VCV Synthesis using Task Dynamics to Animate a Factor-Based Articulatory Model

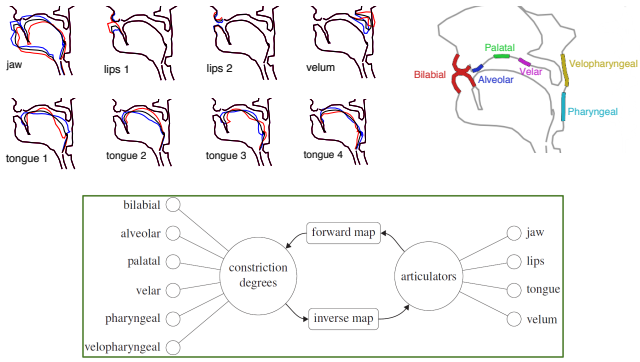
Rachel Alexander, Tanner Sorensen, Asterios Toutios, Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, USA

Introduction

- **Articulatory synthesis:** Synthesis of speech acoustics by simulation of the physics of the propagation of sound in the vocal tract and the dynamics of the vocal tract-shaping. The time-domain simulation was developed by Maeda [1]
- **Real-time MRI** (rtMRI) provides high-speed imaging data of the mid-sagittal vocal tract. Advances in automatic air-tissue segmentation have enabled the development of individualized, speaker-specific, articulatory models [2]
- **Dynamical systems** deform such factor-based models [3] in order to achieve constrictions at various places of articulation. This work extends that framework to animate the articulatory model for **vowel-consonant-vowel sequences**

Methodology

- A **factor analysis** [2] was applied to the segmented rtMRI videos to determine a set of constant factors that correspond to speaker-specific linguistically meaningful articulatory components
- Each configuration is then compactly represented by a vector of **articulatory parameters** that correspond to the degree of deformation of each factor, and can be used to accurately reconstruct the vocal tract.
- Six **constriction degrees** were defined [3] to determine the distance between surfaces of active and passive articulators, and related to the articulatory parameters described above using a locally-linear map



Left: deformations of articulatory components
Right: constriction degrees at six places of articulation
Bottom: forward map relates articulatory parameters and constriction degrees

- **Dynamical systems model:** finds optimal deformation from initial vocal-tract shape represented by appropriate articulatory parameters, to target constriction represented by appropriate tract variables.

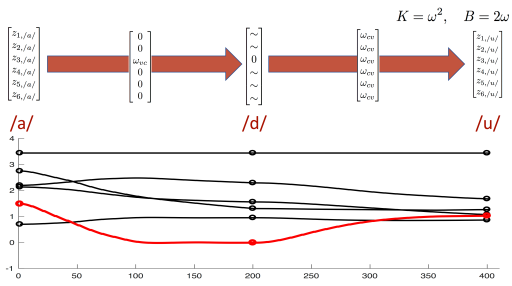
Change in constriction degrees: $\ddot{\mathbf{z}} = -\mathbf{K}(\mathbf{z} - \mathbf{z}_0) - \mathbf{B}\dot{\mathbf{z}}$

(stiffness, target, damping)

Change in articulatory parameters: $\ddot{\mathbf{w}} = \mathbf{J}^{\dagger}[\mathbf{B}\mathbf{J}\dot{\mathbf{w}} - \mathbf{K}(g(\mathbf{w}) - \mathbf{z}_0) - \dot{\mathbf{J}}\dot{\mathbf{w}}]$

VCV Specifications

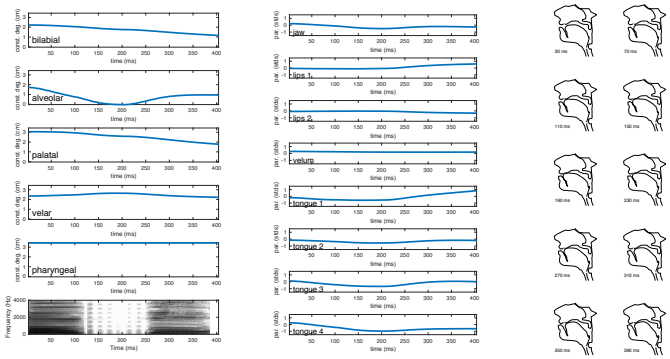
- We use two dynamical systems to create a VCV; the first begins with a **vowel V1**, obtained as the average set of articulatory parameters for all utterances in the dataset and converted to constriction degrees using the linear map
- The dynamical system deforms that configuration to achieve a specified constriction for the **consonant C**
- This second system deforms the consonant configuration into the vocal-tract shape for the **vowel V2** (again found as the average in the dataset)
- For both systems, we specify non-zero stiffness and damping coefficients only for the relevant constriction degrees that must be changed



Specifications to generate constriction degree trajectories

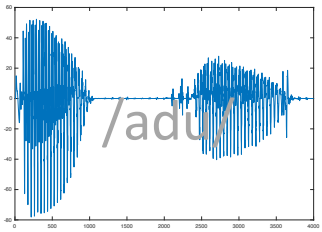
Synthesizing Acoustics

- The combination of both systems animates the articulatory model for the VCV sequence, which is then used to reconstruct the shape of the mid-sagittal slice



Left: Constriction degrees over time and spectrogram of synthesized signal
Middle: Articulatory parameter trajectories over time
Right: Reconstructed midsagittal slice

- **Area function dynamics** are obtained from the mid-sagittal slice dynamics using a simple alpha-beta model ($A=\alpha d^6$) with parameters specified by Maeda [4]
- **Glottal specifications** are developed empirically, consisting of an F0 trajectory, a slow-varying, and a fast-varying component [5]
- The synthesizer calculates the propagation of sound in a time-varying lumped electrical transmission-line network to produce the final speech signal
- We animated VCV sequences with combinations of **3 vowels** (/a/, /i/, /u/) and **3 voiced plosive consonants** (/b/, /d, /g/).



Synthesized speech signal for VCV sequence /adu/

Conclusion and Future Directions

- We have presented an initial architecture for synthesizing VCV sequences based on inputs to a dynamical system that are derived from a factor-based articulatory model and deformed to achieve a target consonantal constriction.
- **Additional consonants**: We can adjust the voicing trajectories and velopharyngeal port to produce voiceless and nasal consonants
- **Improving parameters**: rtMRI is up to 83 fps, which we can use to fit the parameters of the dynamical system and alpha-beta model to individual speakers
- **Speaker variability**: Animating this framework for different speakers provides an avenue to explore speaker variability through analysis by synthesis
- **Text to speech**: In the long term, we are considering the potential for generating the parameters for this architecture directly from text

References

[1] S. Maeda, "A digital simulation method of the vocal-tract system," Speech Communication, vol. 1, no. 3-4, pp. 199–229, 1982.
[2] A. Toutios and S. S. Narayanan, "Factor analysis of vocaltract outlines derived from real-time magnetic resonance imaging data," in International Congress of Phonetic Sciences (ICPhS), Glasgow, UK, Aug. 2015.
[3] T. Sorensen, A. Toutios, L. Goldstein, and S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time MRI," in Interspeech, San Francisco, CA, 2016.
[4] Maeda, S, "Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer", in Sound Patterns of Connected Speech: Description, Models and Explanation, A. Simpson and M. Patzold, Eds., 1996, pp. 145–164.
[5] Toutios, A.; Maeda, S. "Articulatory VCV synthesis from EMA data", Interspeech, Portland, Oregon, 2012.

Work supported by NIH grant R01DC007124 and NSF grant 1514544.
Synthesizer code and examples: http://sail.usc.edu/span/artsyn2017