Speech organ contour extraction using real-time MRI and machine learning method

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Organ Contour Extraction using Real-Time MRI and Machine Learning Method

Hironori Takemoto[1], *Tsubasa Goto*[1], *Yuya Hagihara*[1], *Sayaka Hamanaka*[1], *Tatsuya Kitamura*[2], *Yukiko Nota*[3], *Kikuo Maekawa*[3]

[1]Chiba Institute of Technology, Japan
[2]Konan University, Japan
[3]National Institute for Japanese Language and Linguistics, Japan

hironori.takemoto@p.chibakoudai.jp,
{s1522115qu, s1522246by, s1522251jd}@s.chibakoudai.jp,
t-kitamu@konan-u.ac.jp, ynota@atr.jp, kikuo@ninjal.ac.jp

Abstract

Real-time MRI can be used to obtain videos that describe articulatory movements during running speech. For detailed analysis based on a large number of video frames, it is necessary to extract the contours of speech organs, such as the tongue, semi-automatically. The present study attempted to extract the contours of speech organs from videos using a machine learning method. First, an expert operator manually extracted the contours from the frames of a video to build training data sets. The learning operators, or learners, then extracted the contours from each frame of the video. Finally, the errors representing the geometrical distance between the extracted contours and the ground truth, which were the contours excluded from the training data sets, were examined. The results showed that the contours extracted using machine learning were closer to the ground truth than the contours traced by other expert and non-expert operators. In addition, using the same learners, the contours were extracted from other naive videos obtained during different speech tasks of the same subject. As a result, the errors in those videos were similar to those in the video in which the learners were trained.

Index Terms: real-time MRI, machine learning, speech organs, articulatory movements

1. Introduction

Real-time MRI (rtMRI) can be used to record articulatory movements as a video during continuous speech and singing [1-4]. Unfortunately, the air-tissue boundary of the vocal tract, which is necessary for detailed analysis, is not always clearly identifiable because of issues such as noise. To overcome this problem, many intensive studies have been conducted to automatically or semi-automatically segment these tissues from the air [5-9]. A maximum temporal resolution of 100 frames per second (fps) was reached in 2015 [10], and as a large number of frames are available for use, such segmentation methods may become increasingly important.

In order to develop the rtMRI database of Japanese speech, the articulatory movements of Japanese subjects have been recorded by rtMRI since 2017 [11]. More than 50 videos were obtained for each subject, and approximately twenty bi-morae in a carrier sentence were included in each video. In the beginning, for certain analyses, the contours of speech organs such as the tongue, lips, and palates were manually traced in a limited number of frames. Each contour was represented by an open polygon and the number of points was fixed. This suggested to us that these frames and points could be used as training data sets for extracting contours. Accordingly, a machine learning library, Dlib [12], was introduced to conduct a preliminary contour extraction test. As a result, the extracted contours showed a promising level of accuracy for all the frames in the video, even in the case of small training data sets. Thus, it was expected that the machine learning method could be used to extract the same contours that a human operator could trace, if sufficient training data sets were built.

The purpose of the present study is to evaluate the accuracy of the machine learning method in extracting the contours of the speech organs from a video, for a given subject. Furthermore, the availability of the learners used in the evaluation is also examined, in the case for which they are applied to other raw video data for the same subject.

2. Materials and Methods

2.1. rtMRI videos

In this study, the subject was one male standard-Japanese speaker (63 years old). Fifty-five rtMRI videos were recorded in the midsagittal plane using a 3T MRI scanner (Siemens MAGNETOM Prisma fit 3T) that is installed in the Brain Activity Imaging Center, ATR-Promotions Inc. Note that each video had a serial number and “video N” indicates the video with serial number N.

Each video consisted of 512 frames obtained during 37 s. Thus, the frame rate was approximately 13.8 fps. Each frame size was 256x256 pixels, the pixel resolution was 1 mm, and the slice thickness was 10 mm. During each video recording, the subject uttered approximately 20 bi-morae, embedded in the carrier sentence “Kore wa __ gata. (This is __.)”. For example, video 15 included the following bi-morae: /kuha/, /kuhi/, /kuhu/, /kuhe/, /kuho/, /kusa/, /kusi/, /kusu/, /kuse/, /kuso/, /kusha/, /kushu/, /kusho/, /kuma/, /kumi/, /kumu/, /kume/, /kumo/, /keka/, and /keki/.

2.2. Training data sets

The contours of the following five speech organs were the extraction targets: the tongue (tongue), the lips and lower jaw (lips), the soft and hard palates (palates), the posterior wall from the pharynx to the trachea (p-wall), the anterior wall from the epiglottis to the trachea (a-wall). Hereafter, these five

http://dx.doi.org/10.21437/Interspeech.2019-1593

organs are referenced using the words in parentheses, as shown in Fig. 1.

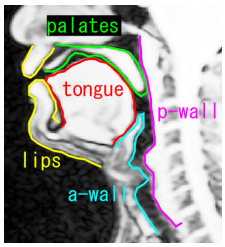

Figure 1: *The targets for extraction: contours of the five speech organs.*

To build training data sets, three expert operators, A, B, and C, manually traced the contours of the five speech organs as open polygons from the frames of video 15, which included palatal plosives, palate-alveolar fricatives, and bilabial nasals. According to a pilot study, these phonemes should be included in the training data sets. These expert operators were undergraduate students, who have learned the anatomy of the speech organs and developed a tracing software [13]. The expert operators A, B, and C traced the contours of the tongue, those of the lips and palates, and those of the p-wall and a-wall, respectively. The tracing of each part required approximately 20 minutes.

The number of points forming the contour, the number of frames to be traced, and the phonemes to which the frames correspond are listed in Table 1 for each target speech organ. The frames were selected independently for each speech organ based on the variety of its shape and the number of patterns in which it was in contact with other speech organs. Thus, the number of frames for the tongue was the largest. In addition, because each phoneme was imaged over a few frames, the phonemes in Table 1 roughly represent the phonemes to which the selected frames correspond. Furthermore, not all frames are represented by the phonemes in Table 1, because certain frames were selected between two phonemes and during quiet breathing.

Table 1: *Training data sets for the five speech organs*

Speech organ	points	frames	phonemes
tongue	40	19	/a/, /i/, /o/, /g/, /h/, /k/, /n/, /r/, /t/, …
lips	40	16	/a/, /e/, /u/, /m/, …
palates	40	17	/a/, /e/, /u/, /h/, /k/, /s/, …
p-wall	30	10	/a/, /e/, …
a-wall	30	11	/a/, /e/, /o/, /r/, …

2.3. Contour extraction by machine learning

Tools for extracting facial landmarks in the machine learning library, Dlib [12], were introduced to extract contours for the five target speech organs from each frame. These tools were originally developed to extract facial landmarks such as eyes, nose, and lips as polygons using the random forest algorithm [14]. In the present study, those facial landmarks were replaced with the five speech organs. Thus, using the training data sets described above, five learners were trained independently for each speech organ. Using these learners, the contours of the five speech organs were extracted for each frame of the video, i.e., 28,160 frames in total.

Using a PC with Intel Core i7-6500U (2.5 GHz, 4 cores) and 16 GB memory, approximately 90 seconds were required to build a learner from a training data set. Approximately 15 seconds were necessary for each learner to extract the contours from 512 frames of a video.

2.4. Accuracy of speech organ contour extraction

In order to examine the accuracy of the contours for the five targets that were extracted through machine learning, the contours that were not included in the training data sets were randomly selected as the ground truth. The following number of contours were used for each speech organ: five for the tongue, three for the lips, five for the palates, six for the p-wall, and five for the a-wall.

In the case of the frames in which the contours of the ground truth were traced, two other expert operators who were not assigned to that speech organ in building the training data sets and three non-expert operators manually traced the contours. Note that these non-expert operators were also undergraduate students, who were only gave basic instructions for this tracing procedure. For each speech organ, the error values relative to the ground truth were calculated for the contours extracted through machine learning and those traced by the expert and non-expert operators. The calculation method for the error value is described in the next paragraph. Hereafter, the mean error values of the two expert operators and the three non-expert operators are simply referred to as ex-operator and nex-operator errors, respectively.

Figure 2 shows two contours for a speech organ: one is the ground truth, and the other is the evaluation target. The evaluation targets are the machine learning, ex-operator and nex-operator results. The shortest (vertical) distance in pixels from each point of the evaluation target to the nearest segment line of the ground truth can be calculated. In the present study, a mean value of the distances ($\sum_{n=1}^{N} d_n/N$, N: number of points) is defined as the error of the evaluation target in a frame for a given speech organ. This error value represents the geometrical distance between the ground truth and the evaluation target. Furthermore, for each speech organ, a mean value was calculated among the frames and operators. This value is defined as the error of the evaluation target for a given speech organ.

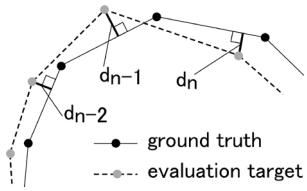

Figure 2: *Calculation of the error value for a given speech organ.*

2.5. Machine learning accuracy

In order to examine the availability of the learners, whose training data sets were obtained only from video 15, the machine learning error values for the five speech organs were calculated for videos 5, 10, 15, 20, 25, 35, 45, and 55. For each video, frames 128, 256, and 384 were used in this evaluation. Note that in video 15, frames 128, 256, and 384 were not used in building the training data sets. Thus, using a

total of 24 frames, three expert operators manually traced the contours of the same speech organs that were used for the training data set, and these contours were defined as the ground truth in this evaluation. For each speech organ in each video, a mean value for the three frames was calculated. Variations in the error values across the videos were then examined.

3. Results and Discussions

3.1. Error values for the five speech organs

Figure 3 shows the error values for machine learning, the ex-operators, and the nex-operators for the five speech organs. For all the speech organs, machine learning provided the lowest error, with an error value that was almost constant across the organs. However, the errors of the ex-operators and nex-operators increased in the following order: the tongue, lips, palates, p-wall, and a-wall. Consequently, the difference in the error between the machine learning and the nex-operators or ex-operators also increased in that order. For all speech organs other than the tongue, the machine learning error was significantly lower than that of the nex-operators. For the tongue, the contour was easy even for nex-operators to identify. In addition, for the a-wall, the error of machine learning was significantly lower than that of the ex-operators.

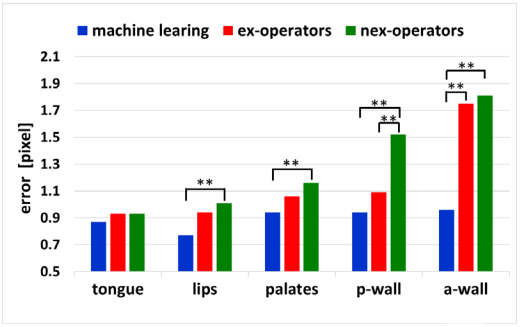

Figure 3: *Errors of machine learning, ex-operators, and nex-operators for the five target speech organs. (**p<0.01, Tukey's multiple test)*

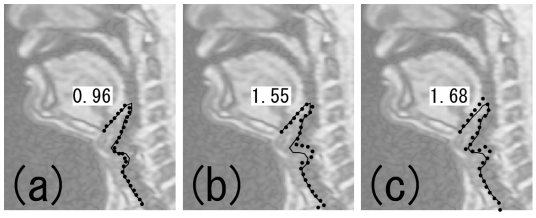

Figure 4: *Contours of the a-wall. The solid line is the ground truth and the dots indicate the a-wall contours extracted by (a) machine learning, traced by (b) an ex-operator, and (c) a nex-operator in frame 112 of* video *15. The error is indicated in each panel.*

For all the speech organs, the error of the ex-operators was lower than that of the nex-operators. According to the examination of the contours, the nex-operators tended to trace anatomically incorrect contours compared with the accuracy of the ex-operators. This could reflect their relative amounts of anatomical knowledge and tracing experience. This difference was particularly clear around the arytenoid region on the p-wall, and thus, the error of the ex-operators was significantly lower than that of the nex-operators for that region.

For the a-wall, even the ex-operators tended to incorrectly trace the contours, as shown in Fig. 4, and thus, the error was rather large. In fact, the contour of the a-wall was difficult to trace. This could be because the larynx tended to be blurred in the midsagittal image because of the partial volume effect [15]. In the present study, the slice thickness was set to 10 mm, which was large relative to the lateral dimension of the larynx. Thus, both the lateral and medial tissues were imaged together in the same frame, and two or more contours were found.

The errors of the ex-operators and nex-operators varied because of the various difficulties related to tracing each organ. However, the machine learning approach was not affected by such difficulties and so the machine learning error was almost constant. This implies that machine learning is especially effective in regions for which the contours are difficult to trace.

3.2. Machine learning error for different speech patterns

Figure 5 presents the variations in machine learning error for each speech organ across eight videos (5, 10, 15, 20, 25, 36, 45, and 55), which contained different speech. Note that the learners were derived only from video 15, and used to extract the contours for all these videos.

For each speech organ, there was almost no significant error variation between the different videos. In comparison with the errors for video 15, only those of the lips in videos 25, 35, and 55 were significantly large ($p<0.01$, t-test). This could be because the error for the lips in video 15 was the lowest (0.69) of the eight videos. These results suggest that the errors associated with the other videos were not significantly higher than those of video 15, even though the learners were trained using only video 15. Although this error analysis concerned only 3 of the 512 frames for each video, the contours were successfully extracted from the remaining frames, by visual judgement. These facts indicate that it is not necessary to build the learners for every individual video.

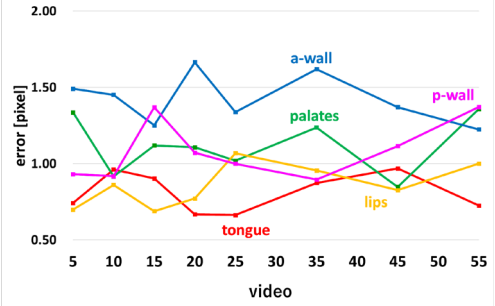

Figure 5: *Changes in machine learning error for each speech organ across the eight videos.*

Furthermore, there was no particular trend indicating an increase in error for videos that were temporally separated from video 15. For example, video 45 was taken 30 minutes and 37 seconds after video 15. This time lag included rtMRI recordings, the rest of the subject, and checking of the images. During this time period, the subject would involuntarily move the head and neck, even if the head was fixed by soft padding. Consequently, the head orientation would slightly change and thus, the learners could not extract the contours correctly. Although such changes were observed, the results indicated that they did not influence the learners.

Figure 6 shows the five contours that were extracted using machine learning (points), the ground truth (lines), and the errors (values) for frames 128, 256, and 384 of video 15. In these three frames, most points are located on or near the lines. Only around the arytenoid region of the p-wall in frame 128 (Fig. 6b), the extracted points diverged from the ground truth line. Here, contours with errors of less than 1.0 pixel provide a good agreement with the ground truth.

Two contours were sometimes overlapped because the machine learning approach involves independent extraction of contours for each speech organ. The contours of the palates and the p-wall were overlapped in this manner, as shown in Fig. 6b. This indicated that if two speech organs were in contact, it was difficult for the learner to extract the contours correctly. This tendency was common among the video frames. An important point is that this incorrect extraction was observed only in the palates, when the soft palate was in contact with the pharyngeal wall, and only in the tongue, when the tongue was in contact with the hard and soft palates. In other words, incorrect extraction occurred in one of the two organs that moved more drastically with speech. This problem could therefore be solved in post processing. One possible solution is that the contour of the two overlapping contours that corresponds to the organ that moves less drastically is fixed, and the overlapped points of the other contour are pushed back to the fixed contour.

Figure 7 presents the extracted five contours, the ground truth, and errors for three frames of video 45. Although the learners were trained using video 15, they appear to extract the contours correctly. As described above, extracted contours with an error of less than 1.0 pixel seemed to well agree with the ground truth.

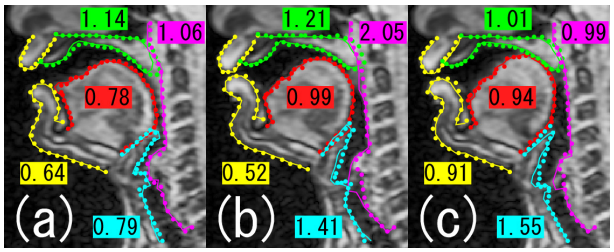

Figure 6: *Contours of the five target speech organs extracted from frames 128 (a), 256 (b), and 384 (c) in* video *15 using machine learning. The solid line indicates the ground truth. The values in each panel indicate the error (in pixels).*

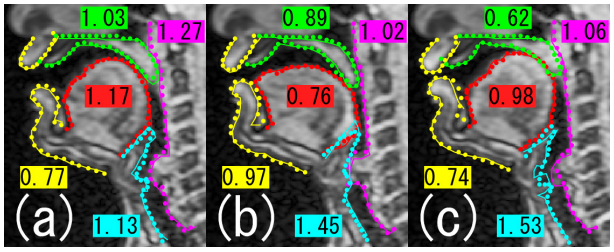

Figure 7: *Contours of the five speech organs extracted from frames 128 (a), 256 (b), and 384 (c) in* video *45 using machine learning. The solid line indicates the ground truth. The values in each panel indicate the error (in pixels).*

4. Conclusions

The present study extracted the contours of the tongue, lips, palates, p-wall, and a-wall, as open polygons from 512 frames of 55 rtMRI videos of a speaking subject using a machine learning library, Dlib [12]. The random forest learners [14] were built for each speech organ from contours traced by an expert operator in the frames of one video.

Contour-extraction accuracy was first examined only for one video using machine learning, two expert operators (the third expert operator provided the training data sets), and three non-expert operators. As a result, the machine learning provided higher accuracy than the expert and non-expert operators for all the speech organs. In other words, the machine learning best reproduced the skills of the specific expert operator who provided the training data sets. The specific expert operator would have developed better tracing skills for an assigned organ compared to the other two expert operators because they traced the contours in more than three times more frames than the others. In addition, the accuracy of machine learning was almost constant among the five speech organs, while that of the ex-operators and nex-operators varied. As discussed above, the errors of the ex-operators and nex-operators could reflect difficulties related to tracing. This implies that machine learning is effective especially in the case of organs for which difficulties associated with manual tracing lead to a wide variation in the contours produced by different operators.

The accuracy of the machine learning approach was investigated, testing whether learners trained using only one video could be applied to the other videos. The extraction accuracy was examined for eight of the fifty-five videos. As a result, there was almost no significant difference in extraction accuracy between the video that was used for training and the others. This indicates that the learners that were trained using one video can be applied to others with the same subject, even if the head orientation slightly changes during long-term rtMRI experiments. However, in the case of a subject whose head orientation changes drastically between videos, it may be better to build the training data sets using multiple videos.

It is true that manual tracing is time consuming; however, the subsequent use of machine learning methods can save time. Although the difficulties related to tracing varied from one organ to another, an expert operator required approximately twenty minutes to trace a contour in a frame. Thus, building the training data sets required approximately twenty-two hours in total. However, once the data sets were obtained, the subsequent procedure required little time: each learner was built within 90 seconds and the learner extracted the contours of the 512 frames in a video in 15 seconds. Therefore, in only 14 minutes, the learner could extract contours from all 55 videos. As these contours consist of only a small number of points, they could be used in various analyses.

The present study did not examine whether the learners could be applied to videos containing other subjects. This could be the subject of future work. This study indicates the effectiveness of the machine learning approach in mapping the contours of speech organs using rtMRI data. Further developments of this approach could allow the accurate tracking of articulatory movements during speech.

5. Acknowledgements

This work supported by KAKENHI grant (17H02339).

6. References

[1] O. Engwall, “A revisit to the application of MRI to the analysis of speech production-testing our assumptions,” Proceedings of the Sixth International Seminar on Speech Production, Sydney, 2003.

[2] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, “An approach to real-time magnetic resonance imaging for speech production,” The Journal of the Acoustical Society of America, vol. 115, no. 4, pp. 1771–1776, 2004.

[3] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y. C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC),” The Journal of the Acoustical Society of America, vol. 136, no. 3, pp. 1307–1311, 2014.

[4] M. Echternach, P. Birkholz, L. Traser, T. V. Flügge, R. Kamberger, F. Burk, M. Burdumy, and B. Richter, “Articulation and vocal tract acoustics at soprano subject's high fundamental frequencies,” The Journal of the Acoustical Society of America, vol. 137, no. 5, pp. 2586–2595, 2015.

[5] E. Bresch and S. S. Narayanan, “Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images,” IEEE Transactions on Medical Imaging, vol. 28, no. 3, pp. 323–338, 2009.

[6] Z. Raeesy, S. Rueda, J. K. Udupa, J. Coleman, “Automatic segmentation of vocal tract MR images,” Proceedings of 10th International Symposium on Biomedical Imaging, pp. 1328–1331, 2013.

[7] A. Toutios and S. S. Narayanan, “Factor analysis of vocal tract outlines derived from real-time magnetic resonance imaging data,” International Congress of Phonetic Sciences (ICPhS), 2015.

[8] S. Asadiabadi and E. Erzin, “Vocal tract airway tissue boundary tracking for rtMRI using shape and appearance priors,” Proceedings of the INTERSPEECH 2017, pp. 636–640, 2017.

[9] A. Koparkar and P. K. Ghosh, “A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5004–5008, 2018.

[10] P. W. Iltis, J. Frahm, D. Voit, A. A. Joseph, E. Schoonderwaldt, and E. Altenmüller, “High speed real-time magnetic resonance imaging of fast tongue movements in elite horn players,” Quantitative Imaging in Medicine and Surgery, vol. 5, no. 3, pp. 374–381, 2015.

[11] K. Maekawa, “A real-time MRI study of Japanese moraic nasal in utterance-final position,” in *International Congress of Phonetic Sciences (ICPhS)*, 2019, (accepted).

[12] D. E. King. “Dlib-ml: A Machine Learning Toolkit,” Journal of Machine Learning Research 10, pp. 1755–1758, 2009.

[13] T. Goto, Y. Hagihara, S. Hamanaka, H. Takemoto, T. Kitamura, and K. Maekawa, “Examination of edge detection of speech organs from real-time MRI movie by a machine learning method,” Proceedings of the 2018 Autumn Meeting of the Acoustical Society of Japan, pp. 814–815, 2018, (In Japanese).

[14] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874, 2014.

[15] M. Angel, G. Ballester, A. P. Zisserman, and M. Brady, “Estimation of the partial volume effect in MRI,” *Medical Image Analysis*, vol. 6, pp. 389–405, 2002.