PHYSICAL MODELS OF THE VOCAL TRACT SOUND DIFFERENT WITH THE SAME SHAPE BUT DIFFERENT TEMPORAL CHARACTERISTICS AND VICE VERSA

Takayuki Arai

Sophia University (Tokyo, Japan)
arai@sophia.ac.jp

ABSTRACT

Two physical models of the human vocal tract have been successfully developed for producing English /l/ and /r/ and Japanese /r/. The first model was originally designed for the alveolar lateral approximant and the retroflex approximant, while the second model was designed to produce English "bunched /r/." With these models, we observed that different configurations of the vocal tract can produce similar sounds in addition to retroflex and bunched /r/. We also observed that the models produce the target sounds when the articulators move with certain patterns of temporal change. However, moving the same articulators with the same movement but different temporal patterns produced less intelligible and/or different sounds. Thus, using the two physical models, we tested and confirmed that different configurations of the vocal tract using similar temporal changes yielded similar sounds, while the same configuration of the vocal tract with different temporal changes yielded different sounds.

Keywords: physical model, vocal tact, articulation of speech, approximant, flap

1. INTRODUCTION

Although computer models of the human vocal tract are now widely available to simulate the articulation of human speech, it is reported that physical models have advantages for both education and research purposes [1-3]. For example, when we teach the acoustic theory of speech production, simple acoustic tubes with different configurations effectively demonstrate source-filter theory when sound sources are fed as inputs. The output sounds are affected by both the source and the filter (vocal-tract configuration), so learners can intuitively understand what is going on without having to understand the mathematical equations backing the theory. Because of this simplicity, physical models of the human vocal tract are being used at workshops and exhibitions in museums to effectively demonstrate to students and even to children how speech is produced.

The same set of physical models are being used for phonetic education, as well [4-6]. Certain sounds, such as English /r/ and Japanese /r/ are difficult even for native children to acquire, and they are also troublesome in non-native acquisition. In such cases, computer models are useful because learners can see a virtual image of the mouth against which they can compare their own articulations. However, physical models have advantages over computer models. For example, when instructors or learners manually change the configurations of the vocal tract to produce different sounds, learners can see, hear, and even touch the difference, in real time. Using computer and physical models together, plus articulatory measurements with ultrasound [9], for example, creates powerful, synergistic effects.

For research and application, we can explore speech science and phonetics/phonology using the physical models. Making a speaking robot is a possible application using the physical models. Even a research question, such as the interaction between the source and filter in vowel production, can be addressed with the physical models. These phenomena, while easily demonstrated with the models, are not so simple to simulate computationally, and there are also the limitations of computer modeling. Of course, we can measure real human articulation; however, we are not able to easily see and/or modify some experimental conditions inside the actual human vocal tract, nor are we able to reproduce the same exact experimental procedure multiple times.

In this paper, looking from both a phonetic education and science perspective, we focused on two types of physical models of the human vocal tract for the sounds of English /l/ and /r/ and Japanese /r/. The first model we used was originally designed for the alveolar lateral approximant and the retroflex approximant [4]. This model can change the length of the tongue, and in addition, we can rotate the first half of the tongue manually to produce the lateral and retroflex approximants. The second model was originally designed to produce the bunched /r/ in English by pushing up the blocks lined up in the oral cavity [6]. In both types of models, we observed that changing the speed of the articulators yielded a different perception of sounds. Therefore, in this study, we ask the following

questions, using the two models mentioned above: Q1) Are we able to produce the same sounds with different configurations of the vocal tract but a similar temporal change? Q2) Are we able to produce different sounds with the same configurations of the vocal tract but different temporal change?

2. TWO PHYSICAL MODELS OF THE HUMAN VOCAL TRACT

2.1. Model 1

Model 1 was originally developed for lateral and retroflex approximants [4]. The flapping tongue enables the front half of the tongue to rotate towards the palate. Shortening the length of the tongue produces alveolar/retroflex approximants, and lengthening it produces lateral approximants.

Figure 1(a) shows Model 1 with the long tongue (Model 1a), and Figure 1(b) shows the same model with the short tongue (Model 1b). Figure 1 displays the lever used to rotate the front half of the tongue. When the tongue is in resting position, a 45-mm-long narrowing in the pharyngeal cavity enables Model 1 to produce the vowel /a/.

The extended tongue of the model in Figure 1(a) enables the tongue blade to touch the palate. Lateral pathways for airflow on both sides of the tongue allow for the production of lateral sounds [13]. The tongue is short with Model 1b so the tongue blade is not touching the palate. This simulates a retroflexed tongue and allows retroflex approximants to be produced [13].

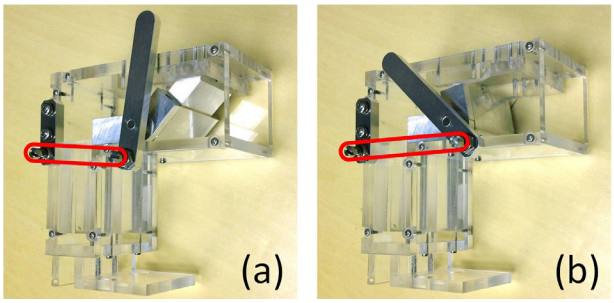

Figure 1: Model 1: (a) The tongue is long and its blade is touching the palate (Model 1a). (b) The tongue is short and its blade is not touching the palate, but the tongue is retroflexed (Model 1b).

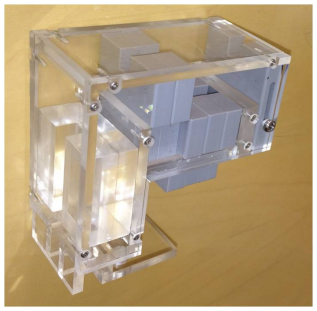

Figure 2: Model 2: In the velar position, the last two blocks are raised for "bunched /r/" in English.

To help the model return to resting position, one can use a spring or a rubber band between the bottom of the lever and the body of the vocal tract, indicated as a red line. Made of transparent acrylic material, the outer frame of the model makes tongue movement visible from the outside.

2.2. Model 2

This model was originally developed for English bunched /r/ (Model 2) [6]. As with Model 1, the vocal tract is bent at a right angle in the middle of its length. Model 2 consists of several blocks in the oral cavity which can be moved up and down. The top of each block has a 9 x 9 mm notch in the center along the length of the vocal tract. The notches of the blocks simulate the groove of the tongue with its concave shape. Placing the blocks in the highest position creates (a) narrow constriction(s) in the oral cavity. Movable blocks simulate dynamic tongue movements. For example, the narrow constriction displayed in Figure 2 simulates bunching of the tongue for "bunched /r/" in English. Additionally, the model is able to simulate vowels, such as the front vowels /i/ and /e/.

Figure 2 shows Model 2 with bunching of the tongue. As you can see in Figure 2, the blocks stand perpendicular to each other. Lined up next to each other as they are, the blocks' own weight enables them to move downward. When the tongue is in the resting position, Model 2 produces the vowel /a/. Again, this was achieved by the 45-mm-long narrowing in the pharyngeal cavity, just as in Section 2.1. In Figure 2, the last two blocks in the velar position are raised (typically, the upward movement of the blocks are done by hands). The first block, or "the lip block" is also raised in this figure, so that the output sound from this vocal tract configuration produces bunched /r/ in English. In this study, the lip block was not raised for the recordings of Section 3. The outer frame of this model is also made of a transparent acrylic material so that the tongue movement is visible from the outside. We distinguish between "Model 2a," when the "alveolar" block is raised and "Model 2b," when the last two blocks in the velar position are raised.

3. ANALYSIS AND EVALUATION

To answer the two questions from Section 1, we focused on the following three sounds: English /l/, English /r/, and Japanese /r/. These three sounds were produced by using the two physical models described in Section 2. We have reported informally that Model 1 can produce a sound perceptually similar to Japanese /r/, despite the fact that Model 1a and Model 1b were originally designed for English

/l/ and /r/, respectively [5]. It turns out that temporal characteristics are key for making one sound over another, as we discovered when we moved the lever of Model 1 quickly, and produced a sound perceived as Japanese /r/.

It is well-known that American English /r/ is not always articulated as a retroflex, with the tongue tip raised [6], as shown in Model 1b. 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/." While both bunched and retroflex /r/ have similar spectral characteristics in the frequency range of the 1st to 3rd formants [7], Zhou et al. [15] have shown that these two versions of /r/ differ considerably in the spacing between the 4th and the 5th formants.

Interestingly, it was also reported that Model 2 can produce English /l/, despite the fact that this model was originally designed for bunched /r/ [6], and the model does not produce lateral pathways for airflow. However, the resultant output is perceived as English /l/, given certain temporal movements.

In this section, we produced sounds with Models 1 and 2 with different temporal characteristics. Then, we analyzed them acoustically, and performed a perceptual experiment involving two experienced phoneticians, a native speaker of American English and a native speaker of Japanese.

Table 1: The temporal movements of the tongue for Models 1 and 2.

Model	Condition	Speed of the tongue	
		when raising	when returning
1a/1b	MM	medium	medium
	MF	medium	fast
	SF	slow	fast
	FF	fast	fast
2a/2b	MM	medium	medium
	SM	slow	medium
	FM	fast	medium

*Please note that the fast return movement for Model 1 and the medium return movement for Model 2 were accomplished with a rubber band (Model 1) and by the blocks' own weight (Model 2).

3.1. Recordings

A reed-type sound source [2] was attached to the glottis end of each model. By blowing an air stream into the sound source, the reed vibrated at approximately 100 Hz, and a glottal sound was produced. The output sounds from the models were recorded by a digital recorder (Marantz, PMD660) with a microphone (Sony, EM-23F5). The original sampling frequency of 48 kHz was retained for the acoustic analysis and evaluation in Section 3.3.

We recorded /aCa/ utterances with Models 1 and 2. In both cases of the models, the temporal movements described in Table 1 were used for the recordings.

3.2. Acoustic analysis

Figures 3 and 4 show the spectrograms of the three or four utterances produced by the models with the conditions described in Table 1. Because each of Models 1a, 1b, 2a, and 2b has the same vocal tract configuration, the only difference being temporal characteristics, each of the panels in Figures 3 and 4 shows similar spectral characteristics. The vertical spikes in these spectrograms were impulsive sounds occurred when the tongue returned in resting position with a certain speed (such impulsive sounds were ignored when evaluating the consonants in Section 3.3).

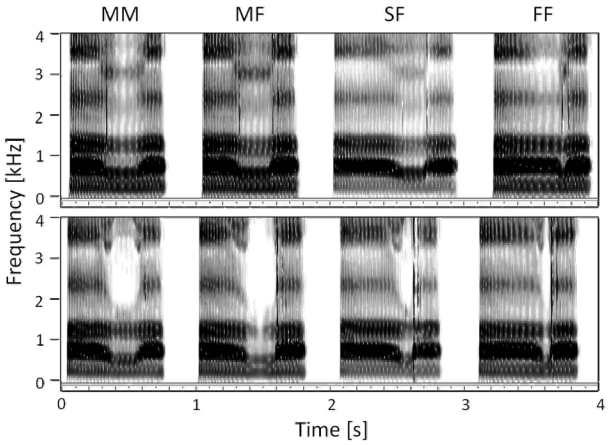

Figure 3: Spectrograms of the four utterances produced by Model 1 with the conditions described in Table 1. Top panel: Model 1a; and bottom panel: Model 1b.

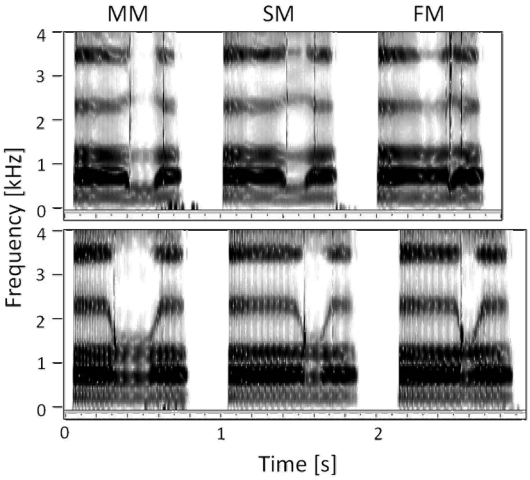

Figure 4: Spectrograms of the three utterances produced by Model 2 with the conditions described in Table 1. Top panel: Model 2a; and bottom panel: Model 2b.

Table 2: The results of the evaluation for 14 utterances by the two phoneticians.

Model No.	Temp. Pattern	IPA		Likeliness score					
				EN				JA	
				/l/		/r/		/r/	
		Pe	Pj	Pe	Pj	Pe	Pj	Pe	Pj
1a	MM	ɺ	l̠	4	4	1	1	1	3
	MF	l	l̠	5	4	1	1	1	3
	SF	l	l̠	5	4	1	1	1	3
	FF	ɺ	ɾ	4	1	1	2	2	5
1b	MM	ɻ	ɻ	1	1	4	4	1	3
	MF	*1	*2	5	4	4	1	1	3
	SF	*3	ɻ	2	1	2	5	1	4
	FF	ɽ	ɾ	1	1	1	2	4	5
2a	MM	l	l	5	5	1	1	1	4
	SM	l	l	5	5	1	1	1	4
	FM	ɾ	l̪	1	5	1	1	5	4
2b	MM	ɻ	ɻ	1	1	5	5	1	2
	SM	ɻ	ɻ	1	1	4	4	1	4
	FM	ɽ	ɻ̞	1	1	2	4	2	4

Notes:
*1) The response was "sequence of /r/ followed by /l/."
*2) The response was "ɺ" with lateral release.
*3) The response was "unclear."

3.3. Evaluation

We asked two phoneticians to evaluate all 14 utterances. One phonetician was a native speaker of American English (Pe), and the other was a native speaker of Japanese (Pj). Each utterance was played as many times as the phoneticians wished. They were asked to transcribe each stimulus phonetically. The likelinesses of English /l/, English /r/, and Japanese /r/ were also evaluated. For the likeliness scores, a 5-point scale was used: 5: very appropriate, 4: appropriate, 3: moderate, 2: not so appropriate, and 1: not accepted at all. Table 2 shows the results of the evaluation.

4. DISCUSSION AND CONCLUSION

For results in Section 3 we can conclude that the same sounds were produced with different configurations of the vocal tract but similar temporal characteristics. That was true not only for the retroflex (Model 1b) and bunched /r/ (Model 2b) as in [6], but also for English /l/ (Models 1a/2a) and Japanese /r/ (Models 1a/1b with the FF condition). When comparing Models 1b and 2b with the MM condition, the spectrograms show the F3 drop below 2 kHz, which is a sign of English /r/ [4,6].

When comparing Models 1a and 2a with the non-FF conditions, the spectrograms show rapid F1 movement just before the onset of the second vowel, which is an acoustic cue for English /l/, while the F2 frequency is more or less steady throughout the utterance [4]. It is interesting to notice that Model 2a sounded lateral, although the blocks only have a center notch. The notches were originally designed to simulate the tongue bracing against the teeth or palate [10,14]. Study [6] also pointed out that the stability for bunched /r/ results from the "saturation effect" [8], and a steady acoustic output is achieved ("quantal theory" by Stevens [11,12]). Since the tongue blade cannot be braced, this sound is a little difficult to achieve in natural speech.

The pairs of sounds above have common temporal and spectral characteristics, even if the articulation patterns are different. Therefore, each pair is considered as the same phone.

When comparing Models 1a and 1b with the FF condition, the spectrograms show both short consonants; this pair has similar temporal characteristics. However, their spectral characteristics are different. In other words, the F3 is steady in Model 1a and the F3 drops below 2 kHz in Model 1b. Thus, this pair is considered as different phones (Model 1a: alveolar lateral flap, Model 1b: short retroflex approximant). Both phones are still different from the typical Japanese /r/ acoustically, which is alveolar flap and has a short but complete gap during the target consonant. All of the phones are considered as the same phoneme in Japanese.

As described in [5], Japanese /r/ has even more allophones. In other words, alveolar lateral and retroflex approximants are also allophones of Japanese /r/. In fact, the likeliness scores by Pj for such sounds are not low. These sounds are phonetically distinct; however they are categorized as /r/ in Japanese because they are acceptable allophones of the Japanese alveolar flap [5].

The results of this study can be applied in phonetic education and speech pathology. Second language learners or patients can use the knowledge to acquire the pronunciation of a certain sound by testing spectral and temporal characteristics of their articulators. The physical models of the human vocal tract should be useful in a classroom as well as in a clinical situation.

5. ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI Grant Numbers 24501063 and 15K00930.

6. REFERENCES

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

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

[3] Arai, T. 2013. Mechanical models of the human vocal tract. *Acoustics Today* 9(4), 25-30.

[4] Arai, T. 2013. Physical models of the vocal tract with a flapping tongue for flap and liquid sounds. *Proc. INTERSPEECH*, Lyon, 2019-2023.

[5] Arai, T. 2013. On Why Japanese /r/ sounds are difficult for children to acquire. *Proc. INTERSPEECH*, Lyon, 2445-2449.

[6] Arai, T. 2014. Retroflex and bunched English /r/ with physical models of the human vocal tract. *Proc. INTERSPEECH*, Singapore, 706-710.

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

[8] Fujimura, O., Kakita, Y. 1979. Remarks on the quantitative description of the lingual articulation. In: Lindblom, B., Öhman, S. (eds), *Frontier in Speech Communication Research*. London: Academic Press, 17-24.

[9] Gick, B., Wilson, I., Derrick, D. 2013. *Articulatory Phonetics*. Wiley-Blackwell.

[10] Gick, B., Allen, B., Stavness, I., Wilson, I. 2013. Speaking tongues are always braced. *J. Acoust. Soc. Am.* 134, 4204.

[11] Stevens, K. N. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In: Denes, P. B., David Jr., E. E. (eds), *Human Communication: A Unified View*. New York: McGraw Hill, 51-66.

[12] Stevens, K. N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17, 3-46.

[13] Stevens, K. N. 1998. *Acoustic Phonetics*. Cambridge: MIT Press.

[14] Stone, M. 1990. A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *J. Acoust. Soc. Am.* 87(5), 2207-2217.

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

Click here for sound demo Model 1a (MM, MF, SF, FF)

Click here for sound demo Model 1b (MM, MF, SF, FF)

Click here for sound demo Model 2a (MM, SM, FM)

Click here for sound demo Model 2b (MM, SM, FM)