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

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

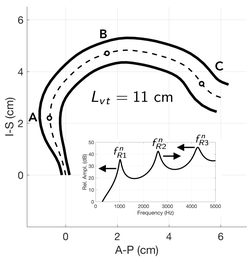
Presented at the 5th Meeting of the ASA and ASJ
30 Nov. 2016, Session 3pSC
Honolulu, HI

1. Acoustically-driven vocal tract model

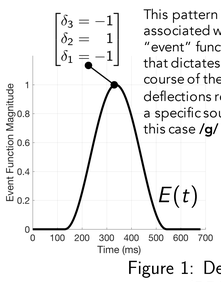
A model of child-like speech production has been developed in which vowels and consonants are specified as resonance deflection patterns, or RDPs. These deflection patterns are denoted as a set of three numbers, each of which can vary between -1 and 1; a negative value implies a downward shift in a resonance 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."

a. Neutral vocal tract configuration ∽ 3 year old "talker"

Example Goal: produce /əgʌ/

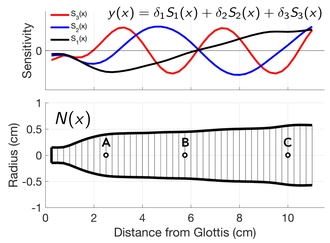


b. Specify a "resonance deflection pattern" (RDP)



This pattern is associated with an "event" function **E(t)** that dictates the time course of the resonance deflections required for a specific sound – in this case **/g/**

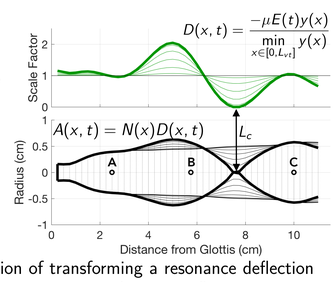
c. Calculate sensitivity functions to relate resonance frequency change to VT configuration



y(x) is a deformation function based on the linear combination of sensitivity functions scaled by the specified RDP.

N(x) is the neutral vocal tract configuration from which S_1, S_2, and S_3 were calculated.

d. Superimpose a time-dependent deformation pattern on the neutral VT configuration



D(x,t) is the time-dependent and normalized deformation function derived from **y(x)**.

μ determines the degree of constriction.

A(x,t) is the time-dependent vocal tract shape (area function) generated by deforming **N(x)** with **D(x,t)**.

Figure 1: Demonstration of transforming a resonance deflection pattern (RDP) to a time-varying vocal tract configuration.

Demonstration of VCV simulation

- Figure 2 demonstrates the steps in simulating a VCV, in this case /igɑ/.
- Three event functions are shown in part (a). Associated with the first (black) is an RDP that specifies that f_{R1} decrease and f_{R2} increase; the zero value for f_{R3} means that it is unspecified. This pattern is intended to produce the vowel /i/. The other two patterns are intended to produce /g/ and /ɑ/, respectively. Note that $\mu = 0.9$ for both vowels, and $\mu = 1.05$ for the stop consonant (i.e., $\mu \geq 1$ produces an occlusion).
- In part (b), the dashed lines are the resonances of $N(x)$, the gray (unbroken) lines are the resonances of $A(x, t)$ if only the vowel event functions were present (i.e., vowel-vowel transition), and the thick black lines are the deflected resonances for all three event functions combined. The red and blue filled regions indicate the direction of resonance shift away from the underlying vowel transition. The shaded gray region denotes the period of time during which the three event functions were overlapped.
- The resulting time-varying vocal tract shape for /igɑ/ is shown in part (c). The speech waveform produced with "TubeTalker" (Story, 2013) and the corresponding narrowband spectrogram of the simulated /igɑ/ utterance are given in part (d).

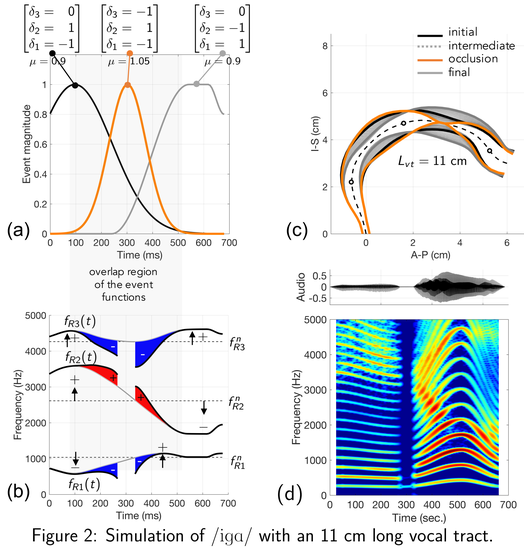


Figure 2: Simulation of /igɑ/ with an 11 cm long vocal tract.

Acknowledgements

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

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 curly 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_{vt}, area function scale factors, α, relative to the initial "Child 1," and fundamental frequency ranges, f_o, prescribed over the course of each VCV are given in Table 1d.

a. RDPs for vowels

	{ə/ʌ}	{i}	{ɑ}
δ_3	0	0	0
δ_2	0	1	-1
δ_1	0	-1	1
μ	0	0.9	0.9

b. RDPs for stop consonants

	{b/p}	{d/t}	{g/k}
δ_3	-1	1	-1
δ_2	-1	1	1
δ_1	-1	-1	-1
μ	1.05	1.05	1.05

c. V_1CV_2 combinations

$V_1V_2 \backslash C$	{b}	{d}	{g}	{p}	{t}	{k}
{ə}{ʌ}	əbʌ	ədʌ	əgʌ	əpʌ	ətʌ	əkʌ
{ə}{i}	əbi	ədi	əgi	əpi	əti	əki
{ə}{ɑ}	əbɑ	ədɑ	əgɑ	əpɑ	ətɑ	əkɑ
{i}{ɑ}	ibɑ	idɑ	igɑ	ipɑ	itɑ	ikɑ

d. VT lengths, α, & f_o range

	L_{vt} (cm)	α	f_o (Hz)
Child 1	11.0	1.0	240→430
Child 2	13.2	2.0	200→360
Child 3	15.4	2.5	155→270
Adult 1	17.5	3.0	80→145
Adult 2	18.5	4.0	70→120

Table 1: Parameters for simulation of VCVs

To demonstrate the structural and acoustic differences across the five "talkers," time-varying vocal tract configurations simulated for each talker's {igɑ} are shown in the top row of Figure 3. Below each vocal tract is a plot of the corresponding resonance frequencies for the {igɑ}. Although there are large differences in the absolute frequencies produced by the vocal tracts, the deflection patterns remain intact regardless of the vocal tract size.

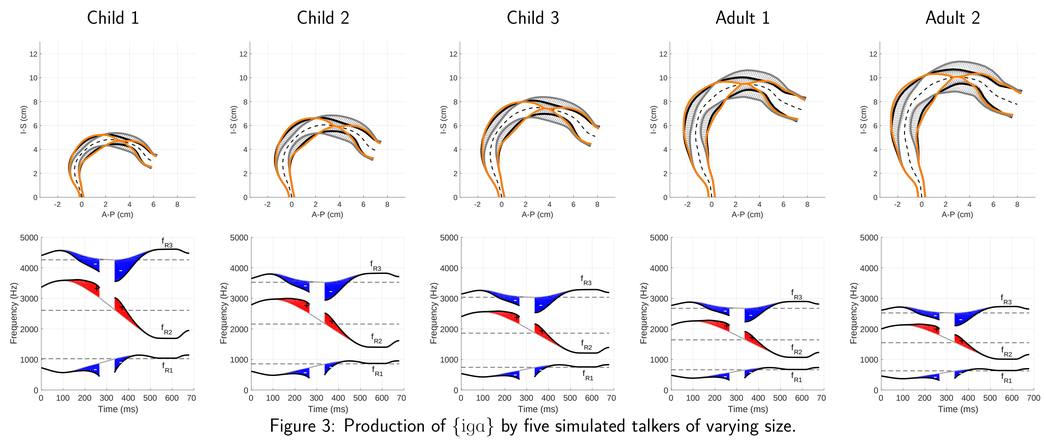


Figure 3: Production of {igɑ} by five simulated talkers of varying size.

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 (Hillenbrand & Gayvert, 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 MSP3) set at a comfortable listening level. After hearing each sample, a listener used a computer mouse to choose "b", "d", or "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, $\overline{\text{age}} = 20.3$ 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.

		Voiced			Unvoiced		
		[b]	[d]	[g]	[p]	[t]	[k]
All Talkers	{b}	**99.7**	0	0.3	**94.2**	5.4	0.4
	{d}	0.3	**95.8**	3.9	1.4	**91.5**	7.1
	{g}	2.5	9.2	**88.3**	9.9	7.8	**82.4**
Child Talkers	{b}	**99.8**	0	0.2	**93.8**	5.6	0.7
	{d}	0.5	**93.1**	6.5	1.9	**91.2**	6.9
	{g}	4.2	14.6	**81.3**	13.7	10.2	**76.2**
Adult Talkers	{b}	**99.6**	0.0	0.4	**94.8**	5.2	0
	{d}	0	**100.**	0	0.7	**92.0**	7.3
	{g}	0	0.7	**99.3**	4.2	4.2	**91.7**

Table 2: Confusion matrices for consonant identification experiment.