Perception of synthetic vowel exemplars of 4 year old children and estimation of their corresponding vocal tract shapes

Richard S. McGowan
CReSS LLC and Haskins Laboratories, 1 Seaborn Place, Lexington, Massachusetts 02420

Formant scalings for vowel exemplars of American 4 year olds who were imitating adult production were used along with published data of American adult male vowel production to synthesize /ɑ, æ, u, i/. Other vowel exemplars were also synthesized. Adult listeners were asked to categorize these synthetic vowels in a forced choice task. With some exceptions, the formant frequencies preferred for the vowels /ɑ, æ, u, i/ were close to the published data. In order to gain insight on children's articulation during imitation of vowels /ɑ, æ, u, i/, a five-tube model was used in an algorithm to infer vocal tract shape from the first three formant frequencies of the adult productions, the formant frequencies derived for 4 year olds by scaling, and formant frequencies for 4 year olds derived based on the listening experiments. It was found that the rear tube length for the children, in proportionate terms, was nearly always greater than that of the adult. The rear tube length was proportionately twice as long in children compared to adults for the vowel /u/. Tongue root flexibility and the oblique angle between the pharynx and mouth may be more important than pharynx length in determining formant scalings for 4 year old children. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2345833]

I. INTRODUCTION

Children generally learn to speak the language that is used in their surroundings. Within the first 2 years, children will have acquired a basic vocabulary that is understandable to many adults in their families. However, the acoustic output of young children's speech cannot completely resemble adults' speech simply because of anatomical differences between adults and young children. One of the most obvious of these differences between children and adults is in vocal tract length, but there are also differences in the relative lengths of the pharynx and mouth (Goldstein, 1980; Kent and Vorperian, 1995; Fitch and Giedd, 1999). Despite anatomical differences, children do learn to speak, and, in particular, they produce vowels that are recognized by adults.

Acoustic studies of vowel production by American men, women, and/or children, such as those of Peterson and Barney (1952), Hillenbrand, Getty, Clark, and Wheeler (1995), and Lee, Potamianos, and Narayanan (1999) have documented the acoustics of vowel production differences between adults and children. A quantitative look at such data indicates that there is more than simple length scaling that must be invoked to account for formant frequency differences between, say, men and children (e.g., Fant, 1975). Some of the differences might be attributed to differences in the pharynx-to-mouth length differences. However, the effect of anatomical differences on articulation is still unclear.

One way to understand the relation between adults' production and children's production of particular utterances is to examine the children's production when they are asked to imitate the speech of adults. Imitation is understood here in the narrow sense of the situation when children are asked to imitate adult phonetic segments in experimental context. Here, we will restrict ourselves to vowel production and ask

the following questions. How do children imitate adult vowels when their anatomies differ? Are there any constraints on children's production of vowels beyond the length differences in fixed vocal tract structures? The answers to these questions will help in understanding how children develop as speakers.

One source of data on children's vowel production when they were asked to imitate adult vowels comes from Kent and Forner (1979). In their experiment, Kent and Forner asked three 4 year old boys and six 4 year old girls to imitate vowels that had been synthesized on a cascade formant (Klatt) synthesizer. Five of these synthetic vowels were modeled on the mean adult male data of Peterson and Barney (1952) for the vowels /i, u, ɑ, æ, ɔɹ/, and five other vowels were used to fill the vowel space, but they were without specific phonemic identity. These ten stimuli were presented five times each for a total of 50 imitations. Children's formant frequencies from the imitation task were made from spectrograms in a procedure described in their paper (Kent and Forner, 1979). From these measures, an average scale factor (=ratio of child formant frequency to adult formant frequency) for each formant frequency and each vowel can be computed across all ten 4 year olds (see Fig. 10 of Kent and Forner, 1979 for a closely related quantity). Based on these formant scaling factors for 4 year old children and the formant frequencies for adult males provided by Olive, Greenwood, and Coleman (1993), a set of formant frequencies for vowels /ɑ, æ, u, i/ as might be spoken by a 4 year old imitating an adult can be found. It should be noted that the scaling derived from Kent and Forner and the formant frequencies provided by Olive *et al.* (1993) could be affected differently by individual and regional dialect differences.

Listening experiments with adult subjects were used to verify that the formant values for the four year-old children's

vowels derived using the Kent and Forner (1979) and Olive et al. (1993) parameters were, indeed, reasonable. Thus, using these data, vowels /ɑ, æ, u, i/ were synthesized, as well as vowels whose formant frequencies were altered from these four vowels. This also served to verify the scaling results of Kent and Forner (1979) when they are applied to a set of published adult formant frequency values of Olive et al. (1993).

In order to begin to answer the questions about how children imitate it is necessary to know something about the articulation employed by children to imitate the vowels of adult speech. It is difficult to obtain articulatory measures during speech production in young children. Therefore, our strategy was to infer vocal tract shape from the speech acoustic data of children's vowel production described above. The vocal tract shapes of both children and adult males in the production of vowels /ɑ, æ, u, i/ were inferred using an analysis-by-synthesis procedure that maps formant frequencies to an acoustic tube model of the vocal tract. While vocal tract shapes are not synonymous with articulatory positions, vocal tract shape can be informative as to the articulatory strategy used by children during speech imitation.

Our previous work in inferring vocal tract shape and dimensions from acoustics has been to use analysis-by-synthesis, in which the synthesis is done using an articulatory synthesizer with a model vocal tract. In this previous work, task-dynamic parameters of the model vocal tract were adjusted in a stochastic optimization algorithm, a genetic algorithm, so that the resulting formant frequency trajectories produced by an articulatory synthesizer matched the formant frequencies of the corresponding data vowel (e.g., McGowan, 1994). The closeness of the match for each utterance was judged by the size of sum of the square differences between the model and the data in the first three formant frequencies. The work here will use static formant frequency values to infer static vocal tract tube shapes. There can be problems of ambiguity (i.e., multiple optimal solutions) in such a procedure, so the analysis-by-synthesis procedure was run several times, and, further, the number of allowed vocal tract tube sections was limited. By running a stochastic optimization procedure several times, it can be expected that a variety of optimal solutions will be found. Also, by attending to grosser features of the vocal tract corresponding to the information that the first three formant frequencies provide, it can be assured that the inferences made about the vocal tract shapes are valid.

II. LISTENING EXPERIMENT: METHOD

Vowels /ɑ, æ, u, i/ were synthesized as monophthongal vowels with a cascade (Klatt) formant synthesizer using the formant frequencies shown in Table I. The first three formant frequencies (F1, F2, F3) were derived from the scaling factors for 4 year olds provided by Kent and Forner (1979, Fig. 10) and the formant frequencies for an adult male provided by Olive et al. (1993, p. 104). The formant frequencies were measured from sentences containing the words "bottle," "bat," "boot," and "beet" spoken by an adult male from Pittsburgh, Pennsylvania (Olive et al., 1993, p. 8). The formant

TABLE I. The first three formant frequencies, without alteration, used to synthesize KFO condition vowels of a 4 year old. The scaling factors from Kent and Forner (1979) and adult formant frequencies from Olive et al. (1993) are in brackets.

	Child F1 (scale, Adult F1)	Child F2 (scale, Adult F2)	Child F3 (scale, Adult F3)
ɑ	1125 Hz (1.5, 750 Hz)	1650 Hz (1.5, 1100 Hz)	4030 Hz (1.55, 2600 Hz)
æ	980 Hz (1.4, 700 Hz)	2475 Hz (1.5, 1650 Hz)	4250 Hz (1.7, 2500 Hz)
u	429 Hz (1.43, 300 Hz)	1080 Hz (1.27, 850 Hz)	4375 Hz (1.75, 2240 Hz)
i	448 Hz (1.6, 280 Hz)	3488 Hz (1.55, 2250 Hz)	4125 Hz (1.5, 2750 Hz)

frequencies from Olive et al. (1993) are all within 10% of the mean values for adult males given in Peterson and Barney (1952), which were the values used by Kent and Forner (1979) in their experiments. Further, with the exception of the three formants for /u/ and F2 for /ɑ/, the formant frequencies are within one standard deviation of the mean values given by Lee et al. (1999). In the cases of these four formant frequency values, the values of Olive et al. (1993) are within 20 Hz of the values given by Peterson and Barney (1952). Because the speech data for the Lee et al. (1999) were collected in St. Louis, Missouri and the data of Peterson and Barney (1952) were collected in New Jersey over 40 years previously, these differences can probably be attributed to dialect differences amongst the various studies. With the values of F1, F2, and F3 for the synthetic children's vowels derived from Kent and Forner and from Olive et al., the values F4 and F5 were scaled from 3500 Hz and 4500 Hz using the scaling factor for F3. This was done because of a lack of data for F4 and F5. These synthetic base vowels, /ɑ, æ, u, i/, are called KFO condition vowels in this paper. These vowels were intended to be synthetic approximations of vowels produced by 4 year olds imitating adults.

Other vowels were synthesized with the first three formant frequencies of the base vowels altered by −20%, 0%, or 20%, and the fourth and fifth formant frequencies were changed the same amount as the third formant frequency from the values for the base vowels. With all possible combinations of these perturbations there were twenty-seven = 3^3 different vowel exemplars for each base vowel, including the base vowel itself. All vowels were sampled at a 20 kHz sampling rate. The vowels were 400 ms long with the fundamental frequency gradually decreasing from 240 Hz to 200 Hz throughout that duration. This was on the low side for 4 year old fundamental frequencies [e.g., Eguchi and Hirsh (1969) show a mean fundamental frequency of 286 Hz], but they were within reason and engendered a less ambiguous vowel quality than would have higher fundamental frequencies. Also, after an initial increase in amplitude for the first 20 ms, the voice amplitude decreased while parameters making the voice sound more breathy, such as random noise amplitude, increased with time into the vowel. These parameter settings made the voice quality child-like, as judged by the author.

TABLE II. The mean of the formant frequencies in each vowel category receiving a score of 50%, or greater, of the highest score. The scaling factor from adult male formants and percent difference from the children's KFO condition formant frequencies are shown in Table I in brackets.

	Child F1 (scaling, % diff.)	Child F2 (scaling, % diff.)	Child F3 (scaling, % diff.)
ɑ	1210 Hz (1.61, +7.5%)	1672 Hz (1.52, +1%)	4128 Hz (1.59, +2%)
æ	1087 (1.55, +11%)	2700 (1.64, +9%)	4173 (1.67, −2%)
u	393 (1.31, −8%)	1080 (1.27, 0%)	4421 (1.97, +1%)
i	386 (1.38, −14%)	3541 (1.57, +2%)	4188 (1.52, +2%)

Fifteen adult students who were recruited by advertisement from the University of Connecticut student body participated in a listening experiment with these synthetic vowels. Each subject was seated at a computer terminal and listened to the stimuli over a set of speakers controlled by the computer. The 108 synthetic vowels were presented in randomized lists, in which each token appeared three times. As soon as a stimulus was played, a list of five hVd words (except for one hVti word) appeared on the screen (see the Appendix). The subject was asked to choose the word with the vowel that most closely matched the vowel presented auditorily. Once the choice was made, the subject was asked for the goodness of the match on a scale from one to five. After the subject responded to this choice, he was given the choice to repeat the trial, continue on to the next stimulus, or to exit the experiment. In this way the subject was able to be as certain as possible about his judgments and less likely to attempt to compensate for a perceived poor judgment in listening to a future stimulus.

III. LISTENING EXPERIMENT: RESULTS

For each synthetic vowel, the goodness scores for each category were totaled. For instance, a certain synthetic vowel would sometimes be heard as /i/ and sometimes as /ɪ/, and the goodness scores for this vowel were totaled separately for each of these phonemic categories. Table II shows the average formant frequencies of the synthetic vowels that received scores of 50% (an arbitrary level), or greater, of the maximum goodness score for the phonemic categories of each of the four base vowels. The percents in brackets are the amount these formant frequencies differ from those in Table I. These average formant frequency values specify vowels that are called *listening condition* vowels in this paper.

With some exceptions, listeners favored the formant frequency values for the 4 year olds that were taken to be representative of children's productions according to the Kent and Forner (1979) scalings and the Olive *et al.* (1993) formant frequencies. The percents in brackets indicate that listeners favored more extreme F1 values, or extreme tongue height positions, than those given in the KFO condition. Further, the listeners favored a higher F2, or more fronted tongue, for /æ/.

TABLE III. The number of times a particular vowel, as determined by the KFO condition, is classified into one of five vowel categories with a goodness score of greater than, or equal, to 3.

KFO condition vowel	scores	Listener Choices				
		ɑ	ɔ	ʌ	æ	oʊ
ɑ	5, 4	9	7	1	0	0
ɑ	3	9	6	2	3	2
		æ	ɛ	ɑ	eɪ	ʌ
æ	5, 4	11	4	0	0	0
æ	3	0	8	1	0	0
		u	ʊ	oʊ	ʌ	ɑ
u	5, 4	10	7	1	0	0
u	3	7	2	2	2	0
		i	ɪ	eɪ	ɛ	æ
i	5, 4	20	1	2	0	0
i	3	11	0	1	1	0

The KFO condition vowels, whose formants are shown in Table I, were examined to find which categories listeners favored. The instances when these vowels received a goodness score of four or five for a particular category were counted, and the instances when these vowels received a goodness score of three were counted separately. Dividing the goodness scores four and five from goodness score three indicates how many times a token was rated as very good or excellent as an exemplar from when it was merely rated as a moderately good exemplar. The total number of times that a KFO condition vowel was rated is 45 (=15 subjects times 3 repetitions). Table III shows these counts for each KFO condition vowel.

Table III accounts for 39 tokens of /ɑ/, 24 tokens of /æ/, 31 tokens of /u/, and 36 tokens /i/, out of 45 tokens each. As indicated by Table III, the four vowels in the KFO condition were acceptable tokens of their target phonemic category in the listening condition. The low vowels /ɑ/ and /æ/ in the KFO condition were often perceived as good examples of slightly higher, neighboring phonemes, that is as /ɔ/ and /ɛ/, respectively. This is consistent with Table II, which indicates that listeners preferred higher first formant frequencies (i.e., a lower vowel) than provided by these vowels in the KFO condition. Also, the KFO condition /u/ was often heard as a good example of /ʊ/, a neighboring lower vowel. This, again, is consistent with Table II: listeners preferred higher vowels (i.e., lower first formant frequencies) for /u/ than provided in the KFO condition. While listeners preferred higher vowels than provided by the KFO condition /i/, they still counted the KFO condition /i/ as an unambiguously good example of that vowel.

IV. VOCAL TRACT DIMENSIONS: METHOD

Children's area functions for the vowels /ɑ, æ, u, i/ were inferred from formant frequency data from two different sources: the KFO condition for 4 year old children (Table I)

and preferences given by adults in the listening experiment (Table II). A five-tube vocal tract model was used, which is justified by the fact that the four adult male vowels could be simulated using as few as four tube sections (Stevens, 1998), thus leaving one extra tube in case the children's vocal tract shapes proved anomalous. Adult male area functions were also inferred for these vowels.

A genetic algorithm was used for each vowel to infer the lengths and areas of the five tubes that could produce the formant frequencies given in Tables I and II. A simple genetic algorithm was used as in previous work in recovering articulatory movement from formant frequencies (e.g., McGowan, 1994), but only static area functions were inferred in the current work. A genetic algorithm is a stochastic optimization technique, in which each variable parameter is coded as a binary number, which is called a gene (Goldberg, 1989). All the genes, each coding a particular parameter, are concatenated to form a chromosome. In this study a gene represented either a tube length or a tube area, for a total of ten genes in a chromosome. For the present study, the fitness of a chromosome was based on the first three formant frequencies produced by the five-tube model specified by the genes of that chromosome. Those three formant frequencies were compared with formant frequency data from one of the conditions in Table I and II and a distance between the model formants and data formants was computed. This distance was related to fitness, so that the smaller the distance, the greater the fitness. The specific form of the relation between distance and fitness is given below.

A simple genetic algorithm starts with a random group, or population, of chromosomes, each coding a different set of parameters. The population is then allowed to progress through generations. For each generation, pairs of chromosomes are selected for possible mating. In the algorithm used in this study, the number of pairs equaled one-half the size of the population. The chromosomes for possible mating are selected according to their fitness, so that the probability of a chromosome being selected in this work was proportionate to its fitness. Each pair of chromosomes selected has a probability of mating. If they do not mate, they are put back into the population unchanged, except for a small probability of mutation, where bits of the chromosome are changed in a process of mutation. Otherwise each of the chromosomes is cut at a randomly selected bit in each chromosome, and the divided strings of bits are swapped to produce new offspring chromosomes. These chromosomes are then subjected to the same small probability of mutation as the unmated pairs of chromosomes. The fitness of each new chromosome is evaluated, and a new generation is thus created. In the algorithm used here, the fittest individual of a generation was always preserved to enter the next generation. The process is stopped after a fixed number of generations. This process is supposed to be analogous to natural selection, and it can be shown to tend to produce chromosomes that are more fit as the number of generations increase (Goldberg, 1989).

In the current study, populations of 240 chromosomes were coded with five bits per tube area with minimum areas of 0.05 cm^2 and maximum areas of 10 cm^2, and with five bits per tube length with minimum lengths of 0.5 cm and

maximum lengths of 8.0 cm. These populations were run through 120 generations with fitness proportionate selection. There was a 0.6 chance for two selected chromosomes to mate, where mating entailed cutting a chromosome at a random bit and swapping the cut substrings to produce children chromosomes. There was a 0.005 chance for any bit to mutate from a one to a zero or vice versa. To compute the fitness function, the square-root of the sum over the first three formant frequencies of the square fraction of the differences between the data formant frequency and the inferred formant frequency was computed. The fitness was the exponential of the inverse of this quantity. The genetic algorithms were run five times for each of the sets of three formant frequencies for the four vowels from both Tables I and II for the children, and from Table I for adult males.

V. VOCAL TRACT DIMENSIONS: RESULTS

The vocal tract dimensions that best fit the adult male acoustic data are shown in Fig. 1. For the adults there were often multiple recovered area functions. Also, while the cross-sectional area near the glottis varies from less than 0.5 cm^2 for /ɑ/ [Fig. 1(a)] to over 3.0 cm^2 for /i/ [Fig. 1(d)], these areas only represent an average cross-sectional area in the pharyngeal region. For the adult /i/ [Fig. 1(d)], there is some ambiguity in the area of the tube in front of the constriction, because there are compensatory area changes in the rear cavity. For the adult /ɑ/ [Fig. 1(a)] there is a compensatory relation between the length of the rear tube and the total vocal tract length. The most striking ambiguity is for the vowel /æ/ [Fig. 1(b)], where, in the first case, there is a narrow rear tube for about one-third of the total length of the vocal tract with a wide front tube, and in the other case there is a more gradual increase in area from the rear to the front of the vocal tract. Figure 2 shows the spectra for the two adult /æ/s. Because the second formant has most of its energy associated with the longer front tube in the first case, it has greater amplitude than the third formant, which is associated with the rear tube [Fig. 2(a)]. In the second case the third formant has greater amplitude than the second formant when there is a more gradual increase in area [Fig. 2(b)]. Thus more of the second formant energy is associated with the rear and more of the third formant energy is associated with the front of the vocal tract compared to the first case. Accounting for relative formant amplitudes would disambiguate these situations.

Comparisons of the tube shapes from the adults' production, the children's KFO condition, and the adult listening preferences of children's synthetic vowels show the same general vocal tract shapes (Figs. 1, 3, and 4). Both /æ/ and /ɑ/ have rear constriction tubes, with the one for /ɑ/ more constricted than the one for /æ/ [Figs. 1(a), 1(b), 3(a), 3(b), 4(a), and 4(b)]. In both the KFO and listening conditions, the children's /æ/ had a less, constricted, but longer rear tube than the adult case with a rear tube length about one-third the vocal tract length. The children's /æ/ is more closely related to the second adult case with a more gradually changing area function. However, there were configurations with more constricted, shorter rear tubes for the children's KFO and listen-
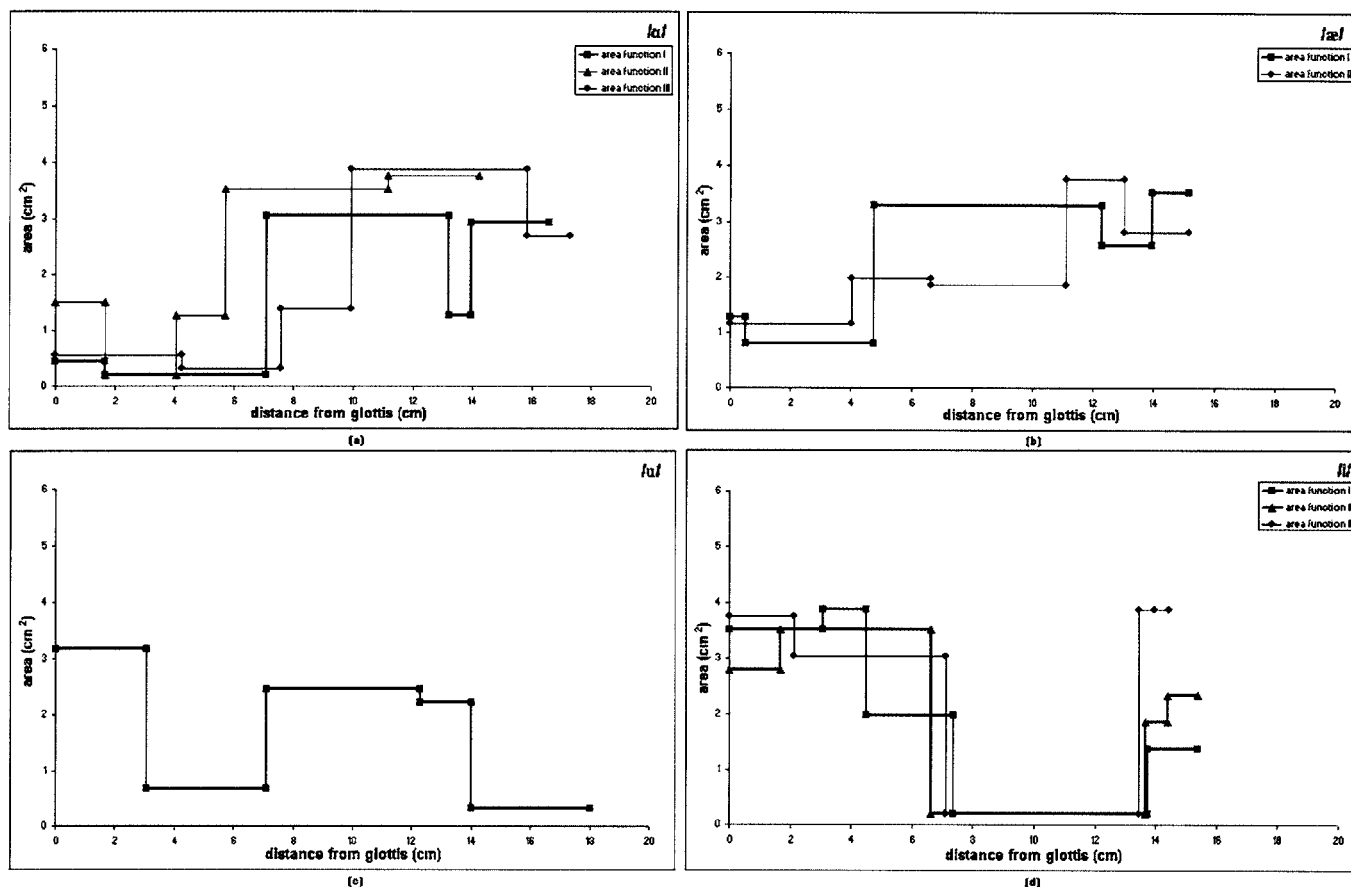
J. Acoust. Soc. Am., Vol. 120, No. 5, November 2006

Richard McGowan: Children's vocal tract shape for vowels   2853

FIG. 1. Tube dimensions recovered from adult formant frequency data (a) /ɑ/, (b) /æ/, (c) /u/, and (d) /i/. Different recovered vocal tracts for a single vowel are possible and are denoted by different symbols and line thickness.

ing conditions that did not attain the fitness of the configurations shown in Figs. 3(b) and 4(b), but were nearby in terms of formant frequency fit. Thus there appears to be a similar ambiguity to the adult males in the vowel /æ/ for children. For /u/ there are two cavities separated by a constriction, and there is a constriction at the lips [Figs. 1(c), 3(c), and 4(c)]. For /i/ there is a back cavity with a front constriction [Figs. 1(d), 3(d), and 4(d)].

Beyond the general shapes of the vocal tract tubes there were differences between the children's and adult's vocal tract tubes. A difference that appeared fairly consistently across the four vowels involves the length dimension of the rear tube, whether or not that tube was a constriction or a cavity[1]. Comparisons of these lengths and their ratios with the total recovered vocal tract lengths are shown in Table IV. (Recall that there were multiple tube configurations for the adult data, as shown in Fig. 1.)

All of the rear tube length ratios in the case of the children's vowels in the listening condition were at least as large as the corresponding adult rear tube length ratios. The same is true for the children's vowels, except for /i/, in the KFO condition. Even for /i/ in the KFO condition, the adults' ratios do not exceed the children's ratios by more than 0.06. In the case of /u/, the ratio is on the order twice as large in children compared to the adults. /u/ is the vowel that has the greatest range of the three formant scalings: 1.27–1.75 for the KFO condition (Table I) and 1.27–1.97 for the listening

condition (Table II). For the KFO condition of /u/, the scaling difference was 1.43 for F1 and 1.27 for F2, which is the greatest difference among the vowels for these two formants. Thus, it is not surprising to find substantial differences in the ratios of tube lengths between adults and children. The recovery algorithm indicates that on a proportionate length basis, the children's /u/s are more "fronted" than the adult /u/. Also, this is the vowel of the four with the shortest rear tube length for both children and adults.

It is interesting to compare the rear tube lengths to the location of the velum. In one method of computation using MRI images, Fitch and Giedd (1999) took the distance from the glottis to the free edge of the velum (uvula) as the length of each of their subject's pharynx. The pharyngeal segment, along with four other line segments spanning the rest of the upper vocal tract, was also used to compute the total length of each vocal tract in their study. From their data, the ratio of a 4 year old's pharynx length to total length is between 0.25 and 0.29. For an adult male, this ratio is 0.37. In reference to Table IV, this means that all the children's rear tubes include the uvula, although for /u/ the uvula is very close to the front of the rear tube. For the adult male however, it is possible, according to these analyses, that the rear tube does not include the uvula for /ɑ/ and /æ/, and is wholly contained in the pharynx below the uvula. Further, the rear tube for the adult male /u/ is wholly contained in the pharynx below the uvula.

Richard McGowan: Children's vocal tract shape for vowels
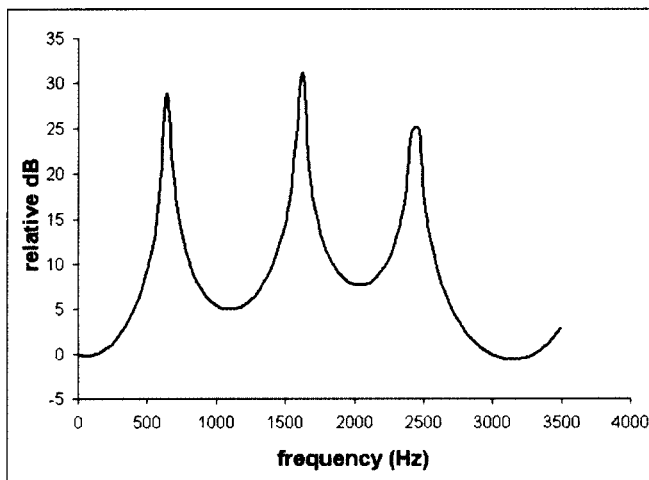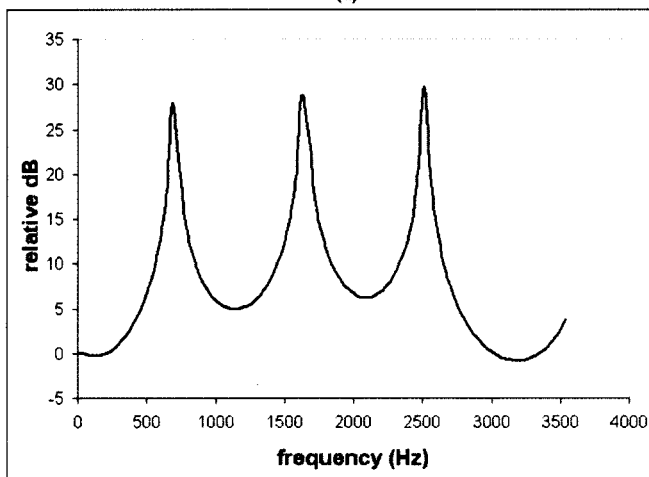
FIG. 2. Transfer functions for two tube configurations for adult /ae/ (a) short rear tube, and (b) long rear tube.

Because the rear tube in /u/ is a part of the volume element in one of the Helmholtz resonators in a double Helmholtz resonator system (Fant, 1960), it is possible to explore the acoustic ramifications of this geometry. F1 and F2 in /u/ are the resonances of this double Helmholtz resonance system, with the rear tube and the constriction closest to the rear constituting one resonator, the back resonator, and the mouth cavity and lip constriction forming the other Helmholtz resonator, the front resonator. The resonant frequencies of these resonators, if they were uncoupled are denoted $F1_0$ and $F2_0$ here, with the $F1_0$ the lower frequency. When these Helmholtz resonances are coupled, as they are in the vocal tract, the resonance frequencies shift from their uncoupled values to F1 and F2, with $F1 < F1_0$ and $F2 > F2_0$. Estimates of the resonant frequencies of the uncoupled back and front resonators are shown in Table V.

It can be seen that $F1_0$ and $F2_0$ are closer together for both of the children's conditions than for the adult's, both in absolute and relative terms. This corresponds to the fact that in both the production and listening condition for children, the scaling factor for F1 is greater than that for F2, which means that F1 is relatively closer to F2 for both children's conditions than for the adult male. (The proximity of F1 to

F2 is not completely accounted for by the proximity of $F1_0$ and $F2_0$, but also depends on the degree of coupling between the resonators.) Further, while the lower frequency $F1_0$ is associated with the front resonator for the adult and the listening condition for the children, it is associated with the back resonator in the KFO condition for the children.

Based on these results, 4 year old children are able to form vocal tract shapes that are appropriate for the vowels /ɑ, æ, u, i/. However, not all of these vowels are related to the adult versions of these vowels by a simple formant scaling factor. This is most apparent with /u/, and to some extent /æ/, in the data from Kent and Forner (1979). This acoustic pattern appears to be the result of a proportionately longer rear cavity for /u/, and, possibly, a longer rear tube for /æ/ in the child's vocal tract, compared to adults.

VI. SUMMARY AND DISCUSSION

The adult formant frequencies for the vowels /ɑ, æ, u, i/ as given by Olive et al. (1993), were scaled to the formant frequencies of 4 year old children imitating adults according to the values given by Kent and Forner (1979). These KFO condition vowels, along with vowels with formant frequencies altered from these values, were synthesized on a cascade formant synthesizer employing a child-like voice quality. Adults were asked to categorize and then to assign goodness scores to these vowels in a forced choice listening task. The average of each formant frequency for each vowel that received at least 50% of the score of the vowel with the largest total score in each phonemic category was computed. The results for the four vowels /ɑ, æ, u, i/ indicated that the formant frequencies of vowel tokens preferred by adult listeners differ little from those computed from adult data and scale factors. The largest differences between the formants in the listening condition and the KFO condition formants were more extreme F1s for all four vowels and the higher F2 for /æ/ in the listening condition. However, the KFO condition vowels, /ɑ, æ, u, i/, were usually perceived as good exemplars of their vowel categories in the listening condition.

Five-tube models were used in an analysis-by-synthesis procedure to infer the lengths and cross-sectional areas of the tubes from the first three formant frequencies. There was some minor ambiguity in the optimum five tubes for the adult male /ɑ/ and /i/. However, there was a more striking two-way ambiguity in the adult male /æ/, which could be differentiated if formant amplitudes had been accounted for along with formant frequencies. While adult males may produce /æ/ in a configuration with a relatively short and constricted rear tube or with an alternative tube with a gradually changing area, the inferred shapes for children show that they may prefer the latter configuration. Informal experimentation shows ambiguity for the mid, front vowel /ɛ/ as well. These observations on low-to-mid front vowels will be pursued in subsequent work, noting here that they are vowels with difficult-to-define constriction locations (e.g., Ladefoged and Maddieson, 1996; p. 284).

In both the listening and KFO conditions, children appear to produce back vowels /ɑ/ and /u/ with a rear tube at least as long, proportionately, as adults. (The rear tube is a

J. Acoust. Soc. Am., Vol. 120, No. 5, November 2006

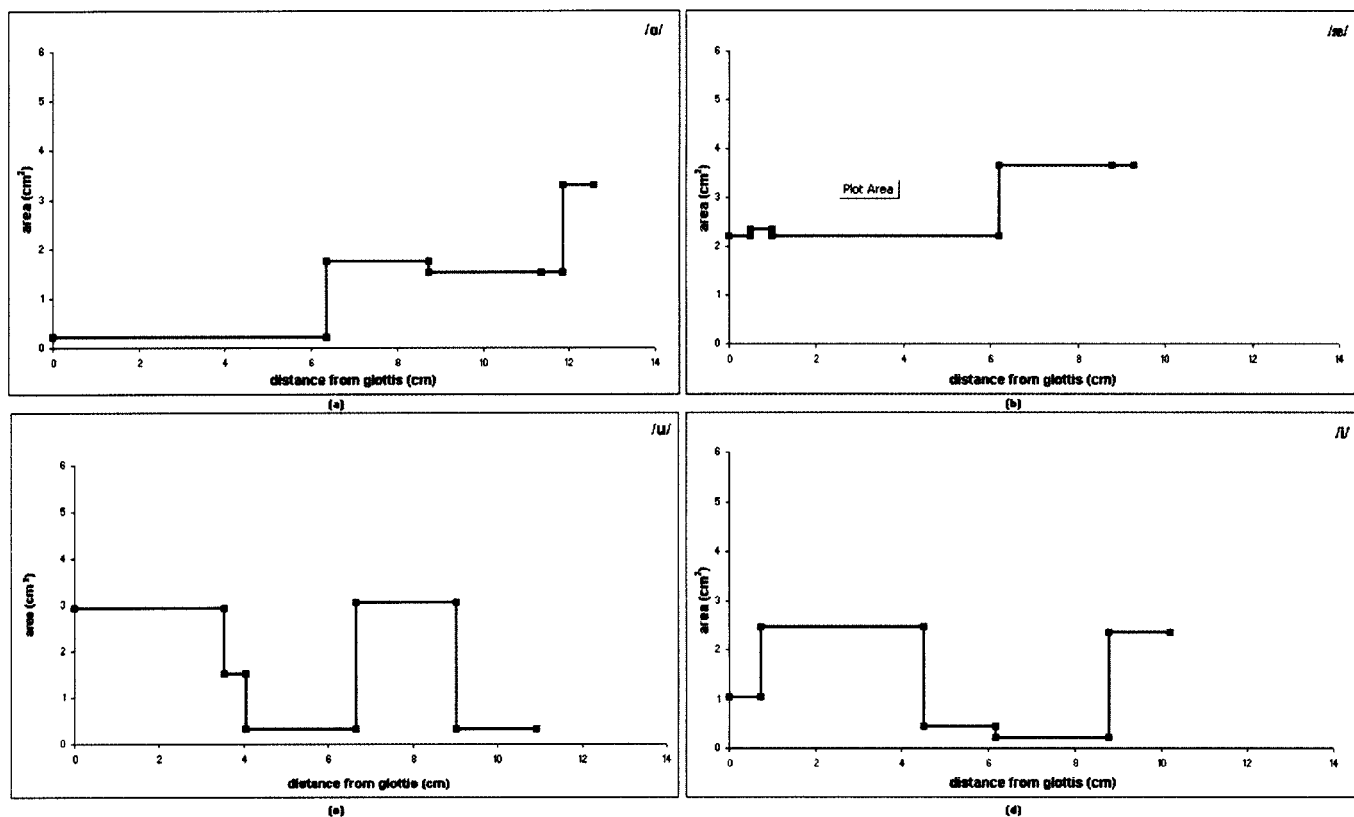Richard McGowan: Children's vocal tract shape for vowels   2855

FIG. 3. Tube dimensions recovered from children's KFO condition formant frequency data (a) /ɑ/, (b) /æ/, (c) /u/, and (d) /i/.
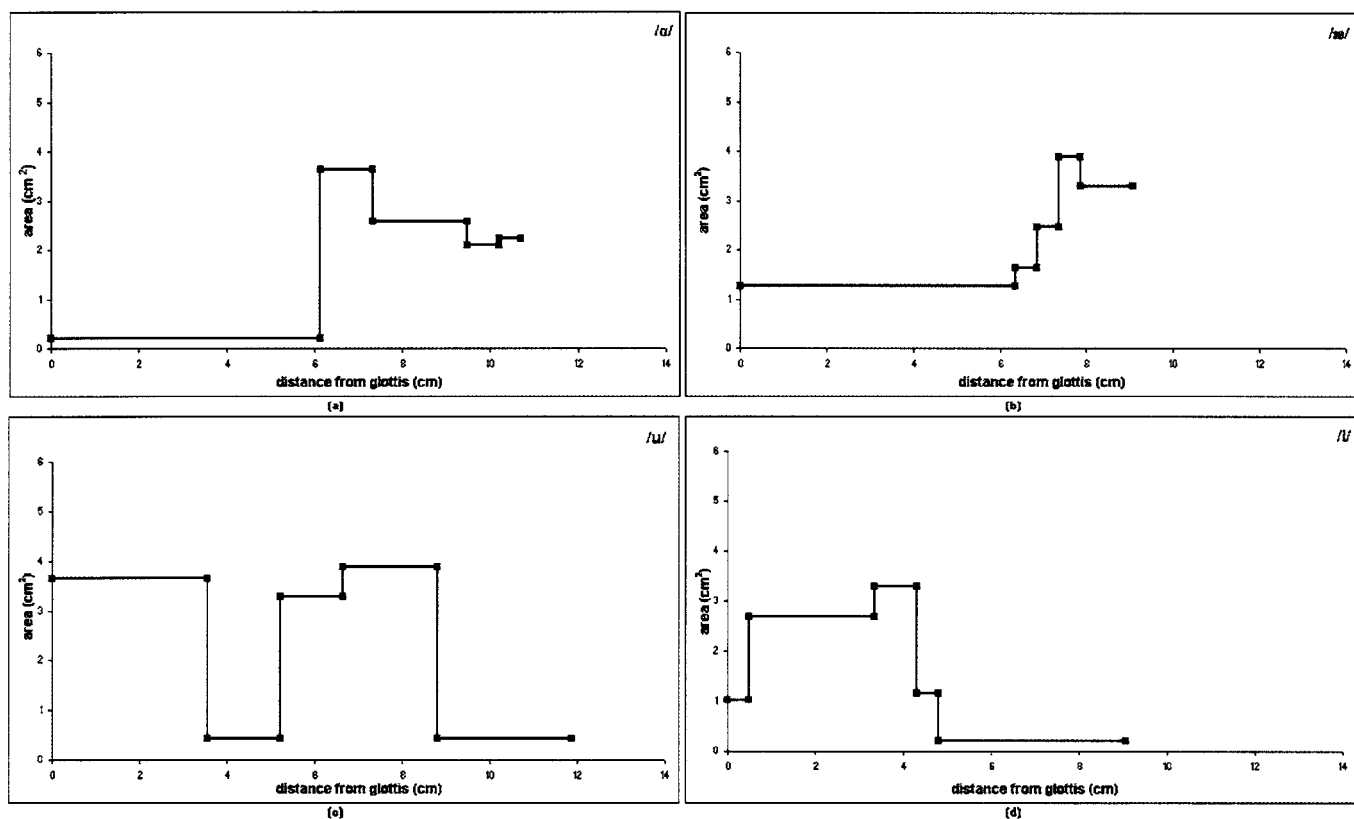

FIG. 4. Tube dimensions recovered from children's listening condition formant frequency data (a) /ɑ/, (b) /æ/, (c) /u/, and (d) /i/.

Richard McGowan: Children's vocal tract shape for vowels

TABLE IV. The length of the recovered rear tube and its ratio with the total vocal tract length. Multiple values appear when there were multiple recovered shapes. Exceptions to when the children's rear tube-to-total length ratio is greater than or equal to the same ratio for adults are shown in italics. (When there was a tube of intermediate area between the rear tube and the third tube from the rear, one half of its length was added to the rear tube length to provide the number in this table.)

	Adult length and ratio	Children (listening) length and ratio	Children (KFO) length and ratio
ɑ	4.89 cm, 0.34	6.13 cm, 0.57	5.89 cm, 0.53
	7.09 cm, 0.43		
	6.28 cm, 0.41		
	8.75 cm, 0.51		
æ	4.75 cm, 0.31	9.06 cm, 0.78	6.19 cm, 0.67
	7.57 cm, 0.50		
u	3.08 cm, 0.17	3.55 cm, 0.30	3.80 cm, 0.35
i	6.62 cm, 0.43	9.30 cm, 0.49	4.51 cm, 0.44
	5.94 cm, 0.39		
	7.11 cm, 0.48		
	6.65 cm, 0.45		
	7.09 cm, 0.49		

constriction for /ɑ/ and an unconstricted cavity for /u/.) In the case of /u/, the difference in proportionate terms can be as large as a factor of 2, and in absolute terms, the rear tube for adults and for children is of similar length. One result of the proportionate difference in rear tube length is the possibility that the natural frequency of the rear Helmholtz resonator in children is less than the frequency of their front Helmholtz resonator, where the opposite is the case for adult's production of a "canonical" /u/.

The following can be offered as one cause for the relatively forward tongue constriction in /u/ for a 4 year old talker. If the child's constriction were moved any further back, it may be that the rear cavity would be so small or totally obliterated, that a double Helmholtz resonance would not be possible. This, in turn, may be due to the limited control and/or flexibility of the tongue body and root, and the fact that the pharynx tube forms an obtuse angle with the mouth cavity, while adult pharynges form right angles with the mouth (Kent and Vorperian, 1995, p.161). (The more obtuse the angle between the pharynx and mouth, the closer the pharynx and mouth are to forming a straight tube, and the more the tongue body needs to bend to form the rear cavity and constriction.) Corroborating research on the production of [ɹ] by young children indicates that it is difficult for them to make two simultaneous constrictions with the tongue, and that the rear constriction for [ɹ] may be missing or not par-

TABLE V. Helmholtz resonance frequencies for the uncoupled back and front resonators.

	Adults resonance freq. (Hz)	Children (listening) resonance freq. (Hz)	Children (KFO) resonance freq. (Hz)
Back resonator	723	784	564
Front resonator	357	575	832

ticularly tight (McGowan, Nittrouer, and Manning, 2004). Thus, the child produces an acceptable /u/ with a proportionately longer rear cavity compared to the adult double Helmholtz resonator.

The results presented here are in accord with the finding that simple area function scaling from adult to child is not sufficient to account for the observed vowel formant frequencies. For instance, Nordström (1979) attempted to reproduce children's formant patterns by scaling the tube lengths for adults' area functions according to the ratio of children-to-adults pharyngeal length and mouth length separately. As this did not provide sufficient fit to children's vowels, the cross-sectional areas in the pharyngeal and mouth regions were, themselves, scaled by the squares of these scaling factors. This also did not provide a good fit to children's formant frequencies.

Scaling based on two anatomical dimensions is not sufficient to explain the differences in adults and children in their vowel production. Possible physiologic causes have been posited in this paper for the differences in /u/ production, coupled with an anatomic difference: flexibility of the tongue, control of the tongue body and root, and the oblique angle between the pharynx and the mouth. There are further factors to consider, including the mandible, which is disproportionately smaller for children four years of age (Kent and Vorperian, 1995). There may be other factors that prevent proportionate or encourage disproportionate formant scaling, as in the case of /u/. Further, /æ/ exhibits nonuniform formant scaling, but a physical reason has not been posited here. It is intriguing that, as Maeda and Honda (1994) point out, /u/ and /æ/ are extreme vowels for the styloglossis-genioglossis anterior antagonist pair. Further work should study /æ/ in relation to other relatively open vowels, such as /ɛ/ and /ʊ/.

The findings here do not corroborate the speculation made by Maeda and Honda (1994) that one needs the right angle of the rear pharyngeal wall and the mouth to produce the vowels /ɑ/ and /i/. (It should be born in mind, however, that the present study does not use articulatory constraints.) Table III shows that listeners judged the children's vowels under the KFO condition to be good exemplars of their phonemic categories. Even without a pharynx at right angles to the mouth, the average 4 year old child can produce these vowels in a ways that are perceptually acceptable to adults. In fact, Kent and Forner (1979) show that the vowels /ɑ/ and /i/ are produced with scale factors that are fairly constant across the three formants (Table I). The results here indicate that in order for children to achieve proportionate formant scaling it is important to form constrictions and cavities whose lengths are not far from proportionate to the adults. (A statement that proportionate length is sufficient cannot be made because of differences in the associated cross-sectional areas.) The vowels /ɑ/ and /i/ show that mismatches in the relative positions and sizes of fixed anatomical structures, such as the pharynx and mouth, do not necessarily hinder a child's ability to form such constrictions and cavities. Indeed, in perceptual tests using speech produced by an articulatory synthesizer, Ménard, Schwartz, and Boë (2004) showed that infants can produce all the French vowels. How-

ever, caution should be used in this kind of experiment, because scaling of an adult vocal tract model to simulate a lengths in certain instances, and when this is impossible, he or she attempts to imitate the general vocal tract shape. It is not possible to say from these data whether children attempt to imitate vocal tract shape or acoustic parameters. This may be impossible in any study that involves only acoustic and/or articulatory data because of the causal relation between the two domains.

ACKNOWLEDGMENTS

APPENDIX

The choices given to the listener when the synthetic vowel was /ɑ/, or a vowel with its formant frequencies up to 20% different, were: "hot," "haughty," "hut," "hat," "how" or /ɑ, ɔ, ʌ, æ, oʊ/.

The choices given to the listener when the synthetic vowel was /æ/, or a vowel with its formant frequencies up to 20% different, were: "had," "head," "hot," "hayed," "hut" or /æ, ɛ, ɑ, eɪ, ʌ/.

The choices given to the listener when the synthetic vowel was /u/, or a vowel with its formant frequencies up to 20% different, were: "whoed," "hood," "hoed," "hut," "hot" or /u, ʊ, oʊ, ʌ, ɑ/.

The choices given to the listener when the synthetic vowel was /i/, or a vowel with its formant frequencies up to 20% different, were: "heed," "hid," "hayed," "head," "had" or /i, ɪ, eɪ, ɛ, æ/.

[1]Rear cavities were defined using the following rules: (1) Rear tubes are constrictions for /ɑ/ and /æ/, and they are cavities for /i/ and /u/, (2) going

from rear to front, a constriction (cavity) never ends at a tube whose neighbor to the rear has a greater (lesser) cross-sectional area, (3) cross-sectional

ㅗgoᴄㅛᴅ, ᵴ., ᴀnᴅ ㅛnᵴ1, ㅛ, ᴊ. (ɪ ɘ0ɟ). "ᴅᴄ ᴠᴇiᴏᴘᴍᴇnt ᴏᴦ ᴤᴘᴄᴄᴄ̈iᴦ ᴤᴜᴅᴦ1ᴜᴤ �111 ᴄ̈111ᴏᴅ ren," Acta Oto-Laryngol., Suppl. Suppl. 257, 5–48.

Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fant, G. (1975). "Non-uniform vowel normalization," Speech Transmission Laboratory, KTH. STL-QPSR 2-3/1975, pp. 1–19.

Fitch, W. T. and Giedd, J. (1999). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. 106, 1511–1522.

Goldberg, D. E. (1989). *Genetic Algorithms* (Addison-Wesley, Reading).

Goldstein, U. (1980). "An articulatory model for the vocal tracts of growing children," Ph.D. dissertation, MIT, Cambridge, MA.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. 97, 3099–3111.

Kent, R. D. and Forner, L. L. (1979). "Developmental study of vowel formant frequencies in an imitation task," J. Acoust. Soc. Am. 65, 208–217.

Kent, R. D. and Vorperian, H. K. (1995). "Development of the craniofacial-oral-laryngeal anatomy: A review," J. Medical Speech-Language Pathology 3, 145–190.

Ladefoged, P. and Maddeison, I. (1996). *The Sounds of the World's Languages* (Blackwell, Oxford).

Lee, S., Potamianos, A., and Narayanan, S. (1999). "Acoustics of children's speech: Developmental changes of spectral and temporal parameters," J. Acoust. Soc. Am. 105, 1455–1468.

McGowan, R. S. (1994). "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary results," Speech Commun. 14, 19–49.

McGowan, R. S., Nittrouer, S., and Manning, C. J. (2004). "Development of [ɹ] in young, Midwestern, American Children," J. Acoust. Soc. Am. 115, 871–884.

Maeda, S., and Honda, K. (1994). "From EMG to formant patterns of vowels: The implications of vowel spaces," Phonetica 51, 17–29.

Ménard, L., Schwartz, J.-L., and Boë, L.-J. (2004). "Role of vocal tract morphology in speech development: Perceptual targets and sensorimotor maps for synthesized French vowels from birth to adulthood," J. Speech Lang. Hear. Res. 47, 1059–1080.

Nordström, P.-E. (1979). "Attempts to simulate female and infant vocal tracts from male area functions," Speech Transmission Laboratory, KTH. STL-QPSR 2-3/1975, pp. 20–33.

Olive, J. P., Greenwood, A., and Coleman, J. (1993). *Acoustics of American English Speech* (Springer-Verlag, New York).

Peterson, G. E. and Barney, H. L. (1952). "Control Methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.

Stevens, K. N. (1998). *Acoustic Phonetics* (MIT Press, Cambridge, Massachusetts).