Identification of stop consonants produced by an acoustically-driven model of a child-like vocal tract

1. Acoustically-driven vocal tract model

A model of child-like speech production has been developed in which vowels and consonants are specified as resonance deflection patterns, or RDPs. These deflection patterns are denoted as a set of three numbers, each of which can vary between -1 and 1; a negative value implies a downward shift in a resonance frequency whereas an upward shift results for positive value. The RDPs are transformed into a time-dependent deformation function that modifies the vocal tract configuration such that the specified RDP is achieved - thus, the vocal tract shape is "acoustically-driven."

Example Goal: produce /igA/

- Neutral vocal tract configuration - 3 year old "talker"

2. Simulation of VCV stimuli

With a goal of generating stimuli for a consonant identification experiment, a set of 24 VCVs were simulated based on the process demonstrated in Figure 2 for the 11 cm long vocal tract, but with various combinations of RDPs representing both vowels and consonants. The RDP settings are given in Tables 1a and 1b, and the VCVs resulting from the combination of RDP settings are listed in Table 1c. The IPA symbols embedded within the unconventional early brackets are used to differentiate vocal tract area functions and calculated resonance frequencies produced by the model from actual prescribed phonetic targets or transcriptions of real or synthetic talkers.

In addition to the 11 cm vocal tract (representative of a 3 year-old talker), four other "talkers" were also simulated producing the same 24 VCVs. For all talkers, vocal tract lengths, $L_t$, area function scale factors, $\alpha$, relative to the initial "Child 1," and fundamental frequency ranges, $f_0$, prescribed over the course of each VCV are given in Table 1d.

3. Consonant identification experiment

- There were 120 files total [3 RDPs x 4 vowel-vowel contexts x 2 voiced/unvoiced x 5 talkers] that were presented to listeners via the Alvin interface (Hilkebrand & Guyvert, 2005).
- The experiment was run in ten blocks. Within each block the condition of voicing and talker was constant.
- Each listener was seated in a sound booth and samples were played over a loudspeaker (Yamaha MSP5) set at a comfortable listening level. After hearing each sample, a listener used a computer mouse to choose "\(\text{b}\), "\(\text{d}\)," or "\(\text{g}\)" from buttons displayed on the computer screen.
- Within each block, the samples were played in random order, and each was repeated three times.
- Twelve listeners (1 male, 11 female, mean age = 20.3 yrs) were recruited to participate in the experiment, thus a total 4320 responses were collected (120 samples x 3 repetitions x 12 listeners). All listeners passed a hearing screening.
- The results are shown as confusion matrices in Table 2, and were collapsed across 1) all talkers, 2) three child talkers, and 3) two adult talkers. Each row indicates the target RDP, and each column the listener response.

Acknowledgements

Research supported by NIH R01-DC011275 and NSF BCS-1145011

Brad Story & Kate Bunting
Speech, Language, and Hearing Sciences
The University of Arizona

Presented at the 5th Meeting of the ASA and ASJ
30 Nov. 2016, Session 3y'SC
Honolulu, HI

Table 1: Parameters for simulation of VCVs

<table>
<thead>
<tr>
<th>Talkers</th>
<th>$L_t$ (cm)</th>
<th>$\alpha$</th>
<th>$f_0$ (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult 1</td>
<td>17.5</td>
<td>10.0</td>
<td>80 - 140</td>
</tr>
<tr>
<td>Adult 2</td>
<td>18.5</td>
<td>6.0</td>
<td>100 - 170</td>
</tr>
<tr>
<td>Child 1</td>
<td>11.0</td>
<td>4.0</td>
<td>240 - 480</td>
</tr>
<tr>
<td>Child 2</td>
<td>13.2</td>
<td>8.0</td>
<td>220 - 280</td>
</tr>
<tr>
<td>Child 3</td>
<td>15.4</td>
<td>1.0</td>
<td>255 - 270</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrices for consonant identification experiment.