw position. One possible Φ matrix, determined after exploring all such conditional independence

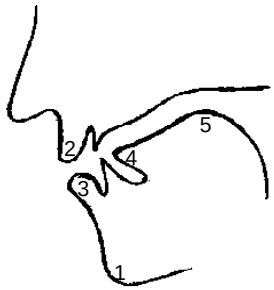

Fig. 2. Measured articulators: 1 - Jaw, 2 - Upper Lip, 3 - Lower Lip, 4 - Tongue Tip, 5 - Tongue Dorsum.

relations, follows:

$$\Phi = \begin{bmatrix} \phi_{00} & \phi_{01} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \phi_{10} & \phi_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{22} & \phi_{23} & \phi_{24} & \phi_{25} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{32} & \phi_{33} & \phi_{34} & \phi_{35} & 0 & 0 & 0 & 0 \\ \phi_{40} & \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44} & \phi_{45} & 0 & 0 & 0 & 0 \\ \phi_{50} & \phi_{51} & \phi_{52} & \phi_{53} & \phi_{54} & \phi_{55} & 0 & 0 & 0 & 0 \\ \phi_{60} & \phi_{61} & 0 & 0 & 0 & 0 & \phi_{66} & \phi_{67} & \phi_{68} & \phi_{69} \\ \phi_{70} & \phi_{71} & 0 & 0 & 0 & 0 & \phi_{76} & \phi_{77} & \phi_{78} & \phi_{79} \\ \phi_{80} & \phi_{81} & 0 & 0 & 0 & 0 & \phi_{86} & \phi_{87} & \phi_{88} & \phi_{89} \\ \phi_{90} & \phi_{91} & 0 & 0 & 0 & 0 & \phi_{96} & \phi_{97} & \phi_{98} & \phi_{99} \end{bmatrix}. \quad (5)$$

Clearly the detailed structure of this matrix will be the subject of continued research. Note that asserting this structure simultaneously reduces the number of parameters and makes the parameter estimation more robust. For the parameter estimation problem it is even more desirable for Φ to be block diagonal, achieved through a change of basis,

$$\mathbf{z} = [Jx,\ Jy,\ ULx,\ ULy,\ LLx - Jx,\ LLy - Jy, TTx - Jx,\ TTy - Jx,\ TDx - Jx,\ TDy - Jy]^T, \quad (6)$$

leading to a revised Φ matrix

$$\Phi = \begin{bmatrix} \phi_{00} & \phi_{01} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \phi_{10} & \phi_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{22} & \phi_{23} & \phi_{24} & \phi_{25} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{32} & \phi_{33} & \phi_{34} & \phi_{35} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{42} & \phi_{43} & \phi_{44} & \phi_{45} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{52} & \phi_{53} & \phi_{54} & \phi_{55} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{66} & \phi_{67} & \phi_{68} & \phi_{69} \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{76} & \phi_{77} & \phi_{78} & \phi_{79} \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{86} & \phi_{87} & \phi_{88} & \phi_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{96} & \phi_{97} & \phi_{98} & \phi_{99} \end{bmatrix}. \quad (7)$$

Finally, the form of h is extremely difficult to visualize, so deducing an appropriate representation is challenging. We have had some success [2] using a mixture linear

model; more recently Gao *et al* [3] used a MLP with one hidden layer and 100 neurons in the layer.

The very final stage of our speech production model is conversion from MFCCs to a speech waveform, which is carried out as a separate step.

3. MODEL PARAMETER LEARNING

The recent availability of the University of Wisconsin X-ray microbeam speech production database (UW-XRMB) allows us to train our model on simultaneously-recorded articulatory and acoustic data. In the event that the amount of such complete data is inadequate, model training can be supplemented using acoustic data alone under a more general EM framework.

We first assume the articulatory phone boundaries are known. Suppose that the collection of state variables $\mathbf{Z} = \{\mathbf{z}(0), \mathbf{z}(1), \ldots, \mathbf{z}(K)\}$, belonging to the same phone, are fully observable. Then the maximum likelihood estimates of Φ and T follow:

$$\hat{\Phi} = \mathbf{B}\mathbf{A}^{-1}, \tag{8}$$

$$\hat{\mathrm{T}} = (\Phi - \mathbf{I})^{-1} \cdot \left\{ \Phi \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k) \right] - \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k+1) \right\}. \tag{9}$$

where

$$\mathbf{A} = \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k) \right] \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k) \right]^T - \frac{1}{K} \sum_{k=0}^{K-1} \left[\mathbf{z}(k)\mathbf{z}(k)^T \right], \tag{10}$$

$$\mathbf{B} = \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k+1) \right] \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k) \right]^T - \frac{1}{K} \sum_{k=0}^{K-1} \left[\mathbf{z}(k+1)\mathbf{z}(k)^T \right]. \tag{11}$$

The above result applies to estimating a general, unconstrained matrix Φ. In the block-diagonal case (7) the above estimator is just applied separately to each block. Notably, the above estimator does *not* apply to the more interesting constrained case (5); we have derived such an estimator, however the result cannot be expressed in a succinct, closed form and the relatively complicated expressions are omitted.

As is typical, the assumption that $\mathbf{A}$ is nonsingular holds with sufficient training data; because of the efficiency of our model and the few degrees of freedom needed to parameterize it, only a very modest amount of data is required.

Articulator Boundary Estimation

Typically we don't know the *articulatory* boundaries where the targets switch values, as required by the parameter estimation of the state equation. In many cases the acoustic phone boundaries may be available, such as in the TIMIT or UW-XRMB databases, or they need to be suboptimally approximated [4]. Since each articulatory boundary normally lies within adjacent acoustic boundaries (especially for tongue, but less so for lips and velum), we can search through all frames within two acoustic boundaries P_{i-1} and P_i for the optimal articulatory boundary γ_i, i.e.,

$$\hat{\gamma}_i = \arg \min_{P_{i-1} < \gamma_i < P_i} \sum_{k=P_{i-1}}^{P_i} |\hat{\mathbf{z}}(k) - \mathbf{z}(k)|^2, \tag{12}$$

where

$$\hat{\mathbf{z}}(k+1) = \hat{\Phi}_i \hat{\mathbf{z}}(k) + (\mathbf{I} - \hat{\Phi}_i)\hat{\mathrm{T}}_i, \ \gamma_i \le k < \gamma_{i+1}, \tag{13}$$

$$\hat{\mathbf{z}}(0) = \mathbf{z}(0). \tag{14}$$

$\hat{\Phi}_i$ and $\hat{\mathrm{T}}_i$ are obtained by assuming a given γ_i, and γ_i itself varies between two adjacent acoustic boundaries.

Parameter Learning from Acoustic Data

The convenience of simultaneously-recorded articulatory and acoustic data is rare. Normally our model has to be based on acoustic data alone. One effective solution is to apply the EM algorithm which iteratively uses a Kalman smoother to compute the likelihood $L(\hat{\Phi}, \hat{\mathrm{T}})$, and then varies the parameters to maximize the likelihood function. Because our observation model is nonlinear, we actually apply the *extended* Kalman smoother, which requires the Jacobian of the nonlinear observation function $h[\,]$. For example, for a MLP that we have implemented with one hidden layer the Jacobian is

$$J_{mn} \equiv \frac{\partial \mathbf{o}_m}{\partial \mathbf{z}_n} = \sum_{i=0}^{I} W_{mi} w_{in} g' \left[\sum_{n=0}^{N} w_{in} \mathbf{z}_n \right], \tag{15}$$

where W and w are the weights of the MLP and g is the sigmoid function.

4. RESULTS AND FURTHER IMPROVEMENTS

Fig. 3 illustrates one example of fitting our linear dynamic state equation to the tongue-position trajectories of a simple sentence. We adopt the block diagonal structure from (7), and the articulatory phone boundaries are also learned automatically, based on (12). The fitted trajectories are produced by (13) and (14) with the optimal boundaries. Our model produces an excellent fit, especially considering that the model, containing only three-hundred degrees of freedom (parameters), is being used to fit a total of 4800 articulatory

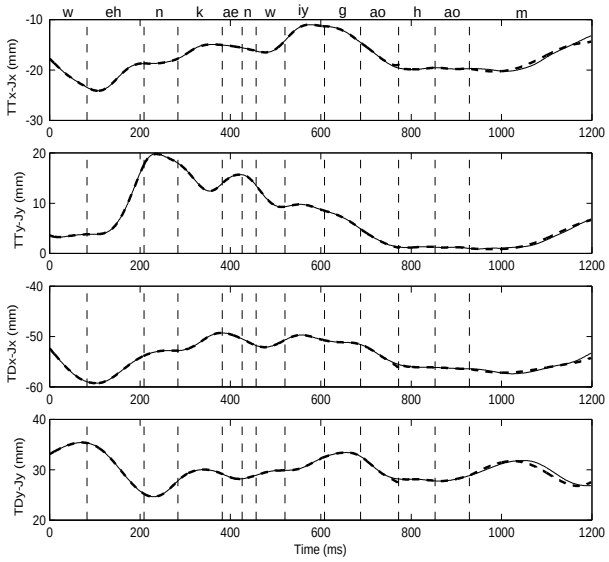

Fig. 3. Articulatory trajectory fitting for: *When can we go home*; solid line: real trajectory; broken line: fitted one.

data points. Two model-data mismatches can be observed: at the end of the sentence, where the articulators start going to the rest position after realizing the last phone, and the small jump at the boundary between /ao/ and /h/, which we postulate stems from having only zeroth-order (i.e., continuity) constraints in $\mathbf{z}(k)$, rather than higher-order derivative constraints. Arguably the model is promising at reflecting the true dynamics of articulatory movements. The following represent future developments:

1. Our state $\mathbf{z}$ contains only articulatory elements, so our first-order model (3) affects the continuity of $\mathbf{z}$, but not of its derivatives. We can enforce higher-order constraints, as appropriate, by including certain articulatory derivatives in the state.

2. Targets of different articulators do not switch synchronously during speech, as can be observed from Fig. 3, consistent with past work on overlapping of articulatory features [5, 6]. Since the articulatory boundary γ_i can be learned separately for each articulator, it should be possible to derive a mechanism to deduce overlapping articulatory features and provide them as input to the model.

3. Our current implementation fixes the articulator target position T to a single value for each phone. To better account for compensatory articulation, it may be more desirable to model the target as a multivariate distribution.

5. DISCUSSIONS

Recently in speech recognition, new models incorporating dynamic properties of human speech have been proposed [2, 3, 7] to better account for coarticulation and phonetic deduction phenomena in casual speech and to overcome some known limitations of HMM-based approaches. The speech production model described in this paper is intended for the same purpose by providing a more accurate model of human speech that directly takes into account articulatory dynamics. We have reached a stage of research where the model structure has been designed, parameter learning algorithms been developed, and the effectiveness of the algorithms been verified. Future work will focus on integrating all components of the model and evaluating the model on speech synthesis and recognition tasks.

6. ACKNOWLEDGMENTS

We would like to thank Dr. Jianwu Dang of ATR Japan for valuable discussions. This work has been supported by NSERC Canada and OGS Ontario.

7. REFERENCES

[1] R. D. Kent, S. G. Adams, and G. S. Turner, "Models of speech production," in *Principles of Experimental Phonetics*, N.J. Lass, Ed. 1996, pp. 2–45, Mosby.

[2] J. Z. Ma and L. Deng, "Spontaneous speech recognition using mixture linear models incorporating the target-directed propoerty," to appear in *IEEE Trans. Speech Audio Process.* in 2001.

[3] Y. Gao, R. Bakis, J. Huang, and B. Xiang, "Multi-stage coarticulation model combining articulatory, formant and cepstral features," in *Proc. ICSLP*, Beijing, 2000, pp. 25–28.

[4] 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.

[5] J. A. S. Kelso, E. L. Saltzman, and B. Tuller, "The dynamical perspectives on speech production: Data and theory," *Journal of Phonetics*, vol. 14, pp. 29–59, 1986.

[6] L. Deng and D. X. Sun, "A statistical spproach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acous. Soc. Am.*, vol. 95, no. 5, pp. 2702–2719, 1994.

[7] H. B. Richards and J. S. Bridle, "Acoustic-phonetic modelling using the hidden dynamic model," in *Proc. IChPS*, San Francisco, 1999, pp. 691–694.