INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Mechanical Production of [b], [m] and [w] Using Controlled Labial and Velopharyngeal Gestures

Takayuki Arai

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

Abstract

As an extension of a series of models we have developed, a mechanical bent vocal-tract model with nasal cavity was proposed for educational and clinical applications, as well as for understanding human speech production. Although our recent studies have focused on flap and approximant sounds, this paper introduced a new model for the consonants [b], [m] and [w]. Because the articulatory gesture of approximants is slow compared to the more rapid movement of plosives, in our [b] and [m] model, the elastic force of a spring is applied to affect the movement of the lower lip block, as was done for flap sounds in our previous studies. The main difference between [b] and [m] is in the velopharyngeal port, which is closed for [b] and open for [m]. In this study, we concluded that 1) a slower manipulation of the lip block is needed for [w], while 2) [b] and [m] require a faster movement, and finally, 3) close-open coordination of the lip and velopharyngeal gestures is important for [m].

Index Terms: speech production, physical models of the human vocal tract, lips, nasal cavity, velopharyngeal coupling

1. Introduction

We have been developing a series of mechanical vocal-tract models for multiple purposes, including educational and clinical applications, understanding speech production through the use of mechanical models, and speech technology applications, such as designing a speaking robot that mimics human speech. Although the majority of our mechanical models produce vowel sounds [1-3], we have gradually designed and implemented mechanical vocal-tract models for consonants, too. Our first successful consonant models were for Japanese /r/, and English /r/ and /l/ [4,5].

The sound that most frequently represents Japanese /r/ is the alveolar flap. (Please see [6] for a detailed discussion of allophonic variations). To implement an alveolar flap sound, we designed a vocal-tract model with a 90-degree bend in the middle and a mechanical tongue, the first half of which can be raised to touch the alveolar ridge by means of a rotating lever [4]. We used a rubber band to increase the speed of the return movement of the tongue. With this model, we were successfully able to produce the short nonsense word /ere/ with an alveolar flap, as in Japanese.

The typical English /r/ sound is a retroflex or alveolar approximant. For these sounds, we applied the model for Japanese /r/ and modified it so that the front half of the tongue could rotate against the palate without touching it. This simulated the movement of an approximant [4]. With this model, we were successfully able to produce a short nonsense word /ara/ with a retroflex /r/, as in English.

To produce the lateral approximant English /l/ sound, we used the English /r/ model, but only modified the tongue length so the tip of the tongue would touch the palate [4]. In this model, the tongue does not completely block airflow in the oral cavity, but leaves lateral pathways on both sides of the tongue. With this model, we were successfully able to produce the short nonsense word /ala/, with a lateral approximant, as in English.

In addition, we also developed a mechanical vocal-tract model for English "bunched /r/" [5]. This model did not have the rotating tongue or the lever. Instead, there were 10-mm-thick plates lined up perpendicularly in the oral cavity such that each plate could be pushed up from the outer bottom to raise it and make a constriction at a particular position in the oral cavity. We were able to produce the nonsense word /ara/ with "bunched /r/" by raising the plates positioned 50-60 mm from the lips (more properly with lip-rounding) [5].

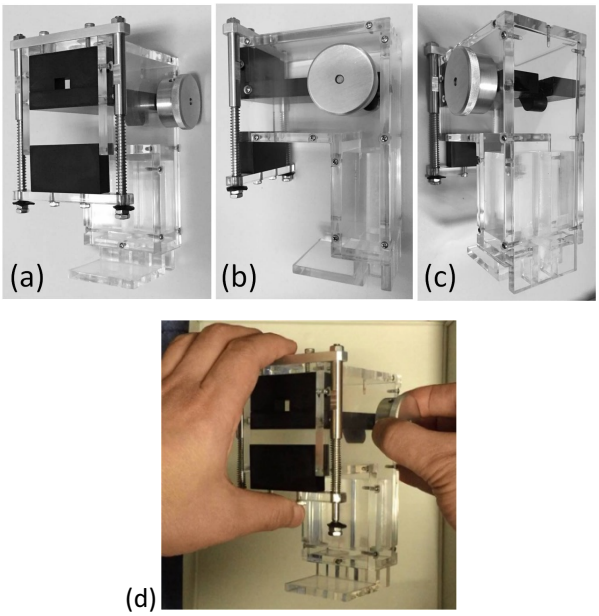
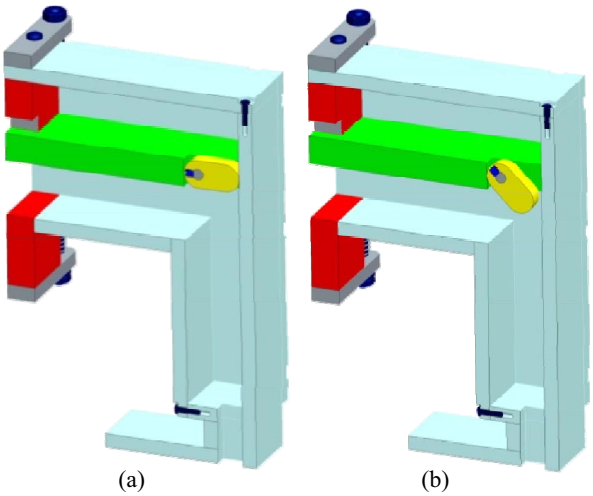


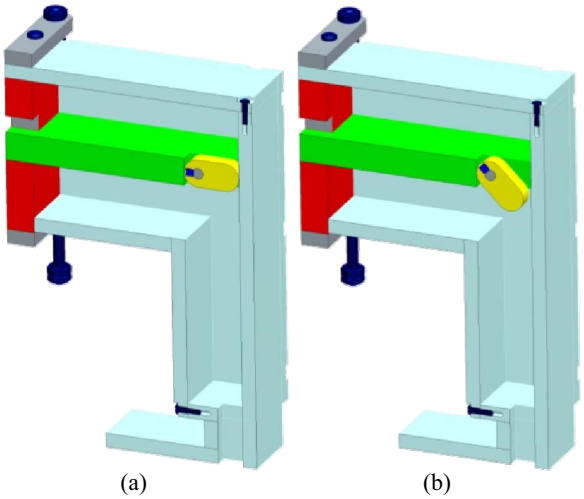
Figure 1: *The proposed vocal-tract model with the nasal cavity. (a) Front view. (b) Side view. (c) Rear view. (d) Left hand manipulates the lip block for the labial gesture and right hand rotates the knob for the velopharyngeal gesture.*

http://dx.doi.org/10.21437/Interspeech.2016-189



(a) (b)

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. The block in the oral or pharyngeal cavity is also temporally removed. The lips are open. The velopharyngeal port is closed in (a) and open in (b).*



(a) (b)

Figure 3: *Schematic illustrations of the proposed model. This view of the model was created by cutting along the midsagittal plane and removing the left portion. The block in the oral or pharyngeal cavity is also temporally removed. The lips are closed. The velopharyngeal port is closed in (a) and open in (b).*

In the present study, we extend and design a new mechanical model that can produce the additional consonants: [b], [m] and [w], in order to increase the number of sounds our models can produce. As a model by Umeda & Teranishi [7], the proposed model has a nasal cavity. Furthermore, it is bent and contains a lip block to control the labial gesture and a velopharyngeal port to control nasality. In previous studies with approximants, we manipulated the models manually, where as with flaps, we needed elastic force to enhance the return movement of the tongue. For [b] and [m], we also apply elastic force to increase the return movement of the lower lip in the present study.

2. Design

Figure 1 shows the proposed vocal-tract model. This model has the nasal cavity on top of the oral cavity, and velopharyngeal coupling is achieved by rotating the knob. The lower lip is moveable, so the area of the lip opening can also be controlled by manually pushing up the lower lip block. In these figures, both the lips and the velopharyngeal port are open.

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, due to the uniform cross-sectional dimension (45 mm wide x 20 mm) in both the oral and pharyngeal cavities located above the larynx. The larynx has a cross-sectional dimension of 9 mm x 9 mm and a length of 20 mm. When there is a constriction in the oral or pharyngeal cavity, different vowel qualities can be achieved. Fig. 1(a) shows the configuration for /a/ with a block located in the pharyngeal cavity. This block has a cross-sectional dimension of 45 mm x 20 mm with a length of 45 mm. Because there is a 9 mm x 9 mm square groove along the length of the block, the sound propagates in the groove. (The block is placed 5 mm above from the bottom of the pharyngeal cavity in Fig. 1a). When the same block is located in the oral cavity, with the edge of the block set back 20 mm from the mouth end of the vocal tract, vowel /i/ can be produced.

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 (height) x 10 mm (depth).

Figures 2 and 3 show 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. The block in the oral or pharyngeal cavity is also temporally removed. In Fig. 2, the lips are open, and the velopharyngeal port is closed in Fig. 2(a) and open in Fig. 2(b). In Fig. 3, the lips are closed, and the velopharyngeal port is closed in Fig. 3(a) and open in Fig. 3(b). 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.

2.1. Labial gesture

As described earlier, this model has a movable lower lip. The lower lip can be pushed up by raising the lip block manually. 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, restoration force is generated by raising the lip block. The sudden release of the holding hand produces the fast lip opening movement necessary for [b] and [m].

2.2. Velopharyngeal gesture

As described above, this model has a rotating piece for the velopharyngeal gesture. The dimensions of the rotating piece are 10 mm (wide) x 10 mm (height) x 15 mm (length). When the rotation is 0 degrees, as shown in Fig. 2(a) and Fig. 3(a), the velopharyngeal port is completely closed. When the rotation is 45 degrees, as shown in Fig. 2(b) and Fig. 3(b), the area of the velopharyngeal port is approximately 70 mm^2. This area is approximately the same size that House & Stevens (1956) discussed in the previous study for nasalized vowels [8].

3. Producing [b], [m] and [w]

Next, we produced a set of short nonsense words with the consonants, [b], [m] and [w] as well as vowel [a] and its nasalized version, by using the proposed model. The nonsense word was /V_1CV_2/, where /C/ was either [b], [m], or [w], and the two vowels /V_1/ and /V_2/ were always [a] in the rest of this paper. As an input signal, a whistle-type sound source was fed into a glottal hole at the larynx. Recordings were done for the produced sounds, which were later used for the acoustic analysis and perceptual evaluation. The output signals from the model were recorded digitally with a digital audio recorder (Marantz, PMD670) with a microphone (Sony, EMC-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.

3.1. Consonant [b]

In general, the plosive [b] manifests a burst at the lips as a result of the build-up and release of air pressure after the closure of the oral cavity [9]. However, it is well-known that only the fast formant transitions, especially the rising transitions of the first and second formants (F1 and F2) yield the perception of the [b] sound [10]. Therefore, we simulated [b] by controlling the labial and the velopharyngeal gestures as follows:

- The velopharyngeal port is closed all the time.
- First, the lip block is open for the initial vowel [a].
- Second, it is closed for [b].
- Finally, the block is suddenly released after a short interval.

Figure 4 shows spectrograms of two repetitions of the nonsense word /aba/. In this case, the velopharyngeal port was always closed, and only the lip block was manipulated. As can be seen, the closure portion is clear in this figure. The sudden release and corresponding formant transitions shown in the figure are the crucial cue for [b].

3.2. Consonant [m]

For [m], the articulatory gestures in the oral cavity are similar to those for [b]. The main difference between [b] and [m] is the velopharyngeal coupling [11]. We simulated [m] by controlling the labial and the velopharyngeal gestures as follows:

- First, the velopharyngeal port is closed and the lip block is open for the first vowel [a].
- Second, the velopharyngeal port is open and the lip block is closed for [m].
- Finally, the velopharyngeal port is closed and the lip block is open, again, for the second vowel [a].

Each of Figs. 5(a) and 5(b) shows spectrograms of two repetitions of the nonsense word /ama/. Two different versions were tested in Figs. 5(a) and 5(b). In the first version (Fig. 5a), we manipulated the velopharyngeal port and the lip block simulataneously, so that the velopharyngeal opening and labial closing occurred synchronously. In the second version (Fig. 5b), however, the velopharyngeal port was open a little bit earlier, and then the lip block was closed for [m]. In both cases, the closure portion is clearly observed, and as we observed for [b], the sudden oral release and corresponding formant transitions were achieved after the nasal murmur.

3.3. Consonant [w]

When producing [w], there are two narrow constrictions in the vocal tract [10]. That is, the tongue dorsum is raised and makes a constriction at the velar position, and the lips are labialized. Although we could have carried out the velar constriction, we mainly controlled the labial gesture to simulate [w] as follows:

- The velopharyngeal port is closed all the time.
- First, the lip block is open for the initial [a].
- Second, it is nearly closed for [w].
- Finally, the block is gradually open after a short interval.

Figure 6 shows spectrograms of two repetitions of the nonsense word /awa/. In this case, the velopharyngeal port was always closed, and only the lip block was manipulated. As shown in this figure, there is no closure portion, and the gradual formant drops of the first three formants were observed as acoustic cues for the labio-velar approximant [12].

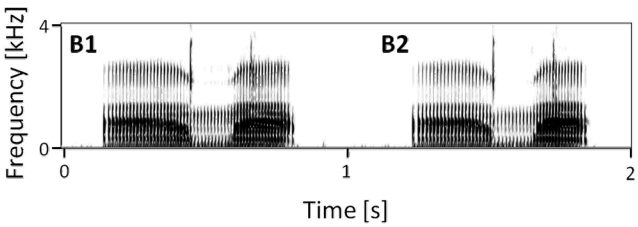


Figure 4: *Spectrographic representation of /aba/.*

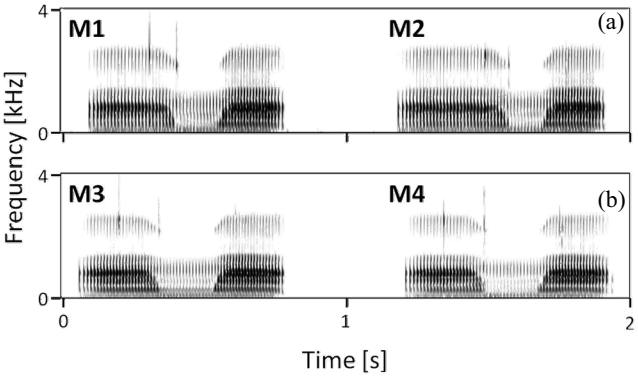


Figure 5: *Spectrographic representation of two versions of /ama/. (a) The velopharyngeal opening and labial closing occurred synchronously. (b) The velopharyngeal port was open a little bit earlier, and then the lip block was closed for [m].*

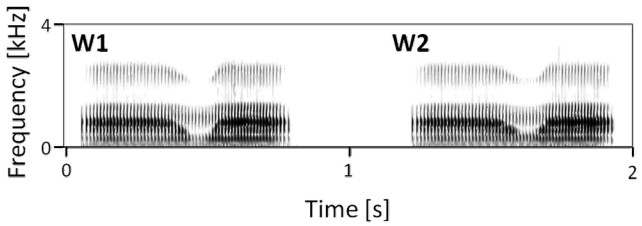


Figure 6: *Spectrographic representation of /awa/.*

4. Perceptual evaluation

A perceptual evaluation test was conducted by two experienced phoneticians. The target stimuli of the evaluation test were the eight speech samples displayed in Figs. 4 through 6. There were two sessions in this test. In the first session, the raters were asked to transcribe the intervocalic consonant of the stimuli phonetically. In the second session, the raters were informed of the target nonsense word for each stimulus before making their judgments. Then, the raters were asked to evaluate from 1 (Very Bad) through 5 (Very Good), based on how good each stimulus was as the target word. The two experienced phoneticians were native speakers of Japanese (Rater1) and American English (Rater2), and their evaluation results are listed in Table 1 (for the detailed descriptions of this table, please see the caption).

For B1, the two raters perceived [m], probably because many of the utterances sounded slightly nasalized. For B2, Rater2 perceived [b] with a relatively high score (4). Rater1 perceived nasalized [b] for B2, although the stimulus did not get as high a rating as Japanese /b/. One of the reasons why B1 and B2 did not sound like a perfect [b] seems to be due to the relatively high intensity during the [b]-closure. The more we can suppress the airflow during the [b]-closure, the more we may achieve the sudden release after the [b]-closure.

The two raters more or less perceived [m] for M1 through M4. Because M1 and M2 were obtained by opening the velopharyngeal port and closing the lips simultaneously, a consonant with a shorter duration was produced. However, the scores of M1 and M2 were lower than those of M3 and M4. For M3 and M4, the velopharyngeal port was open a little bit earlier, and then the lip block was closed for [m]. This time delay simulated what we humans do; as a result, it sounded more like what we produce, and the evaluation scores were high. The consonant tended to be a little bit longer in M3 and M4.

Finally, Rater1 perceived a voiced bilabial approximant and Rater2 perceived a voiced labiodental approximant for W1 and W2. The labial sound is reasonable, since we only manipulated the labial gesture but not the velar gesture. Based on the perceptual test scores, the nonsense word /awa/ sounded reasonably well for Japanese, but it sounded more like English /ava/. This is also reasonable because the proposed model produced a sound of which the most closest one was /v/ in English.

5. Discussion and conclusions

In this study, we designed and developed a new bent vocal-tract model with a nasal cavity. We extended and modified our previous studies, most of which were originally designed for vowels. While our recent models were aimed at producing the consonants /r/ and /l/, the target consonants of the present paper were [b], [m] and [w]. As a result, we were able to produce these three sounds, although we found several points for improvement in the future. Among the three sounds in the present study, the consonant [m] was nearly perfectly simulated, although the gestures were the most complicated. We confirmed that for temporal coordination between the labial and velopharyngeal gestures there needs to be a certain time delay. Because the main difference between [b] and [m] is the status of the velopharyngeal port (close / open), the motion of the labial gesture with the proposed model was fast enough with the support of the elastic force to produce [b] and [m]. One of the problems of [b] might be that the air pressure built-up during the [b]-closure was not sufficient to make a strong burst when releasing the air pressure in Section 3. For [w], the required movement of the labial gesture is relatively slow, and it was thought that this sound would be easy to produce. However, with the proposed model, we were not able to achieve the velar gesture, and the results were language dependent. Thus, we need further discussion on the contributions of different gestures language by language. In the future, we would also like to control the gestures by computer for more reliable reproduction with controlled speeds.

Table 1: *Results of the perceptual evaluation test. Two raters participated in the two sessions. Session 1 was the phonetic transcription for each stimulus. Session 2 was the goodness rating. In Session 2, each rater was asked to give a score as to how well the stimulus sounded as compared to the target nonsense word appearing between the slash marks. The "JP" / "EN" stands for "Japanese" / "English" and means that the stimulus was rated as Japanese / English. The produced sound was rated on a 5-point scale, where 1: Very Bad, 2: Bad, 3: Moderate, 4: Good, and 5: Very Good.*

Stimulus No.	Rater1 (Japanese)		Rater2 (American Eng.)	
	Session 1	Session 2	Session 1	Session 2
B1	[mm]	JP /aba/: 1	[b] / [m]	EN /aba/: 2
B2	[m$\tilde{b}$]	JP /aba/: 2	[b]	EN /aba/: 4
M1	[mm]	JP /ama/: 3	[m]	EN /ama/: 4
M2	[m]	JP /ama/: 4	[m]	EN /ama/: 3
M3	[ŋm]	JP /ama/: 5	[m]	EN /ama/: 4
M4	[m:]	JP /ama/: 5	[m]	EN /ama/: 5
W1	[$\tilde{\beta}_{+}$]	JP /awa/: 3	[ʋ]	EN /awa/: 2
W2	[β_{+}]	JP /awa/: 4	[ʋ]	EN /awa/: 3

6. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 15K00930. I would also like to thank Miho Yamada, Takuya Kimura 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., "Physical models of the vocal tract with a flapping tongue for flap and liquid sounds," *Proc. of INTERSPEECH*, 2019-2023, 2013.

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

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

[7] 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.

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

[11] Krakow, R. A. and Huffman, M. K., "Instruments and techniques for investigating nasalization and velopharyngeal function in the laboratory: An introduction," in *Nasals, Nasalization, and the Velum*, edited by M. K. Huffman and R. A. Krakow, Academic Press, San Diego, 3-59, 1993.

[12] Lisker, L., "Minimal cues for separating /w,r,l,j/ in intervocalic production," *Word*, 13, 257-267, 1957.