INTERSPEECH 2009 BRIGHTON

Sliding Vocal-tract Model and its Application for Vowel Production

Takayuki Arai

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

Abstract

In a previous study, Arai implemented a sliding vocal-tract model based on Fant's three-tube model and demonstrated its usefulness for education in acoustics and speech science. The sliding vocal-tract model consists of a long outer cylinder and a short inner cylinder, which simulates tongue constriction in the vocal tract. This model can produce different vowels by sliding the inner cylinder and changing the degree of constriction. In this study, we investigated the model's coverage of vowels on the vowel space and explored its application for vowel production in the speech and hearing sciences.

Index Terms: vocal-tract model, three-tube model, vowel production, education in acoustics, speech science

1. Introduction

Arai [1] proposed a sliding three-tube (S3T) model, also called a sliding vocal-tract model, based on Fant's three tube model [2]. Like other physical models of the human vocal tract, such as the cylinder and plate-type models [3,4] based on the vocal tract measurements done by Chiba & Kajiyama (1941) [5], the S3T models teach basic concepts in speech production. For example, the concept of the source-filter theory and the relationship between vocal-tract shapes and the quality of vowels can easily be taught using these models.

The S3T model can produce different vowel sounds by sliding an inner cylinder within the outer cylinder as shown in Fig. 1, a schematic view of its mid-saggital cross-section. In this model, the inner cylinder forms a constriction like that of the tongue within the vocal tract. Unlike the cylinder-type models, the S3T model demonstrates that one single degree of freedom can control and change the quality of vowel. If we add another degree of freedom, the constriction area of the tongue, the variety of vowels can be increased.

Because of this simple structure, the rough estimation of the first few resonance frequencies can easily be done [6] by decomposing the whole model into three parts:

a quarter-length resonator with the first resonance frequency of $\frac{c}{4\ell_3}$,

a half-length resonator with the first resonance frequency of $\frac{c}{2\ell_1}$, and

a Helmholtz resonator with the resonance frequency of $\frac{c}{2\pi}\sqrt{\frac{A_2}{A\ell_1\ell_2}}$.

Therefore, this S3T model is useful for explaining the basics of the acoustic theory of vowel production for undergraduate and graduate students. Further, the simple structure of this model makes it possible to use for a science workshop where children make their own vocal-tract models (see details in [7]).

A nomogram can be drawn with this S3T model as a function of the position of a constriction [1]. When the inner cylinder slides inside the outer cylinder, the lengths of the first (back) and the third (front) tubes, or ℓ_1 and ℓ_3, vary from 0 to $L-\ell_2$, where L is the length of the outer cylinder and $L=\ell_1+\ell_2+\ell_3$, and ℓ_2 is the length of the inner cylinder. The dots in Fig. 2 show the measured formant frequencies (up to 3 kHz) of the S3T model (L is 175 mm, in this case) as a function of the back tube length ℓ_1, where ℓ_1 was shifted from 1 to 125 mm in 2 mm steps [1]. This figure also shows the three underlying resonance curves (solid lines). In this figure, the hole diameter D of the outer cylinder was 34 mm, and the length of the constriction was 50 mm. There were two settings for the hole diameter d of the inner cylinder; 10 mm (Setting 1) and 24 mm (Setting 2). The cross-sectional areas A and A_2 were $A=\frac{\pi}{4}D^2$ and $A_2=\frac{\pi}{4}d^2$, respectively. As you can see in Figure 2, the formant frequency estimation using the simple approximation was quite similar to the measured frequencies.

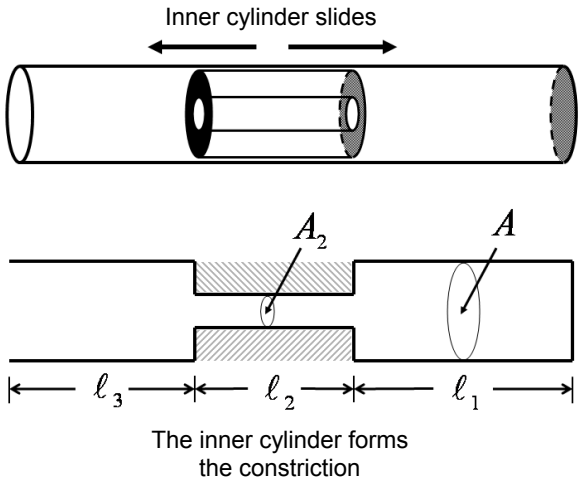

Figure 1: *Schematic figure of the "sliding three-tube model" (adapted from [1]). The inner cylinder (constriction) slides back and forth inside the outer cylinder and the shape of the vocal tract varies, simulating different vowels.*

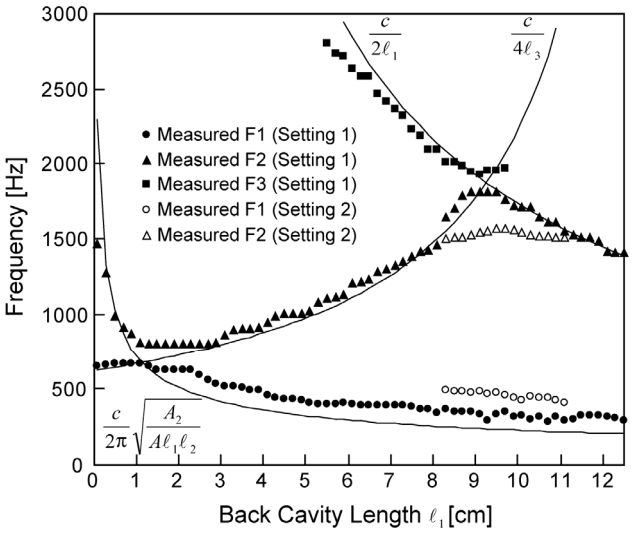

Figure 2: *Measured formants and underlying resonances produced by the S3T model (from [1]).*

In this study, we first investigated how wide a coverage the S3T model had of different vowels on the vowel space. Then we reported several options to increase the coverage, and we discussed the model's possible applications for speech & hearing sciences and acoustic phonetics.

2. Coverage on the vowel space

2.1. Simulation

First, we used Vtcalcs, a speech production simulator by Maeda [8], to investigate the S3T's vowel coverage. This simulator computes the transfer function based on a given vocal-tract shape, and the estimated formant frequencies were used to plot of the first (F1) and the second (F2) formants. In this simulation, L was set to 170 mm, and the position of the tongue constriction was shifted from 0 to $L - \ell_2$ in 10 mm steps.

Figure 3 shows the F1-F2 plot when d = 10 mm and $\ell_2 =$ 40, 50, 60, 70 mm. When d is less than 10 mm, the result becomes a fricative-like consonantal sound, and therefore, d = 10 mm is the minimal diameter [6]. From this figure, the produced vowel travels an outer edge, i.e., /i/ -> /u/ -> /o/ -> /a/ or /ae/. (Note that there are markers "x" indicating the average F1 and F2 frequencies for the major American English vowels. The values are taken from [6]. The same markers were used from Fig. 3 through Fig. 6.) The four trajectories with different ℓ_2 values are overlaid with respect to each other, and differences among them only appeared in the low vowel region, when the constriction was positioned at the glottis end. In this position, the produced vowel is like /ae/ when $\ell_2 =$ 40 and 50 mm, whereas the produced vowel is more like /a/ when $\ell_2 =$ 70 mm.

Figure 4 shows the similar F1-F2 plot when $\ell_2 =$ 40, 50, 60, 70 mm, but d = 20 mm, in this case. From this figure, the trajectories are compressed more towards the central region. The produced vowel travels from the region of /e/, compared with /i/ in the previous case. Again, the four trajectories with different ℓ_2 values are overlaid with respect to each other, and the differences among them only appeared when the constriction was positioned at the glottis end.

In addition to d = 10 and 20 mm, we used several other d values, such as, 15 and 25 mm. As a result, we observed that the trajectories shifted and were compressed towards the central region on the vowel space. Because the resulting tube finally becomes uniform when $d = D$ (34 mm, in this simulation), the produced vowel becomes "schwa," and the F1 and F2 frequencies approximate to 504 and 1418 Hz, respectively.

2.2. Measurement

We recorded the output signals from the S3T model and measured their formant frequencies. For the recordings, the basic parameters were the same as the values used for the measurements in Fig. 2 [1]. Both the inner and the outer cylinders were acrylic resin, and the thickness of the outer cylinder was 3 mm. A driver unit (TOA TU-750) for a horn speaker was attached to the outer cylinder. An impulse train was fed into the driver unit via the digital-to-analog (D/A) converter of a digital audio amplifier (Onkyo MA-500U); the sampling frequency was 16 kHz. 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 hard rubber cylindrical filler inside the neck. We made a hole in the center of the rubber filling with an area of 0.07 cm^2. A flange with the diameter of 25 cm was attached at the open end of the tube. The output sounds were recorded using a microphone (Sony ECM-23F5) and a digital recorder (Marantz PMD670) with a sampling frequency of 16 kHz. The microphone was placed approximately 20 cm in front of the output end in an anechoic room.

Figure 5 shows the F1-F2 plot when d = 10 mm and $\ell_2 =$ 40, 50, 60, 70 mm. The measurement was done by a formant estimation by linear prediction (the order was 24) with the Wavesurfer software. Figure 6 shows the F1-F2 plot when d = 20 mm and $\ell_2 =$ 50 mm. In these figures, the data points were omitted when the formant estimation was not successful.

3. Discussion

3.1. Simulation vs. measurement

When comparing the results of the simulation by Vtcalcs (Figs. 3 and 4) and the physical measurement of the S3T model (Figs. 5 and 6), the F1-F2 plots show a similar tendency. In both results, the coverage spreads out widely on the vowel space. One of the reasons there are some differences between them is the total length of the tube. The tube length in the simulation was 170 mm, while that of the physical measurement was 175 mm. Therefore, the results in Figs. 5 and 6 are less than the ones in Figs. 3 and 4 (approximately 3%).

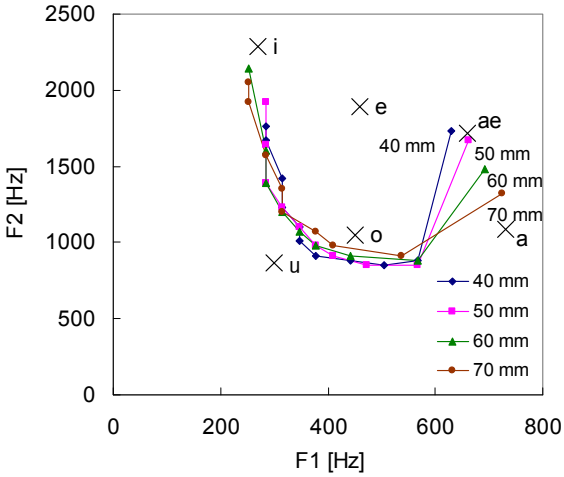

Figure 3: *F1-F2 plot of the simulated vowels when the diameter of the constriction* d *= 10 mm. The length of the constriction* ℓ_2 *= 40, 50, 60, 70 mm was used as a parameter to draw the four trajectories.*

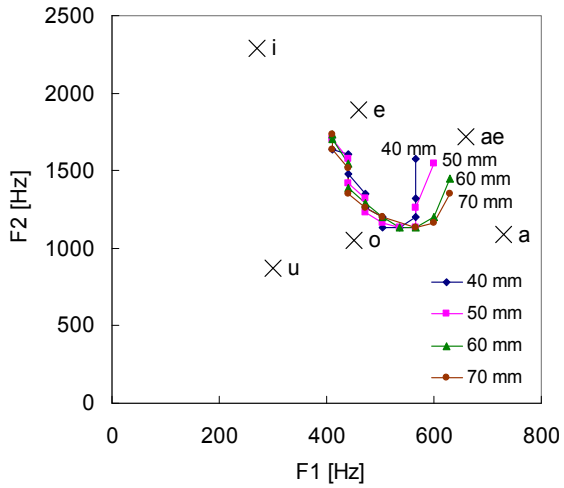

Figure 4: *F1-F2 plot of the simulated vowels when the diameter of the constriction* d *= 20 mm. The length of the constriction* ℓ_2 *= 40, 50, 60, 70 mm was used as a parameter to draw the four trajectories.*

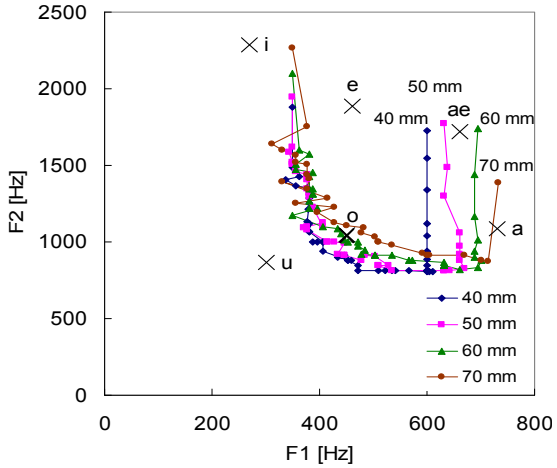

Figure 5: *F1-F2 plot of the measured vowels when the diameter of the constriction* d *= 10 mm. The length of the constriction* ℓ_2 *= 40, 50, 60, 70 mm was used as a parameter to draw the four trajectories.*

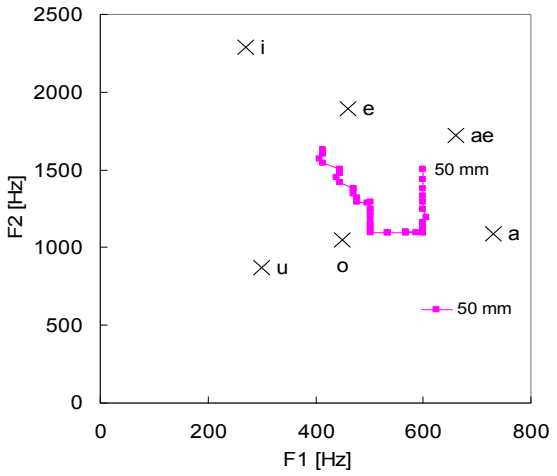

Figure 6: *F1-F2 plot of the measured vowels when the diameter of the constriction* d *= 20 mm. The length of the constriction* ℓ_2 *= 50 mm was used.*

3.2. Laryngeal constriction / lip rounding

As one looks at the models by Chiba & Kajiyama [5] and Arai [3], the laryngeal part is narrower than the main vocal tract. To simulate this laryngeal constriction we put a 20-mm long constriction with a diameter of 10 mm. As a result, we confirmed that the /i/ vowel was improved. According to the perturbation theory [5,6], the volume velocity minimum is located at the glottis end for all resonances, so F1 and F2 frequencies increase when there is a constriction at the glottis end. Thus coverage around the /i/ region was improved with laryngeal constriction.

For /u/, on the other hand, both the F1 and F2 frequencies are low. To improve this region, it is important to consider lip rounding and/or lip protrusion. The perturbation theory tells us that the volume velocity maximum is located at the lip end for all resonances, so F1 and F2 frequencies decrease when there is a constriction at the lip end. Thus the coverage around the /u/ region was improved with lip rounding. Lip rounding is also important for other vowels, such as, /o/. However, the S3T can produce vowel /o/ without the lip rounding, as well. Figure 7 shows the S3T model with laryngeal constriction and lip rounding, which are removable options in this picture.

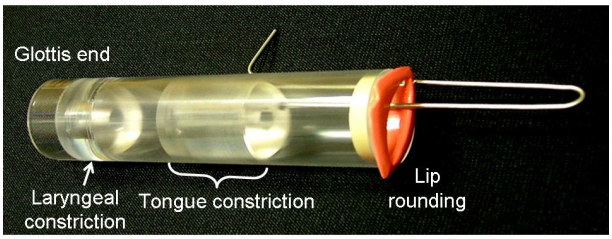

Figure 7: *The S3T model with the laryngeal constriction and the lip rounding.*

Figure 8: *A slider with a mechanism to vary the area of the constriction. A user can rotate and change the angle of the two semicylindrical parts so that the area for the tongue constriction can be changed. The schematic cross-sections are also drawn.*

3.3. Application for speech & hearing sciences / acoustic phonetics

As we have seen, the S3T model mainly has two degrees of freedom; one is the location of tongue constriction, and the other is the constriction area of the tongue. In Section 2.2, we prepared several inner cylinders as a slider with different hole diameters. In this case, we continuously changed the location of the constriction; however, we need to replace a slider to another to change the constriction area. To continuously change these two variables with a single slider, we designed the slider shown in Fig. 8. This slider has a physical mechanism for varying the area of the constriction by rotating the angle of the two semicylindrical parts. With the S3T model and this sliding part, we can quickly and effectively demonstrate the production of many different vowels for pedagogical purposes.

Figure 9 shows a head-shaped version of the S3T model. In this case, the tube is bent and placed in the head model so that learners of acoustics can easily identify the location of the vocal tract.

The S3T model can also be used to investigate how the auditory system segregates vowel-type information from information relating to the size of the speaker [9]. With a shorter tube simulating a child's vocal tract, we can clearly produce a set of different vowels, indicating that the S3T effectively models the essence of human vowel production.

4. Conclusions

A precursor in the attempt to demonstrate the anatomical-physiological origin of vowels by using a cylindrical vocal-tract model with a slider that changed the length of the vocal tract was Robert Willis (1830) [10], although only one characteristic resonance was considered. On the other hand, Arai [1] proposed a sliding vocal-tract model and showed the

Figure 9: *Head-shaped version of the S3T model.*

multiple resonances match well to the modern acoustic theory of vowels. In this study, we reviewed the sliding vocal-tract model proposed by Arai [1] and confirmed that the model produces a wide range of vowels on the vowel space. We also explored further developments and applications for vowel production in speech & hearing sciences and acoustic phonetics.

5. Acknowledgements

I acknowledge the assistance of Kanae Amino, Hinako Masuda, Marie Ogawa, Kimi Tanaka and Misaki Tsuji. I also thank the anonymous reviewers for their helpful comments. This work was partially supported by Grants-in-Aid for Scientific Research (19500758) from the Japan Society for the Promotion of Science, and Sophia University Open Research Center from MEXT.

6. References

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

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

[3] 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.

[4] 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.

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

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

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

[8] Maeda, S., "A digital simulation method of the vocal-tract system," *Speech Communication*, 1:199-229, 1982.

[9] 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.

[10] Willis, R., "On vowel sounds, and on reed organ pipes," *Transactions of the Cambridge Philosophical Society* III, 231-276, 1830.