How physical models of the human vocal tract contribute to the field of speech communication

Takayuki Arai*
Department of Information and Communication Sciences, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan

Abstract: Several types of physical models of the human vocal tract have been developed previously by our group. Even though they were originally designed for education purposes in acoustics and speech science, some of the models can also be applied to research, pronunciation training and clinical purposes. For example, a model for the English /r/ was originally designed to teach how the sound is produced, but we have also found it to be affective when applied to practicing English vowels for non-native speakers. Another model for lateral approximant was originally designed to teach how lateral sounds are produced. The model was then tested to measure differences in sounds radiated from the center and lateral directions with the possibility of evaluating misarticulation in a clinical situation. A recent model with a movable lower lip and rotating tongue to imitate the retroflex gesture was used to simulate the English /br/ cluster, a particularly difficult speech sound for Japanese native speakers. By using this model, users can observe each articulators’ movement visibly with individually adjustable parameters to produce different speech sounds. Thus, the vocal-tract models potentially contribute to the field of speech communication.

Keywords: Vocal-tract models, Speech communication, Speech science, Acoustic phonetics

PACS number: 43.70.+i [doi:10.1250/ast.41.90]

1. INTRODUCTION
Since starting my teaching career at Sophia University, students in my class were always a mixture of engineering and humanities students. To teach speech communication more intuitively, I wanted to use educational tools, such as physical models of the human vocal tract. That was my original motivation to replicate the physical models designed by Chiba and Kajiyama [1]. Once I started using them in my class, I have noticed the effectiveness in using these models in the classroom, where students were able to understand the theories better and showed more enthusiasm in class. Since then, different types of physical models of the human vocal tract and related models were developed and described in [2–4].

Although our set of the vocal-tract models were originally designed for educational purposes, we have found them to be useful for other purposes: basic research purposes, pronunciation training purposes, and clinical application purposes. In this article, we will review such applications of the vocal-tract models and discuss how they contribute to the field of speech communication.

2. ARTICULATION SIMMULATOR
Figure 1 shows a bent-type physical model of the human vocal tract [5]. It is bent at 90 degrees in the middle of the vocal tract. In the oral cavity, there is a mechanical tongue, of which the front half can be rotated with the lever. The lower lip is movable, so that the lips can make a complete closure. There is also a dial for the velopharyngeal port, so that the nasal cavity couples with the main vocal tract, that is, the oral and pharyngeal cavities, when the port is opened.

Because this model can produce certain consonants, such as [b], [m], [w], [r] and [l] as well as vowels, such as [a], the model can be manipulated to produce consonant clusters, such as [br]. In addition to that English [r] is problematic for Japanese native speakers, the consonant clusters in English are often difficult for English learners including Japanese native speakers, as Japanese does not allow a sequence of two consecutive consonants, with few exceptions.

To simulate the consonant cluster [br], the model was manually manipulated with different timings by the author. Forty-two repetitions were tested with different timings between the close-open gesture of the lower lip for [b] and the lingual retroflexion for [r]. The output signals from the
were recorded digitally. We recorded video images simultaneously with sound recordings for each utterance. We used a digital camera with the ability for high-speed recording (the speed of the video imaging was 300 frames-per-second). Subsequently, the four dots (indicated as ‘O,’ ‘T,’ ‘L,’ and ‘R’ in Fig. 1) were traced for extracting gestural trajectories.

As a result, several patterns of [br] clusters could be produced and they were perceived as native-like sounds by a phonetician (Table 1). In addition, a typical misarticulation of this cluster (e.g., [brə] with the epenthetic schwa) can also be produced, when the timings of the two gestures were longer than a certain lag, simulating what a native speaker of Japanese might produce instead. The average delays between lingual and labial release gestures were approximately 14 ms for [br] and 28 ms for [brə]. The standard deviations were approximately 8 ms and 16 ms for [br] and [brə], respectively (Fig. 2). A two-sided t-test indicates that mean delay for [br] is significantly less than the mean delay for the one with schwa ([brə]) ($p = 0.0117$).

Thus, the model can be used as an articulation simulator, allowing users to observe each articulators’ movement visibly with individually adjustable parameters to produce different possible speech sounds.

### 3. PRONUNCIATION TRAINING

Figure 3 shows another bent-type physical model of the human vocal tract [6]. In the oral cavity, several 10-mm wide blocks lined up next to each other. One can slide each block up and down, so that a so-called “bunched-/r/” in English can be simulated with this model. In addition, high vowels can also be simulated.

In our previous study [7], we tested whether this model can be applied to pronunciation training. We asked participants to repeat the sounds that they heard through a loudspeaker, either by using their own speech organs and/or this physical model of the human vocal tract with their fingers.

Twelve participants were divided into two groups: control/target. The control group was asked to repeat each vowel they heard based on what they perceived through a loudspeaker. The target group was asked to repeat each
vowel, but in addition, they were asked to “repeat the same vowel” by manipulating the vocal-tract model with their fingers.

The formant analysis shows as follows. For the control group, F2 peaks often move around the target formants, but without consistency. For the target groups, F2 peaks seem to approach the target formants as the session progresses for /e/, a sound that does not exist in the speakers’ native language, Japanese.

Thus, the participants’ pronunciation for an English open-mid front vowel were observed to be slightly better in terms of formant frequencies than without using the models in the training sessions. There is a potential for applying such models of the human vocal tract to pronunciation training based on the consideration that tongue and finger movements are related in terms of motor control.

4. CLINICAL APPLICATION

In a clinical situation, a speech-language pathologist need to train patients for their articulation. In that case, a vocal-tract model is going to be a useful tool. Figure 4 is a model developed, of which one of the goals is such clinical applications.

Lateral articulation, a type of articulation disorders, is a common misarticulation of certain sounds, such as /shi/, /chi/, /ji/, /hi/, /ki/, /gi/, /ri/, in Japanese. Kato (1991) reported that the tongue came into complete contact with the hard palate and all occlusal and palatal surfaces on one side and the tongue did not come into contact with any molar occlusal surfaces but only with the palatal surface of the preterminal molar on the other side [9]. Akagi et al. (2001) investigated the relationship between perceptual diagnosis and the physical correlates, and they found that the spectral envelope characteristic has a peculiar variation at approximately 3.2 kHz [10]. Speech pathologists often diagnose this type of misarticulation by observing patients and listening to their pronunciations, as well as using the mirror test to visualize the air leakage from the lateral directions.

Because the air leakage coming from the lateral side(s) of the mouth during lateral articulation, we have recently developed a method to measure the directivity of sound radiations through the mouth [11]. Figure 5 shows a picture of the proposed system. This system divides the space in front of the mouth into three directions: midline, left and right. In each of three divisions, we can setup a microphone, so that we can record a speaker’s utterance with three microphones simultaneously.

Prior to measuring utterances produced by an actual human speaker, we are able to test the system with a vocal-tract model as an articulation simulator, we can adjust individual parameters in the system in various ways, which otherwise would not be easily controllable with an actual human speaker.

5. SUMMARY

This paper showcases a series of vocal-tract models developed by the author, and how they were originally used for educational purposes. In addition to their original purposes, we also showed that they are useful in other areas such as basic research purposes, pronunciation training purposes, and clinical application purposes, in this article. In the future, we will continue to develop and fine tune more types of vocal-tract models to suit different purposes.

ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI Grant Number 18K02988. The author would like to thank the members of Arai Lab. The keynote lecture at the UAC 2018 Seminar was a part of the ISCA Distinguished
T. ARAI: HOW VOCAL TRACT MODELS CONTRIBUTE TO COMMUNICATION

Lecture series by the International Speech Communication Association (ISCA).

REFERENCES

[1] T. Chiba and M. Kajiyama, The Vowel: Its Nature and Structure (Tokyo-Kaiseikan Pub. Co., Ltd.), pp. 1941–1942, Tokyo.

[2] T. Arai, “Education system in acoustics of speech production using physical models of the human vocal tract,” Acoust. Sci. & Tech., 28, 190–201 (2007).

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

[4] T. Arai, “Vocal-tract models and their applications in education for intuitive understanding of speech production,” Acoust. Sci. & Tech., 37, 148–156 (2016).

[5] T. Arai, “Integrated mechanical model of [r]-[l] and [b]-[m]-[w] producing consonant cluster [br],” Proc. Interspeech 2017, pp. 979–983 (2017).

[6] T. Arai, “Retroflex and bunched English /r/ with physical models of the human vocal tract,” Proc. Interspeech 2014, pp. 706–710 (2014).

[7] T. Arai, “Hands-on tool producing front vowels for phonetic education: Aiming for pronunciation training with tactile sensation,” Proc. Interspeech 2015, pp. 1695–1699 (2015).

[8] T. Arai, “Flexible tongue housed in a static model of the vocal tract with jaws, lips and teeth,” Proc. Interspeech 2018, pp. 171–172 (2018).

[9] M. Kato, “Lateral articulation: Observation of articulation movement using electro tooth-palatography,” Jpn. J. Logop. Phoniatr., 32, 18–31 (1991).

[10] M. Akagi, N. Suzuki, K. Hayashi, H. Saito and K. Michi, “Perception of lateral misarticulation and its physical correlates,” Folia Phoniatr. Logop., 53, 291–307 (2001).

[11] T. Arai and H. Kihara, “Detecting lateral articulation with partitions considering its directivity,” Proc. Spring Meet. Acoust. Soc. Jpn., pp. 773–774 (2019) (in Japanese).