PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Inverse estimation of the vocal tract shape from speech sounds including consonants using a vocal tract mapping interface

Kohichi OGATA[1]; Takayuki TANAKA[1]

[1]Kumamoto University, Japan

Abstract

This paper describes the inverse estimation of vocal tract shape from speech sounds including consonants using a vocal tract mapping interface. The interface determines vocal tract shape based on formant frequencies of speech sounds via the interface window, which maps pixel points onto assigned vocal tract shapes. Change in vocal tract shape for a sound sequence can be observed by the trajectory of pixel points on the window corresponding to the estimated vocal tract shape. This study attempted to inversely estimate vocal tract shapes from VCV speech sound sequences; estimated trajectories on the window during the VC transition were evaluated to investigate articulatory behavior. The results showed that voiced consonants had longer trajectory patterns on the window than unvoiced consonants. Additionally, differences in the following vowel (i.e., the final vowel V) affected the route of the trajectories for the VC transition. These results suggest that the interface's inverse estimation function can be a useful tool for analyzing and investigating vocal tract behavior for speech sounds including consonants.

Keywords: Speech production, Vocal tract, Inverse estimation, VCV, Coarticulation

1 INTRODUCTION

Effectively representing the vocal tract shape for speech production is useful in understanding the speech production process and for developing speech synthesis systems based on speech production mechanisms. A mapping interface for setting vocal tract–related parameters with minimal effort was thus proposed to enable effective modeling and easy control of the vocal tract shape. Moreover, a method of inverse estimation of vocal tract shape from vowel sounds was proposed using the mapping interface [1]. The interface represents an entire vocal tract shape as a point on the mapping interface window. By moving the point on the window, change in vocal tract shape can be controlled; inversely, vocal tract shape can be estimated based on a formant frequency data set. This inverse estimation function provides a new perspective for understanding articulatory behavior via estimated points on the interface window that symbolize entire vocal tract shapes.

This study attempted to estimate vocal tract shapes from speech sounds including consonants. Although the mapping interface was developed for vowels and cannot obtain vocal tract shapes for consonants, inverse estimation results for the transition from a vowel to a consonant may provide insight into articulatory behavior. Trajectory patterns on the interface window can thus reveal the characteristics of articulatory behavior depending on consonants, and the related findings could prove useful for mimicking articulatory behavior using a speech synthesis model. Therefore, this study investigated the behavior of vocal tract shapes obtained by inverse estimation during VCV utterances.

2 MAPPING INTERFACE [1]

This section provides an overview of the vocal tract mapping interface. The interface window display uses a pentagonal chart. Vocal tract shapes for the five vowels /a/, /i/, /u/, /e/, and /o/ are located at the vertices of the pentagonal chart, and the average vocal tract shape for the five shapes is located at the center (origin). The pur-

ogata@cs.kumamoto-u.ac.jp

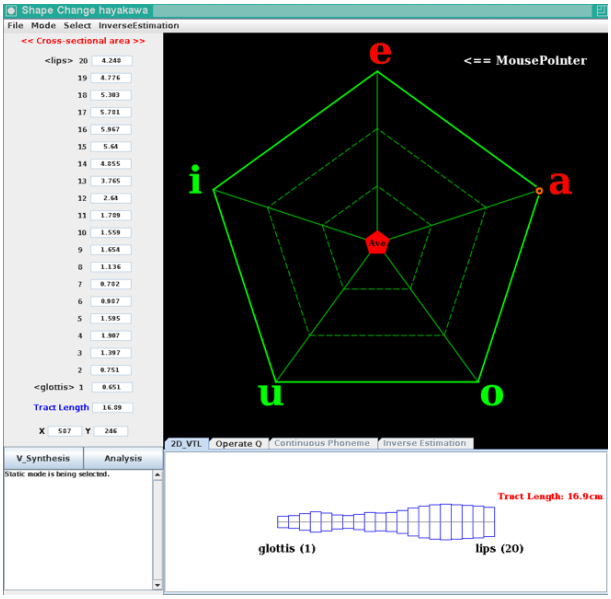


Figure 1. A screen capture of the mapping interface window.

pose of the mapping interface is to generate various vocal tract shapes corresponding to arbitrary points on the interface window using those at the vertices and the center. The mapping interface also has an inverse function (i.e., acoustic to articulatory mapping). Using this function, change in vocal tract shape can be observed by the trajectory of the estimated points on the mapping interface window. The first and second formant frequencies of vowel sounds are used as a data set to estimate the vocal tract.

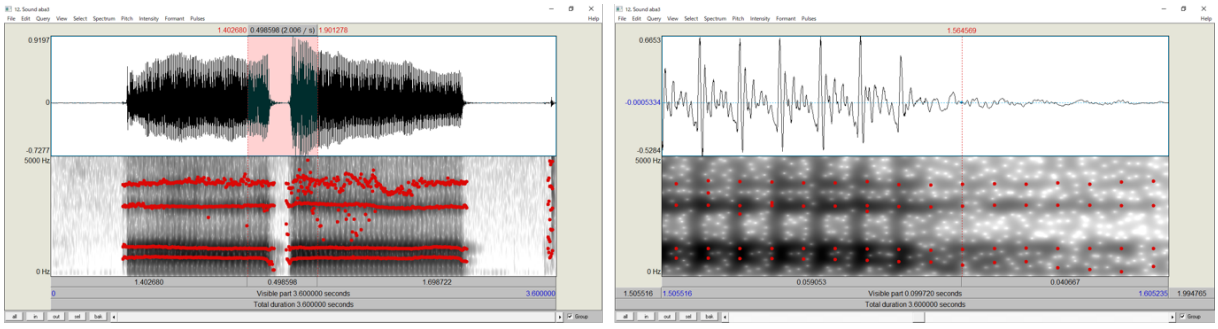
Figure 1 shows a screen capture of the mapping interface window. In this example, the point on vertex /a/ is chosen by mouse click, and the corresponding vocal tract shape, glottis (left) and lips (right), is displayed as an acoustic tube under the interface window.

3 METHOD

A digital recorder (Marantz PMD671) with a microphone (SONY ECM-77B) was used to record speech sounds. The sampling frequency of the digital recorder was 48 kHz. Speech sound files were processed using Praat [2], and the time sequence of the formant frequencies was obtained as 160 data sets per second for each utterance.

Figure 2 shows an example of speech processing using Praat; Figure 2 (a) displays a speech waveform and its sound spectrogram with extracted formant frequencies, and Figure 2 (b) shows a magnified portion of (a). This portion includes the transition of the /ab/ sequence in this example. An extracted formant frequency data set corresponding to the time where the speech waveform of /a/ was attenuated was used to obtain the end point of the trajectory pattern. In the figure, a vertical red line indicates the end point time. This study opted to analyze the VC transition because of its stable formant extraction.

Figure 3 shows an example of the trajectory patterns for the utterance /aba/. The position of the estimated point for the initial /a/ is close to vertex /a/, and it moves in the direction of vertex /u/. The color of the points gradually changes from red to white during the utterance. Vocal tract shapes during consonants cannot be obtained because the mapping interface is designed for vowels, and formant frequencies cannot be obtained during the closure for consonants. However, although this study focused on the VC transition, estimated points for the CV transition after the consonant closure are also displayed as a trajectory trend from near vertex /u/ to /a/.



(a) An example of speech processing for /aba/.

(b) Extraction of the endpoint.

Figure 2. An example of speech processing using Praat.

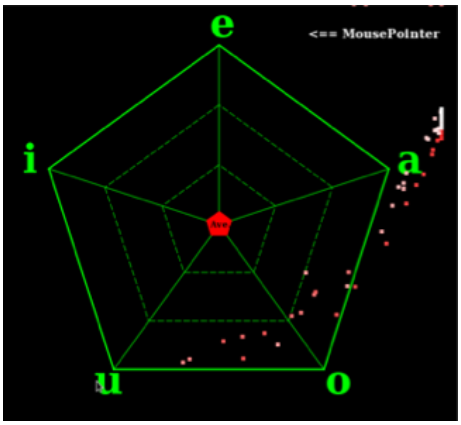


Figure 3. An example of inverse estimation. The trajectory pattern consists of points on the mapping interface window corresponding to estimated vocal tract shapes during the utterance /aba/. Vocal tract shapes are determined based on formant frequency data sets input into the system.

4 RESULTS

Figure 4 shows the trajectory patterns for /abV/ (V = a, i, u, e, and o) for subject 1. In each figure, trajectory patterns for the transition part of /ab/ for three trials are shown. Each trajectory pattern is displayed as a line by connecting the estimated points corresponding to the vocal tract shapes obtained by inverse estimation. The solid green circles show the vertices and the center (origin) of the pentagonal chart. Because the initial vowel is /a/, trajectories start their patterns around vertex /a/ for all figures, and the patterns for all three trials are similar. This similarity suggests high repeatability of the change in vocal tract shape for the same utterance. To form the lip closure for consonant articulation, the cross-sectional area of the vocal tract around the lips decreases from initial vowel /a/ to consonant /b/. Because this change is similar to vowel articulation for the /au/ transition, the estimated trajectory point appears to move toward vertex /u/.

Figure 5 shows the trajectory patterns for /apV/ (V = a, i, u, e, and o) for the same subject 1. High repeatability of the trajectory patterns for the same utterance over the three trials can also be seen for these utterances. However, the position of the end points of the trajectory patterns is different from /abV/ (V = a, i, u, e, and o) utterances. As shown, the end points for the voiced consonant /b/ in Figure 4 are closer to vertex /u/ than those of the unvoiced consonant /p/ in Figure 5.

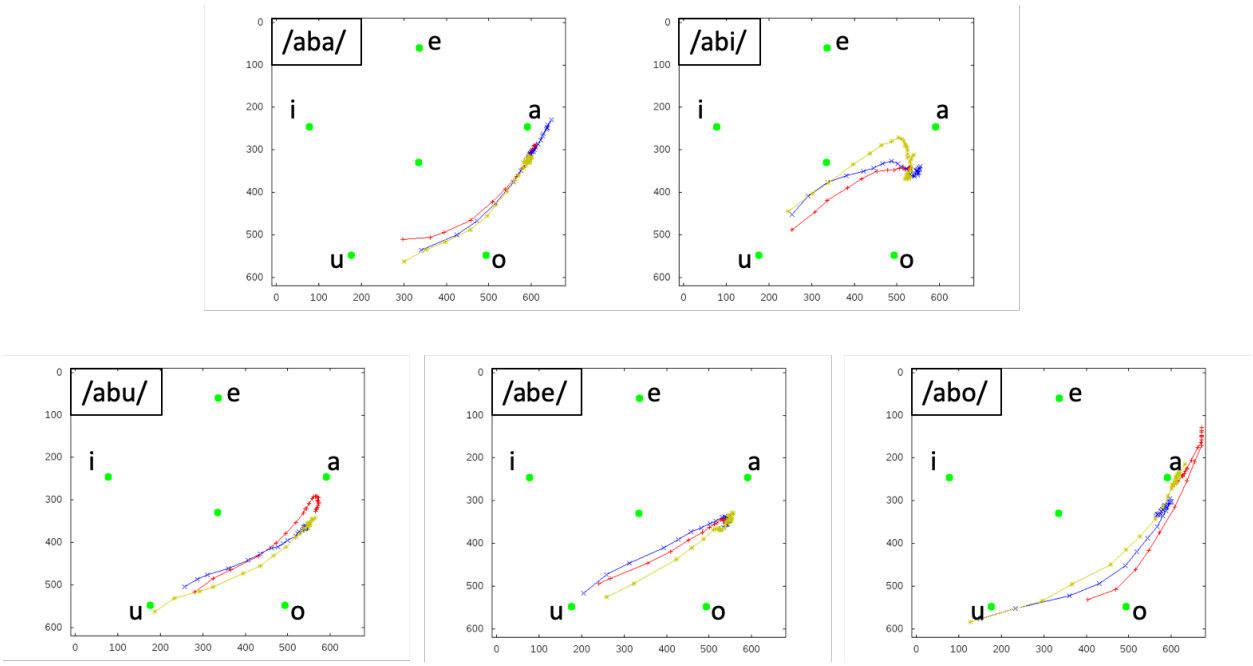


Figure 4. Trajectory patterns for /abV/ (V = a, i, u, e, and o) for subject 1. Patterns corresponding to estimated vocal tract shapes for the /ab/ transition before the consonant closure are shown by connecting the estimated points. The results for each of the three trials are indicated by different colors.

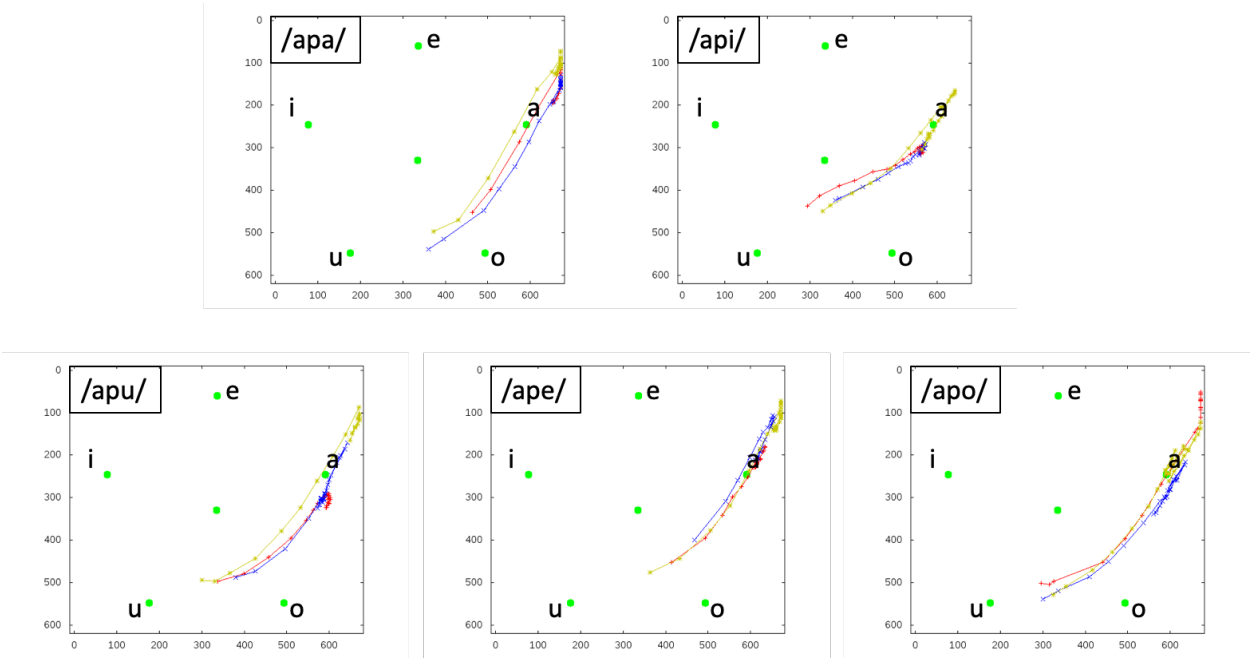


Figure 5. Trajectory patterns for /apV/ (V = a, i, u, e, and o) for subject 1. Patterns corresponding to the estimated vocal tract shapes for the /ap/ transition before the consonant closure are shown by connecting the estimated points. The results for each of the three trials are indicated by different colors.

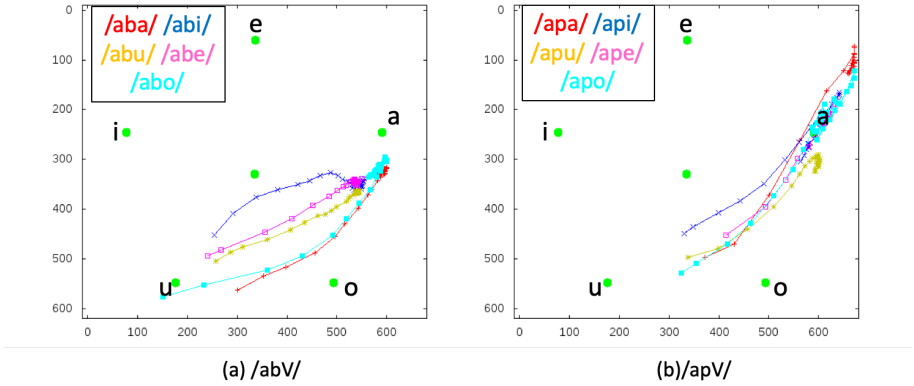


Figure 6. Typical trajectory patterns for /abV/ and /apV/ (V = a, i, u, e, and o) for subject 1. Patterns for the /ab/ and /ap/ transition before the consonant closure are shown by connecting the estimated points.

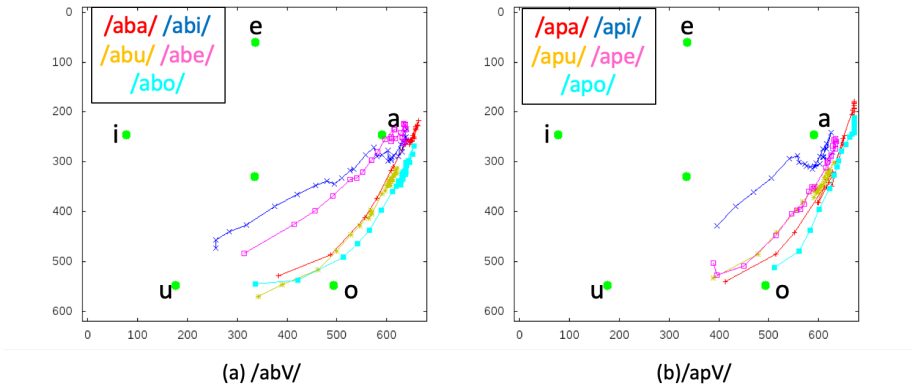


Figure 7. Typical trajectory patterns for /abV/ and /apV/ (V = a, i, u, e, and o) for subject 2. Patterns for the /ab/ and /ap/ transition before the consonant closure are shown by connecting the estimated points.

Figure 6 shows the trajectory patterns for /abV/ and /apV/ (V = a, i, u, e, and o) for subject 1. In each figure, a typical trajectory pattern was chosen from among the three trials for each utterance, and the five patterns are displayed in different colors. Comparison of (a) /abV/ and (b) /apV/ clearly shows the difference in the position of trajectory pattern end points between the voiced and unvoiced consonants. Moreover, differences in trajectory route can be seen according to the final vowel V. For instance, the pattern for /abi/ (blue) is close to the center (origin), whereas the patterns for /aba/ (red) and /abo/ (light blue) are close to vertex /o/. This dependency on the following vowel suggests preparation for articulation of the vowel. In other words, anticipatory coarticulation effects [3] can be clearly seen in the variation of the trajectory patterns.

Figure 7 shows the trajectory patterns for /abV/ and /apV/ (V = a, i, u, e, and o) for subject 2. Comparison between (a) /abV/ and (b) /apV/ also shows that the trajectory pattern end points for the voiced consonant /b/ are closer to vertex /u/ than those for the unvoiced consonant /p/.

Figures 8 and 9 show the trajectory patterns for /agV/ and /akV/ (V = a, i, u, e, and o) for the subjects 1 and 2. Comparison of (a) /agV/ and (b) /akV/ similarly shows that trajectory pattern end points for the voiced consonant /g/ are closer to vertex /u/ than those for the unvoiced consonant /k/. The trajectory patterns in each figure also show dependency on the following vowel. The sequences with /i/ as the following vowel

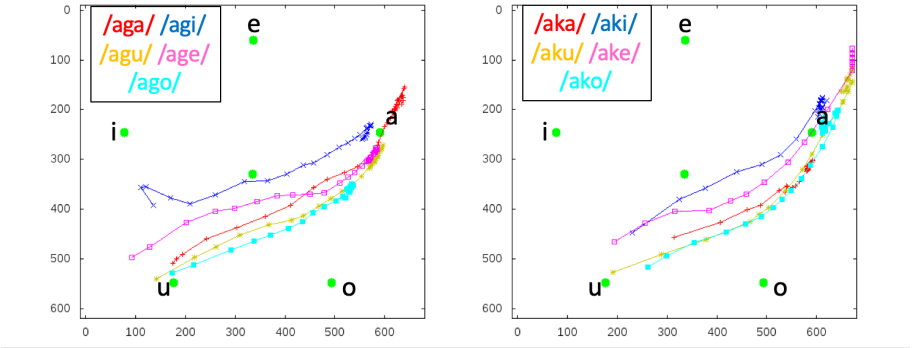


Figure 8. Typical trajectory patterns for /agV/ and /akV/ (V = a, i, u, e, and o) for subject 1. Patterns for the /ag/ and /ak/ transition before the consonant closure are shown by connecting the estimated points.

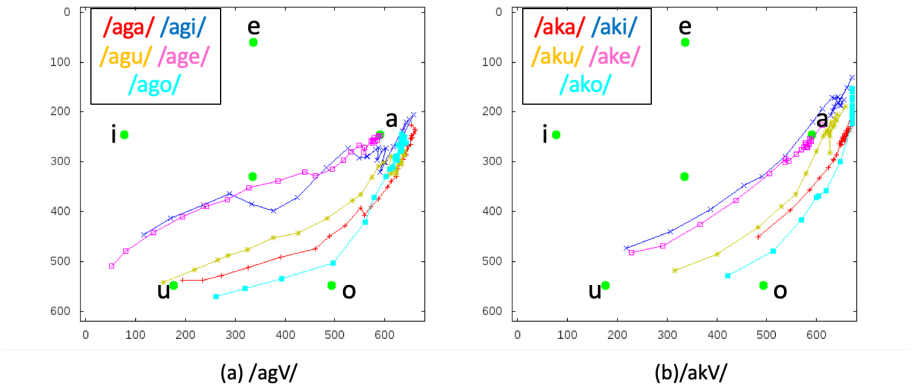


Figure 9. Typical trajectory patterns for /agV/ and /akV/ (V = a, i, u, e, and o) for subject 2. Patterns for the /ag/ and /ak/ transition before the consonant closure are shown by connecting the estimated points.

show trajectories with a route closer to the center (origin) of the chart, as compared to the sequences with other following vowels. In contrast, the trajectories of sequences with /o/ as the following vowel pass through routes close to vertex /o/. Compared to subject 2, trajectory patterns for subject 1 are more distant from vertex /o/, especially for utterances with /a/ and /o/ as following vowels. Because approaching vertex /o/ from /a/ causes a decrease in lip area [1], this tendency demonstrates a difference in the degree of lip constriction between the subjects.

Figures 10 and 11 show the trajectory patterns for /adV/ and /atV/ (V = a, i, u, e, and o) for subjects 1 and 2. Comparison of (a) /adV/ and (b) /atV/ shows that in most cases, the trajectory pattern end points for the voiced consonant /d/ are closer to vertex /u/ than those for the unvoiced consonant /t/. The distribution of trajectories on the interface window according to following vowel is wider for subject 2 than for subject 1, demonstrating a different tendency between the subjects—namely, individual differences are shown compared to the other utterances of /abV/, /apV/, /agV/, and /akV/. However, both subjects show a tendency in which trajectories for utterances with following vowel /i/ are located in the upper part of the interface window compared to utterances with following vowel /o/.

Thus, on the interface window, all consonants used in this study showed a trajectory from vertex /a/ to

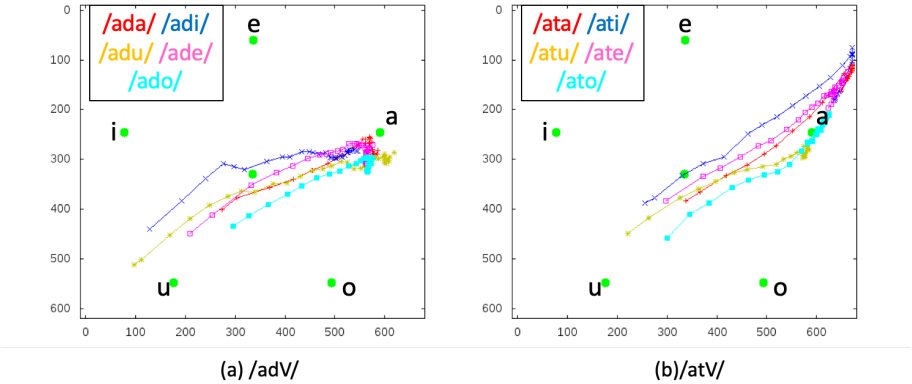


Figure 10. Typical trajectory patterns for /adV/ and /atV/ (V = a, i, u, e, and o) for subject 1. Patterns for the /ad/ and /at/ transition before the consonant closure are shown by connecting the estimated points.

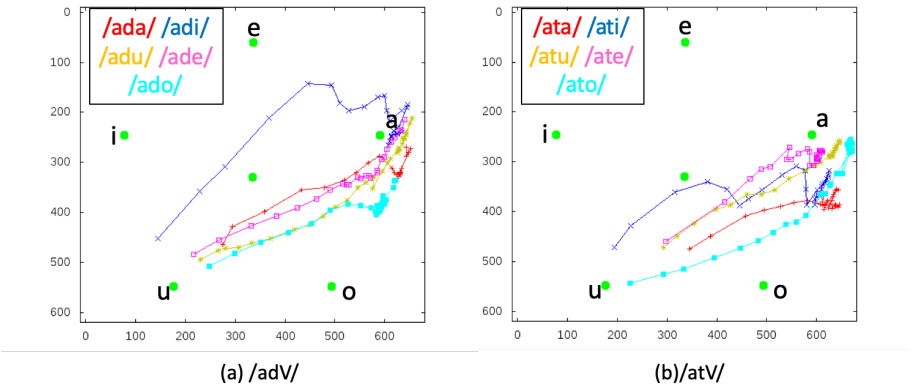


Figure 11. Typical trajectory patterns for /adV/ and /atV/ (V = a, i, u, e, and o) for subject 2. Patterns for the /ad/ and /at/ transition before the consonant closure are shown by connecting the estimated points.

vertex /u/ as a main trend, with route deviations depending on the following vowel.

Figure 12 shows examples of the vocal tract shapes for the trajectory end points for the utterances /aga/ (red) and /agi/ (blue) in Figure 9 (a). In this system, vocal tract shape is parameterized by 20 cross-sectional areas and the vocal tract length. To produce the consonant closure for /g/, the vocal tract changes its shape from the vowel /a/ to /g/. For each utterance, the figure illustrates the constriction of the vocal tract at the front cavity just before the consonant closure. As shown, the vocal tracts have unique shapes and positions of constriction depending on whether the following vowel is /a/ or /i/ (9th and 6th section from the lips for /aga/ and /agi/, respectively). This difference suggests an effect of the following vowel on articulatory movements before the intervocalic consonant closure.

5 CONCLUSIONS

This paper described inverse estimation of vocal tract shape from speech sounds including consonants using a vocal tract mapping interface. Changes in vocal tract shape during the VC transition in VCV sequences were estimated, and the corresponding trajectory patterns on the interface window were evaluated. Each utterance

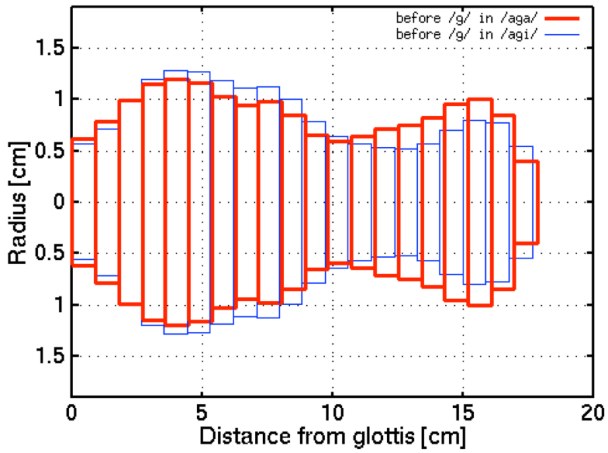


Figure 12. Vocal tract shapes for trajectory end points for the utterances /aga/ and /agi/ in Figure 9 (a). The diameter of the vocal tracts, consisting of 20 cascaded acoustic tubes, is displayed: glottis (left) and lips (right).

demonstrated high repeatability of trajectory pattern over three trials. Sequences with voiced consonants had longer trajectory patterns on the window than those with unvoiced consonants. The results also showed that differences in the following vowel (i.e., the final vowel V) affected the route of the trajectory patterns for the VC transition. These results suggest that the inverse estimation function of the mapping interface can be a useful tool to analyze and investigate vocal tract behavior for speech sounds including consonants.

ACKNOWLEDGMENTS

This work was partly supported by Grant-in-Aid for Scientific Research JP17K06464 from the Japan Society for the Promotion of Science.

REFERENCES

[1] Ogata K, Kodama T, Hayakawa T, Aoki R. Inverse estimation of the vocal tract shape based on a vocal tract mapping interface. J Acoust Soc Am. 2019;145(4):1961–1974.

[2] Boersma P, Weenink, D. Praat: Doing phonetics by computer. [accessed 29 Jun 2019] Available from: http://www.fon.hum.uva.nl/praat/

[3] Öhman S. E. G. Coarticulation in VCV utterances: Spectrographic measurements. J Acoust Soc Am. 1966;39(1):151–168.