INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Integrated mechanical model of [r]-[l] and [b]-[m]-[w] producing consonant cluster [br]

Takayuki Arai

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

Abstract

We have developed two types of mechanical models of the human vocal tract. The first model was designed for the retroflex approximant [r] and the alveolar lateral approximant [l]. It consisted of the main vocal tract and a flapping tongue, where the front half of the tongue can be rotated against the palate. When the tongue is short and rotated approximately 90 degrees, the retroflex approximant [r] is produced. The second model was designed for [b], [m], and [w]. Besides the main vocal tract, this model contains a movable lower lip for lip closure and a nasal cavity with a controllable velopharyngeal port. In the present study, we joined these two mechanical models to form a new model containing the main vocal tract, the flapping tongue, the movable lower lip, and the nasal cavity with the controllable velopharyngeal port. This integrated model now makes it possible to produce consonant sequences. Therefore, we examined the sequence [br], in particular, adjusting the timing of the lip and lingual gestures to produce the best sound. Because the gestures are visually observable from the outside of this model, the timing of the gestures were examined with the use of a high-speed video camera.

Index Terms: speech production, mechanical models of the human vocal tract, flapping tongue, lips, consonant cluster

1. Introduction

Our earlier physical models of the human vocal tract were mainly designed for vowels [1-4]. More recently, we have developed additional mechanical models which produce not only vowels but consonants, as well [5-7]. In 2013 we designed a model [5] for the retroflex approximant [r] and the alveolar lateral approximant [l]. This model consisted of a main vocal tract and a flapping tongue. The front half of the tongue can be rotated against the palate with a lever, and the tongue can vary in length from short (normal) to long. When the tongue is short and the rotation is approximately 90 degrees, the retroflex approximant [r] is produced. When the tongue is long, the tongue tip is able to touch the alveolar ridge if the front part of the tongue is rotated approximately 45 degrees. In this position, there are lateral pathways for the airstream, and the lateral approximant [l] is produced.

* Please note that the correct IPA symbol for the retroflex approximant is [ɻ]. However, the symbol [r] is used for the retroflex approximant through this paper.

Another model designed in 2014 [6] was for bunched [r]. There are several 10-mm thick plates lined up next to each other in the oral cavity, which can be moved up and down by pushing up and releasing each plate from the bottom. By pushing the plates up around 50-60 mm from the lips, we can clearly hear the bunched [r] sound.

Our recent model, designed in 2016 [7] was for [b], [m], and [w]. Besides the main vocal tract, there is a movable lower lip for lip closure and a nasal cavity with a controllable velopharyngeal port. The area of the lip opening can be controlled by manually pushing up the lower lip block. Velopharyngeal coupling is achieved by rotating the knob. When the lips are open and the velopharyngeal port is closed, with no oral or pharyngeal block, the output sound is more or less similar to schwa. When there is a constriction in the oral or pharyngeal cavity, different vowel qualities can be produced. When the lip block is raised completely, oral closure is achieved at the lip end. The sudden release of the block produces the quick lip opening movement necessary for [b] and [m] with and without the proper velophreayngeal gesture.

In the present study, the two mechanical models in [5] and [7] are integrated, and a new model is designed consisting of the main vocal tract, a flapping tongue, a movable lower lip, and a nasal cavity with a controllable velopharyngeal port. With this model, more combinations of consonant sequences are available, including the cluster [br]. For this study, [br] is tested with different timings of lip and lingual gestures.

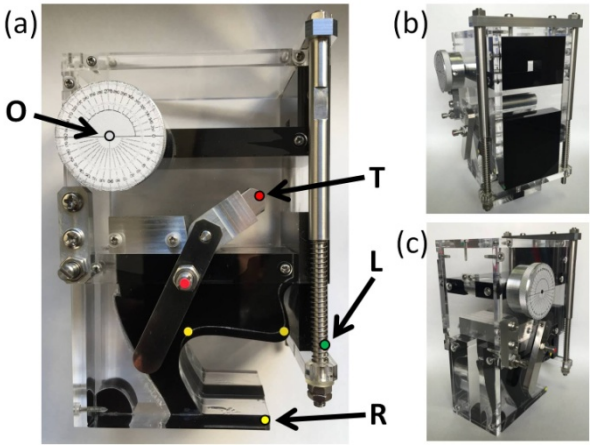

Figure 1: *The proposed vocal-tract model designed for [r], [l], [b], [m], and [w]. (a) Side view. (b) Front view. (c) Rear view.*

http://dx.doi.org/10.21437/Interspeech.2017-617

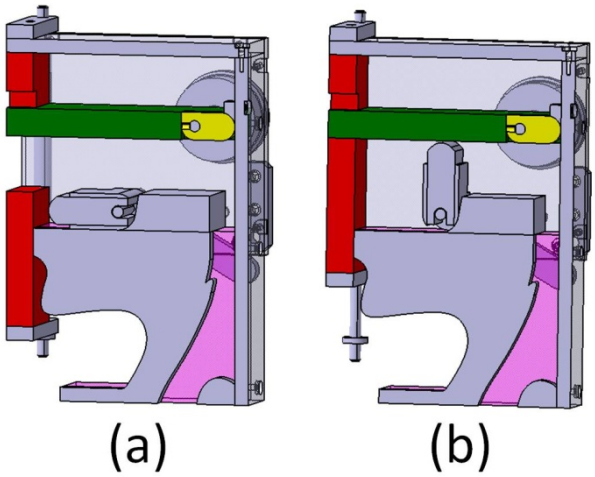

Figure 2: *Schematic illustrations of the proposed model. This view of the model was created by cutting along the midsagittal plane and removing the left portion. (a) The short tongue is at resting position; the lips are open. (b) The short tongue is rotated at 90 degrees; the lips are closed.*

2. Design

Figure 1 shows the proposed vocal-tract model. In Fig. 1, the lips are open, the velopharyngeal port is closed, the tongue is short, and it is in resting position. The design of this model is based on the combination of the two mechanical models in [5] for sounds [r]-[l] and in [7] for sounds [b]-[m]-[w].

This model has the nasal cavity on top of the oral cavity, and velopharyngeal coupling is achieved by rotating the knob. When the lips are open and the velopharyngeal port is closed, the output sound is more or less similar to the vowel [a], due to the narrow constriction in the pharyngeal region and the wide oral cavity with a cross-sectional dimension of 45 mm x 20 mm. The nasal cavity has the same cross-sectional dimension as the oral cavity, i.e., 45 mm x 20 mm. The length of the nasal cavity is 75 mm. The rotating part for the velopharyngeal gesture is located at the velum. The front-end block of the nasal cavity has a single nostril, with a dimension of 10 mm wide x 6 mm high x 10 mm deep. The dimensions of the rotating piece are 10 mm wide x 10 mm high x 15 mm long. When the rotation is 0 degrees, as shown in Fig. 2, the velopharyngeal port is completely closed. When the rotation is 45 degrees, the area of the velopharyngeal port is approximately 70 mm^2. This area is approximately the same size that House & Stevens (1956) discussed in a previous study for nasalized vowels [8, 9].

The lower lip is moveable, and the area of lip opening can be controlled by manually pushing up the lower lip block. Because the mouth end dimension has a maximum opening of 45 mm wide x 20 mm high, the lip block can be raised from 0 mm to 20 mm. When the lip block is raised completely (20 mm), complete oral closure is achieved at the lip end. When releasing the oral closure, one can either gradually reduce the force applied to the lip block from the bottom or suddenly release the hand holding up the lip block because a pair of springs are attached to both sides of the lip block, and restoration force is generated by raising the lip block. The sudden release of the lip block produces the fast lip opening movement necessary for [b] and [m].

The first half of the tongue can be rotated from 0 degrees (resting position) to approximately 90 degrees with the short length of the tongue. To rotate the tongue, we manipulate a lever attached to the rotation axis. When the length of the tongue is long, the maximum rotation is approximately 45 degrees, because the tongue tip makes contact with the alveolar ridge. When the tongue is short, the length of the rotating part is approximately 24 mm, while it is approximately 32 mm with the long length.

Figure 2 shows schematic illustrations of the same model. In these figures, the model is viewed by cutting along the midsagittal plane and removing the left portion of the model. In Fig. 2(a), the short tongue is at resting position, and the lips are open. In Fig. 2(b), the short tongue is rotated 90 degrees, and the lips are closed. In both figures, the lip block of the oral cavity and the end block of the nasal cavity are red (the thickness of these blocks is 10 mm), while the rotating part for the velopharyngeal opening is yellow.

3. Producing [br] cluster

Next, we produced a set of short nonsense words using the proposed model with labial and retroflex gestures. As an input signal, a reed-type sound source [3] was fed into a glottal hole at the larynx. The produced sounds were recorded and later used for a perceptual evaluation, acoustic analysis and gestural trajectory extraction.

3.1. Recordings

The output signals from the model were recorded digitally with a digital audio recorder (Marantz, PMD670) with a microphone (Sony, EMC-23F5). The original 48 kHz sampling frequency for the recordings was retained for the perceptual evaluation.

We recorded video images simultaneously with sound recordings for each utterance. We used a digital camera with the ability for high-speed recording (Casio, Exilim Pro EX-F1). The speed of the video imaging was 300 frames-per-second. Subsequently, the four dots shown in Fig. 1(a) were traced for extracting gestural trajectories.

The author manipulated the model manually and a total of 42 utterances were recorded. In each utterance, two gestural motions were produced: labial and retroflex. For the labial motion the lower lip was initially at resting position, it was then raised by pushing the lip block upwards for complete lip closure, and finally the lips were suddenly released. For the retroflex motion the tongue was initially at resting position, then the front half of the tongue was rotated by manipulating the lever, and finally, the tongue was returned to its original position again. The timing of these motions varied by utterance.

3.2. Perceptual evaluation

The recorded utterances were perceptually evaluated by an experienced phonetician who is a native speaker of American

English. The evaluation results are listed in Table 1. The phonetician was asked to transcribe each utterance phonetically. The major transcriptions in Table 1 are categorized into the following patterns: [ara], [abra], [arbra], [arbəra], [arbɚ], and [arba] (the transcriptions that only appear once in this table were omitted). As shown in this table, 13 out of 42 utterances contain the [br] cluster. This low rate of "13/42" was expected, because various timings between labial and retroflex motions were tested.

3.3. Gestural trajectories

One of the major causes of variation in the transcriptions in Table 1 is the timing of the labial and retroflex motions. To measure the timing of these motions, we can observe the articulatory motions directly on the proposed mechanical model with relatively low degrees of freedom. Because the proposed models have transparent side plates, the inside of the oral cavity is visible. Before the measurement, we placed several colored markers on the right side of the model as shown in Fig. 1(a). Dot "O" is located at the center of the knob and used as the origin. Dot "R" is located at the front end of the base plate and used as a reference point. Dot "L" is located at the lowest end of the lower lip block. Dot "T" is located at the tongue tip.

The x- and y-coordinates of the four dots were all extracted manually on a PC monitor screen frame by frame (the frame rate was again 300 fps). Then, the extracted (x, y) data were adjusted in the following three steps: 1) scaling, 2) shifting, and 3) rotation. After this adjustment, dot "O" became the origin and the (x, y) data were in millimeters.

With the vertical motion of the labial gesture in this model, we tracked the temporal trajectory of dot "L". We only focused on the y-coordinate of dot L, or Ly, for the labial motion. We also tracked the temporal trajectory of dot "T" but we only focused on the y-coordinate of dot T, or Ty, for retroflexion.

The left panels of Figure 3 show the temporal trajectories of Ly and Ty for the four utterances: (a) No. 7 ([abra]), (b) No. 15 ([arbəra]), (c) No. 19 ([arbra]), and (d) No. 33 ([arba]). The red (thick) lines in these plots are the labial motion and they drop steeply when the labial closure is released for the sound [b]. The black (thin) lines show retroflexion, and the timings vary among the four utterances. A nine-point median filter was applied. The right panels of the same figure show the spectrograms of the utterances.

4. Discussion and conclusions

In this study, we joined the [r]-[l] model and the [b]-[m]-[w] model to form a new model and were able to produce consonant sequences, including [br]. This model has only low degrees of freedom in terms of articulatory gestures. This makes it simple to manipulate and effective for educational purposes. The low degrees of freedom increase replicability, which makes this model particularly suited for research purposes as well.

Table 1: *Results of the perceptual evaluation test. A phonetician transcribed each utterance phonetically. The timings of retroflex motion was also measured relative to the timing of the labial closure release, where "A — B" shows the onset and offset times of the return motion of retroflexion.*

No.	IPA	Timings of Retroflex [ms]		No.	IPA	Timings of Retroflex [ms]
1	ara	37 — 153		22	arbra	7 — 143
2	ara	27 — 207		23	arbra	0 — 150
3	ara	20 — 173		24	arbəra	17 — 160
4	ara	17 — 207		25	arbəra	30 — 197
5	ara	10 — 173		26	arbəra	27 — 183
6	ara	(unclear)		27	arbɚ	(unclear)
7	abra	27 — 197		28	arbəra	27 — 173
8	abra	20 — 193		29	arbɚ	(unclear)
9	abara	27 — 213		30	arbəra	7 — 173
10	abra	23 — 190		31	arbra	10 — 153
11	arbra	20 — 183		32	arbəra	20 — 150
12	arbəra	20 — 207		33	arba	-37 — 117
13	arbra	7 — 147		34	arbəra	20 — 167
14	arbra	23 — 180		35	arba	-57 — 117
15	arbəra	23 — 183		36	arbəra	70 — 223
16	arbra	10 — 173		37	ara	-53 — 100
17	arbra	17 — 160		38	ara	20 — 160
18	arbəra	47 — 380		39	a a	(unclear)
19	arbra	7 — 153		40	ara	3 — 133
20	arbra	10 — 160		41	ar a	(unclear)
21	arbəra	33 — 190		42	arbəra	20 — 160

For research, it is important to know which timings of the labial and lingual gestures are suitable in order to produce the [br] consonant cluster. Therefore, in the present study, we acoustically and visually recorded 42 utterances with the [br] cluster.

With reference to the starting point at which the labial closure is released, let us examine the timing and duration of the retroflex motion. For utterances 7 and 19, the retroflex and labial motions began almost simultaneously and took approximately 100 ms to return to resting position. For both of these utterances, the [br] cluster was heard. For utterance 33, the retroflex motion had already begun before the labial release. In this case, [br] was not perceived. For utterance 15, the retroflex motion started to move approximately 30-40 ms after the labial release. In this case, the utterance sounded like schwa [ə] was inserted between the [b] and [r]. It seems that this schwa is "the targetless schwa" in Brownman & Goldstein [10, 11]. Table 1 also shows the timings of retroflex motion relative to the timing of the labial closure release for each utterance. The notation of "A — B" in the last column of this table shows when the return motion of retroflexion started and ended ("0 ms" is the timing of the labial closure release).

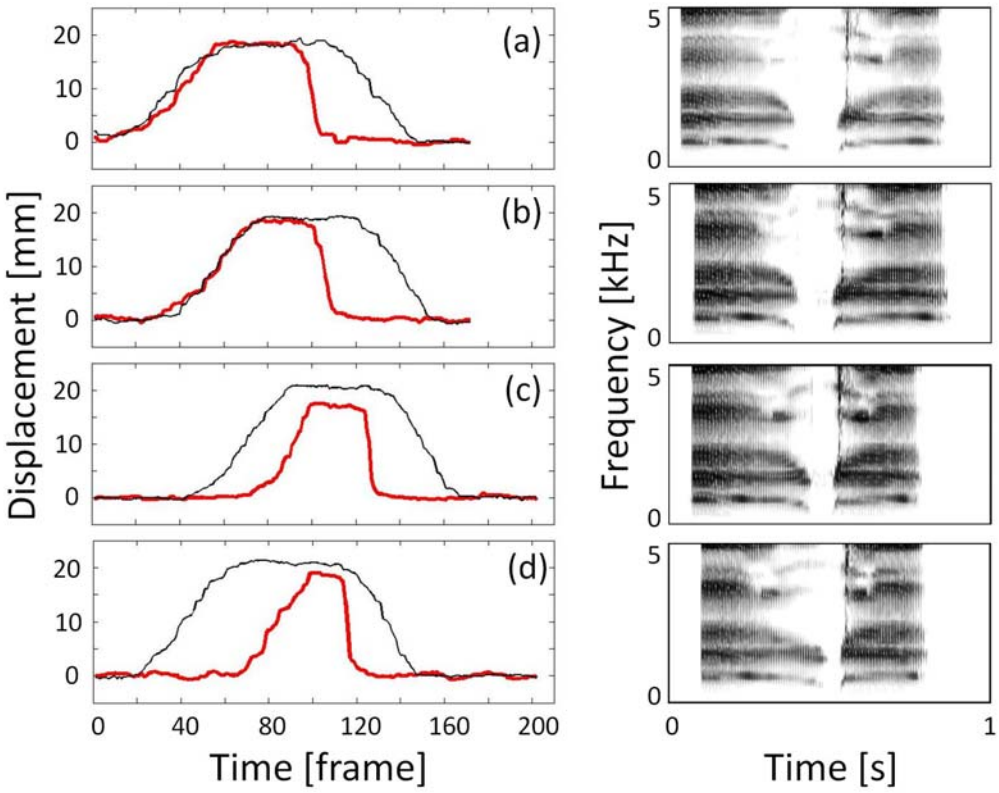

Figure 3: Left: temporal trajectories of Ly (red/thick) and Ty (black/thin) for the four utterances. The vertical axis is in mm, whereas the horizontal axis is in frame of 300 fps. Right: spectrograms of the utterances. (a) No. 7, (b) No. 15, (c) No. 19, (d) No. 33.

The average delays of the starting points of the return motion of retroflexion were 13.9 ms and 27.8 ms for [br] and [bər], respectively. The standard deviations were 8.18 ms for [br] and 15.81 ms for [bər]; a two-sided t-test indicates that mean delay for [br] is significantly less than the mean delay for [bər] ($p = 0.0117$).

Thus, this study showed that although the model was designed as an educational tool, it is also useful for research purposes. In the future, we can continue to discuss issues, such as the "in-phase" coproduction of [b] and [r] constriction onsets in Articulatory Phonology (the "in-phase" phasing relationship is well illustrated for utterances No. 7 and No. 19. Furthermore, we can mechanically control the articulatory movements by actuators as in [12-14].

5. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 15K00930. I would also like to thank Rion Iwasaki and Terri Lander for their support.

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., "Vocal-tract models and their applications in education for intuitive understanding of speech production," *Acoust. Sci. Tech.*, 37(4):148-156, 2016.

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

[7] Arai, T., "Mechanical Production of [b], [m] and [w] using controlled labial and velopharyngeal gestures," *Proc. of INTERSPEECH*, 1099-1103, 2016.

[8] House. A. S. and Stevens, K. N., "Analog studies of the nasalization of vowels," *J. Speech and Hearing Disorders*, 21, 218-232, 1956.

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

[10] Browman, C. P. and Goldstein, L., "Articulatory phonology: An overview," *Phonetica*, 49, 155-180, 1992.

[11] Moore, J. and Arai, T., "Articulation of English consonant clusters by native English speakers and Japanese speakers," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 259-260, 2015.
[12] Fukui, K., Kusano, T., Mukaeda, Y., Suzuki, Y., Takanishi, A. and Honda, M., "Speech robot mimicking human articulatory motion," *Proc. of INTERSPEECH*, 1021-1024, 2010.
[13] Arai, T., "Mechanical vocal-tract models for speech dynamics," *Proc. of INTERSPEECH*, 1025-1028, 2010.
[14] Brady, M. C., "Prosodic timing analysis for articulatory re-synthesis using a bank of resonators with an adaptive oscillator," *Proc. of INTERSPEECH*, 1029-1032, 2010.