Retroflex and Bunched English /r/ with Physical Models of the Human Vocal Tract

Takayuki Arai

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

Abstract

It is known that American English /r/ can be produced as a retroflex or bunched /r/, but it can be challenging to teach students how to articulate both. We already developed a physical model for retroflex /r/ and demonstrated that the model produces the /r/ sound. However, almost no studies have reported a physical model for bunched /r/. We developed a new physical model using sliding blocks for the lips and tongue to help teach students how to produce bunched /r/. We recorded several sets of sounds produced by the models, analyzed the output signals, and used them for perceptual experiments. Acoustic analysis and perceptual experiments confirmed that the retroflex and bunched /r/ models produced clear American /r/ sounds, and that the narrow constriction placed between 5-7 cm from the lips seems to be the key in producing these sounds. Furthermore, bunched /r/ with lip rounding produced the most clear /r/ sound. Both models are helpful for practicing pronunciation because learners can readily see there are two ways to produce /r/, they can see and alter the tongue position manually, and they can hear the output sounds.

Index Terms: speech production, physical models of the human vocal tract, tongue, retroflex /r/, bunched /r/

1. Introduction

Since the year 2000, we have developed different types of vocal-tract models, mainly for education in acoustics and speech science [1-4]. We demonstrated the usefulness of the models for phonetic education for both native and non-native speakers [e.g., 5]. For native speakers who are very young, or for patients with speech disorders, the approximants /r/ and /l/ may be difficult to pronounce. Also for non-native language learners, the Japanese, for example, production of native-sounding American /r/ and /l/ can be problematic.

Therefore, Arai (2003) developed a mechanical vocal-tract model with a movable tongue, specially designed to produce American English approximants [5]. We call this Model A. In model A, half of the tongue can be raised around a pivot located in the middle of the tongue. Model A can produce both the retroflex and lateral approximants.

However, American English /r/ is not always produced as a retroflex, with the tongue tip raised. Some American speakers produce /r/ with the tongue tip down and the sides of the tongue bunched up against the top back teeth. This is called the “bunched /r/.” Bunched /r/ is often easier for students having difficulty with /r/ to learn to produce, perhaps because one can feel the sides of the tongue bunch up and touch the teeth. For some, it would seem, having a solid boundary against which to set the tongue is easier to imitate than the more nebulous instruction to curve the tongue tip back, as for retroflex /r/. Both bunched and retroflex /r/ have similar spectral characteristics for the frequency range of the 1st to 3rd formants [6] (Zhou *et al.* [7] have shown that these two versions of /r/ differ considerably in the spacing between the 4th and the 5th formants).

In the present study, we designed a dynamic model for bunched /r/, which we call Model B. This new model was partially based on Umeda & Teranishi's model with sliding strips, where the space created between the inner wall and the strips forms an arbitrary vocal tract shape [8]. Model B also has sliding strips but only in the oral cavity. Furthermore, unlike Umeda & Teranishi's model, the vocal tract of Model B is bent in the middle, as in Model A. We compared the acoustic and perceptual characteristics of Models A (retroflex) and B (bunched) and evaluated the degree to which the output sound exemplified American English /r/.

2. Design

2.1. Model A for retroflex /r/ [5]

During production of retroflex /r/, the tongue tip is curved upward but does not make contact with the palate. Figure 1 shows two photographs of Model A from [5], which produces the retroflex approximant. Figure 2 shows two corresponding schematic illustrations of the same model. The tongue on the model is made of aluminum. In these figures, when the model is in resting position (Figs. 1a and 2a), the vocal tract is configured for the vowel /a/.

The original model A was designed to produce both the retroflex approximant /r/ and the alveolar lateral approximant /l/. For the alveolar lateral approximant, the tongue blade touches the palate. This can be done with an extended tongue. For the retroflex approximant, a shortened tongue is used, so that the tongue does not touch the palate as shown in Figures 1b and 2b. These figures show the tongue bent toward the middle portion of the palate to produce a retroflex approximant. There is a lever to control the movement of the tongue. Pushing the lever toward the back causes the tongue to rise. By moving the lever back and forth one can move the tongue back and forth as shown in Figures 1a/2a to 1b/2b and back to 1a/2a.

2.2. Model B for bunched /r/

During production of bunched /r/, the tongue dorsum is in a concave shape, but is "bunched" toward the palate, and the tongue touches the back top teeth. This movement is often accompanied by lip rounding [7]. To simulate bunched /r/, we designed a new model, Model B, shown in Figure 3. Figure 3 shows the vocal tract bent at a right angle in the middle of its length. This model consists of several blocks and an outer frame made of transparent acrylic material. The standard

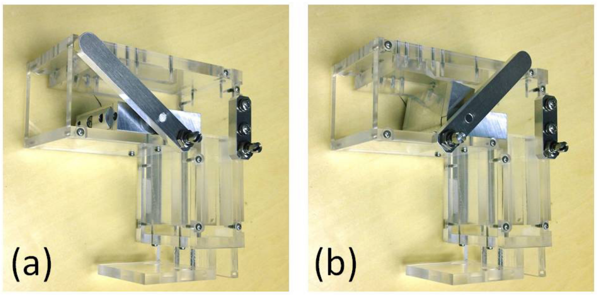

Figure 1: *Model A for retroflex /r/: (a) the tongue is in resting position; (b) the tongue blade is not touching the palate, but the tongue is retroflexed (from [5]).*

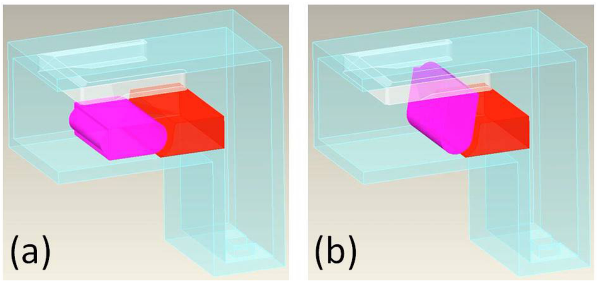

Figure 2: *Schematic illustrations of Model A for retroflex /r/: (a) the tongue is in resting position; (b) the tongue blade is not touching the palate, but the tongue is retroflexed.*

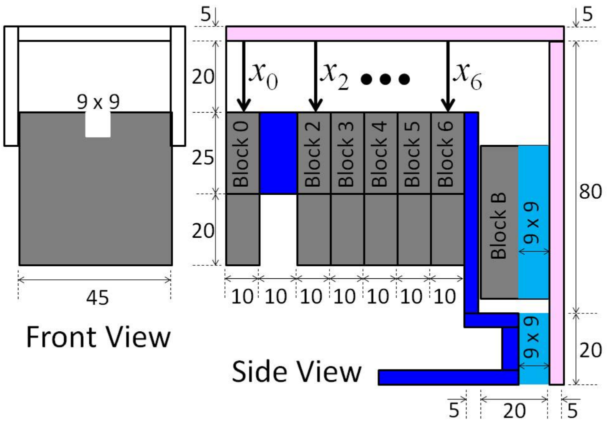

Figure 3: *Design of Model B for bunched /r/.*

dimension of the cross section of the inside of the outer frame is 45 x 20 mm. The detailed dimensions of this model are indicated in mm by the numbers in this figure.

Block 0 and Blocks 2 through 6 are movable, perpendicularly. In resting position, the distance x_i (i = 0, 2, ..., 6) from the top plate is maximally 20 mm. As each block is pushed up, the distance x_i decreases to 0 mm as the block touches the top plate. The blocks return to resting position by their own weight. Block 0 simulates the lips, while Blocks 2 through 6 simulate the top surface of the tongue. The top of each block has a 9 x 9 mm notch at its center along the vocal tract length. The notches of Blocks 2 through 6 simulate the groove of the tongue and its concave shape.

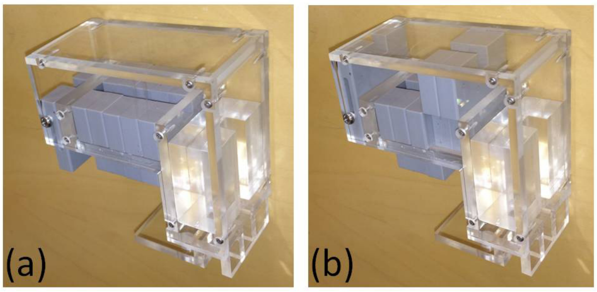

Figure 4: *Model B for bunched /r/: (a) the tongue blocks and the lip block are in resting position; (b) the tongue blocks and the lip block are raised.*

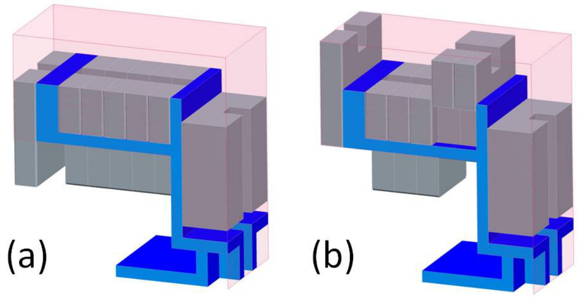

Figure 5: *Schematic illustrations of Model B for bunched /r/: (a) the tongue blocks and the lip block are in resting position; (b) the tongue blocks and the lip block are raised.*

The dimension of the laryngeal cavity of this model is 9 x 9 x 20 mm, as indicated in light blue at the bottom right corner of Figure 3. There is another block, Block B, placed 5 mm above the laryngeal cavity. Block B is in position for the narrow pharyngeal cavity characteristic of the vowel /a/. Block B also has a 9 x 9 groove on the side facing the pharyngeal wall, also indicated in light blue in Figure 3.

Figures 4 and 5 portray photographs and schematic illustrations of Model B for bunched /r/. In Figures 4a and 5a, the tongue is in resting position, and the vocal tract is configured for vowel /a/. Figures 4b and 5b show the tongue dorsum raised, as well as the lip block raised for lip rounding.

3. Perceptual evaluation

3.1. Stimuli

We conducted a perceptual test on recordings of output sounds produced from Models A and B. A driver unit (TOA, TU-750) for a horn speaker was attached to the glottis end of the model. The input signal for the recordings was a lowpass-filtered impulse train with a sampling frequency of 16 kHz. The duration of the input signal was 800 ms, and its fundamental frequency started at 110 Hz and ended at 135 Hz. The input signal was fed into the driver unit via an audio interface (Onkyo, MA-500U). 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 placed the glottis end of each vocal-tract model on top of a thin metal plate, directly connected to the metal filling inside the neck. There was a hole in the center of the metal filling and plate with an area of 0.13 cm^2.

Table 1: *Perceptual evaluation of Model A. The numerical values in this table show the degrees of tongue rotation.*

Angle (deg.)	Transcription
50	not clear
60	not clear
70	r
80	r
90	r
100	r / l

Table 2: *Perceptual evaluation of Model B. The numerical values in this table show the vertical position of each block, x_i, in mm.*

x_0	x_2	x_3	x_4	x_5	x_6	Transcription
20	20	3	20	20	20	l
0	20	3	20	20	20	l
20	20	20	3	20	20	r / l
0	20	20	3	20	20	not clear
20	20	20	20	3	20	r
0	20	20	20	3	20	r
20	20	20	20	20	3	r
0	20	20	20	20	3	r
20	20	3	3	20	20	l
0	20	3	3	20	20	r / l
20	20	20	3	3	20	r
0	20	20	3	3	20	r
20	20	20	20	3	3	r
0	20	20	20	3	3	r
20	20	3	3	3	20	l
0	20	3	3	3	20	r(l)
20	20	20	3	3	3	r
0	20	20	3	3	3	r

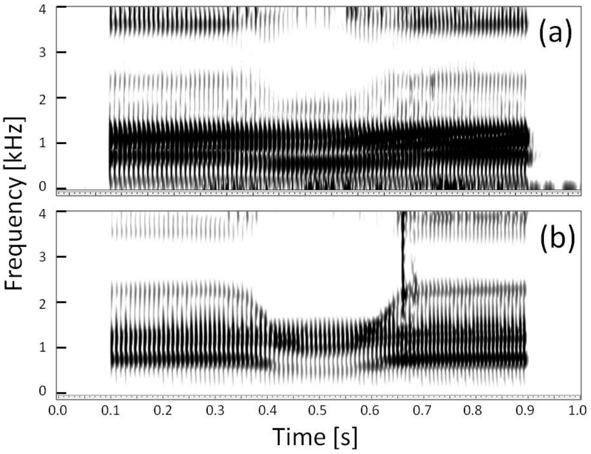

Figure 6: *Spectrograms of the output sounds from (a) Model A and (b) Model B.*

The stimulus set was all /aCa/ utterances, where /C/ was the target consonant surrounded by the vowel /a/. There were 6 stimuli from Model A and 18 stimuli from Model B. The recordings were conducted in a sound-treated room. The output signals from Models A and B were recorded digitally with a digital audio recorder (Marantz, PMD660) with a microphone (Sony, EM-23F5). The original sampling frequency of 48 kHz for the recordings was retained for the perceptual evaluation but converted into 8 kHz for the acoustic analysis.

3.2. Procedure

In the perceptual evaluation test, stimuli were presented diotically through headphones (Sennheiser, HD595). One experienced phonetician, a native speaker of American English, participated in the test. The 24 stimuli were ordered randomly. The phonetician transcribed each stimulus phonemically, identifying which phoneme each output sound most closely resembled.

3.3. Results

3.3.1. Model A

We manipulated Model A to make /aCa/ utterances, where the target for C was retroflex /r/. The lever was rotated to raise the first half of the tongue, and the final angle of the tongue rotation was controlled from 50 to 100 degrees in 10 degree steps. The lever was rotated back to resting position as soon as it reached the maximum. This series of movements was done manually by the author after a training session. Figure 6(a) shows the spectrogram of the utterance from Model A when the degree of the tongue was 90 degrees. This figure shows the third formant (F3) drop below 2 kHz, which is an acoustic cue of American English /r/ [9].

The column "Transcription" in Table 1 shows the results of the perceptual evaluation for Model A. The results show that Model A was able to produce an English /r/ sound when the degree of tongue rotation was 70 or more. This finding supports the results in [5]. The reason that the phonetician perceived the sound with the rotation degree of 100 as "r / l" might be due to the speed of tongue movement. Because we tried to minimize the difference of the duration of the consonant among utterances as much as possible, the speed of tongue rotation became faster as the rotation degree increased. At the maximum rotation of 100 degrees, the speed was highest, and the output sound yielded /l/-like acoustic cues, one of which is the fast movement of the first formant (F1) [9].

3.3.2. Model B

We manipulated Model B for making /aCa/ utterances where the target of C was bunched /r/. Blocks 0 and 2-6 were pushed up in different combinations as listed in Table 2. For block 0, when x_0 = 0 mm the model simulates lips rounded, and when x_0 = 20 mm the model simulates lips open. For Blocks 2-6, x_i (i = 2, ..., 6) the distance was either 3 or 20 mm. A narrow constriction is simulated when x_i = 3 mm, and a wide constriction is simulated when x_i = 20 mm. All combinations of movements were done manually by the author after a training session. Figure 6(b) shows the spectrogram of the utterance from Model B when Blocks 5 and 6 make a narrow constriction in the oral cavity with lip rounding. This figure clearly shows again the F3 drop below 2 kHz, which is an acoustic cue of American English /r/ [9].

The column "Transcription" in Table 2 shows the results of the perceptual evaluation of Model B. The results show that Model B was able to produce an English /r/ sound with a bunched /r/ configuration. This was especially true when Blocks 5 and 6 were raised. This finding supports the evidence of the previous study using magnetic-resonance imaging [7].

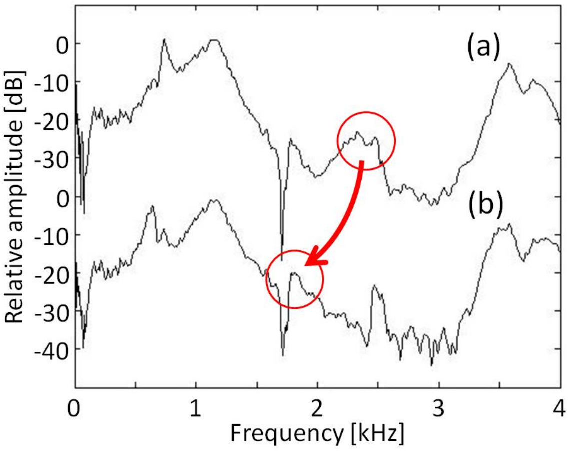

Figure 7: *Spectra of the impulse responses of Model A: (a) when the tongue is in resting position and (b) when the first half of the tongue is rotated to 90 degrees.*

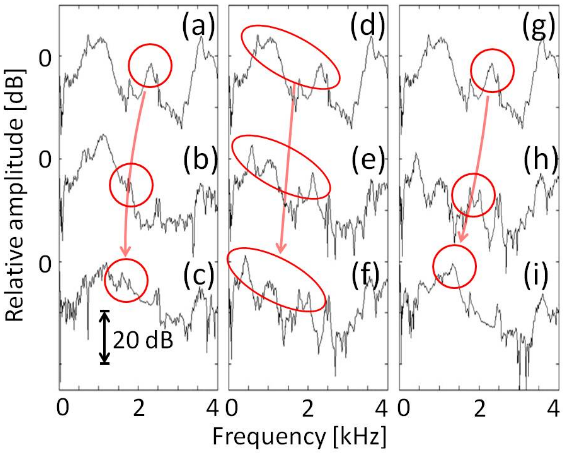

Figure 8: *Spectra of the impulse responses of Model B: (a)-(b)-(c) when the tongue dorsum is rising without lip rounding; (d)-(e)-(f) when the tongue is in resting position with lip rounding; and (g)-(h)-(i) when the tongue is forming bunched /r/ with lip rounding.*

4. Measuring impulse responses

4.1. Experimental setup

We measured impulse responses of Models A and B for retroflex and bunched /r/. The input signal for this measurement was a swept-sine signal with a sampling frequency of 48 kHz. The length of the swept-sine signal was 65536 samples. The input signal was fed into the driver unit (TOA, TU-750) via an audio interface (ECHO, AudioFire2) and a power amplifier (FOSTEX, AP1020). The close-fitting metal cylindrical filler was again inserted inside the neck of the driver unit. We placed the glottis end of each vocal-tract model on top of a thin metal plate. The output sounds were recorded using a microphone from the sound level meter (RION, NL-18) and an audio interface (ECHO, AudioFire2) 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 signals recorded were synchronously averaged multiple times to gain the signal-to-noise ratio.

4.2. Results

4.2.1. *Model A*

Figure 7 shows spectra obtained from the impulse responses of Model A. Figure 7(a) is when the tongue is in resting position, whereas Figure 7(b) is when the front half of the tongue is rotated to 90 degrees. As shown in these plots, the peak frequency corresponding to F3 decreases as the tongue rotates.

4.2.2. *Model B*

Figure 8 shows spectra obtained from the impulse responses of Model B. In Figure 8, spectra (a), (d), and (g) are all the same, when the tongue is in resting position. In Figure 8, spectra (b) and (c) are when $x_5 = x_6 = 3$ mm and 0 mm, respectively (the tongue dorsum is rising). As shown in these plots, the peak frequency corresponding to F3 decreases as the tongue dorsum rises. Also in Figure 8, spectra (e) and (f) are when $x_0 = 3$ mm and 0 mm, respectively (the lips are rounding). As shown in these plots, the formant frequencies decrease as the lips are rounding. In Figure 8, spectrum (h) is the same as (f), where $x_0 = 0$ mm (the lips are rounding). Figure 8(i) is when $x_0 = 0$ mm (the lips are rounding) and $x_5 = x_6 = 3$ mm (the tongue dorsum is rising). The peak frequency corresponding to F3 decreases the most in this latter case of Figure 8(i), when the lips are rounded and the tongue dorsum is rising.

5. Discussion and conclusions

In the current study, we successfully produced English bunched /r/ with a new model. The quality of the /r/ sound was even clearer when lip rounding was associated with the tongue movement. We produced English /r/ with Models A and B for a group of listeners who are phoneticians and/or engaged in phonetic education in Japan. They were asked to evaluate whether they wanted to use the models in phonetic education for English /r/ using a five-point scale from 1 (strongly negative) through 5 (strongly positive). The average score was 3.3, and we received both positive and negative comments from the participants.

We received the following positive comments: the models help one to understand that English /r/ can be produced with more than one configuration; that visualization helps one to understand the phenomenon; that the models help one imagine what is going on inside the oral cavity; and that instructors in phonetic education should be trained with the models.

We received the following negative feedback: that it is still not 100% clear how the models can be used in phonetic education because second language learners often struggle with mapping vocal-tract configurations with their own speech organs; also that it was difficult to manipulate the lip and tongue blocks simultaneously by hand.

In this study, Block B was left in the pharynx during the whole utterance /ara/. This was partly because of the narrow pharyngeal characteristics for the vowel /a/. However, the pharyngeal narrowing is also important for the production of these /r/ sounds [10]. In the future, we will change the size of Block B to investigate the effect of the pharyngeal narrowing.

6. Acknowledgements

I acknowledge the anonymous reviewers for their helpful comments. This work was partially supported by JSPS KAKENHI Grant Number 24501063. I would also like to thank Keiichi Yasu and Terri Lander for their support.

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

[4] Arai, T., “Gel-type tongue for a physical model of the human vocal tract as an educational tool in acoustics of speech production,” *Acoust. Sci. Tech.*, 29(2):188-190, 2008.

[5] Arai, T., “Physical models of the vocal tract with a flapping tongue for flap and liquid sounds,” *Proc. of Interspeech*, 2019-2023, 2013.

[6] Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S. and Alwan, A., “Acoustic modeling of American English /r/,” *J.Acoust. Soc. Am.*, 108(1), 343-356, 2000.

[7] Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C. and Choe, A., “A magnetic resonance imaging-based articulatory and acoustic study of 'retroflex' and 'bucnehd' American English /r/,” *J.Acoust. Soc. Am.*, 123(6), 4466-4481, 2008.

[8] Umeda, N. and Teranishi, R., “Phonemic feature and vocal feature: Synthesis of speech sounds, using an acoustic model of vocal tract,” *J. Acoust. Soc. Jpn.*, 22(4):195-203, 1966.

[9] Kent, R. D. and Read, C., *Acoustic Analysis of Speech*, Singular Publishing, San Diego, CA, 2001.

[10] Delattre, P. and Freeman, D. C., “A dialect study of American English r's by x-ray motion picture,” *Linguistics*, 44, 28-69, 1968.