Physical Models of the Vocal Tract with a Flapping Tongue for Flap and Liquid Sounds

Takayuki Arai

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

Abstract

Certain sounds are difficult for children to produce, even if the sounds are in their native language. For example, Japanese /r/ can be difficult for Japanese children to learn. Second language learners can also have difficulty acquiring certain sounds. For example, Japanese speakers learning English often have difficulty with English /r/ and /l/. To address this problem, we have developed two new physical models of the vocal tract: one for flap sounds (Model A) and another for liquid sounds (Model B). Each of them has a flapping tongue, and for Model B, the length of the tongue is variable. When the tongue is short, we can produce alveolar/retroflex approximants, and when the tongue is long we can produce lateral approximants. We recorded several sets of sounds produced by these models, analyzed the speech data, and used them for perceptual experiments. From the acoustic analysis and the perceptual experiments, we confirmed that the sounds produced by Model A were heard as Japanese /r/, and the sounds produced by Model B were heard as English /r/ and /l/. Furthermore, the models are helpful for practicing pronunciation because learners can see the tongue, alter tongue position manually, and hear the output sounds.

Index Terms: speech production, physical models of the human vocal tract, tongue, flap/liquid sounds

1. Introduction

The vocal-tract models we have developed so far [e.g., 1-5] may be used for a wide range of purposes, such as research (including physical measurements), education, and self-learning by users. Our early models [1] were based on Chiba and Kajimaya's measurements and simplifications [6]. These initial models consisted of five straight cylinders with step-wise (0th-order) and polyline (1st-order) approximations. Today, our original models are still used in classrooms at several institutions, as well as at a number of exhibitions in science museums.

We later modified the models, taking their form in two directions. We created both more simplified models, and more complex models, broadening their scope to more closely approximate the human vocal tract. Some models have a 90-degree bend in the middle, and some do not. An example of a simplified, straight model is the sliding three-tube (S3T) model [3]. With the S3T model, one can check the relationship between constriction location and vowel quality. The S3T model can also be used to demonstrate the source-filter theory, as can the other vocal-tract models.

Other models of the more complicated form include the MRI model and the computer controlled Umeda and Teranishi model. The MRI model is based on magnetic resonance imaging, where vocal-tract shape is measured by MRI [7]. With this model, one can compare phonemically same vowel sounds which are produced from slightly different vocal-tract configurations, to highlight inter-speaker differences.

The other complicated model we would like to cover here, is our extension of Umeda and Teranishi's vocal-tract model [5, 8]. Our extension adds computer control to the model. The design is as follows: the vocal tract is straight, and a set of movable plastic bars are inserted from the side. One can manipulate the position of the bars to simulate an arbitrary shape of the vocal tract. Furthermore, each of the plastic bars is connected to an actuator that controls the position of the bar by computer command. With this model, one can simulate/demonstrate complicated movements of the human vocal tract in time.

Thus, all of our models can be divided into two categories: static and dynamic. Static models produce steady-state vowels, while dynamic models approximate more realistic tongue movement, with sounds that change over time. Dynamic models produce vowels, diphthongs, consecutive vowels, and they can even produce sonorant consonants, such as flaps and liquids.

In the present study, we focus our attention on the sonorants, and to that end, we have designed two new dynamic models that employ a flapping tongue to produce flap and liquid sounds.

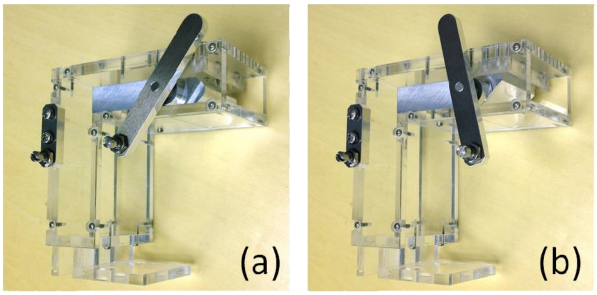

Figure 1: *Model A for flap sounds: (a) the tongue is in resting position; and (b) the tongue blade is touching the palate.*

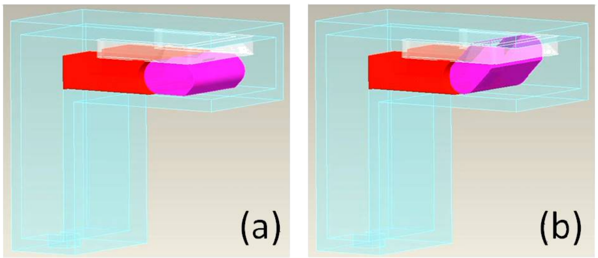

Figure 2: *Schematic illustrations of Model A for flap sounds: (a) the tongue is in resting position; and (b) the tongue blade is touching the palate.*

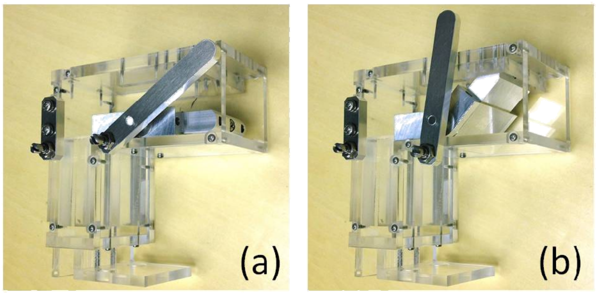

Figure 3: *Model B for lateral approximants: (a) the tongue is in resting position; (b) the tongue blade is touching the palate.*

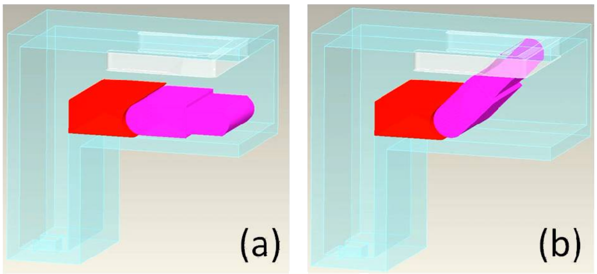

Figure 4: *Schematic illustrations of Model B for lateral approximants: (a) the tongue is in resting position; (b) the tongue blade is touching the palate.*

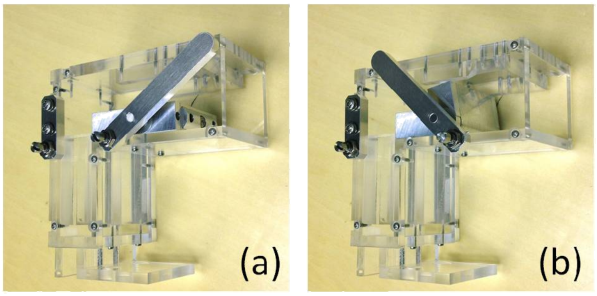

Figure 5: *Model B for retroflex approximants: (a) the tongue is in resting position; (b) the tongue blade is not touching the palate, but the tongue is retroflexed.*

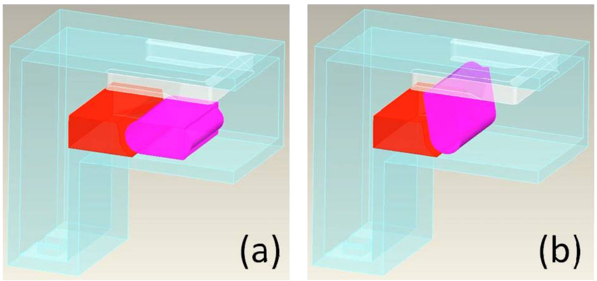

Figure 6: *Schematic illustrations of Model B for retroflex approximants: (a) the tongue is in resting position; (b) the tongue blade is not touching the palate, but the tongue is retroflexed.*

2. Design

2.1. Model A produces the flap

Both English and Japanese languages have a flap as a part of their phonetic inventory. The flap is produced at the alveolar ridge with a short flip of the tongue blade. Complete closure is made between the tongue and palate, but the closure is very short, around 20-50 ms. In American English, the flap is an allophone of the phonemes /t/ and /d/, as in the words “writer” and “rider”. The Japanese flap is the normal form of the phoneme /r/.

Figure 1 shows two photographs of Model A, which produces the flap sound. 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), its vocal tract is configured for the vowel /e/. The reason /e/ was chosen for resting position has to do with tongue height. Since the tongue is already relatively high when producing /e/, the tongue has a shorter distance to travel in order to produce the flap.

Figures 1b and 2b show the tongue blade touching the palate. Pushing the lever towards the back causes the tongue to rise. To produce a flap, one moves the lever back and forth very quickly. To assist the return movement of the tongue, a spring or a rubber band may be used between the bottom of the lever (i.e., the short head at the bottom of the lever in Fig. 1) and the body of the vocal tract (the small aluminum plate with a short head on the left of Fig. 1).

2.2. Model B produces approximants/liquids

In discussing Figures 3 and 4, a brief linguistic review of /l/ and /r/ may be helpful. English /l/ is called a lateral alveolar approximant. English /r/ is called a retroflex alveolar approximant. Both of these sounds are liquids. During production of /l/, the tongue blade retains contact with the alveolar ridge but the air stream is not fully blocked because there are lateral pathways through which air escapes on each side of the tongue. During production of English /r/, the tongue blade is curved upwards, but does not make any contact with the palate, and, as with /l/, the air steam is not fully blocked at any point along the vocal tract.

Figure 3 shows two photographs of Model B, with the tongue extended. This extended tongue position is used to produce the lateral alveolar approximant. Figure 4 shows schematic illustrations corresponding to the same model. In this model, the vocal tract is configured for the vowel /a/ when the tongue is in resting position (Figs. 3a and 4a). Figures 3b and 4b show the tongue blade touching the palate. Movement from Figures 3a/4a to 3b/4b is done by moving the lever back. To move the tongue rapidly, one moves the lever back and forth quickly. To assist the return movement of the tongue, one can again use a spring or a rubber band between the bottom of the lever and the body of the vocal tract.

Figure 5 shows two photographs of Model B with the tongue shortened. The shortened tongue position is used to produce the alveolar retroflex approximant (English /r/). Figures 5a and 6a show when the tongue is in resting position. Figures 5b and 6b show the tongue bent towards the middle portion of the palate to produce a retroflex approximant. By moving the lever back and forth one can cause the movement from Figures 5a/6a to 5b/6b and back to 5a/6a.

3. Experiments

3.1. Recordings

A reed-type sound source [3] was attached to the glottis end of each model. By blowing an air stream into the sound source, the reed was set into vibration at approximately 100 Hz, and a glottal sound was produced. The output sounds from the models were recorded by a digital recorder (SONY, PCM-D1) with internal microphones. The original sampling frequency of 48 kHz was retained for the perceptual experiments, but converted into 8 kHz for the acoustic analysis.

3.1.1. Flaps (Model A)

We manipulated model A for flap sounds with and without a rubber band. Also, we manipulated the lever so the tongue did or did not make contact with the alveolar ridge. Initially, we considered using just four vocal tract configurations to produce the output sounds. However, the design has a small bump simulating the upper teeth, enabling a complete closure when the tongue makes contact with the alveolar ridge. Since this bump is removable, eventually we added four more configurations. The added conditions account for the presence or absence of the upper teeth. Thus, there were eight conditions in all. In each condition, the recordings were done three times.

Figure 7(a) shows the spectrogram of the third repetition when there was the upper teeth, the rubber band was used, and there was tongue contact with the alveolar ridge. This figure clearly shows a short gap of less than 50 ms, which is typical for the flap sound.

3.1.2. Approximants/Liquids (Model B)

We manipulated model B, with both an extended and shortened tongue. A rubber band was always used for the recordings. In each case, the recordings were done three times.

Figures 7(b) and 7(c) show the spectrogram of the second repetition for each tongue length. Figure 7(b) shows rapid F1 movement just before the onset of the second vowel, which is an acoustic cue of English /l/, while the F2 and F3 frequencies are more or less steady throughout the utterance. Figure 7(c) shows the F3 drop below 2 kHz, which is a sign of English /r/.

3.2. Perceptual experiment 1 (Exp. 1)

We conducted the first perceptual experiment based on the recordings in Section 3.1.1. The stimulus set was all /eCe/ utterances, and there were 24 stimuli (2 teeth conditions x 2 rubber band conditions x 2 tongue contact conditions x 3 repetitions). The experiment was conducted in a sound-treated room. Stimuli were presented monaurally through a loudspeaker (NAE NESmini) connected to an audio interface (RME Babyface) via an amplifier (FOSTEX AP1020). The five participants were seated 3-4 m from the loudspeaker. The sound level was approximately 70 dBA on average. There was a training session with eight stimuli prior to the main session. Twenty young listeners with normal-hearing (10 males and 10 females, ages 20 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 session, the stimuli were presented in random order. There were 24 trials in total. A stimulus was presented in each trial, and the listeners were instructed to select one answer for the question displayed on a computer screen by means of a graphical user interface (GUI). The question was as follows:

How well suited is the sound to Japanese /r/:
100%, 75%, 50%, 25%, or 0%?

The column of "Likeliness" in Table 1 shows the average scores as a percentage for the stimulus set. Each percentage was averaged over 20 participants. In this table, we observed that some of the conditions produce what is perceived to be an acceptable Japanese /r/ sound.

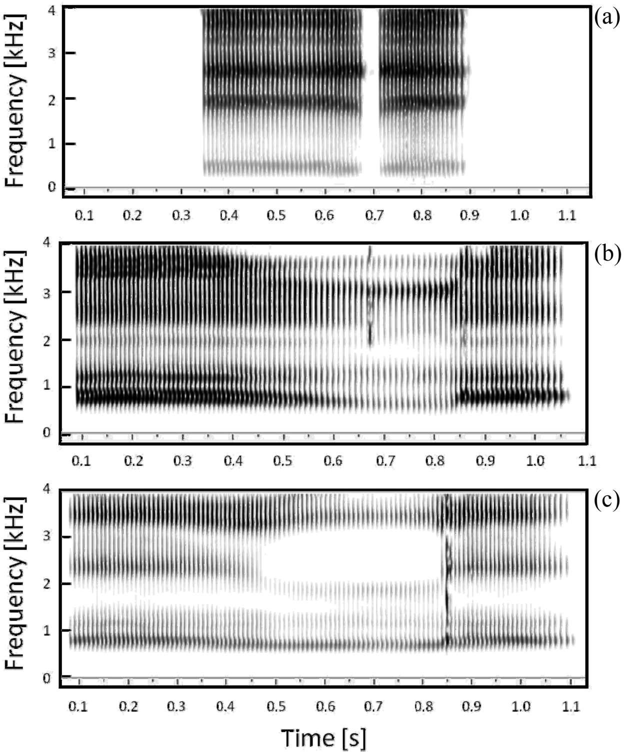

Figure 7: *Spectrograms of the output sounds from Models A and B: (a) Model A; (b) Model B with tongue extension; and (c) Model B with shortened tongue.*

3.3. Perceptual experiment 2 (Exp. 2)

We conducted the second perceptual experiment based on the recordings in Sections 3.1.1 and 3.1.2. The first stimulus set was the same as the one recorded in Section 3.1.1 and used in Exp. 1. The second stimulus set was all /aCa/ utterances recorded in Section 3.1.2, and there were 6 stimuli (2 tongue length conditions x 3 repetitions). The experiment was conducted in a quiet room. Stimuli were presented diotically through headphones (Sennheiser, HD595). One experienced phonetician, a native speaker of American English, participated in the experiment. During the main session, the icons of a loudspeaker were displayed on a PC screen. The icons were distributed randomly. Each icon corresponded to one stimulus. The sound was played as many times as the phonetician double clicked on the icon from the PC. The phonetician was asked to transcribe each stimulus phonetically.

The column of "Phonetic transcription" in Table 1 shows the results for the first stimulus set. The results show that Model A was able to produce a flap sound. Table 2 shows the results for the second stimulus set. The results show that Model B was able to produce lateral and retroflex approximants comparable to English /l/ and /r/.

4. Discussion and conclusions

When children learn to pronounce speech sounds, they often mimic someone's speech production. However, certain sounds are difficult for children to acquire, even if the sounds are in their native language. Japanese /r/ is one such example. Trouble with acquisition is true as well for second language learners. English /r/ and /l/, for example, are difficult for native Japanese speakers to acquire.

Table 1: *The average scores of the likeliness of Japanese /r/ sounds in Exp. 1 (in %).*

									Likeliness of /r/	Phonetic transcription
Teeth	without	Rubber band	without	Contact	without	Repetition	1	25.0	ɾ (alveolar flap)
							2	35.0	ɾ (alveolar flap)
							3	18.8	ɾ (alveolar flap)
					with	Repetition	1	37.5	l (alveolar lateral approximant)
							2	45.0	l (alveolar lateral approximant)
							3	41.3	l (alveolar lateral approximant)
			with	Contact	without	Repetition	1	62.5	- (no consonants)
							2	58.8	- (no consonants)
							3	43.8	- (no consonants)
					with	Repetition	1	66.3	l (alveolar lateral approximant)
							2	72.5	ɾ (alveolar flap)
							3	62.5	l (alveolar lateral approximant)
	With	Rubber band	without	Contact	without	Repetition	1	62.5	- (no consonants)
							2	70.0	l (alveolar lateral approximant)
							3	71.3	- (no consonants)
					with	Repetition	1	50.0	d (alveolar plosive)
							2	56.3	θ (dental fricative)
							3	56.3	ɾ (alveolar flap)
			with	Contact	without	Repetition	1	33.8	l (alveolar lateral approximant)
							2	33.8	l (alveolar lateral approximant)
							3	28.8	l (alveolar lateral approximant)
					with	Repetition	1	35.0	ɾ (alveolar flap)
							2	28.8	ɾ (alveolar flap)
							3	31.3	ɾ (alveolar flap)

Table 2: *The phonetic symbols that the phonetician transcribed in Exp. 2 (Model B).*

				Phonetic transcription
Tongue length	Short	Repetition	1	ɻ (retroflex
			2	ɻ approximant)
			3	ɻ
	Long	Repetition	1	ɭ (retroflex
			2	ɭ lateral
			3	ɭ approximant)

In this study, we have developed two new physical models of the vocal tract: one for flap sounds (Model A) and the other for liquid sounds (Model B). From perceptual experiments, we confirmed that Model A, in certain conditions, produces sounds recognized as Japanese /r/. This is true, even if the model is used without upper teeth. Without teeth, complete closure, and hence the flap, is not possible with Model A, in a strict sense. However, the Japanese /r/ has allophonic variations, including plosives (such as, alveolar plosive [ɖ,]), flaps (such as, alveolar flap [ɾ], retroflex flap [ɽ], and alveolar lateral flap [ɺ]), and approximants (alveolar approximant [ɹ], alveolar lateral approximant [l], and retroflex lateral approximant [ɭ,]). The presence in Japanese of this broad range of acceptable allophonic variations may account for why, even under multiple conditions, native Japanese listeners judged the stimuli to be acceptable Japanese /r/ sounds. Please note that the likeliness of the sounds labeled as flaps by the phonetician was not necessarily high. This might be because Japanese /r/ and the flapped variant of American English /t/ are not the same sound.

We also confirmed that Model B produced sounds that were heard as retroflex and lateral approximants. In the case of the long tongue condition in Table 2, the IPA symbol [ɭ] is used. However, the phonetician actually reported that this sound "started out like /r/ and released like /l/." This kind of observation might be related to how the lever was manipulated. For English /l/, quick tongue movement is needed to produce the signature rapid F1 transition noted at the release of the tongue from the alveolar ridge. The rubber band helped with the quick release. However, during the first half of the consonant, there might have been a chance that the movement of the lever was not appropriate, which caused the output to sound more like /r/ than /l/.

Model B was able to produce an alveolar retroflex approximant perceived to be equivalent to the English /r/. English /r/ has a wide variation, and allophones include the so-called bunched /r/. In the future, we would like to design another model for the bunched /r/ (e.g., [9]) as well.

Finally, the models may be helpful for practicing pronunciation because learners can see the tongue, alter tongue position manually, and hear the output sounds. We plan to test the effectiveness of the models in an actual pedagogical setting in the future.

5. 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 supports.

6. 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., “Mechanical vocal-tract models for speech dynamics,” *Proc. of Interspeech*, 1025-1028, 2010.

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

[7] Honda, K., Takemoto, H., Kitamura, T., Fujita, S. and Takano, S., “Exploring human speech production mechanisms by MRI,” IEICE Trans. on Information and Systems, E87-D(5), 1050-1058, 2004.

[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] 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.