Simulating anticipatory coarticulation in VCV utterances with a gestural articulatory synthesizer

Asterios Toutios and Shrikanth Narayanan, University of Southern California
{toutios,shri}@sipi.usc.edu

In [1] we described a framework for articulatory synthesis wherein articulatory models derived from real-time MRI data [2] were animated using dynamical systems [3]. With the present work we discuss an improved implementation of these dynamical systems which allows multiple articulatory gestures to operate simultaneously on the same vocal tract component in an efficient and stable way. We demonstrate the utility of this implementation by simulating anticipatory vowel-to-vowel coarticulation across an intervening consonant, as classically described by Öhman [4].

The articulatory models in [1] represent the vocal tract shaping at any point in time as a linear combination of 8 speaker-specific components, weighted by a (dynamically changing) array of *parameters* $\mathbf{w}$. These give rise to measurable *constrictions* in the vocal tract. We consider constrictions at 6 places of articulation (bilabial, velopharyngeal, alveolar, palatal, velar, and pharyngeal), with their degrees (distances between articulators) forming an array $\mathbf{z}$. Analysis of several minutes of real-time MRI data (at 83 frames per second) determines a set of clusters wherein the mapping from parameters to degrees of constriction is linear, i.e. $\mathbf{z} = \mathbf{G}(\mathbf{w}) = F * \mathbf{w} + \mathbf{z_c}$ where F is a 6×8 matrix. Given the linearity, the jacobian J of the mapping within each cluster, with its derivative $\dot{J}$ and pseudoinverse J^*, are readily available.

The dynamical system [1, 3] that governs the trajectories of the articulatory parameters is:

$$\ddot{\mathbf{w}} = J^*(-BJ\dot{\mathbf{w}} - K(\mathbf{G}(\mathbf{w}) - \mathbf{z}_0)) - J^*\dot{J}\dot{\mathbf{w}} - (I_8 - J^*J)B_N\mathbf{w} - G_N(-B_N\mathbf{w} - K_N\mathbf{w})$$

where G_N, B_N and K_N are parameters of the *neural attractor* (see [1, 3] for details). The stiffness K and damping B matrices are to be set dynamically as functions of the target utterance. In practice, setting an array of 6 *natural frequencies* $\omega_\mathbf{o}$ (each corresponding to a place of articulation) fully determines K and B. The array $\mathbf{z}_o$ of *target* constriction degrees is also a function of the utterance. At each point in time then, the system can be characterized by an array of 6 tuples (ω_o, z_o). These can be visualized in a way that is reminiscent of a *gestural score* [5], as shown in Fig. 1 (note that blocks when ω_o is zero are omitted – the corresponding target values in these cases need not be set).

So far, we had been implementing this dynamical system using MATLAB's `ode45` functions. This led to unstable solutions when, at any given time, more than one element of the natural frequency array were non-zero. To address this problem, assuming a time-step h between two consecutive samples of (ω_o, z_o), replace the derivatives by finite differences:

$$\dot{\mathbf{w}} \leftarrow (\mathbf{w}[n] - \mathbf{w}[n-1])/h \qquad \ddot{\mathbf{w}} \leftarrow (\mathbf{w}[n] - 2\mathbf{w}[n-1] + \mathbf{w}[n-2])/h^2$$

After some algebra, we get a linear system of the form[1]

$$(I_8 + hA_1 + h_2A_2)\mathbf{w}[n] = \mathbf{w}[n-2] + 2\mathbf{w}[n-1] + hA_1\mathbf{w}[n-1] + h^2J^*K(\mathbf{z}_o - \mathbf{z}_c)$$

where the matrix sum on the left side is 8×8 and in general invertible. Thus, given a (ω_o, z_o) specification and two initial samples of parameter arrays, subsequent samples can be calculated step-wise by solving the above linear system[2].

As a first test of this method, we tried to simulate anticipatory vowel-to-vowel coarticulation across an intervening consonant. Öhman [4] has suggested that in a V_1C_{V2} utterance, V_2 will influence the vocal

[1] with $A_1 = -J^*BJ - J^*\dot{J} - B_N + J^*JB_N - G_NB_N$, $A_2 = -G_NK_N - J^*KF$

[2] At each step, $F, \mathbf{z}_c, J, J^*$ and $\dot{J}$ are set to the values corresponding to the cluster wherein $\mathbf{w}[n-1]$ lies.

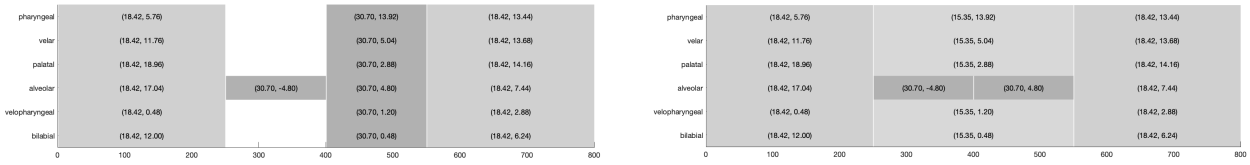

Figure 1: (ω_o, z_o)-specifications for synthesis of /adu/ without (left) and with (right) considering anticipatory co-articulation. (Entries where $\omega_o = 0$ are left blank)

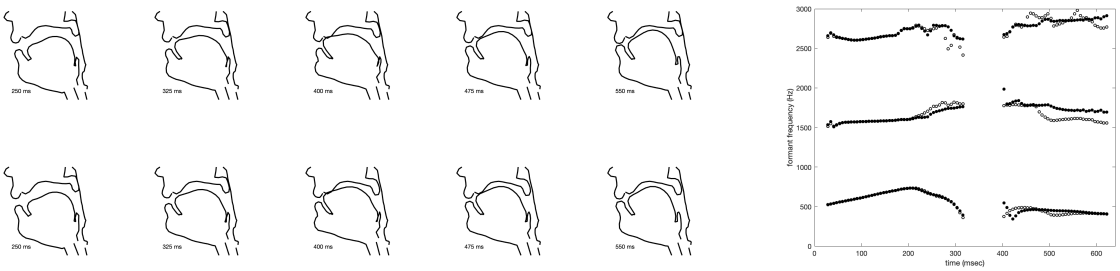

Figure 2: Left: Snapshots of vocal tract dynamics at the time points indicated, synthesized without (top sequence) and with (bottom) implementing anticipatory coarticulation. Right: Formant trajectories without (open circles) and with (filled circles) anticipatory coarticulation.

tract shaping even before C is produced. We considered the two (ω_o, z_o) configurations show in Fig.1. Both of them aim to synthesize /adu/ starting and ending at a scwha position, with only the second configuration including anticipatory coarticulation. (Note that targets for vowels are based on MRI measurements, while natural frequencies are calculated as functions of the duration of the gesture [1]). The resulting vocal tract dynamics are shown in Fig.2; note the important differences in vocal tract shaping at the second (V_1) and third (C) time points between the two sequences.

We followed the rest of the pipeline described in [1] to synthesize audio corresponding the two described dynamic vocal tract configuration. Formant trajectories extracted with Praat are shown in Fig.2, suggesting that the inclusion of anticipatory coarticulation may improve their smoothness. Audio files can be found online.[3].

References

[1] R. Alexander, T. Sorensen, A. Toutios, and S. Narayanan. A modular architecture for articulatory synthesis from gestural specification. *The Journal of the Acoustical Society of America*, 146(6):4458–4471, 2019.

[2] A. Toutios, D. Byrd, L. Goldstein, and S. Narayanan. Advances in vocal tract imaging and analysis. In W. Katz and P. Assmann, editors, *The Routledge Handbook of Phonetics*. Routledge, London and New York, 2019.

[3] E. L. Saltzman and K. G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382, 1989.

[4] S. Öhman. Coarticulation in VCV utterances: spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1):151–168, 1966.

[5] C. P. Browman and L. Goldstein. Articulatory Phonology: An overview. *Phonetica*, 49:155–180, 1992.

[3] http://sail.usc.edu/span/issp2020vcv/. It is the authors' impression that the inclusion of vowel-to-vowel anticipatory coarticulation improves significantly the naturalness of synthesis