Usefulness of inverse estimation using a vocal tract mapping interface for estimating articulatory behavior

Kohichi Ogata, Kento Yamamoto, and Masami Ito

Kumamoto University, 2-39-1 Kurokami, Chuo-ku, Kumamoto 860-8555, Japan
ogata@cs.kumamoto-u.ac.jp

In order to effectively describe the shape of an entire vocal tract with fewer parameters, a vocal tract mapping interface was developed [1]. In this interface, a pentagonal chart on a computer display window is used and vocal tract shapes for five vowels are located at the vertices of the chart. The purpose of the interface is to generate various vocal tract shapes corresponding to arbitrary points on the interface window using an interpolation method based on the five vocal tract shapes located on the vertices. Moreover, an attempt was made to apply the interface to the inverse estimation of vocal tract shapes from formant frequencies, or rather, acoustic to articulatory mapping, because the inverse estimation of vocal tract shapes is a problem of long-standing interest in speech production [2-5]. The usefulness of the inverse estimation based on the interface was shown for vowels and their sequences [6].

In this paper, we are interested in whether inverse estimation based on the interface can capture articulatory behavior such as differences in movement timing. As mentioned earlier, vocal tract shapes are generated based on the vocal tract shapes for the five vowels, and our method describes the overview of the vocal tract shape with fewer parameters, as in a point on the two-dimensional interface window. The behavior of the vocal tract's shape during speech can be observed as a trajectory of the estimated points on the interface window using the inverse estimation function. If the interface can provide a useful estimation of articulatory behavior, such as movement timing in spite of the vocal tract-related map based on the five vowel vocal tract shapes, it suggests the versatility of the interface in understanding speech production from speech sounds.

In this paper, the obtained vocal tract shapes were analyzed in terms of the pseudo articulatory velocity to estimate articulatory behavior. Figure 1 shows an example of inverse estimation for /ab/ transition in /aba/: inversely estimated points on the mapping interface window from formant frequencies (Left) and the vocal tract shapes corresponding to the initial vowel /a/ and before /b/ (Right). Figure 2 shows modeling of the change in the diameter of one of the acoustic tubes describing the vocal tract shape. The upper and lower sides of the tube correspond to the non-moveable hard palate and the surface of the moveable tongue, respectively, and this example illustrates opening the mouth. The vocal tract shape is parameterized by 20 cross-sectional areas and the vocal tract length in the model. The inverse estimation used here provides not articulatory movement itself, but an area function. Therefore, the change in the diameter of each acoustic tube as a function of time was treated as the pseudo velocity of articulatory movement for the sake of simplicity.

Figure 3 shows the pseudo velocity patterns of the vocal tract partial sections of the front cavity for /aba/. Of the 20 sections, data for sections 17, 18, 19, and 20 from the glottis are shown because the behavior of the front cavity is of key interest for /b/ in /aba/. Although the interface cannot inversely estimate the vocal tract shape for /b/, it can provide estimations for the transitions between a vowel and a consonant [7]. Therefore, all pseudo velocity patterns and vocal tract shapes are shown except the data for /b/ in Fig. 3. As indicated by the first arrow, Section 20 (yellow) corresponding to the lip area has an earlier change in velocity than the other sections. This suggests that the lips move before the jaw to close the mouth as quickly as possible for the /ab/ transition in /aba/. In contrast, a synchronized opening gesture can be seen in the four sections at the

/ba/ transition indicated by the second arrow. Thus, it was revealed that our inverse estimation using the mapping interface provides useful information about articulatory behavior, such as the timing of movements.

[Part of this work was supported by JSPS KAKENHI Grant Number JP17K06464.]

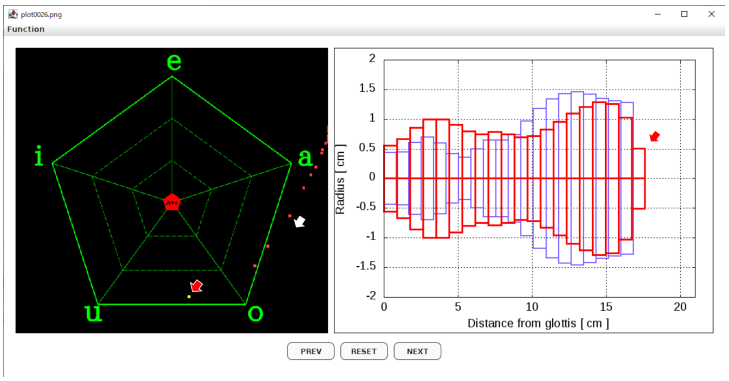

Fig. 1 Example of inverse estimation: estimated points on the window for the /ab/ transition in /aba/ and its vocal tract shapes for the initial vowel /a/ and before /b/.

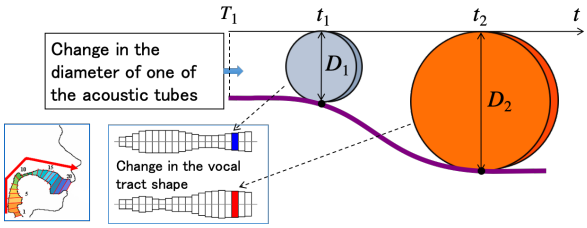

Fig. 2 Modeling of the change in the diameter of one of the acoustic tubes describing the vocal tract shape.

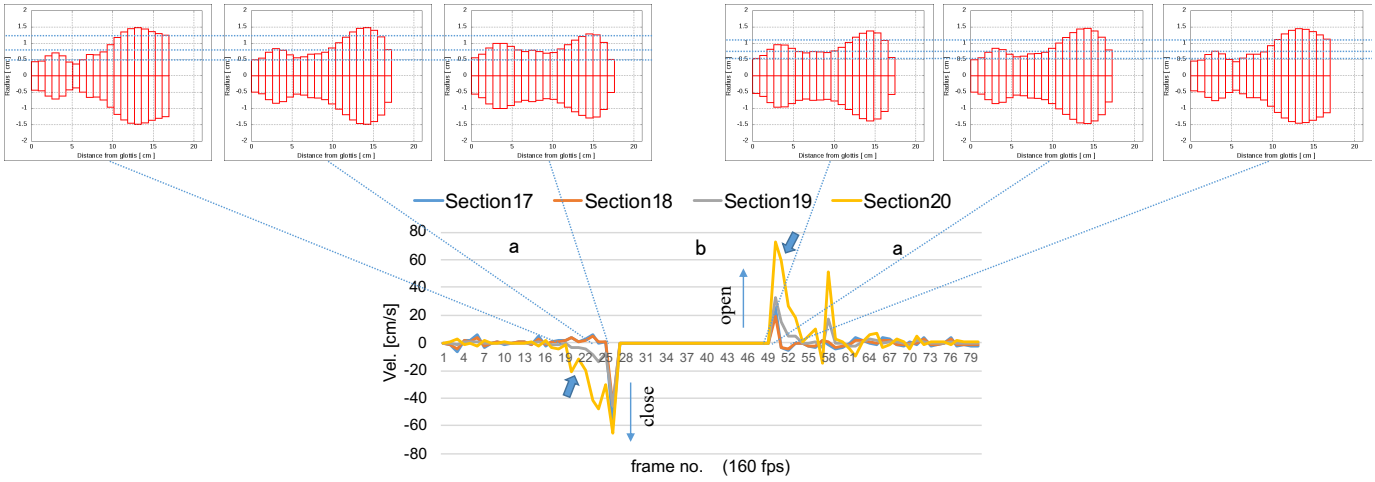

Fig. 3 Pseudo velocity patterns of the partial vocal tract corresponding to sections 17, 18, 19, and 20 (lips) for /aba/.

[1] K. Ogata and K. Yamashita, "One-click vocal tract mapping interface and its application to signal conversion," IEICE Trans. **J96-A**, 529-540, 2013 (in Japanese).

[2] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," J. Acoust. Soc. Am., **41**, 1283-1294, 1967.

[3] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," J. Acoust. Soc. Am., **41**, 1002-1010, 1967.

[4] B. H. Story, "Technique for tuning vocal tract area functions based on acoustic sensitivity functions," J. Acoust. Soc. Am., **119**, 715-718, 2006.

[5] T. Kaburagi, "A method for estimating vocal-tract shape from a target speech spectrum," Acoust. Sci. Technol., **36**, 428-437, 2015.

[6] K. Ogata, T. Kodama, T. Hayakawa, and R. Aoki, "Inverse estimation of the vocal tract shape based on a vocal tract mapping interface," J. Acoust. Soc. Am., **145**, 1961-1974, 2019.

[7] K. Ogata and T. Tanaka, "Inverse estimation of the vocal tract shape from speech sounds including consonants using a vocal tract mapping interface," Proc. 23rd ICA, Germany, 6887-6894, 2019.