Producing Vowels with Sliding Three-tube Model:

Effects of Vocal-tract Length and Fundamental Frequency [*]

○Takayuki Arai（Sophia University）

1 Introduction

Based on Fant's three tube model [1], the sliding three-tube (S3T) model was proposed by Arai [2] as a part of a set of educational tools with other physical models of the human vocal tract. The S3T model consists of an outer and inner cylinder, as shown in Fig. 1. By sliding the inner cylinder within the outer cylinder, we can produce different vowel sounds. 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 the quality of vowel drastically.

Although we have mainly examined vowels by adult males with the S3T models in our previous studies, we knew empirically that the S3T model could also produce children's vowels. From both an educational and academic point of view it is useful to have both a child and adult version of the S3T model. Therefore, in this study, we made smaller versions of the model and combined them with a variety of sound sources to see the effects of vocal-tract length and fundamental frequency on vowel production.

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. Let L and ℓ_2 be the lengths of the outer and inner cylinders, respectively, and let ℓ_1 and ℓ_3 be the lengths of the back and front cavities, respectively (Fig. 1). When the inner cylinder slides inside the outer cylinder, ℓ_1 and ℓ_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 diameters of the holes of the outer and the inner cylinders, respectively. Table 1 shows the values of the parameters of the four versions designed in this study. Both the inner and outer cylinders were made of acrylic resin, and the outer cylinder was 3 mm thick.

2.2 Producing, recording, and analyzing vowels made by the four versions of the S3T model

We produced and recorded vowels made by the four versions of the S3T model described in Section 2.1. For the recordings, 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).

We used six different input signals. 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 f0 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.

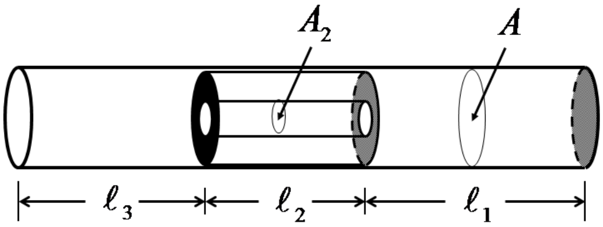

Fig. 1. Schematic figure of the S3T model (adapted from [2]).

[*] スライド式３音響管モデルによる母音の生成 －声道長と基本周波数の影響－ 荒井隆行（上智大・理工）.

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 ℓ_1 varies from 0 to $L - \ell_2$ in 2 mm steps. The signals recorded were synchronously averaged multiple times to gain the signal-to-noise ratio.

3 Perceptual experiment

We conducted a perceptual experiment using vowel sounds recorded from the four versions of the S3T model, which have different f_0 frequencies. In this experiment, we asked listeners to identify the five Japanese vowels and speaker groups.

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 between the four versions of the S3T model and the four rise-fall input signals with different starting f_0 frequencies.

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

	L [mm]	ℓ_2 [mm]	D [mm]	d [mm]
Ver. 1	170	60	34	10
Ver. 2	140	50	34	10
Ver. 3	110	40	34	10
Ver. 4	80	30	34	10

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]	100	200	300	400
Ver. 1	M	M/F	F	B
Ver. 2	M	M/F	F	B
Ver. 3	M	F	F/E	B/E
Ver. 4	M	E	E	B

3.2 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.

3.3 Procedure

The experiment was conducted in a sound-treated room. Stimuli were presented monaurally through a 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 dB (A) on average. There was a training session with eight stimuli prior to the main experiment.

In the main experiment, a stimulus was presented in each trial and the listeners were instructed to select one of the options for each of three questions displayed on the 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 with 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 2 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. The vowel that had the highest average score was plotted at the midpoint of the inner cylinder as it is located along the vocal-tract center line in Fig. 2. Note that the center lines are bent at right angles halfway through the length of the vocal-tract. The locations of the turning points were based on the average lengths of the oral and pharyngeal cavities [3,4]. If the average satisfaction score was less than 75% for the particular stimulus, the vowel was not plotted but marked as “.” in this figure.

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 from children through 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 more “children/babies” for shorter VTLs.

We can also see the importance of the relation between VTL and f_0 in terms of vowel quality. From Fig. 2, 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 /e/ and /a/ 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. Thus, we looked at all 16 combinations to see whether the same set of four vowels were produced. As a result, the following combinations can stably produce these four vowels: 1) L = 170 cm, f_0 = 100, 200 Hz; 2) L = 140 cm, f_0 = 200, 300 Hz; 3) L = 110 cm, f_0 = 300, 400 Hz; and 4) L = 80 cm, f_0 = 400 Hz. The relation 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 the previous study [5]. 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 [6].

Figure 2 also shows the relation between vowel distribution and the turning point of the vocal tract. After normalizing the VTL, vowel distributions are more or less similar among the combinations 1) through 4) above. This yields relatively proportional increase in terms of the formant frequencies as a function of VTL. However, the pharyngeal cavity witnesses a greater change in growth than the oral cavity as children age. Therefore, some vowels shift in articulation from the oral cavity to the pharyngeal cavity as children grow. Similar discussions were pointed out in [7].

Thus, we have seen that the S3T models can produce vowels from children through adults, and they are useful for educational purposes as well. In a classroom demonstration we can easily compare vowels of children and adults by using acoustic tubes with different lengths. Such a demonstration might also be incorporated into an exhibition at a science museum. We have also 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 need a sound source with higher f_0 frequency. In the science workshop, we used a reed-type whistle as a sound source. 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 often created a slide whistle together with the S3T model, but the mouthpiece for the slide whistle can be made easily when the tube length is short, as for a child model.

5 Conclusions

In this study, we confirmed that the S3T models can produce vowels from children through adults when the size of the S3T model is changed and the models are combined with a variety of sound sources. From the perceptual experiment, we were able to obtain good vowel quality when the vocal-tract length was shortened and the fundamental frequency was increased simultaneously. In the future, we would like to use the child model more frequently at a science workshop as well as in a classroom demonstration.

Acknowledgements

This work was partially supported by Grants-in-Aid for Scientific Research (21500841) from the Japan Society for the Promotion of Science, and Sophia University Open Research Center from MEXT.

References

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

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

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

[4] Borden, G. J. and Harris, K. S., *Speech Science Primer*, Williams & Wilkins, Baltimore, 1984.

[5] 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.

[6] 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.

[7] Kasuya, H., Suzuki, H. and Kido, K., “Changes in pitch and frist three formant frequencies of five Japanese vowels with age and sex of speakers,” *J.Acoust. Soc. Jpn.*, 24(6), 355-364, 1968.

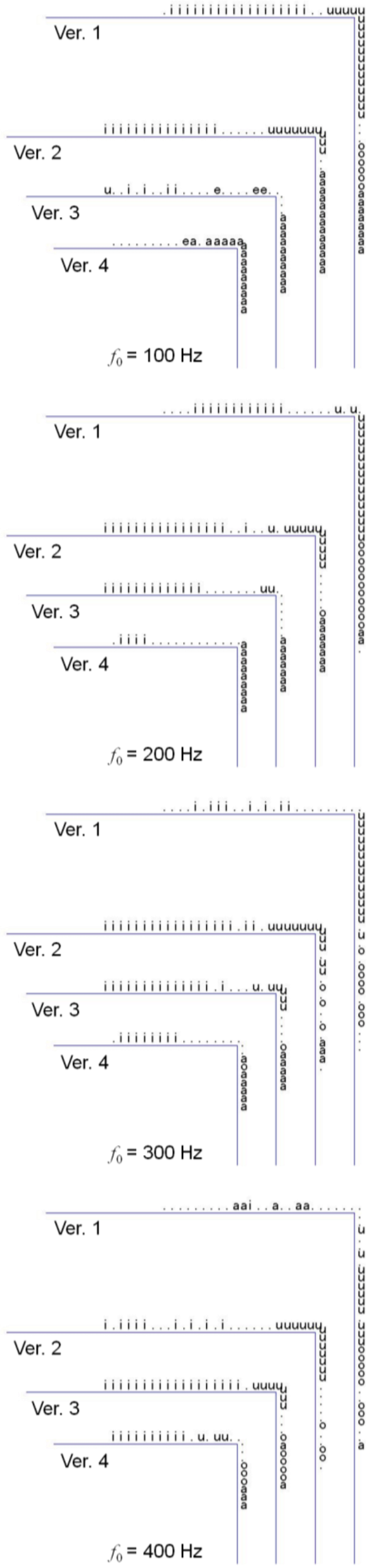

Fig. 2. Identified vowels in the perceptual experiment for each of the 16 combinations between the four versions of the S3T model and the four f_0 frequencies.