Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 2aSC: Linking Perception and Production (Poster Session)

2aSC10. Estimation of vocal tract area functions in children based on measurement of lip termination area and inverse acoustic mapping

Kate Bunton*, Brad H. Story and Ingo Titze

***Corresponding author's address: Speech, Language, and Hearing Sciences, University of Arizona, University of Arizona, Tucson, AZ 85721, bunton@u.arizona.edu**

Although vocal tract area functions for adult talkers can be acquired with medical imaging techniques such as Magnetic Resonance Imaging (MRI), similar information concerning children's vocal tracts during speech production is difficult to obtain. This is largely because the demanding nature of the data collection tasks is not suitable for children. The purpose of this study was to determine the feasibility of mapping formant frequencies measured from the [i, ae, a, u] vowels produced by three children (age range 4 to 6 years), to estimated vocal tract area functions. Formants were measured with a pitch-synchronous LPC approach, and the inverse mapping was based on calculations of acoustic sensitivity functions [Story, J. Acoust. Soc., Am., 119, 715-718]. In addition, the mapping was constrained by measuring the lip termination area from digital video frames collected simultaneously with the audio sample. Experimental results were augmented with speech simulations to provide some validation of the technique. [Research supported by NIH R01-DC011275]

Published by the Acoustical Society of America through the American Institute of Physics

©2013 Acoustical Society of America [DOI: 10.1121/1.4799532]
Received 18 Jan 2013; published 2 Jun 2013

Introduction

Vocal tract area functions for adult talkers can be acquired with medical imaging techniques such as Magnetic Resonance Imaging (MRI) (cf. Baer et al., 1991; Narayanan et al. 1995; Story et al., 1996). Similar information concerning children's vocal tracts during speech production is, however, difficult to obtain. This is largely because the data collection environment (e.g., MR scanner) can be intimidating for children, and the tasks required to obtain the data are fairly demanding. Acoustic reflectometry (Kamal, 2004) is a viable alternative technique but it requires the talker's lips to be constrained by a mouthpiece, thus making vowel production somewhat unnatural.

The purpose of this study was to determine the feasibility of mapping formant frequencies measured from the [i, æ, ɑ, u] vowels produced by three children (age range 4 to 6 years), to estimated vocal tract area functions. Because of the high fundamental frequency (F0) of children's vowels, formants were measured with a pitch synchronous LPC approach applied to recorded audio signals. The inverse mapping consisted of iteratively perturbing an initial vocal tract shape based on calculations of acoustic sensitivity functions such that the resonance frequencies match the measured values (Story, 2006). The mapping was also constrained by measuring the lip termination area from digital video frames collected simultaneously with the audio sample. A preliminary validation of the technique was first performed based on an adult male talker whose vocal tract shapes had been previously measured. The technique was then applied to a 6 year old child talker.

Method

The approach to estimating a vocal tract shape is shown schematically in Fig. 1. There are three main steps: 1) measurement of lip termination area and formant frequencies, 2) generating the initial vocal tract shape (the "seed" function) based on lip area, and 3) perturbing the seed function until the resulting shape produces resonance frequencies that match the measured formants. Each step is explained in more detail in the following sections making use of data collected from an adult male talker for whom area functions have been previously measured using MRI (i.e., the same talker as in Story et al., 1996). Thus, the area functions derived with the proposed methods can be compared to area functions measured with an independent and more direct technique.

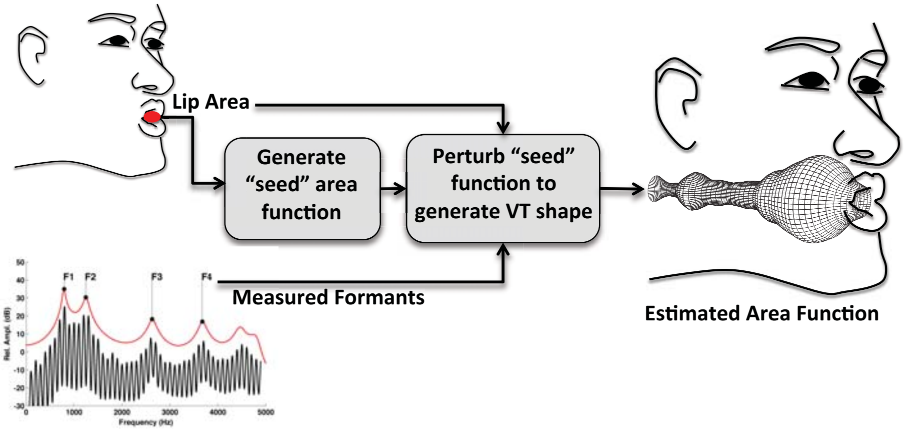


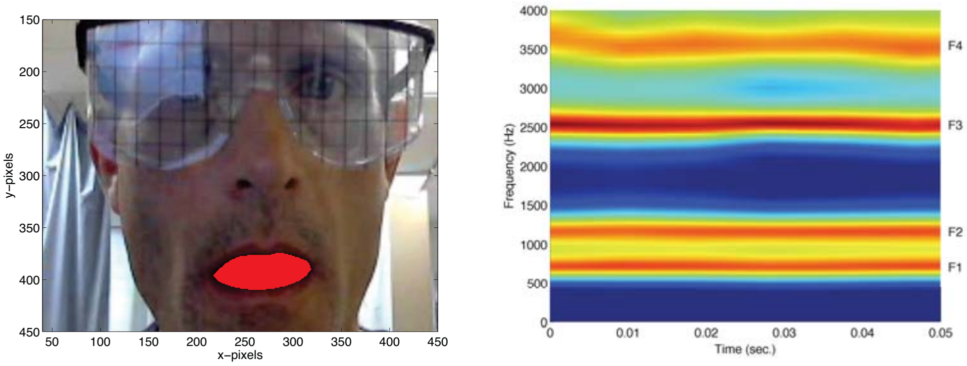
Figure 1: Schematic diagram of lip area and formants-to-area function mapping procedure.

Step 1: Measurements

Audio and video data were collected simultaneously with an AD Instruments data acquisition system (PowerLab 8/35) and LabChart software (2012). The audio channel was recorded with an AKG CS1000 condenser microphone and sampled at 40 kHz. A Logitech webcam was used to record video at 30 frames per second. To calibrate the dimensions of each video frame, the talker wore a pair of safety glasses on which a grid of 1 cm squares had been inscribed (see Fig. 2a).

A video frame was selected that coincided with the portion of the audio signal chosen for a particular vowel, and read into Matlab (Mathworks, 2011). A custom script was used that guides the user through calibration using the grid on the safety glasses, and choosing the lip termination region by a series of mouse clicks along the lip borders using the Matlab function "roipoly". The lip termination area is computed by determining the number of pixels enclosed by the resulting polygon and multiplying it by the calibration factor. The red region of the video frame in Fig. 2a shows the lip termination area of an adult male talker producing the vowel [ɑ]. In this case, the area is 4.6 cm^2.

The portion of the audio signal that corresponded to the video frame was analyzed with a pitch-synchronous formant tracker (Bunton & Story, 2011) and displayed in spectrographic form as shown in Fig. 2b for an [ɑ] vowel. This type of formant tracking is generally unnecessary for adult speech as used for demonstrating the technique, but is useful for accurately finding the formant frequencies in high-F0 speech signals and will be needed when the technique is applied to children's speech. The mean formant frequencies measured for the [ɑ] vowel in this example are $\mathscr{F}_n = [717, 1156, 2530, 3544]$ Hz.



(a) (b)

FIGURE 2: Analysis of lip termination area and formant frequencies. (a) Video frame for production of vowel [ɑ] by an adult male; the grid on the safety glasses is used to calibrate the image. (b) Spectrographic representation of the pitch-synchronous formant tracking method.

Step 2: Generating the Initial "Seed" Area Function

An initial vocal tract area function is a required starting point for the perturbation algorithm described in the next section. The simplest initial tract shape would be a uniform tube; i.e., a constant cross-sectional area A_t along the entire length of the vocal tract. It is known, however, that vocal area functions typically are narrow at the glottal end, in the region of the laryngeal vestibule. For adult males, this part of the tract, referred to as the *epilarynx*, is about 2.5 cm in length and roughly 0.5 cm2 in cross-section (Story et al., 1996, 2001). Thus, the initial area function is generated by imposing a constriction of length L_e and cross-sectional area A_e on a uniform tube of area A_t. In addition, the lip termination area A_m is used to modify the lip end of the initial tract shape such that the A_m is smoothly interpolated with a Gaussian function to A_t over a length L_m. The total length of the vocal tract L_{vt} also must determined. For this study, all vocal tract lengths were chosen *a priori* based on published values; for the adult male example $L_{vt} = 17.6$ cm. This is a limitation of the method, but more accurate means of estimating tract length could be employed in future developments (e.g. Fitch, 1997).

An example is shown in Fig. 3 for an adult male. The values of L_e, A_e and A_t were chosen with respect to previously reported MR-based vocal tract measurements (Story et al., 1996), and the lip termination area A_m was the value measured from the video frame in the previous section. This area function serves as the initial vocal tract configuration for the algorithm in the next section, and is referred to as the *seed* function $a_0(x)$.

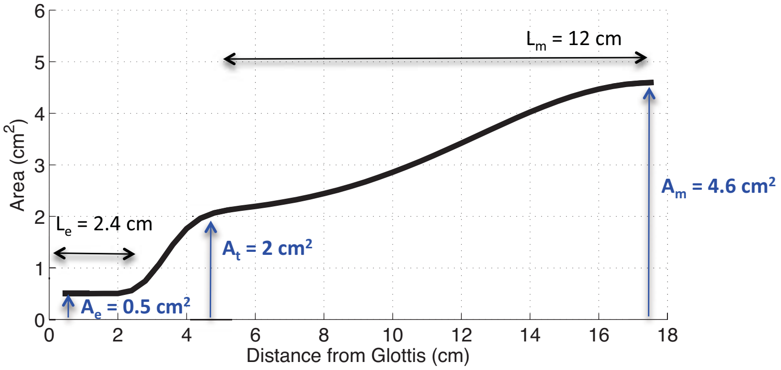


FIGURE 3: Seed function $a_0(x)$ for vowel [ɑ]. The area and length parameters are defined in the text.

Step 3: Estimation of the Vowel Area Function

The area function for a given vowel is estimated by an algorithm that perturbs the shape of the seed function until its resonance frequencies match the measured formants for that vowel. The algorithm used in the present study is identical to that reported in Story (2006) except that the lip termination area is constrained to the value measured from the video frame.

A change in cross-sectional area along the length of the vocal tract can be related an increase or decrease in the resonance frequencies (F_n) of the overall tract shape by an acoustic sensitivity function $S_n(x)$,

$$\frac{\Delta F_n}{F_n} = \sum_{i=1}^{N_{areas}} S_n(x) \frac{\Delta a(x)}{a(x)} \tag{1}$$

where $a(x)$ is the area function, $\Delta a(x)$ is a change in area, and x is the distance from the glottis. The sensitivity function can be calculated based on the difference of the kinetic (KE) and potential (PE) energies that are present along the vocal tract length for a given resonance n, and normalized by the total energy (TE),

$$S_n(x) = \frac{KE_n(x) - PE_n(i)}{TE_n} \quad n = 1,2,3\ldots \quad \text{and} \quad x = [0\ldots L_{vt}] \tag{2}$$

The sensitivity functions can be used to perturb the seed function such that,

$$a_1(x) = a_0(x) + \sum_{n=1}^{N_{fmts}} z_{n_0} S_{n_0}(x) \tag{3}$$

where $a_1(x)$ is a first perturbation of the seed function $a_0(x)$ by a scaled summation of the sensitivity functions of $a_0(x)$. The scaling factors z_{n_0} are calculated as the difference of the measured formants $\mathscr{F}_n$ and the calculated formants F_{n_0} of the seed function $a_0(x)$,

$$z_{n_0} = \left[\frac{\mathscr{F}_n - F_{n_0}}{F_{n_0}} \right] \tag{4}$$

In a strict sense, the sensitivity functions relate only small changes in area to small changes in a resonance frequency. Hence, the perturbation algorithm is an iterative process in which new sensitivity functions are calculated for each new perturbed area function such that,

$$a_{k+1}(x) = \begin{cases} a_k(x) + \sum_{n=1}^{N_{fmts}} z_{n_k} S_{n_k}(x) & \text{for } a_k(x) > 1 \text{ cm}^2 \\ \\ \exp(\ln(a_k(x)) + \ln(\sum_{n=1}^{N_{fmts}} z_{n_k} S_{n_k}(x) + 1)) & \text{for } a_k(x) \leq 1 \text{ cm}^2 \end{cases} . \tag{5}$$

The logarithmic function for areas less than 1cm^2 is needed to prevent the algorithm from generating unrealistically small areas. The scaling factors are also calculated at each iteration by,

$$z_{n_k} = \alpha \left[\frac{\mathscr{F}_n - F_{n_k}}{F_{n_k}} \right] \tag{6}$$

where α is an acceleration factor greater than one that can be used to increase the speed of the algorithm (see Story, 2006). The iterations continue until the differences between the measured and calculated formants are minimized.

Shown in Fig. 4 is the transformation of the seed function (Fig. 3) to an area function representative of an [ɑ] vowel based on the measured formants from Step 1. The thick black line is $a_0(x)$ and the red line is the area function at the 247th iteration. At this point, the calculated resonance frequencies are equal to the measured formants. The frequency values shown in the upper left portion of the plot indicate the change in each formant from the initial to final iteration.

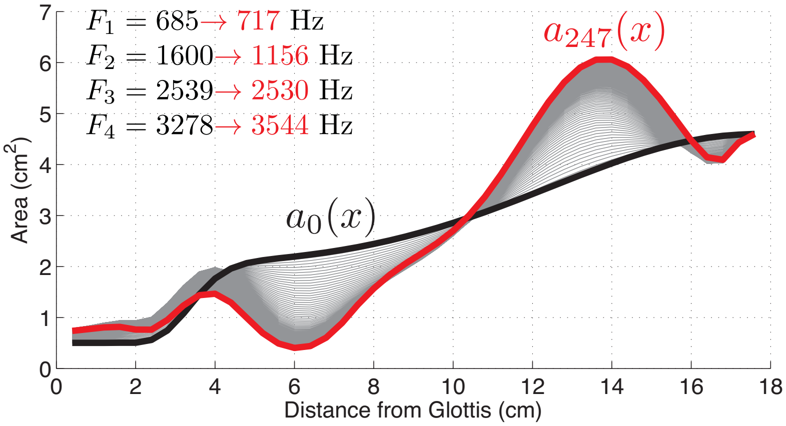
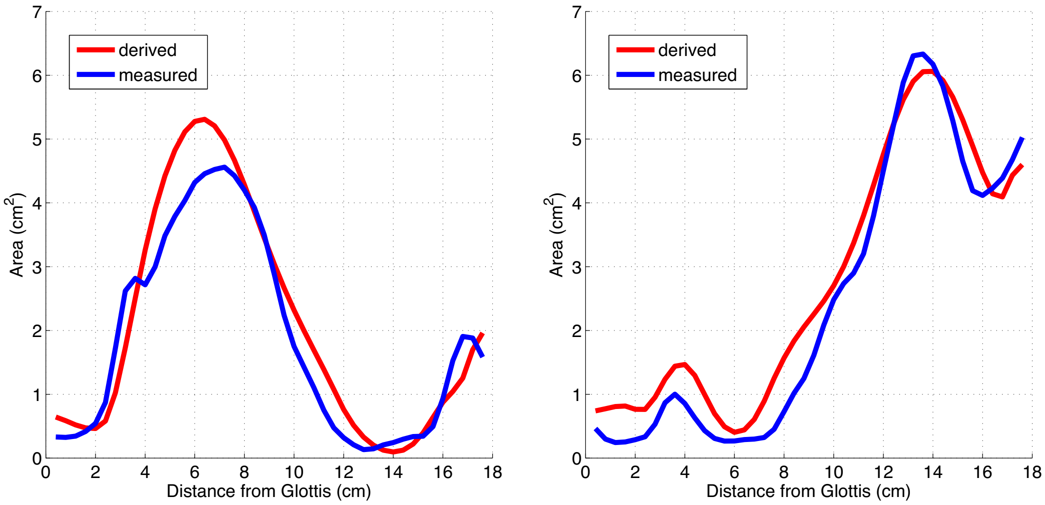


FIGURE 4: Iterative perturbation of the seed function for vowel [ɑ]. The thick black line is seed function and the red line is the area function at the final iteration.

Comparison of Derived and Measured Area Functions

Area functions derived for [i] and [ɑ] vowels produced by the adult male talker are compared to the same talker's previously measured area functions (based on MRI) in Fig. 5. The red line in Fig. 5a is the same as the final iteration plotted previously in Fig. 4, whereas the blue line is the MRI-based measurement from Story et al. (1996).

Although there are some obvious differences in the two area functions, particularly in the lower part of the vocal tract, the similarities are fairly encouraging considering that these are



[i]

[ɑ]

FIGURE 5: Comparisons of derived and measured area functions for two vowels .

based on entirely independent methods. The comparison of the area functions for the [i] vowels are equally encouraging. This provides some degree confidence for applying the method to children's vowel productions, as will be described in the next section.

APPLICATION OF THE METHOD TO CHILD SPEECH

Audio and video data were collected from a 6 year-old female talker in the same manner as described previously in "Step 1." The talker was asked to produce vowels in isolation, in "hVd" and "sVd" contexts, and vowel-vowel transitions. Analysis of three isolated vowels (i, æ, ɑ) will be presented here.

Figure 6 includes the analysis of video frames to obtain lip area for each vowel, the formant measurements based on the pitch-synchronous formant tracking method, and the results of mapping lip area and formants to vocal tract area functions; each column of plots corresponds to one of the vowels. Each image in the top row indicates the lip area as the region filled with red. The mean formants over the 0.2-0.25 second spectrographic segment shown for each vowel in the middle row were used as the target formants $\mathscr{F}_n$ in Eqn. 6 (for the [ɑ], the initial 0.06 seconds was not included in determining the mean F3). The seed area functions were generated by setting L_e, A_e and A_t to be 0.5 cm, 0.3 cm^2, and 1.5 cm^2, respectively. The overall vocal tract length was chosen to be $L_{vt} = 11.4$ cm and the lip area A_m was measured from the corresponding video frames.

The plots in the bottom row of Fig. 6 show the initial seed area functions for each vowel (thick black line), the final area functions (red line) whose resonances match the measured targets, and the iterations between the initial and final configurations. The area functions appear to be reasonably shaped for the specific vowels. It is interesting, however, that the cross-sectional areas, especially in the oral cavity are of the same magnitude as the adult male speaker in Fig. 5. Although this could be an artifact of the technique, the lip areas are known values since they were measured from the calibrated video frames, and are contiguous with the adjacent portion of each area function. This needs additional study with an independent method of imaging, but does suggest that the small size of the child's vocal tract system, relative to an adult, may be primarily due to shorter pharyngeal and oral cavity length, rather than decreased cross-sectional area.

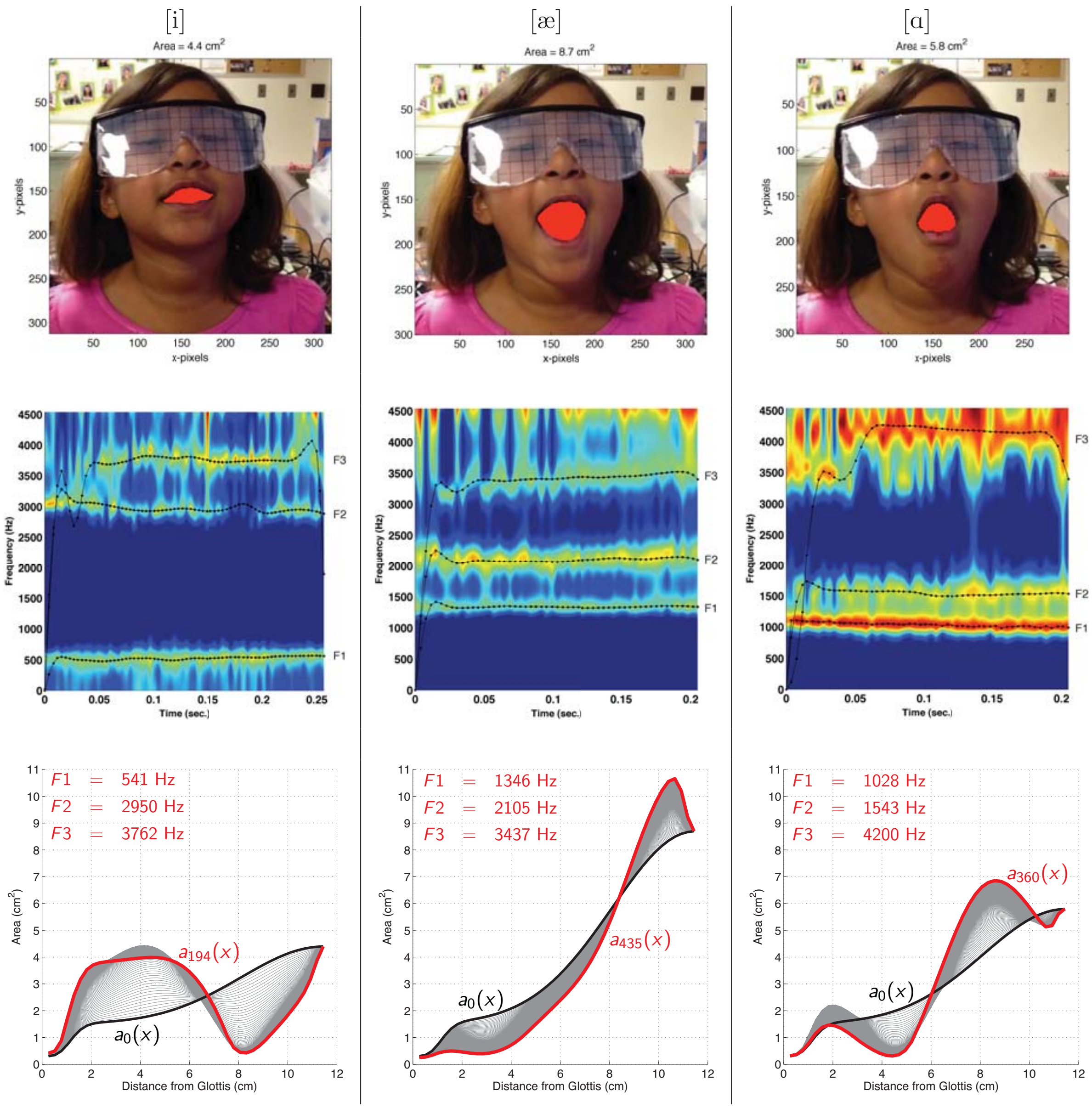


FIGURE 6: Analysis of three vowels produced by a 6 year-old female talker. In the top row are video frames from which lip area was measured (red regions). Pitch synchronous spectrograms are plotted in the middle row where the formant tracks are shown as black dots. The bottom row shows the area functions generated by the mapping technique, where the black line in each plot is the initial seed function, the red line is the final area function whose resonance frequencies are matched to the measured formant frequencies (listed in the upper left hand corner of each plot). The subscripts on the labels indicate the number of iterations required to match the target formants.

CONCLUSION

A technique has been introduced to measure lip area during vowel production, extract formant frequencies from high-fundamental frequency speech, and map this information to vocal tract area functions. The method was partially verified based on an adult male talker and associated MRI-based area functions. It was then applied to the speech of a 6 year-old female talker for isolated vowels [i, æ, ɑ]. The resulting area functions were plausible representations of the talker's vocal tract shapes. Future directions include applying the method to vowels in consonant context, and to vowel-vowel transitions to obtain time-varying area functions.

ACKNOWLEDGMENTS

Research supported by NIH R01-DC011275

REFERENCES

Baer, T., Gore, J.C., Gracco, L.C., Nye, P.W. (1991). “Analysis of vocal-tract shape and dimensions using MRI: Vowels”, J. Acoust. Soc. Am. 90, 799-828.

Fitch, W.T. (1997). “Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques”, J. Acoust. Soc. Am. 102(2 part 1), 1214-1222.

Kamal, I. (2004). “Acoustic Reflectometry of the nose and pharynx” Brown Walker Press, Boca Raton, FL.

LabChart version 7 Pro. Colorado Springs, Colorado: AD instruments, 2012

MATLAB version R2011b. Natick, Massachusetts: The Mathworks, 2011.

Narayanan, S.S., Alwan, A.A., Haker, K. (1995) “An articulatory study of fricative consonants using MRI”, J. Acoust. Soc. Am. 98(3), 1325-1347.

Story, B.H. (2006) “A technique for “tuning" vocal tract area functions based on acoustic sensitivity functions”, J. Acoust. Soc. Am. 119(2), 715-718.

Bunton, K., Story, B.H. (2011). “A test of formant frequency analyses with simulated child-like vowels”, J. Acoust. Soc. Am. 129(4) 2626.

Story, B.H., Titze, I.R., Hoffman, E.A. (1996) “Vocal tract area functions from magnetic resonance imaging”, J. Acoust. Soc. Am. 100(1), 537-554.

Story, B.H., Titze, I.R., Hoffman, E.A (2001) “The relationship of vocal tract shape to three voice qualities”, J. Acoust. Soc. Am. 109, 1651-1667.