# Vowels Produced by Sliding Three-tube Model with Different Lengths

*Takayuki Arai*

Department of Information and Communication Sciences
Sophia University, Tokyo, Japan
arai@sophia.ac.jp

## Abstract

The sliding three-tube (S3T) model, based on Fant's acoustic theory and proposed in our previous studies, has a simple structure, enabling it to produce human-like vowels useful for education in acoustics and speech science. In this study, we changed the size of the S3T model and combined it with sound sources with different fundamental frequencies. We confirmed that the models could produce vowels of different speaker types. We were able to retain good vowel quality for a perceptual study when we simultaneously shortened vocal-tract length and increased fundamental frequency. We also discussed the models in a new way, comparing children's and adults' vowels, especially for educational purposes.

**Index Terms**: physical models of the human vocal tract, vowel production, education in acoustics, speech science

## 1. Introduction

Since we developed the cylinder and plate-type models of the human vocal tract [1,2] based on measurements done by Chiba & Kajiyama (1941) [3], we have proposed a set of educational tools, including more physical models of the human vocal tract. We have found these models to be extremely useful for education in acoustics and speech science [2]. One tool we developed is the sliding three-tube (S3T) model [4,5], which is based on Fant's three tube model [6]. Like all our models, the S3T model is useful when teaching basic speech production concepts such as source-filter theory and the relationship between vocal-tract shape and vowel quality. The S3T model is unique in that this single model can produce a variety of vowels with its simple structure.

Figure 1 shows a schematic figure of the S3T model. It consists of an outer and inner cylinder. When we slide the inner cylinder within the outer cylinder, different vowel sounds can be produced by feeding an input sound source at the glottis end. In this model, the inner cylinder forms a constriction like that of the tongue in the vocal tract. Sliding the inner cylinder is the primary degree of freedom. Even with this single degree of freedom we can change vowel quality dramatically. Additionally, two more degrees of freedom are possible: the constriction area of tongue and lip rounding, which further increases the variety of vowels produced by the model [5,7].

Because of its simple structure, the first two or three formant frequencies of the S3T model can be estimated easily [4,5,8]. This is done by decomposing the whole model into three parts: a quarter-length resonator, a half-length resonator, and a Helmholtz resonator. The S3T model may be used to teach basic acoustic theory of vowel production for undergraduate and graduate students. Furthermore, the model's simple design makes it possible to use in a science workshop where children make their own sliding vocal-tract models [5,9].

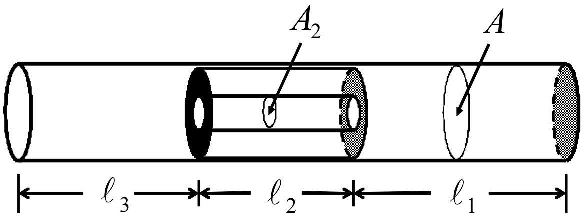

Figure 1: *Schematic figure of the S3T model (adapted from [4,5]). The inner cylinder (the constriction) slides back and forth inside the outer cylinder mimicking a change in the shape of the vocal tract and simulating different vowels.*

Although we have mainly examined vowels by adult males with the S3T models in our previous studies, we knew empirically that the S3T model with a shorter tube could also produce children's vowels. From both an educational and academic point of view, it is useful to have both child and adult versions of the S3T model. Therefore, we have recorded vowels by combining the long and short tubes with sound sources of different fundamental frequencies [10]. In [10] we also conducted a perceptual experiment to see how the recorded sounds were perceived as Japanese vowels. Detailed acoustic analysis of the recorded vowels and the data of the perceptual experiment has not been completed yet, so for this paper we further analyzed the data in the present study.

## 2. Producing vowels by changing vocal-tract length

### 2.1. Four different lengths of the S3T model

We designed four versions of the S3T model by changing the lengths of the outer and inner cylinders [10]. Let $L$ and $\ell_2$ be the lengths of the outer and inner cylinders, respectively, and let $\ell_1$ and $\ell_3$ be the lengths of the back and front cavities, respectively (Fig. 1). When the inner cylinder slides inside the outer cylinder, $\ell_1$ and $\ell_3$ vary between 0 and $L - \ell_2$ under the condition of $L = \ell_1 + \ell_2 + \ell_3$. $A$ and $A_2$ are the cross-sectional areas of the front/back cavities and the constriction, where $A = \pi D^2/4$ and $A_2 = \pi d^2/4$, and $D$ and $d$ are the inner diameters of the outer and the inner cylinders, respectively.

Table 1 shows the values of the parameters of the four versions designed in this study. Figure 2 shows the actual models. Both the inner and outer cylinders were made of acrylic resin, and the outer cylinder was 3 mm thick.

Table 1: *Values of the parameters of the four versions of the S3T model designed in this study.*

| | Ver. 1 | Ver. 2 | Ver. 3 | Ver. 4 |
|---|---|---|---|---|
| $L$ [mm] | 170 | 140 | 110 | 80 |
| $\ell_2$ [mm] | 60 | 50 | 40 | 30 |
| $D$ [mm] | 34 | 34 | 34 | 34 |
| $d$ [mm] | 10 | 10 | 10 | 10 |

### 2.2. Recording the vowels made by the four versions of the S3T model

The four versions of the S3T model were combined with six different input signals for the recordings [10]. A driver unit (TOA TU-750) for a horn speaker was attached to the outer cylinder. Input signals were fed into the driver unit via an audio interface (RME Multiface) and a power amplifier (FOSTEX AP1020).

The six input signals were as follows. The first five signals were based on an impulse train with an original sampling frequency of 16 kHz; later, the signals were upsampled to 48 kHz. For the first signal, the fundamental frequency, $f_0$, increased from 100 to 125 Hz within the first 100 ms, and then decreased to 100 Hz within the next 200 ms. The second, third and fourth signals were similar impulse trains, whose $f_0$ contours were 2, 3, and 4 times higher than the that of the first signal, respectively. The fifth signal was also an impulse train but its $f_0$ was constant at 100 Hz. The durations of the first five signals were 300 ms. The sixth signal was a swept-sine signal with a sampling frequency of 48 kHz. The length of the swept-sine signal was 65536 samples.

To avoid unwanted coupling between the neck and the area behind the neck of the driver unit and to achieve high impedance at the glottis end, we inserted a close-fitting metal cylindrical filler inside the neck. We made a hole in the center of the metal filling with an area of 0.13 cm$^2$. A flange with a diameter of 25 cm was attached at the open end of the tube. The output sounds were recorded using a microphone from the sound level meter (RION NL-18) and an audio interface (RME Multiface) with a sampling frequency of 48 kHz. The microphone was placed approximately 20 cm in front of the output end in a sound-treated room. The recordings were done as we slid the inner cylinder, so that $\ell_1$ varies from 0 to $L - \ell_2$ in 2 mm steps. The signals recorded were synchronously averaged multiple times to increase the signal-to-noise ratio.

### 2.3. Analysis of the recorded sounds

In the present study, we analyzed the recorded sounds, paying particular attention to the spectral characteristics of each configuration of the model. Figure 3 shows spectrographic representations of the impulse responses calculated from the sounds produced by the swept-sine signal. In these figures, the horizontal axis is $\ell_1$ instead of time. Each vertical slice in grayscale represents a spectral plot obtained from the impulse response, which was calculated from the output signal when the swept-sine signal was used as the input signal. The horizontal axes are scaled, so that the different maximum lengths of the back cavity are plotted as the same full scale for all four versions. As shown in these figures, the spectral peaks corresponding to the resonances of the cavities shift as the inner cylinder moves [4,5].

In Fig. 3, the underlying three resonance curves (solid lines) [4,5] based on Stevens [8] are also overlaid. These resonance curves are as follows:

$$1)\ \frac{c}{2\ell_1}, \qquad 2)\ \frac{c}{4\ell_3}, \text{ and} \qquad 3)\ \frac{c}{2\pi}\sqrt{\frac{A_2}{A\ell_1\ell_2}},$$

where $c$ is the velocity of sound (the average temperature was 21.4 degrees and 344.3 m/s was used as $c$ to plot Fig. 3). Line 1 is the monotonic downward resonance curve of the back cavity, sloping down to the right. Line 2 is the monotonic upward resonance curve of the front cavity, sloping up to the right. The end correction of 0.82 $r$ with a flange was applied at the open end, where $r$ is the radius of the third tube (i.e., $D/2$ = 17 mm). Line 3 is the monotonic downward resonance curve of the Helmholtz resonator formed with the back cavity and constriction, plotted at the bottom of this figure.

From Fig. 3, we can observe that

- the peaks of the measured spectrograms match the theoretical resonance curves well; and
- the overall patterns are similar among the four versions, although the resonance peaks shift toward higher frequencies as $L$ and $\ell_2$ of the S3T model decrease.

## 3. Perceptual experiment

A perceptual experiment using vowel sounds recorded from the four versions of the S3T model was conducted with different $f_0$ frequencies [10]. In this experiment, we asked listeners to identify which of the five Japanese vowels they heard, as well as the speaker group.

### 3.1. Stimuli

The stimuli used in this experiment were a subset of the vowel sounds recorded in Section 2. There were 656 vowels resulting from the combinations of the four versions of the S3T model, the four rise-fall input signals with different starting $f_0$ frequencies, and the position of the constriction.

### 3.2. Procedure

The experiment was conducted in a sound-treated room. Stimuli were presented monaurally through a single loudspeaker (NAE NESmini) connected to an audio interface (RME Multiface) via an amplifier (NAE NES500). The five participants were seated 3.4-3.8 m from the loudspeaker. The sound level was 70.3 dBA on average. There was a training session with eight stimuli prior to the main experiment.

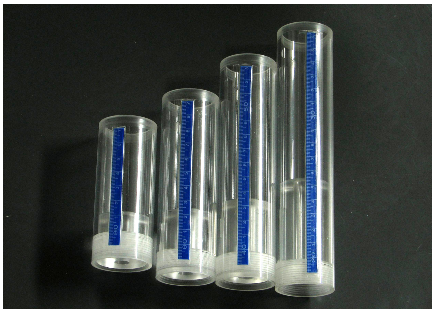

Figure 2: *The four versions of the S3T model with the four different combinations of outer and inner cylinder lengths used in this study.*

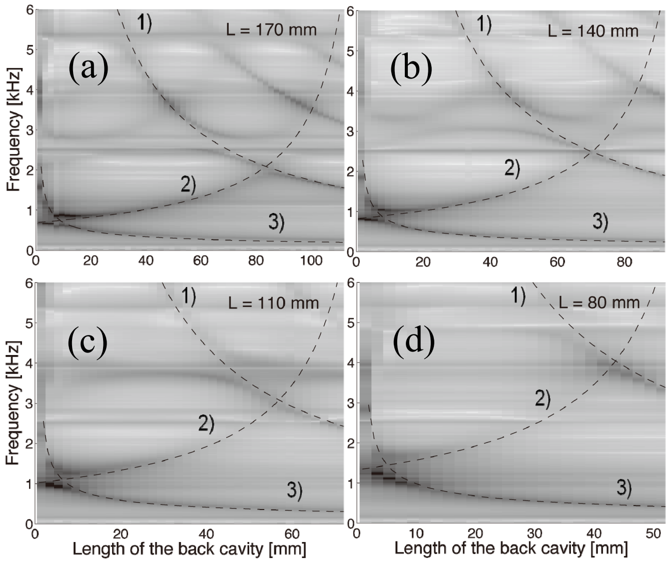

Figure 3: *Spectrographic representations of the vowels produced by each of the four S3T models (frequency vs. the length of the back cavity). The horizontal axes are scaled. (a) Ver. 1 (L = 170 mm), (b) Ver. 2 (L = 140 mm), (c) Ver. 3 (L = 110 mm), and (d) Ver. 4 (L = 80 mm).*

### 3.3. Participants

Twenty young listeners with normal-hearing (11 males and 9 females, ages 21 to 29 years) participated in the experiment. All were native speakers of Japanese and they were divided into four listener groups. Five participants from the listener group took part in the experiment simultaneously.

In the main experiment, a stimulus was presented in each trial, and the listeners were instructed to select one answer for each of the three questions displayed on a computer screen by means of a graphical user interface. The three questions were as follows:
Q1) Which vowel did you hear: "a, i, u, e, o"?
Q2) How well are you satisfied with your previous answer: 100%, 75%, 50%, 25%, or 0%?
Q3) Which would you say would be the most likely speaker: an adult male, an adult female, an elementary-school child or a baby?

For each listener group, 656 stimuli were presented randomly in 16 sessions, so that there were 41 trials in each session. After every five sessions the participants took a 15-minute break. Each session averaged 7-8 minutes in length.

### 3.4. Experimental results

Figure 4 shows the combined results for Q1 and Q2. For Q1, each stimulus was identified as one of the five vowels. For each stimulus, the satisfaction scores from Q2 were accumulated when a particular vowel was identified in Q1, and the accumulated score for each vowel was divided by the total number of participants, i.e., 20, to obtain the average score among participants. Figure 4 consists of the 16 combinations between the four versions of the S3T model and the four $f_0$ frequencies. The columns correspond to the four versions of the S3T model, and the rows correspond to the four $f_0$ frequencies. For each one of the 16 plots, the horizontal axis is the length of the back cavity ($\ell_1$) and the vertical axis is the five vowels. For each step of $\ell_1$, the average scores for the five vowels are displayed in grayscale on the plot. (The darkest color corresponds to the maximum value of the average score, i.e., 1.0).

Table 2: *The most frequently answered speaker group for each combination of L and $f_0$ (M: adult male, F: adult female, E: elementary-school child, and B: baby).*

| $f_0$ [Hz] | Ver. 1 | Ver. 2 | Ver. 3 | Ver. 4 |
|---|---|---|---|---|
| 100 | M | M | M | M |
| 200 | M/F | M/F | F | E |
| 300 | F | F | F/E | E |
| 400 | B | B | B/E | B |

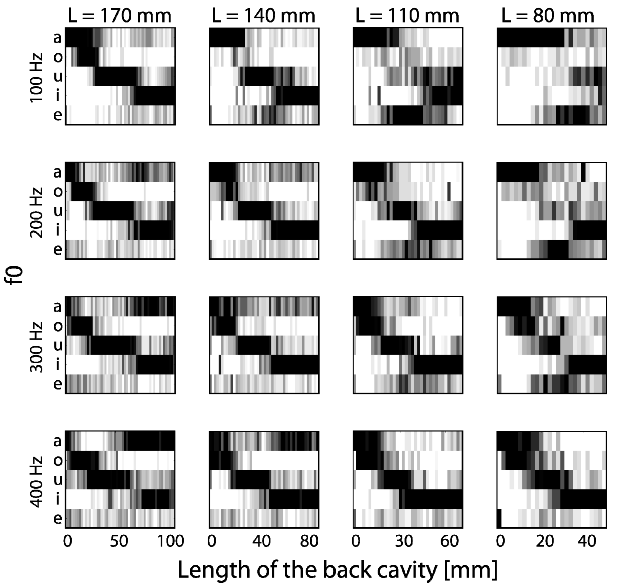

Figure 4: *Experimental results of the perceptual experiment for each of the 16 combinations between the four versions of the S3T model and the four $f_0$ frequencies. Within each plot, the horizontal axis is the length of the back cavity; the vertical axis is the five vowels; and the grayscale corresponds to the average score (see more details in the text).*

Table 2 shows the most frequently selected options for Q3 for each combination of *L* and $f_0$. From this table, we observed a primary tendency: listeners tended to identify the speaker as a female or child as $f_0$ increased. Also, listeners tended to identify a speaker as a female or child as *L* decreased.

## 4. Discussion

Infants often babble and are able to produce vowel-like sounds at an early age. The vowel /a/ is relatively easy to produce, because one need only open the oral cavity widely. When the tongue is raised, the vowel sounds more like /e/. When the mouth is more closed, the vowels /u/ or /o/ may be produced. Vowel /i/, may seem more difficult for babies to produce; however, based on the author's personal observation, it is even reasonable that they produce this vowel. In fact, we often hear /i/-like sounds produced by infants. One hypothesis for why infants can produce /i/ is the similarity in vocal tract configuration when producing /i/ and breastfeeding. When infants nurse, the center line of the tongue surface forms a grove, and the nipple is set between the grove and the "sucking fossa" at the hard palate. This configuration is similar to the configuration for the vowel /i/. We do see that even infants are able to produce a wide variety of vowels, and it is useful to investigate vowels spoken by small children as well as adults within a single framework.

In this study, we confirmed that the S3T models can produce vowels of both children and adults. In general, children's voices have higher $f_0$. Therefore, raising $f_0$ is crucial for making a child-like vowel. However, Table 2 shows that the vocal-tract length (VTL) is also important because the listeners answered "children/babies" when the VTLs were shorter.

We can also see the importance of the relationship between VTL and $f_0$ with vowel quality. In Fig. 4, we see that the same short VTL yields a different perception depending on $f_0$. The combination of the shortest VTL ($L$ = 80 cm) and the lowest $f_0$ (100 Hz) can only produce the vowels /a/, /e/ and /u/ along the vocal-tract length. However, at that same VTL, the vowels /i/, /u/, /o/, and /a/ are produced when $f_0$ = 400 Hz. These four vowels were the ones obtained when the VTL of 170 cm and $f_0$ = 100 Hz were combined for an adult male.

We looked at all 16 combinations to see whether the same set of four vowels were produced. We computed the correlation coefficients for all combinations between each of the 16 matrices and the matrix of VTL = 170 cm and $f_0$ = 100 Hz. Because the number of columns in each matrix in Fig. 4 is not the same, each row of the matrices was re-sampled to the minimum number, that is 26, so that the size of each matrix became 5 x 26. Then, the correlation coefficients were computed as shown in Table 3. As a result, the following combinations can stably produce the four vowels /i/, /u/, /o/, and /a/ with high correlation coefficients: 1) L = 170 cm, $f_0$ = 100, 200 Hz; 2) L = 140 cm, $f_0$ = 100-300 Hz; 3) L = 110 cm, $f_0$ = 200, 300 Hz; and 4) L = 80 cm, $f_0$ = 400 Hz. This shows that good vowel quality can be kept when the VTL is shortened and the $f_0$ frequency is increased simultaneously. This also explains why speech of boys at the age of puberty is less intelligible due to the voice mutation; there is a sudden drop in the fundamental frequency, whereas the VTL changes more slowly.

The relationship between VTL and $f_0$ frequency seems to be monotonic, although the change in VTL is greater with a high $f_0$ than with a low $f_0$. This observation is consistent with a previous study [11]. Furthermore, we hear the same vowels even though the VTL and $f_0$ frequencies are different. This is because our auditory system can extract and separate information about the size and the shape of the vocal tract [12].

In this study as well as in [10], we have seen that the S3T models can produce vowels from children and adults, and we've seen that the models are useful for educational purposes as well. By using acoustic tubes with different lengths, we can easily compare vowels of children and adults in a classroom demonstration. We can also incorporate such a demonstration into an exhibition at a science museum.

We have used the S3T model in a science workshop, where participants created their own vocal-tract models and produced vowel sounds. For a child model, we needed a sound source with higher $f_0$ frequency. We used a reed-type whistle as a sound source in the science workshop. To raise the $f_0$ frequency, we can change the length of the reed: 20 mm for an adult male, 10 mm for an adult female, and 5 mm for a child, etc. In the workshop, we created a slide whistle together with the S3T model. The mouthpiece for the slide whistle is easily made when the tube length is short, as for a child model.

Table 3: *The correlation coefficients for all combinations of each of the 16 matrices and the matrix of VTL = 170 cm (Ver. 1) and $f_0$ = 100 Hz in Fig. 4.*

| $f_0$ [Hz] | Ver. 1 | Ver. 2 | Ver. 3 | Ver. 4 |
|---|---|---|---|---|
| 100 | 1.000 | 0.873 | 0. 772 | 0. 586 |
| 200 | 0.827 | 0.927 | 0.853 | 0. 724 |
| 300 | 0.509 | 0.831 | 0.813 | 0.714 |
| 400 | 0.195 | 0.544 | 0. 763 | 0.833 |

# 5. Conclusions

In the present study we confirmed that the S3T models can produce vowels from children and adults when the size of the S3T model is changed and when it is combined with different $f_0$ frequencies. From the perceptual experiment we were able to obtain good vowel quality when the vocal-tract length was shortened as the $f_0$ frequency was increased. In the future, we would like to use the child model more frequently in science workshops and classroom demonstrations.

# 6. Acknowledgements

I acknowledge the anonymous reviewers for their helpful comments. This work was partially supported by Grants-in-Aid for Scientific Research (21500841, 24501063) from the Japan Society for the Promotion of Science, and Sophia University Open Research Center from MEXT. I would also like to thank Keiichi Yasu for his assistance.

# 7. References

[1] Arai, T., "The replication of Chiba and Kajiyama's mechanical models of the human vocal cavity," *J. Phonetic Soc. Jpn.*, 5(2):31-38, 2001.

[2] Arai, T., "Education system in acoustics of speech production using physical models of the human vocal tract," *Acoust. Sci. Tech*., 28(3):190-201, 2007.

[3] Chiba, T. and Kajiyama, M., *The Vowel: Its Nature and Structure*, Tokyo-Kaiseikan Pub. Co., Ltd., Tokyo, 1941.

[4] Arai, T., "Sliding three-tube model as a simple educational tool for vowel production," *Acoust. Sci. Tech*., 27(6):384-388, 2006.

[5] Arai, T., "Education in acoustics and speech science using vocal-tract models," *J.Acoust. Soc. Am.*, 131(3), Pt. 2, 2444-2454, 2012.

[6] Fant, G., *Acoustic Theory of Speech Production*, Mouton, The Hague, Netherlands, 1960.

[7] Arai, T., "Sliding vocal-tract model and its application for vowel production," *Proc. of Interspeech*, 72-75, 2009.

[8] Stevens, K. N., *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1998.

[9] Arai, T., "Science workshop with sliding vocal-tract model," *Proc. of Interspeech*, 2827-2830, 2008.

[10] Arai, T., "Producing vowels with sliding three-tube model: Effects of vocal-tract length and fundamental frequency," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 299-302, 2011.

[11] Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A. and Johnson, K., "Formants of children, women, and men: The effects of vocal intensity variation," *J. Acoust. Soc. Am.*, 106(3):1532-1542, 1999.

[12] Irino, T. and Patterson, R. D., "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised Wavelet-Mellin transform." *Speech Commun.*, 36:181-203, 2002.