



Mechanical Vocal-tract Models for Speech Dynamics

Takayuki Arai

Department of Information and Communication Sciences

Sophia University, Tokyo, Japan

arai@sophia.ac.jp

Abstract

Arai has developed several physical models of the human vocal tract for education and has reported that they are intuitive and helpful for students of acoustics and speech science. We first reviewed dynamic models, including the sliding three-tube (S3T) model and the flexible-tongue model. We then developed a head-shaped model with a sliding tongue, which has the advantages of both the S3T and flexible-tongue models. We also developed a computer-controlled version of the Umeda & Teranishi model, as the original model was hard to manipulate precisely by hand. These models are useful when teaching the dynamic aspects of speech.

Index Terms: vocal-tract model, speech dynamics, speech production, education in acoustics, speech science

1. Introduction

Historically, in modern speech science, researchers have often made mechanical models of the human vocal tract. Chiba & Kajiyama (1941) made static clay models of five Japanese vowels and showed that their three-dimensional measurements were realistic by comparing artificially produced sounds with naturally produced counterparts [1]. Later, Umeda & Teranishi (1966) implemented a mechanical vocal-tract model with sliding plastic strips. The space created between the inner wall and the strips forms an arbitrary vocal tract shape [2]. With this model, Umeda & Teranishi [2] investigated phonemic and vocal features of speech.

Recently, several models have been developed [3-5]. Honda *et al.* [3] developed “high-fidelity” models based on magnetic resonance imaging (MRI) technology. With these models, highly detailed vocal tract shapes were implemented including the piriform fossa. Honda *et al.* [4] developed a series of talking robots. They [4] replicated the human speech production process from phonation through articulation with

these models. Sawada and Hashimoto [5] also developed a mechanical model of the human vocal system having vocal folds and a vocal tract with auditory feedback control.

Arai [6-12] used mechanical models of the human vocal tract for education purposes and reported that they are intuitive and helpful for students of acoustics and speech science. Figure 1 shows the various models, grouped into two categories: static and dynamic.

For the static models, we first replicated Chiba & Kajiyama’s physical models [1] of the human vocal tract [6, 7]. There were two types of models: cylinder-type and plate-type (Fig 1). The cylinder-type models are a precise reproduction of the original models based on Chiba & Kajiyama’s measurements and simplification [1]. An acrylic column was carved, so that the hole along the rotation axis forms the vocal tract shape of each of the five Japanese vowels.

The plate-type model consists of several 10-mm thick plates situated next to each other. Each plate has a hole at the center with a different diameter such that when the plates are lined up, the holes form an arbitrary vocal tract shape.

In both types of models, a vowel-like sound is emitted when a sound source is fed through the glottis end. We have used the cylinder and plate-type models for education purposes in speech science classes and found that they are extremely useful for teaching basic concepts in vowel production, such as, the relationship between vocal tract shape and vowel quality as well as source-filter theory.

The cylinder-type models were subsequently simplified to highlight those aspects of vocal tract shape which account for differences among vowels. The connected-tube models [8], shown in Fig. 1, are just as quick and effective at demonstrating vowel production as the cylinder-type models. Furthermore, the connected-tube models are so simple that their resonance frequencies are easily estimated by simple approximation.


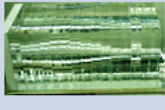


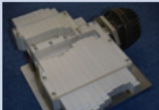


Static				Dynamic		
Cylinder-type	Plate-type	Connected-tube	Head-shaped	Umeda-Teranishi	Sliding three-tube	Gel-type tongue
						
straight	straight	straight	bent	straight	straight	bent

Figure 1: Vocal tract models developed/used by Arai [6-12]. The four models on the left are static, whereas the three on the right are dynamic. This figure also indicates whether each model is straight or bent.

10.21437/Interspeech.2010-338

The cylinder-type, plate-type and connected-tube models are all static and straight in shape, whereas the human vocal tract is both dynamic and bent at a right angle approximately in the middle of its length. Learners often struggle with where the straight models correspond anatomically in the head. To address the latter issue, we developed the head-shaped models [7] as shown in Fig. 1. With these models, we can see the cross section of each vowel visually and check vowel quality by listening to the output sounds through the models.

One drawback to the static design is that we are not able to demonstrate speech dynamics. For example, a different model is needed for each vowel when using the static head-shaped models. In contrast, the dynamic models, shown on the right in Fig. 1, allow us to change vocal tract shape.

In the present study, we first review these dynamic models. Then, we develop new versions of the dynamic model based on the head-shaped model and Umeda & Teranishi's model [2]. These models can effectively be used for teaching speech dynamics.

2. Dynamic models

2.1. Moving-tongue models

2.1.1. Sliding three-tube model

The sliding three-tube (S3T) model [9, 10] is based on the concept of Fant's three-tube model [13]. We developed the straight model in Fig. 1 as well as the curved version described in [9]. In both cases, the short slider slides inside the long tube simulating moving constriction by the tongue. This model can produce different vowels by sliding the inner slider (the 1st degree of freedom) and changing the degree of constriction (the 2nd degree of freedom). Lip rounding is optional (the 3rd degree of freedom), which is needed for certain vowels, such as /u/ and /o/. The structure is so simple, that it is an effective educational tool for teaching vowel production. Furthermore, due to the simplicity, it is an appropriate handicraft for a children's science workshop [7, 14].

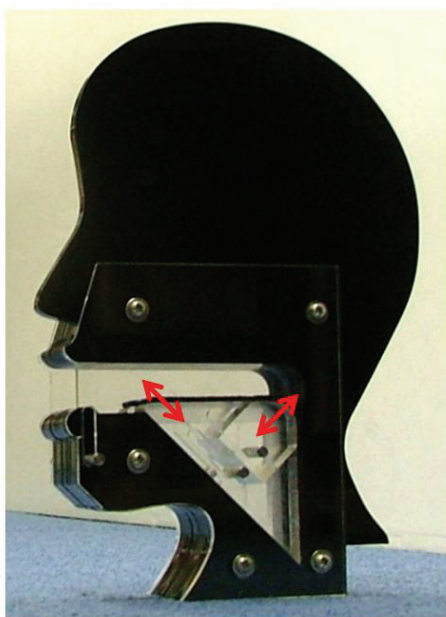


Figure 2: Head-shaped model with the sliding tongue.

2.1.2. Gel-type tongue model

A flexible-tongue model, or “gel-type tongue” model, as shown in Fig. 1, was developed [11, 12], because a more realistic vocal tract model was needed for classroom education. We designed a single bent model with a movable tongue made from a gel-type material. With this model, one may manipulate the position of the tongue and hear the change in output sound simultaneously.

2.1.3. New head-shaped model with the sliding tongue

One of the great advantages of the gel-type tongue model is the tongue's flexibility, which mimics natural tongue movement and allows us to produce many different vowels and their dynamics. However, to master the positioning of the tongue, one needs a certain amount of practice with this model. In contrast, with the S3T model, there are only three degrees of freedom, so it is simpler to produce the target vowels. The drawback of the S3T is the path the model moves in the vowel space. For example, you would have the vowel progression: /i/ -> /y/ -> /u/ -> /o/ -> /a/. In this case, the tongue constriction is fixed to its minimum area, the lip is not rounded, and only the first degree of freedom is changed. As you can see, not all regions or paths are covered in the vowel space.

Thus, we developed a new head-shaped model with a sliding tongue, which has the advantages of both the S3T and the gel-type models. As shown in Fig. 2, the vocal tract is bent in the middle at a right angle. The degrees of freedom for this model are the same as the S3T model, but the movable parts are different. The 1st degree of freedom is “tongue height.” The tongue height in this model is not real height, because the “tongue” does not slide vertically but diagonally. However, this is actually a more realistic movement of the tongue. The 2nd degree of freedom is the position of tongue constriction. This movement is especially important for the vowel /u/. The 3rd degree of freedom is lip rounding.

2.2. Umeda & Teranishi's model

2.2.1. Original model

As mentioned earlier, Umeda & Teranishi [2] developed a simple device that acoustically simulates the human vocal tract [2]. One can change the cross-sectional areas of their model by moving plastic strips closely inserted from one side. (In Fig. 1, the strips are 15 mm thick.) Various vowels and other sustained sounds can be produced by configuring their model differently. A vowel-like sound is emitted from the mouth end by inputting a glottal sound from the glottis end of the model.

2.2.2. Computer-controlled model

The original model can produce not only sustained sounds but also dynamic sounds, including diphthongs. However, it is usually very hard to precisely and rapidly manipulate the multiple strips by hand. Therefore, we have developed a computer-controlled version of this model. Figure 3 shows Umeda & Teranishi's model [2] with the control system. As shown in this figure, each one of the eleven strips is hard-wired to each actuator, and the position of each time frame is controlled by the programmable logic controller (PLC). The vocal tract shape changes in time to produce speech dynamics by sending a time-position data matrix for the actuators.

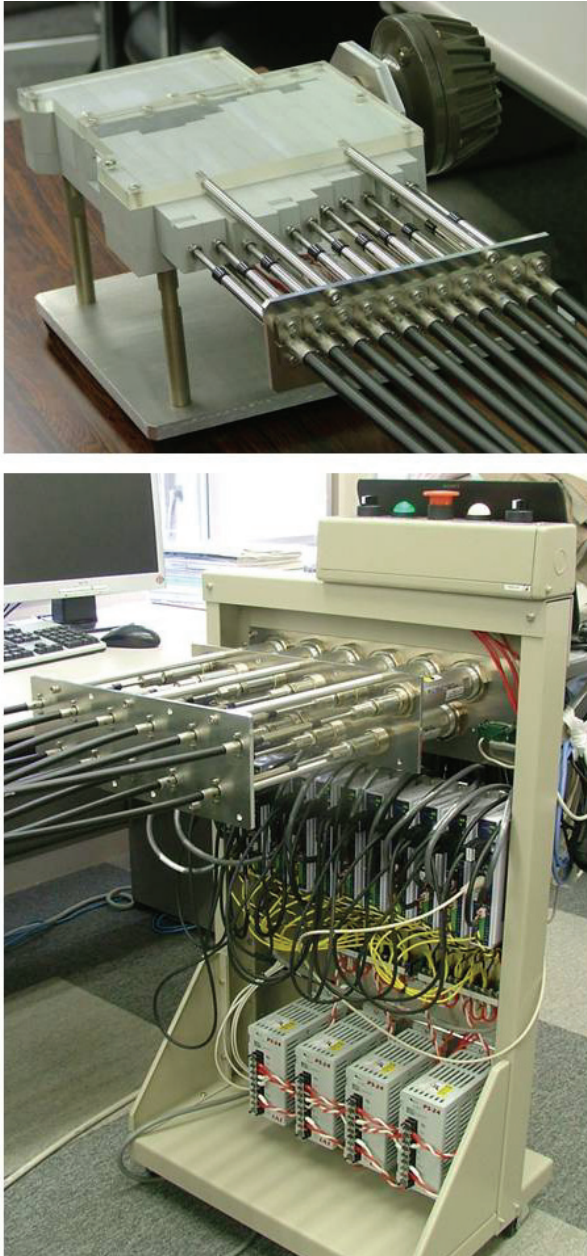


Figure 3: Computer-controlled Umeda & Teranishi model. Each one of the eleven strips is hard-wired to each actuator, and the position of each time frame is controlled by a programmable logic controller.

3. Experiments

3.1. Experiment 1

The newly proposed head-shaped model with the sliding tongue is tested in Experiment 1. Especially, the movement of the vowel sequence /aia/ is tested by observing the spectrogram.

3.1.1. Recordings

A driver unit (TOA TU-750) for a horn speaker was attached to the glottis end of the model. An impulse train with the fundamental frequency of 100 Hz was fed into the driver unit

via the digital recorder (Marantz PMD570) and a power amplifier (Marantz DA04). The sampling frequency was 16 kHz. 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, we inserted a close-fitting cylindrical filler inside the neck. A hole was drilled in the center of the filler with an area of approximately 0.3 cm². The output sounds were recorded using a recorder (Sony PCM-D1) with a sampling frequency of 48 kHz. The microphone was placed approximately 20 cm in front of the output end.

3.1.2. Results

The new head-shaped model has three degrees of freedom. We recorded an output sound and obtained its spectrographic representation (Fig. 4) by changing the main sliding part. As shown in this figure, we can observe the formant transitions moving from /a/ to /i/ and /i/ to /a/. More natural movement of the tongue is achievable, because the vocal tract is bent. With this model, one can make the natural vowel progression from /a/ to /i/ (and vice versa), without passing through /u/, as with the S3T model. With the new model, the path on the F1-F2 space is more direct. Thus, the newly proposed head-shaped model produces a more natural vowel progression than can be achieved with the S3T model.

3.2. Experiment 2

In this experiment, we produce several sequences of vowels using the computer-controlled Umeda & Teranishi model. Especially, the movement of the vowel sequence /aeiuo/ is tested by observing the spectrogram.

3.2.1. Recordings

The same settings were used for the recordings in Exp. 2 as in Exp. 1, except the driver unit; a different type (TOA TU-50) was used in Exp. 2. The vocal tract shape was changed in time as shown in Fig. 5 (these configurations were for the vowel sequence of /aei/). From /a/ through /o/, 11 cross-sectional area functions were used to make a position matrix for the eleven actuators. After each positioning, 10 ms interval was inserted.

3.2.2. Results

As shown in Fig. 6, we can observe the formant trajectories for the vowel sequence. As the speed of the movement increases, the vowel sequences of /iu/ and /uo/ sound more /ju/ and /wo/, respectively, just as in natural speech. Rapid movements of the vocal tract produce consonant-like sounds, and as a result, we can hear a whole sentence with a dynamic sequence from the model.

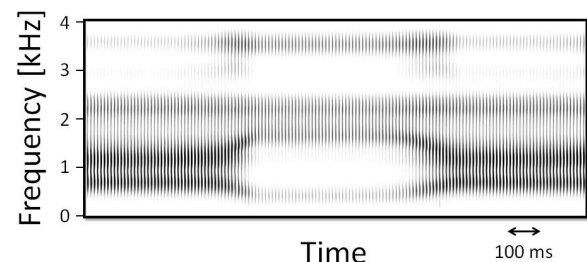


Figure 4: Spectrogram of the vowel sequence /aia/ (head-shaped model with the sliding tongue).

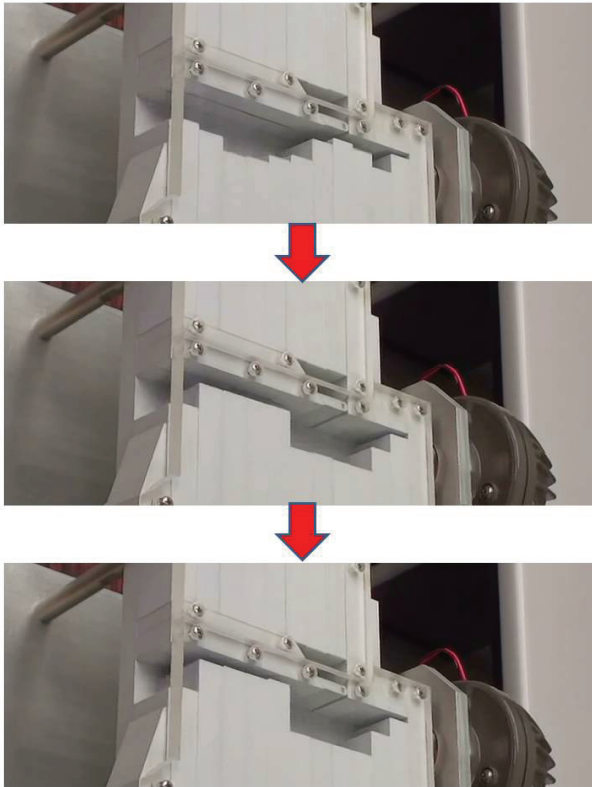


Figure 5: Vocal tract shapes of the vowel sequence /aei/ with the Umeda & Teranishi model.

4. Conclusions

In this paper, we first reviewed dynamic models of the human vocal tract, including the S3T and flexible-tongue models. We then developed a new head-shaped model with a sliding tongue and found that this model combines the advantages of both the S3T and flexible-tongue models. We further developed a computer-controlled version of the Umeda & Teranishi model [2]. We would like to evaluate their usefulness objectively in a pedagogical situation.

5. Acknowledgments

I would like to thank Noriko Umeda, Kenzo Itoh and Tasuke Takahashi for Umeda & Teranishi's model. This work was partially supported by Grant-in-Aid for Scientific Research (21500841) from the Japan Society for the Promotion of Science, and Sophia University Open Research Center from MEXT.

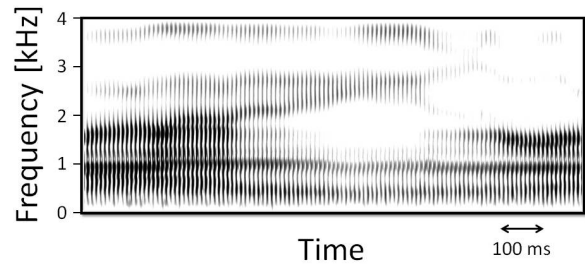


Figure 6: Spectrogram of a vowel sequence with the computer-controlled Umeda & Teranishi model.

6. References

- [1] Chiba, T. and Kajiyama, M., *The Vowel: Its Nature and Structure*, Tokyo-Kaiseikan Pub. Co., Ltd., Tokyo, 1941.
- [2] 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.
- [3] 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.
- [4] Mochida, T., Honda, M., Hayashi, K., Kuwae, T., Tanahashi, K., Nishikawa, K. and Takanishi, A., "Control system for talking robot to replicate articulatory movement of natural speech," *Proc. of Interspeech*, 1533-1536, 2002.
- [5] Sawada, H. and Hashimoto, S., "Mechanical model of human vocal system and its control with auditory feedback," *JSME International Journal, Series C*, 43(3), 645-652, 2000.
- [6] 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.
- [7] 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.
- [8] Arai, T., "Simple physical models of the vocal tract for education in speech science," *Proc. of Interspeech*, 756-759, 2009.
- [9] Arai, T., "Sliding vocal-tract model and its application for vowel production," *Proc. of Interspeech*, 72-75, 2009.
- [10] Arai, T., "Sliding three-tube model as a simple educational tool for vowel production," *Acoust. Sci. Tech.*, 27(6):384-388, 2006.
- [11] 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.
- [12] Arai, T., "Physical models of the human vocal tract with gel-type material," *Proc. of Interspeech*, 2651-2654, 2008.
- [13] Fant, G., *Theory of Speech Production*, Mouton, The Hague, Netherlands, 1960.
- [14] Arai, T., "Science workshop with sliding vocal-tract model," *Proc. of Interspeech*, 2827-2830, 2008.