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 introduced a new speech production model aiming at synthesizing natural speech in real-time by modeling the key dynamic properties of the articulators in a nonlinear state-space framework. The goal-oriented movement of the tongue tip, tongue dorsum, upper lip, lower lip and jaw are described in a linear state equation. The so produced articulatory trajectories combined with the effects of velum and larynx are mapped into acoustic features in the nonlinear observation equation. The input and output of the model are time-aligned phone sequence and speech waveform respectively. This speech production model can also be directly applied to speech recognition to better account for coarticulation and phonetic reduction phenomenon with considerably less parameters than the traditional HMM based approaches.

1. INTRODUCTION

The development of this model is motivated by many previous studies about human speech production in speech science. Although increasingly detailed and sophisticated models about how speech is generated in human speech production system has been developed in the past thirty years [1], they have made little impact on the progress of computer synthesized human speech. From a practical point of view, these models are either too complicated to implement, or lack the comprehensiveness in covering all classes of sounds, or both. On the other hand, the current *cut and paste* approach used in commercial speech synthesizers cannot provide the natural transitions between phonemes as the human articulatory system does.

In this paper, our modeling effort is concentrated on describing the key dynamic features of the articulators that are crucial to natural speech. We did not describe the underlying physiological mechanism that governs the movement of the articulators in our model, but rather choose a target-oriented linear state equation with its parameters learned from real speech data (thus the name *functional* in the title). Obviously this results in a fairly simple implementation. Such simplicity also provides the convenience of applying it directly to speech recognition, which will be discussed at the end.

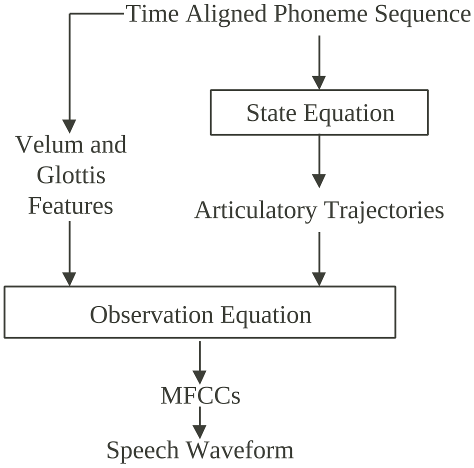
Time Aligned Phoneme Sequence
State Equation
Velum and Glottis Features
Articulatory Trajectories
Observation Equation
MFCCs
Speech Waveform

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

Since our model is still at an early stage of development, the focus of this paper is to present the new ideas we have. The remaining of the paper is organized as follows: The model is described in some detail in Section2. In Section 3, methods for learning parameters in the model are discussed. Section 4 points out some possible further improvements of our model. And finally the potential application of the model in speech recognition are discussed in Section 5.

2. MODEL DESCRIPTION

A block diagram of our speech production model is shown in figure 1. Both the underlying articulatory dynamics and the generated acoustic features are described by the follow-

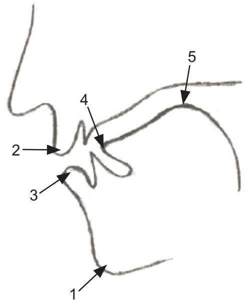

Fig. 2. Articulators described in the state equation: 1. Jaw; 2. Upper Lip; 3. Lower Lip; 4. Tongue Tip; 5. Tongue Dorsum.

ing state-space model:

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

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

Here the state variable $\mathbf{z}(k)$ is the position of the key articulators at time k, and these articulators are shown in figure 2. $\mathbf{o}(k)$ is the acoustic feature generated at the same time. The acoustic feature used in our model is the Mel-frequency cepstral coefficients (MFCCs). Φ is a matrix describing articulatory dynamics. Examples of how to incorporate our prior knowledge about speech into this matrix will be seen shortly. T is the target position of the articulators, and Ψ describes the control effect of the targets on the articulatory movement. These three parameters are all phone-dependent, i.e., they switch values at each phone boundary (Φ and Ψ may be tied for broader classes of phones in the actual implementation). Due to the well-known forward-anticipation property of the articulators, the boundaries for these parameters (especially T) should happen earlier than the actual acoustic boundaries. Exactly how early it should be will be learned from real articulatory data. The nonlinear function h in the observation equation represents the articulatory-to-acoustic mapping. Both $\mathbf{w}(k)$ and $\mathbf{v}(k)$ are discrete-time white Gaussian noise, with time-invariant covariance matrix $\mathbf{Q}$ and $\mathbf{R}$ respectively.

The state equation (1) must satisfy the asymptotic, target-oriented property, i.e., when $k \to \infty, \mathbf{z}(k) \to \mathrm{T}$. This requires some special relationship between Φ and Ψ. An easy and convenient choice is to let $\Psi = \mathbf{I} - \Phi$, and this is used in our current implementation. For later reference, we rewrite the state-space model for this special choice as follows:

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

$$\mathbf{o}(k) = h\left[\mathbf{z}(k)\right] + \mathbf{v}(k). \quad (4)$$

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

$$\begin{aligned}\mathbf{z} = {} & [Jx,\ Jy,\ ULx,\ ULy,\ LLx,\ LLy, \\ & TTx,\ TTy,\ TDx,\ TDy]^T. \end{aligned} \quad (5)$$

The function of matrix Φ is mainly to describe the relationship between $\mathbf{z}(k)$ and $\mathbf{z}(k+1)$. We can preset some elements to zero by noticing the approximate conditional independence among articulators. For example, the movement of the upper lip is related to that of the lower lip, but is largely independent of the jaw position given the position of the lower lip. The resulting Φ matrix after exploring all the conditional independent relations is shown as follows:

$$\Phi = \begin{bmatrix} \phi_{00} & \phi_{01} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \phi_{10} & \phi_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{22} & \phi_{23} & \phi_{24} & \phi_{25} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{32} & \phi_{33} & \phi_{34} & \phi_{35} & 0 & 0 & 0 & 0 \\ \phi_{40} & \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44} & \phi_{45} & 0 & 0 & 0 & 0 \\ \phi_{50} & \phi_{51} & \phi_{52} & \phi_{53} & \phi_{54} & \phi_{55} & 0 & 0 & 0 & 0 \\ \phi_{60} & \phi_{61} & 0 & 0 & 0 & 0 & \phi_{66} & \phi_{67} & \phi_{68} & \phi_{69} \\ \phi_{70} & \phi_{71} & 0 & 0 & 0 & 0 & \phi_{76} & \phi_{77} & \phi_{78} & \phi_{79} \\ \phi_{80} & \phi_{81} & 0 & 0 & 0 & 0 & \phi_{86} & \phi_{87} & \phi_{88} & \phi_{89} \\ \phi_{90} & \phi_{91} & 0 & 0 & 0 & 0 & \phi_{96} & \phi_{97} & \phi_{98} & \phi_{99} \end{bmatrix}. \quad (6)$$

Doing so not only reduces the number of parameters but also make the parameter estimation more robust. It will be even more desirable to have Φ as a block diagonal matrix. We approximately achieve this by choosing a slightly different state variable:

$$\begin{aligned}\mathbf{z} = {} & [Jx,\ Jy,\ ULx,\ ULy,\ LLx - Jx,\ LLy - Jy, \\ & TTx - Jx,\ TTy - Jx,\ TDx - Jx,\ TDy - Jy]^T, \end{aligned} \quad (7)$$

and the resulting new Φ matrix is:

$$\Phi = \begin{bmatrix} \phi_{00} & \phi_{01} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \phi_{10} & \phi_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{22} & \phi_{23} & \phi_{24} & \phi_{25} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{32} & \phi_{33} & \phi_{34} & \phi_{35} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{42} & \phi_{43} & \phi_{44} & \phi_{45} & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{52} & \phi_{53} & \phi_{54} & \phi_{55} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{66} & \phi_{67} & \phi_{68} & \phi_{69} \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{76} & \phi_{77} & \phi_{78} & \phi_{79} \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{86} & \phi_{87} & \phi_{88} & \phi_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_{96} & \phi_{97} & \phi_{98} & \phi_{99} \end{bmatrix}. \quad (8)$$

Currently the dynamics of the velum and larynx are not considered. They are treated as binary variables and included when doing the articulatory-to-acoustic mapping via nonlinear function h. The exact form of h is not decided at the moment, and it will be chosen to be whatever nonlinear function approximator that works the best. We have previous experience of using a mixture linear model [2] to map from the vocal-tract-resonance (VTR) to MFCCs with good results. More recently, a similar mapping was successfully carried out by Gao et al [3] using a MLP with one hidden layer and 100 neurons in the layer. The output of the nonlinear mapping are also MFCCs in our model, while the conversion from MFCCs to speech waveform is carried out independently.

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 articulatory and acoustic data recorded at the same time. When this kind of complete data is not adequate, e.g., when the model has to be adjusted to meet some specific requirements, model training can also be supplemented by acoustic data alone under the generic EM framework.

3.1. Parameter learning in the state equation

We assume that phone boundaries are available in the following derivation. Methods for determining phone boundaries under various conditions are discussed in 3.3. Assume in (3) the state variable $\mathbf{Z} = \{\mathbf{z}(0), \mathbf{z}(1), \ldots, \mathbf{z}(K)\}$ belonging to the same phone are fully observable, the maximum likelihood (ML) estimates (or equivalently the MSE estimates in the case of Gaussian noise) of Φ and T under some reasonable matrix nonsingular assumptions can be obtained as follows:

$$\hat{\Phi} = \mathbf{B}\mathbf{A}^{-1}, \tag{9}$$

$$\hat{\mathrm{T}} = (\Phi - \mathbf{I})^{-1} \cdot \left\{ \Phi \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k) \right] - \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k+1) \right\}. \tag{10}$$

where

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

$$\mathbf{B} = \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k+1) \right] \left[\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}(k) \right]^T - \frac{1}{K} \sum_{k=0}^{K-1} \left[\mathbf{z}(k+1)\mathbf{z}(k)^T \right]. \tag{12}$$

The above result is for general unconstrained Φ matrix case. When we have more than one training token corresponding to the same phone, a double summation has to be used in calculating matrix $\mathbf{A}$ and $\mathbf{B}$. Roughly speaking, $\mathbf{A}$ will be nonsingular as long as we have a reasonable amount of training data.

3.2. Parameter learning in the observation equation

The relationship between articulatory positions and acoustic features is treated to be a nonlinear static mapping. The parameter estimation problem is the same as training the nonlinear function approximator of the choice. For example, if we choose to use a MLP, then we just train the parameters of the MLP at this stage. The only requirement is that the Jacobian of the function approximator must be computable so that EM algorithm can be used to train the model parameters with acoustic data alone.

3.3. Determination of the phone boundaries

When training the parameters in the state equation, we require the knowledge of *articulatory* boundaries, i.e., where the articulatory targets switch their values. Usually this is not available, but sometimes the acoustic phone boundaries are available, such as in the well-known TIMIT database. For the UW-XRMB database, we also hand-labeled some phone boundaries. Since the articulatory boundary must lie within two acoustic boundaries, and the number of frames within a phone is relatively small, the articulatory boundary can be determined by exhaustive search, i.e., we search through all the possible boundaries and pick up the one with maximum likelihood, or equivalently the minimum MSE. When the acoustic boundaries are not available, this becomes a very difficult problem and we have to use some approximations to search for suboptimal solutions, such as the algorithms used in our previous study of a VTR dynamic model for speech recognition [4].

3.4. Model training with acoustic data alone

It is a real luxury to have simultaneously recorded articulatory and acoustic data available and inevitably we have to adjust/adapt our model based on acoustic data alone in practice. The solution is the general EM algorithm that has been used extensively especially in speech recognition. The steps are outlined as follows for our model:

- Initialize the parameters with those trained previously when both articulatory and acoustic data are available, estimate $\mathbf{z}(k)$ based on $\mathbf{o}(k)$ using (extended) Kalman smoother.
- Reestimate all the parameters based on $\mathbf{o}(k)$ and the estimated $\mathbf{z}(k)$.
- Do iteration until convergence occurs (measured by the likelihood of the parameters) or maximum number of iterations are reached.

4. FURTHER IMPROVEMENTS

The model we presented here is still fairly crude from various points of view, and some possible further improvements are as follows:

1. The continuity of the articulatory trajectories are ensured in our model, but in order to have smooth trajectories, we have to impose continuity on the first order derivative. This can be done by including the difference of successive articulatory positions in the state variable and force them to change smoothly.

2. Different articulators are not moving synchronously during speech production, while we have forced them to do so in the current control strategy. A better one is to use time aligned overlapping articulatory features [5, 6] as the input to our model. Since the relative timing information about the articulators can also be learned from the UW-XRMB database, a mechanism to convert phone sequence to overlapping articulatory features automatically is possible in the future.

3. We have fixed the target position of the articulators to a single value corresponding to each phone. To better account for phenomena such as compensatory articulation, modeling the target position as a probability distribution is more desirable.

The key thing to keep in mind is that we have to keep a balance between model accuracy and model complexity so that this model can be of practical significance.

5. APPLICATIONS IN SPEECH RECOGNITION

As people all start to realize the limitations of HMM-based approach for speech recognition, new models incorporating some dynamic properties of speech has been proposed in recent years [2, 3, 7] to better account for coarticulation and phonetic deduction phenomenon in spontaneous and casual speech. The speech production model described here fits this purpose perfectly. It not only overcomes many inherent inaccuracies of the HMM in modeling human speech, which is reflected by its ability to generate natural speech, but also provides a parsimonious set of parameters comparing to those of HMM (expected to be one to two orders of magnitude less).

It is not hard to see that the training process for both speech production and speech recognition is exactly the same. Of course we don't have articulatory data available for speech databases that are designed for speech recognition purpose, but we can use the values trained from the UW-XRMB speech production database as the initial values. Since we have good initial values with much less parameters than HMM, the performance of the speech recognizer is expected to depend much less on the quality and quantity of the training data, which makes it appealing for many practical applications.

However, the recognition phase of speech recognition is much harder than the voice generation phase of speech production. This problem can be simplified if N-best evaluation is used, which is based on the recognition result of a HMM recognizer. In generally, we have to solve the problem of searching for optimal phone boundaries for a test utterance. Some approximate methods of doing so has been developed in our previous work [4].

6. ACKNOWLEDGMENTS

Leo Lee would like to thank Dr. Jianwu Dang of ATR Japan for valuable experience and insights gained from previous cooperation on developing a 2D model for articulatory movement.

7. REFERENCES

[1] Ray D. Kent, Scoot G. Adams, and Greg S. Turner, *Principles of Experimental Phonetics*, chapter Models of Speech Production, pp. 2–45, Mosby, 1996, Editied by N.J. Lass.

[2] Jeff Z. Ma and Li Deng, "Spontaneous speech recognition using mixture linear models incorporating the target-directed propoerty," Submitted to *IEEE Trans. Speech Audio Process.* in December 1999.

[3] Yuqing Gao, Raimo Bakis, Jin Huang, and Bing Xiang, "Multistage coarticulation model combining articulatory, formant and cepstral features," in *Proc. ICSLP*, Beijing, 2000.

[4] Jeff Z. Ma and Li 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] Li Deng and Don 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 John S. Bridle, "Acoustic-phonetic modelling using the hidden dynamic model," in *Proc. IChPS*, San Francisco, 1999, pp. 691–694.