一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2019-40(2019-12)

[Poster Presentation] Estimation of Three-Dimensional Tongue Shape from Midsagittal Tongue Contour using Regression Models

Tatsuya KITAMURA†, Hisanori MAKINAE††, and Masashi ITO†††

† Faculty of Intelligence and Informatics, Konan University
8–9–1 Okamoto, Higashinada-ku, Kobe-shi, Hyogo, 658-8501 Japan
†† Fourth Department of Forensic Science, National Research Institute of Police Science
6–3–1 Kashiwanoha, Kashiwa-shi, Chiba, 277–0882 Japan
††† Department of Electrical and Electronic Engineering, Tohoku Institute of Technology
35–1 Kasumi-cho, Yagiyama, Taihaku-ku, Sendai-shi, Miyagi, 982-8577 Japan

Abstract In this study, we investigated methods to estimate the tongue contours of the outer sagittal planes from a midsagittal tongue contour using linear and machine-learning-based regression models. To validate the method, tongue contours were extracted from each frame of a 3D MRI movie of the three sagittal planes, including the midsagittal plane, recorded while a male adult speaker produced the Japanese vowel sequence /aiueo/. The extracted tongue contours were then converted into harmonic amplitude profiles (HAPs) using a 2D Fourier transformation to reduce the number of dimensions. Finally, we calculated the coefficients of a multiple regression model and trained a random forest regression model to map the HAP of the midsagittal plane to those of the outer sagittal planes. In comparison with the multiple regression model, the random forest regression model demonstrated a higher estimation accuracy and exhibited an average distance error of less than 2.0 mm in a 10-fold cross-validation test.

Key words tongue contour, three-dimensional MRI data, linear regression, random forest regression, data-driven approach

1. Introduction

As magnetic resonance imaging (MRI) and machine learning technologies have advanced, it has become desirable to develop new methods of articulatory modeling to improve our understanding of the speech production mechanisms.

In the past, both geometric and biological models of the articulatory organs have been proposed [1]. In the former, the shape of the vocal tract was determined from the position of each of the articulatory organs [2] [3], while in the latter, the temporal pattern of the muscle contraction force was used [4] [5] [6] [7]. In other research, Maeda [8] proposed a model based on the statistical characteristics of the articulatory movements that was classified into an intermediate category of the geometric and biological models.

MRI technology has contributed greatly to the field of speech science over the past few decades by providing the means to measure static and dynamic features of the speech organs as well as real-time scans of its movement, which can be used in two- or three-dimensional (2D or 3D) visualizations of the inside of the body. While access to 3D articulatory data is essential for simulating and investigating acoustic phenomena in the vocal tract during speech, few studies to date have accomplished real-time 3D dynamic imaging of articulatory movements and those were conducted using a relatively low temporal resolution [9].

The frame rate of 2D MRI imaging has advanced to the point where 2D MRI movies can be imaged at over 100 fps in real-time, which is sufficient for analyzing rapid articulatory movements during the production of consonants (e.g., plosives) [10] [11] [12]. Hiroya and Kitamura [13] demonstrated even higher frame rates of 250-fps in 2D MRI movies that had been interpolated temporally using electromagnetic articulography data, although their approach required off-line processing.

If a 3D MR image could be accurately generated from a 2D MR image of, for example, the midsagittal plane via a deep-learning approach [14], this would enable a high-frame-rate 3D MR movie to be produced from a high-frame-rate 2D MR movie. However, this is not currently possible as training deep neural networks on the relationships between the 2D and 3D MR images would require enormous amounts of 3D MRI data, which is too time-consuming and costly to acquire with existing methods.

To overcome this challenge, rather than pursuing methods of generating 3D images, we developed a data-driven model that did not assume a particular internal structure of the articulatory organs or the nature of the interactions between them, as was the case in Maeda's model, and estimated the 3D articulatory shape directly from the data itself.

We then explored methods for estimating the 3D shapes of the articulatory organs from their cross-sectional (2D) contours. As the hard tissues of these organs, such as the upper and lower jaws, do not change shape during speech, their spatial configuration can be uniquely identified from their cross-sectional contours. By contrast, the soft tissues of the articulatory organs, such as the tongue and lips, deform elastically during speech and therefore require additional methods to be developed for estimating their 3D shape from their cross-sectional contours.

Such estimation problems can be classified as a regression problem in which the response of the dependent variables (i.e., the cross-sectional shape parameters of an arbitrary plane) can be estimated from other explanatory or independent variables (i.e., those of another plane). Due to their simplicity, a popular class of regression models is linear regression models, which include both single and multiple regression variants. On the other hand, machine-learning-based regression models have advanced rapidly in recent years and are now widely used in various applications. Breiman [15] proposed the random forest model as a type of ensemble machine learning method that can be applied to both regression and classification problems. This algorithm achieves high performance using a "forest," which is an aggregation of a quantity of decision trees, in which each decision tree corresponds to a weak classifier of ensemble learning methods. We surmise that machine-learning-based methods may provide better results than linear approaches when estimating the elastic deformation of soft tissues.

In this study, we employ the tongue shape during vowel production as a first step in developing methods to estimate the 3D tongue shape from its midsagittal cross-sectional contour via multiple regression and random forest regression models. In Section 2, we describe techniques for obtaining the 3D tongue data required to build the regression models. In Sections 3 and 4, we illustrate the proposed method of estimating the 3D tongue shape and highlight several encouraging results. Then, in Section 5, we suggest that the proposed methods can be seen as a new data-driven approach for articulatory modeling. Finally, the paper is concluded in Section 6.

2. Datasets

2. 1 MRI data

A 3D MRI movie of an adult male native Japanese speaker was recorded while the participant uttered the Japanese vowel sequence /aiueo/. This movie was captured using the Siemens 3.0 T MAGNETOM Verio installed at the Brain Activity Imaging Center, ATR-Promotions Inc. (Kyoto, Japan). The participant gave written informed consent for the experimental procedures, which had been approved prior to the experiment by the Konan University human subject review committee.

In the experiments, imaging data of the five sagittal planes was acquired at a frame rate of 60 fps via a multi-plane synchronized sampling method. The method was an improved version of the synchronized sampling method [16] and enabled multi-plane movies to be captured by MRI scanners. In this experiment, the participant was asked to pronounce the vowel sequence /aiueo/ aloud repeatedly while in the MRI scanner. A guide sound that was synchronized by way of external trigger pulses to control the scan timing was presented to the participant via earphones to provide an acoustic cue for his utterance. The guide sound was a 0.1-s pure tone and three 0.1-s white noise intervals, each separated by 0.4-s intervals of silence. The participant uttered /ai/ at the first white noise, /ue/ at the second one, /o/ at the third (last) one, and had a break once per ten utterances.

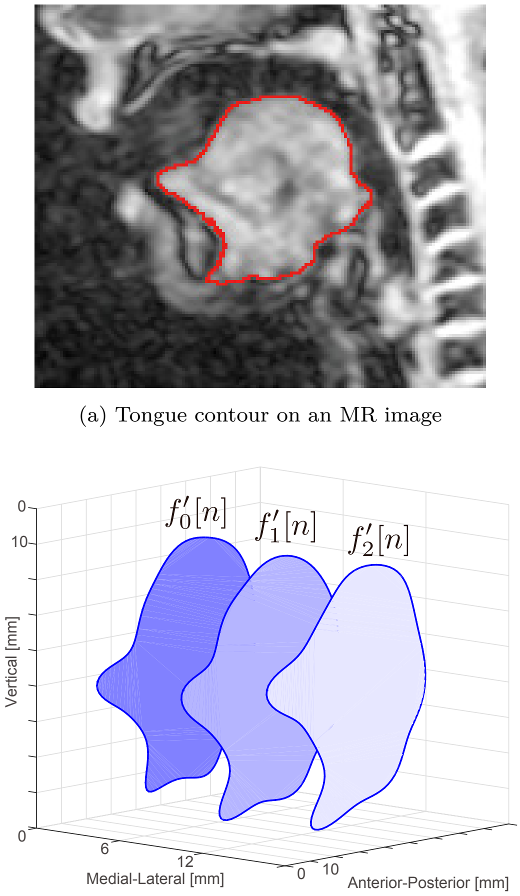
The imaging parameters used in this study are listed in Table 1. MRI data was acquired for the five sagittal planes from the midsagittal plane to the right external planes. The slice thickness and interval were 5 mm and 1 mm, respectively, and the distance between the centers of the slices was 6 mm. The MRI movie acquisition process required about 4 min during which the speaker repeated the utterances. A total of 113 frames (approximately 1.89 s) of volumetric data describing a 256 mm × 256 mm × 29 mm region were obtained, 94 of which captured articulatory movements from /a/ to /o/ that were then used in the following sections.

2. 2 Extraction and smoothing of tongue shape

Hereafter, the midsagittal plane is referred to as S_0 and the sagittal planes that were integer multiples of 6 mm away from S_0 are referred to as S_1, S_2, S_3, and S_4, respectively. The edge of the tongue was distinct and could be identified in the images of S_0 and the two external planes, S_1 and S_2, while the boundary between the tongue and the palate or the pharyngeal wall was unclear in the images of the two most external planes, S_3, and S_4. Thus, we focused on methods that of estimating the tongue contour in S_1 and S_2 from that in S_0.

Table 1 MRI scanning parameters used in the multiplane synchronized sampling method.

Parameter	Value
Echo time	1.28 ms
Repetition time	857.3 ms
Flip angle	15 degree
Field of view	256 mm × 256 mm
Pixel size	1 mm × 1 mm
Slice thickness	5.0 mm
Slice gap	1.0 mm
Number of averages	1



(a) Tongue contour on an MR image

(b) Smoothed tongue contours

Figure 1 (a) Tongue contour traced on an MR image (red line) and (b) example of smoothed tongue contours of the three sagittal planes ($f'_0[n]$, $f'_1[n]$, and $f'_2[n]$).

Initially, the boundary of the tongue was manually extracted using a pen and electronic tablet from the S_0, S_1, and S_2 images of each frame. This is depicted in Fig. 1(a). Next, the tongue contour was resampled using $N = 128$ points at even intervals and the x-y coordinate of each point on the resampled contour at an arbitrary frame was represented by a complex number $f_p[n]$,

$$f_p[n] = x_p[n] + jy_p[n] \quad (n = 1, \ldots, N), \tag{1}$$

where $x_p[n]$ and $y_p[n]$ are the nth x- and y-coordinates on the tongue contour of the sagittal plane S_p $(p = 0, 1, 2)$, respectively, and j is the imaginary unit. Here, $f_p[n]$ denotes the tongue contour plotted on the complex plane.

Let $F_p[k]$ be the Fourier coefficients of $f_p[n]$ where $\mathcal{F}$ is the Fourier transform operator.

$$F_p[k] = \mathcal{F}(f_p[n]) \quad (k = 1, \ldots, N), \tag{2}$$

where $F_p[k]$ is the harmonic amplitude profile (HAP) described by Park and Lee [17] and represents the amplitudes of the harmonic frequencies of a contour $f_p[n]$. The inverse Fourier transformation of $F_p[k]$ after replacing $F_p[k]$ $(k = m+2, \ldots, N-m)$ with zero is equal to a lowpass filtering of the contour, which produces a smoothed contour $f'_p[n]$. Here, m was set to 10. Figure 1(b) depicts smoothed ($f'_p[n]$) tongue contours of the three sagittal planes.

In the following sections, we describe methods of estimating the contours $f'_1[n]$ and $f'_2[n]$ from the midsagittal contour $f'_0[n]$. Note that the position of the tongue in the MR images was shifted due to the movement of the lower jaw; however, as we did not normalize the tongue position, our estimates of the tongue contour included the shift.

3. Tongue contour estimation via linear regression

3.1 Estimation method

We first examined a method of estimating the tongue contours of the planes S_1 and S_2 based on the tongue contour of the midsagittal plane S_0 using the HAP based on the following linear regression model:

$$\mathbf{Y} = B\mathbf{X}, \tag{3}$$

where $\mathbf{X}$, $\mathbf{Y}$, and B are respectively defined as:

$$\mathbf{X} = (1, F_0[1], \ldots, F_0[m{+}1], F_0[N{-}m{+}1], \ldots, F_o[N])', \tag{4}$$

$$\mathbf{Y} = (\hat{F}_p[1], \ldots, \hat{F}_p[m{+}1], \hat{F}_p[N{-}m{+}1], \ldots, \hat{F}_p[N])', \tag{5}$$

$$B = \begin{pmatrix} \beta_{(1,1)} & \beta_{(1,2)} & \cdots & \beta_{(1,2m+2)} \\ \beta_{(2,1)} & \beta_{(2,2)} & \cdots & \beta_{(2,2m+2)} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{(2m+1,1)} & \beta_{(2m+1,2)} & \cdots & \beta_{(2m+1,2m+2)} \end{pmatrix}, \tag{6}$$

where $\hat{F}_p[k]$ is the estimated HAP of the plane S_p $(p = 1, 2)$ obtained from $F_0[k]$ as per the linear regression model. Note that $F_0[k]$ and $\hat{F}_p[k]$ for $n = m+2, \ldots, N-m$ were excluded from vectors $\mathbf{X}$ and $\mathbf{Y}$ because these values were set to zero in order to smooth the contour. The elements of the multiple regression coefficient matrix B were calculated from $F_0[k]$ and $F_p[k]$ $(p = 1, 2)$ of the frames for training using a least squares method. As the HAP was is a complex number, the real and imaginary parts of the vector $\mathbf{Y}$ were estimated separately by calculating matrix B for each part. The estimated tongue contour $\hat{f}_p[n]$ was then obtained via the inverse Fourier transformation of $\hat{F}_p[k]$.

$$\hat{f}_p[n] = \mathcal{F}^{-1}(\hat{F}_p[k]), \tag{7}$$

$$\hat{f}_p[n] = \hat{x}_p[n] + j\hat{y}_p[n] \quad (n = 1, \dots, N). \tag{8}$$

3.2 Evaluation methods

The estimation accuracy was evaluated by closed and 10-fold cross-validation tests based on the distance error e, which is defined as the averaged distance between the ground truth $f'_p[n]$ and the estimated tongue contour $\hat{f}_p[n]$.

$$e = \frac{\sum_{n=1}^{N} |\hat{f}_p[n] - f'_p[n]|}{N}. \tag{9}$$

3.3 Results

The distance errors in the closed and cross-validation tests are shown in Fig. 2. The median values of e were 1.43 mm and 2.41 mm, respectively, in the sagittal plane S_1 and 1.02 mm and 1.40 mm, respectively, for the sagittal plane S_2. The original and estimated tongue contours in the cross-validation test when the error corresponded to the median are illustrated in Fig. 3. As shown, the original (blue) and estimated (red) closed curves in the figure are quite similar. These results suggest that the tongue contours of the sagittal planes can indeed be estimated from the tongue contour of the midsagittal plane; however, it should be noted that a few of the S_1 frames had large errors ($e > 5$ mm) in the cross-validation test.

4. Tongue shape estimation via the random forest regression model

4.1 Estimation method

We next examined a method of estimating of the tongue contours via the random forest regression model [15] as implemented in WEKA 3 [18], which was developed in at the University of Waikato, New Zealand. The HAP $F_p[k]$ was estimated using the default parameters in which the number of trees was set to 100 and the tree depth was unlimited.

4.2 Evaluation method

The estimation accuracy was evaluated using the same configuration of closed and 10-fold cross-validation tests as that in the previous section.

4.3 Results

As shown in Fig. 4, when the distance errors in the closed and cross-validation tests were evaluated, the median values of e for S_1 were 0.64 mm and 1.84 mm, respectively, and those for S_2 were 0.50 mm and 1.34 mm, respectively. Based on these results, the use of the machine-learning-based regression model improved the estimation accuracy and the magnitude of the outliers was much smaller than in the linear regression method (Fig. 2). The original and estimated tongue contours in the cross-validation test for the median error are shown in Fig. 5. These results suggest that the random forest regression model can estimate the 3D tongue shapes with adequate accuracy for vowel production.

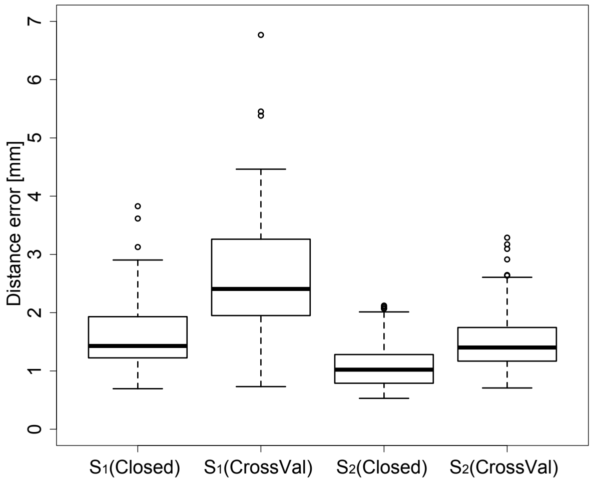
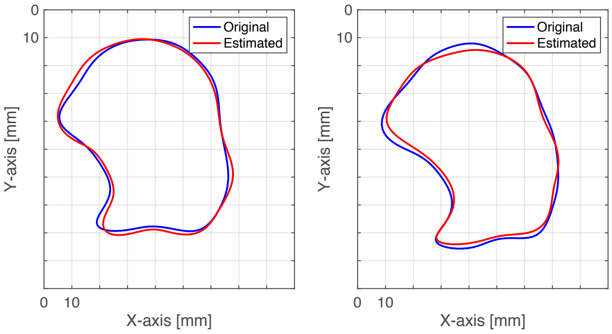


Figure 2 Estimation distance error of the sagittal planes S_1 and S_2 measured by the closed and 10-fold cross-validation tests of the linear regression model. The error was defined as the averaged distance between the original and estimated tongue contours.



(a) $S_0 \to S_1$ ($e = 2.41$) (b) $S_0 \to S_2$ ($e = 1.40$)

Figure 3 Original ($f'_p[n]$) and estimated ($\hat{f}_p[n]$) tongue contours of sagittal planes S_1 and S_2 obtained in the cross-validation test of the linear regression model. The cases shown are when the distance error corresponds to the median of the all frames.

5. Discussion

Based on the results, given the tongue contour of the midsagittal plane, the proposed methods were shown to be capable of estimating the contours of the outer sagittal planes with considerable accuracy from the 3D MRI dataset, and once the regression model was trained using a 3D MRI dataset, it was then able to estimate the 3D tongue move-

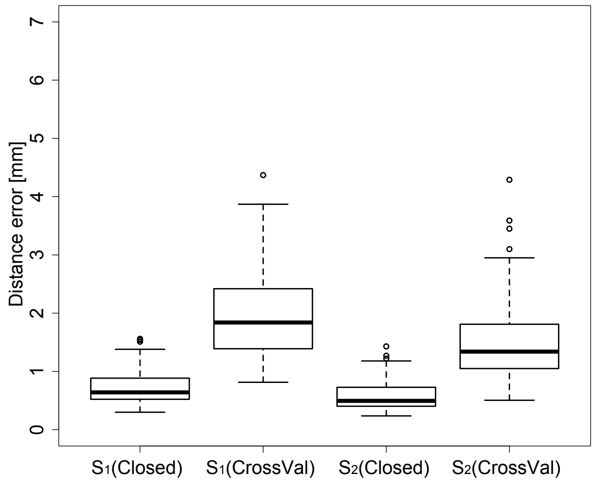
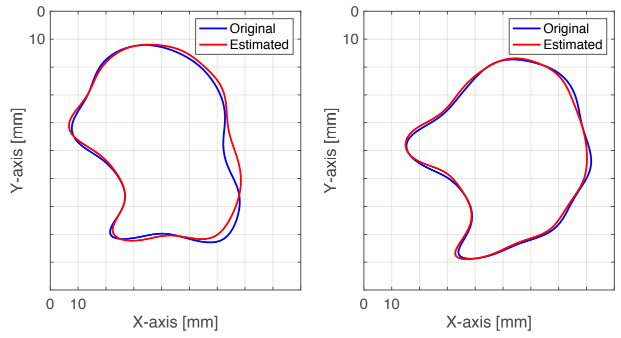


Figure 4 Estimation distance error of sagittal planes S_1 and S_2 measured by the closed and 10-fold cross-validation tests of the random forest regression model.



(a) $S_0 \rightarrow S_1$ ($e = 1.84$) (b) $S_0 \rightarrow S_2$ ($e = 1.34$)

Figure 5 Original ($f'_p[n]$) and estimated ($\hat{f}_p[n]$) tongue contours of the sagittal planes S_1 and S_2 obtained in the cross-validation test of the random forest regression model. In these cases, the distance error corresponds to the median of the all frames.

ment from 2D MRI movies. When tested, the random forest regression model exhibited higher performance than the linear regression model and the median of the average distance error was less than 2.0 mm in the 10-fold cross-validation test for the random forest regression model, which indicates that the model could estimate each point of the tongue contour with an average uncertainty of 2.0 mm. In conclusion, our results indicate that the proposed data-driven approach holds promise for modeling the shape and movement of the articulatory organs.

The performance of the regression models was highly dependent on the training data, which implied that trained models could only be applied to 2D MRI data obtained from the same speaker. Additionally, as the deformation pattern of the 3D tongue shape is dependent on the phoneme environment, it is expected that the regression model would not perform well for 2D tongue contours obtained while speaking vowels in a different order (e.g., /oeuia/). With these limitations in mind, an area of future work is to improve the estimation accuracy for other vowel sequences by developing methods to overcome the data dependency of the models.

6. Conclusions

In this study, we proposed methods to estimate the 3D tongue shape from its cross-sectional shape using multiple and random forest regression models trained on a 3D MRI dataset. Our results show that the original and estimated tongue contours were a close match and we also found that the random forest regression model exhibited better performance than the linear regression model.

The present study highlighted the potential for estimating 3D tongue shapes from its cross-sectional shape. This suggests it should be possible to estimate the 3D shape of the articulatory organs, including the lips, hard and soft palates, and pharyngeal wall from their midsagittal contours. Thus, the proposed method is promising as a new data-driven approach to articulatory modeling.

Acknowledgements

This study was partly supported by JSPS KAKENHI (Nos. 25280066, 17H02339, and 17K02769).

References

[1] K. Honda, "Models of speech production mechanisms," *The Japanese Journal of Behaviormetrics*, 22(1), 11–21, 1995.

[2] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, 53(4), 1070–1080, 1972.

[3] P. Birkholz, "Articulatory synthesis of singing," *Proc. of INTERSPEECH 2007*, 4001–4004, 2007.

[4] J. S. Perkell, "A physiologically-oriented model of the tongue activity during speech production," Ph. D. Thesis, MIT, 1974.

[5] R. Whilhelms-Tracario, "Physiological modeling of speech production: Methods for modeling soft-tissue articulators," *J. Acoust. Soc. Am.*, 97(5), 3085–3098, 1995.

[6] J. Dang and K. Honda, "A physiological articulatory model for simulating speech production process," *Acoust. Sci. & Tech.*, 22(6), 415–425, 2001.

[7] S. Fels, F. Vogt, K. van den Doel, J. Lloyd, I. Stavness, and E. Vatikiotis-Bateson, "ArtiSynth: A biomechanical simulation platform for the vocal tract and upper airway," *Tech. Rep. Computer Science Dept., University of British Columbia*, 22(6), 415–425, 2001.

[8] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Speech Modeling*, W. J. Hardcastle and A. Marchal, Eds. Kluwer Academic Publishers, 131–149, 1995.

[9] Y. Lim, Y. Zhu, S. G. Lingala, D. Byrd, S. Narayanan, and K. S. Nayak, "3D dynamic MRI of the vocal tract during natural speech," *Mag. Reson. Med.*, 81(3), 1511–1520, 2018.

[10] M. Fu, B. Zhao, C. Carignan, R. K. Shosted, J. L.

Perry, Z. P. Liang, and B. P. Sutton, “High-resolution dynamic speech imaging with joint low-rank and sparsity constraints,” *Magn. Reson. Med.*, 73(5), 1820–1832, 2015.

[11] P. W. Iltis, J. Frahm, D. Voit, A. A. Joseph, E. Schoonderwaldt, and E. Alternmüller, “High speed real-time magnetic resonance imaging of fast tongue movements in elite horn players,” *Quant. Imaging Med. Surg.*, 5(3), 374–381, 2015.

[12] K. Nunthayanon, E. Honda, K. Shimazaki, H. Ohmori, M. S. Inoue-Arai, T. Kurabayashi, and T. Ono, “Use of an advanced 3-T MRI movie to investigate articulation,” *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.*, 119(6), 684–694, 2015.

[13] S. Hiroya and T. Kitamura, “Generation of a vocal-tract MRI movie based on sparse sampling,” *Proc. of International Seminar of Speech Production 2011*, 2011.

[14] Y. LeCun,Y. Bengio, and G Hinton, “Deep learning,” *Nature*, 521, 436–444, 2015.

[15] L. Breiman, “Random Forests,” *Machine Learning*, 45(1), 5–32, 2001.

[16] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura, and N. Ninomiya, “MRI-based speech production study using a synchronized sampling method,” *J. Acoust. Soc. Jpn.* (*E*), 20(5), 375–379, 1999.

[17] K. S. Park and N. S. Lee, “A Three-Dimensional Fourier Descriptor for Human Body Representation/Reconstruction from Serial Cross Sections,” *Comput. Biomed. Res.*, 20, 125–140, 1987.

[18] E. Frank, M. A. Hall, and I. H. Witten, The WEKA Workbench. Online Appendix for *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016.