MECHANICAL MODELS OF THE HUMAN VOCAL TRACT

Takayuki Arai
Department of Information and Communication Sciences
Sophia University
Tokyo, Japan

Introduction

This article discusses the use of mechanical models of the human vocal tract in teaching students about how the vocal tract works in the formation of speech.

The idea of having a mechanical model that simulates the human vocal tract dates back at least as far as to Wolfgang von Kempelen (see Fig. 1), whose manually operated speed synthesizer was developed between 1769 to 1791. An account can be found at the Wikipedia site http://en.wikipedia.org /wiki/Wolfgang_von_Kempelen's_ Speaking _Machine

"Having a variety of different types of vocal-tract models is important because individual models can address the different set of configurations that occur in the formation of speech."

Highly recommended reading is the paper in the *Journal of the Acoustical Society of America* (1950) by Dudley and Tarnozy [1] that reviews von Kempelen's development of his mechanical synthesizer and which also reviews the invention of analogous devices in later years. Von Kempelen summarized his work in a 451 page book [2] "Mechanismus der menschlichen Sprache bevst Bescreibung eincer sprechende Maschine (The mechanism of human speech, with a description of a speaking machine).

In 1837, Charles Wheatstone (Fig. 2) resurrected the work of Wolfgang von Kempelen, creating an improved replica of his Speaking Machine. Using new technology developed over the previous 50 years, Wheatstone was able to further analyze and synthesize components of acoustic speech, giving rise to the second wave of scientific interest in phonetics. After viewing Wheatstone's improved replica (Fig. 1) of the Speaking Machine at an exposition, a young Alexander Graham Bell set out to construct his own speaking machine with the help and encouragement of his father. Bell's experiments and research ultimately led to his invention of the telephone in 1876, which revolutionized global communication.

Over the past few years, the present author has developed and built several physical models of the human vocal tract. The reader will find these described in references 3-6. These models have been found to be helpful in teaching students who are studying acoustics and speech science. Having a variety of different types of vocal-tract models is important because individual models can address the different sets of configurations that occur in the formation of speech. All of the models that the author has developed demonstrate (1) the relationship between vocal-tract configuration and vowel quality, and (2) the source-filter theory of speech production [5]. Even with the *connected-tube* (CT) model [2], which is one of the simplest types, it is possible to demonstrate these two points. However, if one wants to teach students about the effects of tongue position and tongue movement, one needs models where the simulated vocal tract is not straight (loosely referred to here as *bent models*) and one also needs adjustable models of the vocal tract. The present article focuses on these types of models.

In more recent years, a talking robot has been developed that can change the vocal tract configuration dynamically [7]. The author and his colleagues, for the purpose of having appropriate demonstration apparatus for students, have also developed several adjustable dynamic models of the human vocal tract for educational purposes. The models that have been developed include (1) a *sliding-three-tube* (S3T) model [4, 10], (2) Umeda and Teranishi's computer-controlled model [11, 12], (3) a gel-type tongue model [5, 6] , and (4) head-shaped models [3, 12]. The first two models have a straight vocal tract, and the actual area variation with length is roughly the same as for the human vocal tract. In these cases, one can demonstrate that changing the vocal-tract configuration yields different vowel qualities in real time. Therefore, learners can compare the changes in the configuration visually (as seen by their eyes) as well as audibly hearing the changes in the output sounds with their ears. These designs are nevertheless relatively simple, so the simulation of the dynamic movements with these models are less realistic than for the other two models.

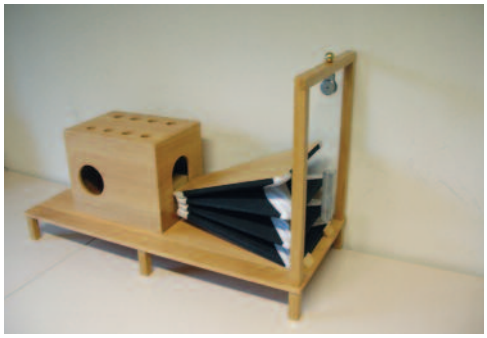


Figure 1. Replica of Von Kempelen's Speaking Machine (taken from the Wikipdia site)

Figure 2. Charles Wheatstone

Bent vocal-tract models are suitable in demonstrations of how one produces vowels with the use of one's speech organs. A common question that learners often ask is why the vocal tract isn't placed in our heads. Static bent

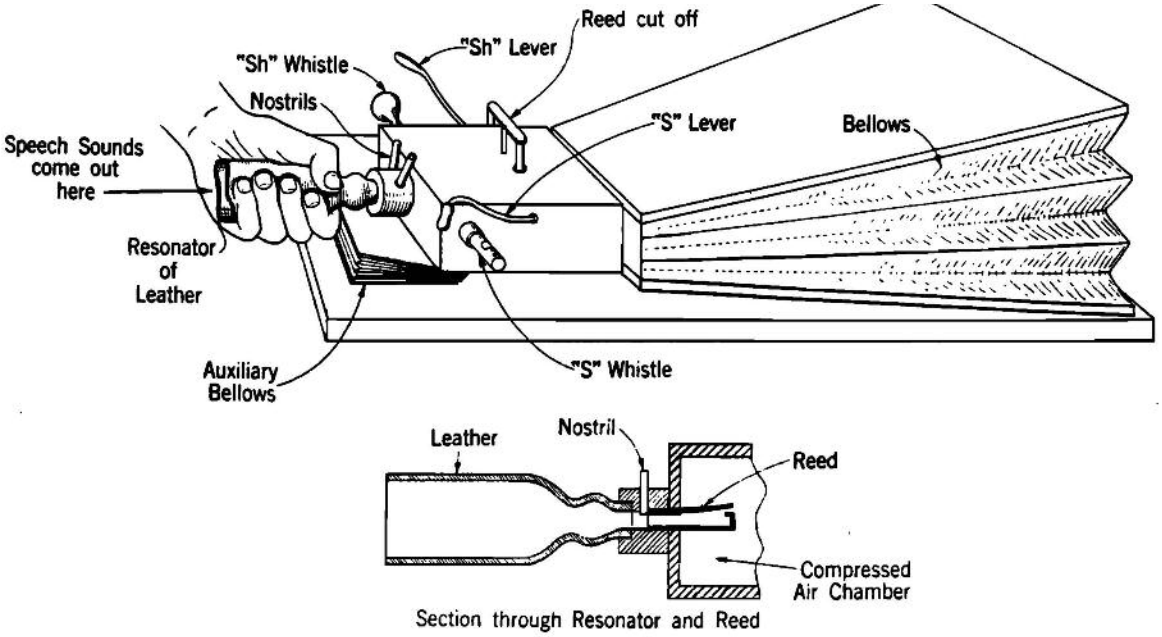


Figure 3. Wheatstone's speaking machine. (Dudley and Tarnoczy, 1959)

models with head shapes [1] often give a rudimentary answer to that question. However, static bent models which were originally developed were limited to the formation of only the vowels /a/ and /i/. The work reported here describes successful efforts to find a more nearly appropriate set of vocal-tract configurations for static bent models for classroom demonstrations.

The dynamic straight models are useful for simple demonstrations, but they do not directly simulate the movements of tongue as mentioned above. Therefore, we have developed dynamic bent models, such as the gel-type tongue model [5, 6] and the head-shaped model with the sliding tongue [12]. With these models, relatively realistic tongue movements can be simulated. The dynamic bent models are useful when one demonstrates a tongue movement between the formation of the vowels /a/ and /i/. With the dynamics bent models , learners can see that the downward / backward tongue movement is needed for formation of vowel /a/, and that the upward / forward tongue movement is needed for the formation of vowel /i/. The gel-type tongue model [5, 6] has many advantages, including the flexibility of the tongue, enabling one to produce many different vowels. One disadvantage of the gel-type tongue model is that it is difficult to manipulate, so that there is considerable challenge in reproducing the same configuration repeatedly. Because of this difficulty, this model is mainly used in classroom demonstrations when the author demonstrates the vowel production to learners in a class or workshop, but the individual learners are never asked to manipulate this model.

The head-shaped model with the sliding tongue [10] has the advantages of both the S3T (sliding three-tube) and the gel-type tongue models. The sliding tongue model has a limited number of degrees of freedom, so that it is simpler for one to produce the target vowel. In addition, the vocal tract is bent in the middle at a right angle, so that one can move the tongue more realistically. The degrees of freedom for this model are as follows: (1) the 1st degree of freedom is the diagonal movement of the tongue; (2) the 2-nd degree of freedom is the protrusion of the tongue dorsum; (3) the 3-rd degree of freedom is lip rounding. This model is able to produce the vowel sequence between /a/ and /i/ relatively easily; however, the other vowels were difficult to produce in a sequential manner. Consequently, the mechanical bent-type models were redesigned, so that a single bent-type model with sliding blocks would cover all of the vowels, /i/, /e/, /a/, /o/, and /u/.

In the study reported in the present article, the CT (connected-tube) model, which was originally designed with cylindrical tubes, was redesigned using square tubes. Then, two bent-type models with sliding blocks were designed: one of these was for the formation of front vowels and the other for back vowels. A final design resulted in a single bent-type model with sliding blocks that produces all five vowels.

Bent-type models with sliding blocks

To find an appropriate set of the vocal-tract configurations for static bent models for classroom demonstration, two bent-type models with blocks were designed The redesign of the CT (connected-tube) models with square tubes resulted in bent-type models with a rectangular cross-section. The basic dimension of the cross-section was 45 mm x 20 mm for the neutral vowel, schwa. The length of the oral and pharyngeal cavities was 90 mm and 70 mm, respectively. There was a narrow constriction at the larynx, the length of which was 20 mm. The dimension of its cross-section was 9 mm x 9 mm. Two bent-type models with sliding blocks: one was for front vowels (Model A) and the other was for back vowels (Model B). In parts a to e of Fig. 4, the left panel shows the three-dimensional representations of the vocal-tract shape (the numbers are the lengths of sections in mm along the vocal- tract length) and the right panel is a picture of the actual model (the front plate was removed in the photographing of the model).

Front Vowels (Model A). The models shown in parts (a) and (b) of Fig. 4 are based on the bent-type model A for front vowels. A block (shown in yellow) inserted from the floor of the oral cavity controls tongue height for the front vowels /i/ and /e/. On the top surface of the block there is a groove running along the vocal-tract length. The cross sectional dimension of the groove is 9 mm x 9 mm (the location of the groove is indicated by the red dashed line). When the surface of the block reaches the roof of the oral cavity, the area of the constriction becomes minimal and it simulates the vowel /i/ (Fig. 4a). When the block shifts 6-8 mm downwards, it simulates the vowel /e/ (Fig. 4b); in this case, the downward shift is 8 mm. From Figs. 4 (a) and (b) one can observe that the tongue constriction for /i/ and /e/ is in the same position, but the area of the constriction is wider for /e/ than for /i/.

Back Vowels (Model B). The models shown in Figs. 4 (c)-(e) are based on the bent-type model B for back vowels. In Figs. 4 (c) and (d), the block shown in yellow was placed inside the pharyngeal cavity to control tongue height for the back vowels /a/ and /o/. There is a groove on the surface of the block facing the pharyngeal wall, the cross-sectional dimension of which is 9 mm x 9 mm, running along the length of the vocal tract. When the block is placed 5 mm above the bottom, as in Fig we. 4 (c), we can produce /a/. When the block is shifted up to 11 mm, one uses an additional block for lip rounding (leftmost yellow block), then one can produce the vowel /o/, as shown in Fig. 4 (d). The block for lip rounding has a square hole, of which the dimension is 18 mm x 18 mm (the location of the hole is also indicated by the red dashed line). By replacing the straight sliding block for /o/ with a right-angle shaped block and positioning it at the corner of the vocal tract as shown in Fig. 4 (e) (rightmost yellow block), one can produce the vowel /u/. In this case, there is also a groove, 9 mm x 9 mm, on the surface of the block facing the pharyngeal wall and the palate. The block for lip rounding is also used in Fig. 4 (e), just as it was for /o/ in Fig. 4 (d).

A Single Bent-type Model with Sliding Blocks. The final design was that of a single bent-type model with sliding blocks that could simulate all five vowels (Model C). The design was achieved by combining the former two bent-type models into one model. In this latter design, the following four points were taken into consideration:

1) Front vowels have a block for tongue constriction

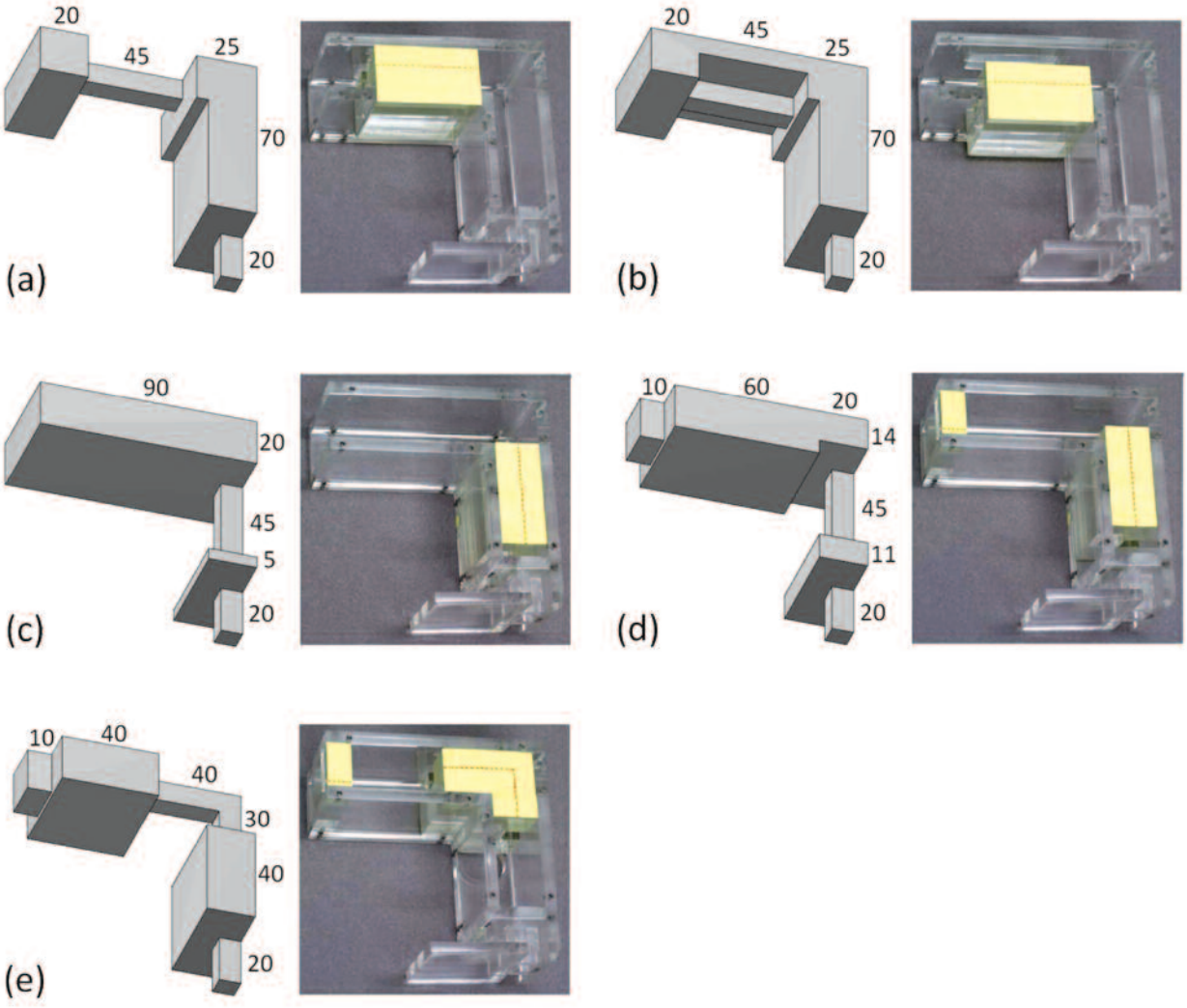


Figure 4. Bent-type modes A (a,b) and B (c-e): (a) vowel /i/, (b) vowel /e/, (c) vowel /a/, (d) vowel /o/ , and (e) /vowel/u

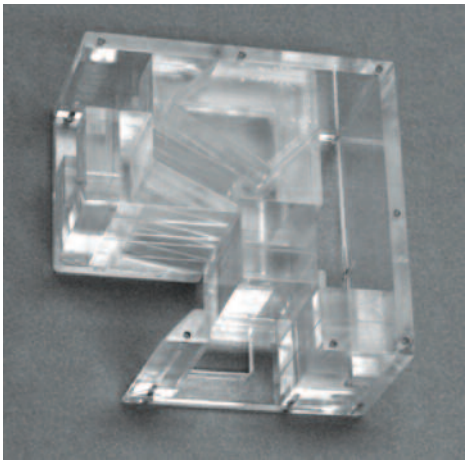


Figure 5 Single bent-type model with sliding blocks (Model C)

located in the palatal region, but back vowels do not. Vertical movement of the block takes care of this.

2) Back vowels must have a block for tongue constriction in the pharyngeal cavity, but front vowels do not. Horizontal movement of the block takes care of this.
3) Diagonal protrusion of the tongue dorsum is necessary for the vowel /u/.
4) An extra block at the mouth end is necessary for lip rounding.

Figure 5 shows a picture of this single bent-type model with sliding blocks. As shown in this figure, there are four sliding blocks. Figure 6 shows the configuration for all five vowels using this single model. (The front plate was removed for the photographs in Figs. 5 and 6.)

Measurements

Sounds were recorded using Models A, B and C for the formation of the five vowels and these sounds were analyzed by inspecting the spectrograms. The sounds produced by these models are also used for informal listening tests.

Two Bent-type Models with Sliding Blocks (Models A and B). A driver for a horn speaker was attached to the glottis end of the model. Input signals were fed into the driver unit via an audio interface and a power amplifier. There were two types of input signals. The first signal was an impulse train with an original sampling frequency of 16 kHz; later, the sampling frequency was increased to 48 kHz. Its fundamental frequency, f0, increased from 100 to 125 Hz within the first 100 ms, and then decreased to 100 Hz within the next 200 ms. The total duration of this signal was 300 ms. The second type of input signal was a swept-sine signal with a sampling frequency of 48 kHz. The length of the swept- sine signal was 65536 samples.

To avoid unwanted coupling between the neck and the area behind the neck of the driver unit and to achieve high impedance at the glottis end, a close-fitting metal cylindrical filler was inserted inside the neck. A hole in the center of the metal filling with an area of 0.13 cm2 was created.The output sounds were recorded using a microphone from the sound level meter and an audio interface with a sampling frequency of 48 kHz. The microphone was placed approximately 20 cm in front of the output end in a sound-treated room (Fig. 4). The signals recorded were synchronously averaged multiple times to gain the signal-to-noise ratio.

Figure 5 is a sound spectrogram of output signals recorded with the first type of input signal. The output signals from the five configurations were concatenated for this analysis. As shown in this figure, one can observe clear formants, especially the first and second formants (F1 and F2) in the lower frequency region. The five vowels were clearly distinguishable during an informal listening test.

A Single Bent-type Model with Sliding Blocks (Model C). A driver unit for a horn speaker was again attached to the glottis end of the model; a close-fitting metal cylindrical filler was also used. The first input signal for Models A and B was fed into the driver unit via a USB audio amplifier .

The output sounds were recorded using a microphone and a digital audio recorder with a sampling frequency of 48 kHz. The microphone was placed approximately 15 cm in front of the output end in a sound-treated room.

The vocal-tract configuration shown in Fig. 3 was used for the recordings of each vowel. For the vowel /u/, the groove of the main block (the largest one) was, unfortunately, connected to the central hole that was created when sliding the block for the tongue dorsum protrusion diagonally. Therefore, small thin plates were used to block the two connections (the short red lines in Fig. 3).

Figure 6 is a sound spectrogram of output signals concatenated for this analysis and recorded from the five configurations. As shown in this figure, we can observe a similar spectrogram to the one in Fig. 5. However, F1 is less clear in vowels /u/ and /o/. From an informal listening test, vowels /i/, /e/ and /a/ were clearly heard. The vowels /o/ and /u/ had reasonable quality but were a bit less intelligible.

Discussion and conclusions

In the activities reported in the present paper, mechanical bent-type models were designed. Two bent-type models

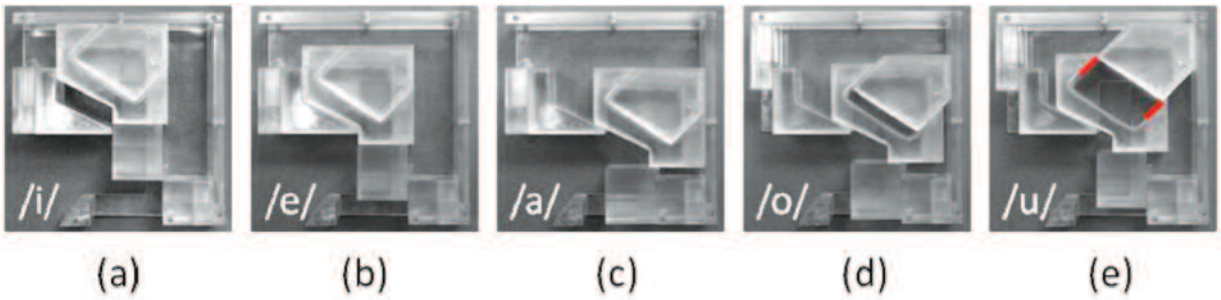


Figure 6. Single bent-type model (Model C) for five vowels

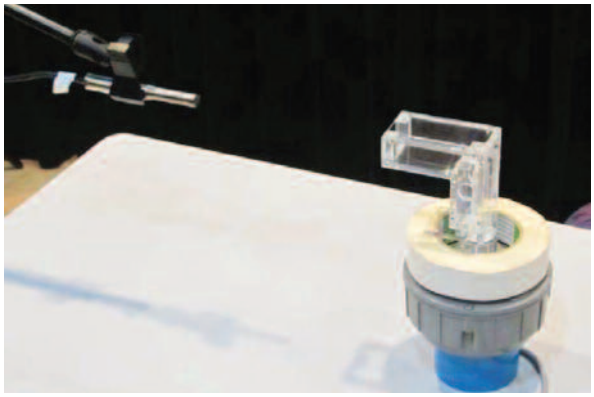


Figure 7. Experimental set-up for making recordings using bent-type models with sliding bocks (Models A and B)

with sliding blocks were tested: one for front vowels and one for back vowels. As a result, it was confirmed that the five vowels were produced clearly. A single bent-type model with sliding blocks was subsequently designed and it was confirmed that the five vowels were again produced with reasonable quality. However, it is not easy to slide the blocks in the current model, so improving the design for better usability is a future goal. Ultimately, users should be able to simultaneously manipulate for tongue height and advancement and to check the actual sounds. One should then objectively evaluate the usefulness of this new model in a pedagogical situation in acoustics, speech science, phonetics, etc.

Based on the current author's teaching experience, one needs both straight-type and bent-type models for different purposes for education in acoustics and speech science. The straight models are much simpler than the bent models, and one can demonstrate that the area as a function of length is the most crucial factor which determines the vowel quality, but not bending itself.

The CT (connected-tube) model, one of the simplest models, is a static model, but if you compare the shapes and the sounds of different types of the CT (connected-tube) model, one can demonstrate that vocal-tract configuration is associated with the quality of vowels. The S3T model, another simple model that can achieve similar shapes comparing to the different types of the CT model, is a dynamic model, so that a single S3T model can produce different vowels by changing the position and the size of the slider. Thus, this model partially simulates the tongue movement in our vocal tract.

However, the straight models have disadvantages: 1) it is not easy for learners to imagine how the vocal tract is placed in our head; and 2) the tongue movement is less realistic. The bent models, on the other hand, can cover these two points. 1) The vocal tract starts from the throat and ends at the lips, and it is very natural to understand for learners that the vocal tract is bent in between them. Furthermore, 2) more realistic tongue movement can be achieved by the bent-type models. As a result, they can produce a vowel sequence, such as /aia/, with a natural tongue movement. The flexible tongue model with a gel-type material is also a bent-type model, but it is very difficult to reproduce the same vocal-tract configuration multiple times with such model. The bent-type models with

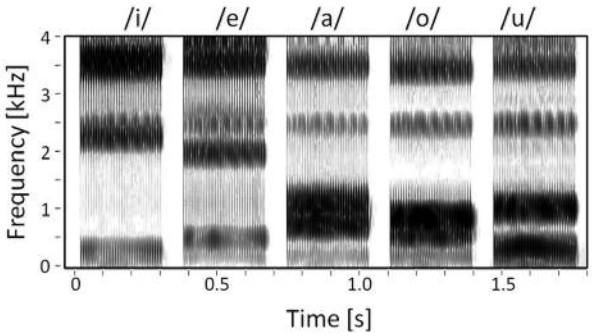


Figure 8. Spectrogram of output signals from the bent-type models with sliding blocks (Models A and B)

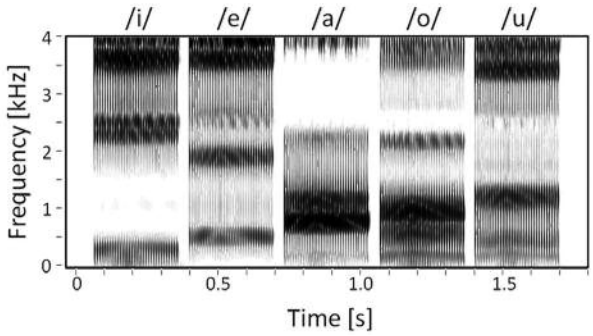


Figure 9. Spectrogram of output signals from the bent-type models with sliding blocks (Model C)

blocks as proposed in this study, however, can easily reproduce the same configuration repeatedly. In addition, the bent-type models can extend the coverage of sounds into some types of consonant, such as "glides" and "liquids." In any case, one should select what types of models that one uses for educational purposes depending on what one wants to teach. Therefore, educators will need to have several different types of vocal-tract models that they can use for different purposes

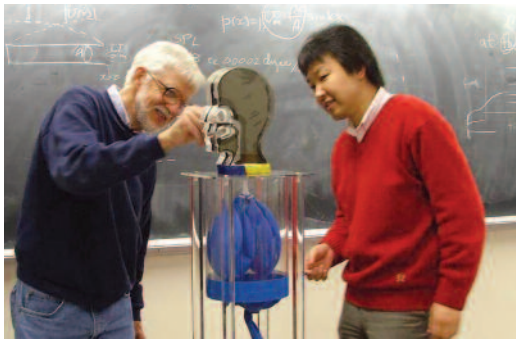
Acknowledgements

This work was partially supported by Grant-in-Aid for Scientific Research (No. 24501063) from the Japan Society for the Promotion of Science. The author would also like to thank Keiichi Yasu for his assistance.AT

References

1. H. Dudley and T. H. Tarnoczy, "The speaking machine of Wolfgand von Kempelen," J. Acoust. Soc. Am. 22, 151-166 (1950).
2. W. von Kempelen, *Mechanismus der menschlichen Sprache und Beschreibung einer sprechenden Machine*, (The mechanism of human speech and the description of a speaking machine), Wien, Austria (1791). (Reprint published in 1970.)
3. T. Arai, "Education system in acoustics of speech production using physical models of the human vocal tract," Acoust. Sci.Tech. 28, 190-201 (2007).
4. T. Arai, "Education in acoustics and speech science using vocal-tract models," J. Acoust. Soc. Am. 131, 2444-2454 (2012).
5. T. Arai, "Gel-type tongue for a physical model of the human vocal tract as an educational tool in acoustics of speech produc-

tion," Acoust. Sci. Tech. 29, 188-190 (2008).
6. T. Arai, "Physicalmodels of the human vocal tract with gel-type material," Proc. of Interspeech, 2651-2654 (2008).
7. G. Fant, *Theory of Speech Production*, Mouton, The Hague, Netherlands (1960).
8. T. Mochida, M. Honda, K. Hayashi, T. Kuwae, K. Tanahashi, K. Nishikawa, and A. Takanishi, "Control system for talking robot to replicate articulatory movement of natural speech," Proc. of Interspeech, 1533-1536 (2002).
9. T. Arai, "Sliding three-tube model as a simple educational tool for vowel production," Acoust. Sci. Tech. 27, 384-388 (2006)
10 . N. Umeda and R. Teranishi, "Phonemic feature and vocal feature: Synthesis of speech sounds, using an acoustic model of vocal tract," J. Acoust. Soc. Jpn. 22, 195-203 (1966).
11. T. Arai, "Mechanical vocal-tract models for speech dynamics," Proc. of Interspeech, 1025-1028 (2010).



Takayuki Arai is a Professor at Sophia University in Japan. The attached photograph was taken several years ago when he was a guest lecturer in Professor Kenneth Stevens's class at M.I.T.

ASA Prizes and Fellowships

The Acoustical Society of America invites applications for prizes and scholarships which it administers. Deadlines for those listed below were imminent at the time this issue went to press. For additional details, deadline dates, and application forms, please visit the ASA home page at **http://acousticalsociety.org/** or write to asa@aip.org.

The F. V. Hunt Postdoctoral Fellowship in Acoustics. The F. V. Hunt Postdoctoral Research Fellowship was established by the Society to carry out Professor Hunt's wish that his estate be used to further the science of, and education in acoustics. Fellows receive a stipend, provided jointly by the Hunt estate and a fund established by the Acoustical Society, to support their research on a topic in acoustics at an institution of their choice. One Fellowship is usually awarded each year.

Robert W. Young Award for Undergraduate Research in Acoustics. A gift to the Acoustical Society Foundation made by the family of the late Robert W. Young in his honor has been established to grant undergraduate student research awards.

Medwin Prize in Acoustical Oceanography. The Medwin Prize in Acoustical Oceanography was established in 2000 from a grant made to the Acoustical Society Foundation by Herman and Eileen Medwin to recognize a person for the effective use of sound in the discovery and understanding of physical and biological parameters and processes in the sea.

William and Christine Hartmann Prize in Auditory Neuroscience. The William and Christine Hartmann Prize in Auditory Neuroscience was established in 2011 through a generous donation by Bill and Chris Hartmann to the Acoustical Society of America to recognize and honor research that links auditory physiology with auditory perception or behavior in humans or other animals.

The deadlines for several ASA Prizes and Fellowships are rapidly approaching.

These include the F. V. Hunt Postdoctoral Research Fellowship in Acoustics (3 September); Robert W. Young Award for Undergraduate Research in Acoustics (16 September); Medwin Prize in Acoustical Oceanography (16 September); William and Christine Hartmann Prize in Auditory Neuroscience (1 October).

Applications for the Hunt Fellowship and the Young Award can be downloaded at **acousticalsociety.org/funding_resources/fellowships_scholarships** and the applications for the Hartmann and Medwin Prizes can be downloaded at **http://acousticalsociety.org/funding_resources/prizes**.

Please consider applying or submitting nominations for these funding opportunities as appropriate or encourage your students or colleagues to apply.