



Combined cine- and tagged-MRI for tracking landmarks on the tongue surface

Honghao Bao¹, Wenhuan Lu¹, Kiyoshi Honda^{2,*}, Jianguo Wei¹, Qiang Fang³, Jianwu Dang^{2,4}

¹ School of Computer Software, Tianjin University

² School of Computer Science and Technology, Tianjin University

³ Chinese Academy of Social Science

⁴ School of Information Science, JAIST, Japan

bao3009218057@126.com, Jianguo@tju.edu.cn, khonda@sannet.ne.jp

Abstract

Magnetic resonance imaging (MRI) techniques have been a promising way in recent speech production studies, and dynamic magnetic imaging with repetitive or real-time MRI scans has been widely used to acquire motions of all the articulators and measure their deformation during speech. While MRI is capable to visualize the entire surfaces of those organs, it lacks landmarks for motion tracking that are available with other techniques such as magnetic sensing methods. One possible solution to have both surface contours and landmarks of the articulators is to combine different imaging techniques. In this paper, we propose a new method to add surface markers on the dynamic MRI of the tongue by combining cine- and tagged-MRI data together. To do so, analysis was done on the images from the two types of scans conducted in the same session. The intersection points of the tag lines with the tongue surface contour were extracted from tagged-MRI data and they were mapped onto cine-MRI data. After minimizing the minute mapping errors, the result showed that marker tracking on both oral and pharyngeal surfaces of the tongue was successful.

Index Terms: MRI data standardization, surface, landmark, cine-MRI, tagged-MRI, registration

1. Introduction

Speech is the most preferred way to communicate with each other. To convey a message, various linguistic sounds are produced by controlling the configuration of the vocal tract. The articulators determine the resonance characteristics of the vocal tract during speech production. While articulatory data provide a stream of information that underlies speech signals, visualization of the movement of the articulators is technically difficult.

Since most of speech articulators lie in the human body, techniques that have been used to measure articulatory movement need customization as seen in previous work using x-ray microbeam system (X-ray) [1], electromagnetic articulography (EMA) [2] and ultrasonography (USG) [3]. These techniques are capable of capturing articulatory information at high sampling rates, even though they are often invasive. None of these modalities, however, offers the complete view of all the vocal-tract articulators at a sufficiently high spatial resolution, and articulation information is limited on the anterior half of the vocal tract.

Recently, development of MRI techniques has allowed for examinations of the entire vocal tract during speech production and provides a powerful means for quantifying the configuration of the articulators, including morphological

characteristics of speakers in conjunction with their articulation and acoustics. Among those, it is common to find single-modality studies [2, 4, 5], rather than multi-modality ones [6, 7]. The analysis of place of articulation using motion images has been a technical issue in experimental phonetics because effective methods for motion analysis are not available in the data. Partly because of this problem, many experimental studies on speech articulation have employed techniques to track markers that are placed on the articulators' surface, as seen in the studies with the X-ray microbeam system or magnetic sensing system. However, those techniques only measure the oral surface of the tongue among the hidden organs, and information from the pharyngeal surface is left unmeasurable. In contrast, MRI motion imaging excels at imaging the entire shape of the tongue surface, while it lacks the functionality for marker tracking. As a result, image analysis of the tongue surface has often been limited to the classical method to track the highest point of the tongue.

This paper proposes a solution to realize tongue-surface motion analysis using the combined technique of synchronized cine- and tagged-MRI. The cine-MRI [8] is good at visualizing each component of the system during speech. The tagged-MRI [7] is one of the motion imaging techniques to track tissue deformation by visualizing tag lines marked on the soft tissue. The tagged motion images can also provide surface marker points by detecting intersection points of the tag lines with the tongue surface outline. Then, those marker points can be mapped onto the cine-MRI data frame-by-frame to obtain motion images of the articulators with marker points on the tongue.

This paper is organized as follows. Section 2 introduces the related material and methods. Section 3 describes the results obtained with some remarks. The conclusion is in Section 4.

2. Materials and methods

2.1. Subjects and stimuli

MRI datasets from two subjects (23-year-old male, and 24-year-old female) were used in this study. Both subjects are native Mandarin speakers and reported no history of speech or language disorders.

Each speaker took the supine posture in the MR scanner, and the speaker's head is padded with foam rubber to minimize head movement. The participants repeated the utterances of two-syllable Chinese words, such as /midu/ or /mune/, during data acquisition.

2.2. Data acquisition

Both cine-MRI and tagged-MRI datasets were obtained during the same scan session using the Siemens Verio 3T installed at

*Corresponding author

the ATR Brain Activity Imaging Center (ATR-BAIC), Japan. The slice thickness of cine-MRI data were 3 mm. Each original image was converted into 8bit gray scaled with 729×729 resolution in the sagittal plane (each pixel size is 0.352778×0.352778 mm). Both the cine- and tagged-MRI data contained 17 frames with a playback frame rate of 10 fps.

2.3. Data standardization

Since the subjects take different head orientations in the MR scanner, data standardization is essential for the purpose of comparative data analysis across subjects [9]. In this study, the standardization procedure is the horizontal alignment of the palatal plane defined by the line from the anterior nasal spine (ANS) to the posterior nasal spine (PNS). The steps of the procedure include: 1) image expansion to 400% with lanczos3 interpolation; 2) measurement of the inclination of the palatal plane to the image's horizontal plane; and 3) rotation of the image to nullify the inclination angle of the palatal plane. Then, the image is further processed by cropping of ROI, resizing to 50%, and brightness correction (using the gamma = 2). Figure 1 shows the data before and after the standardization.

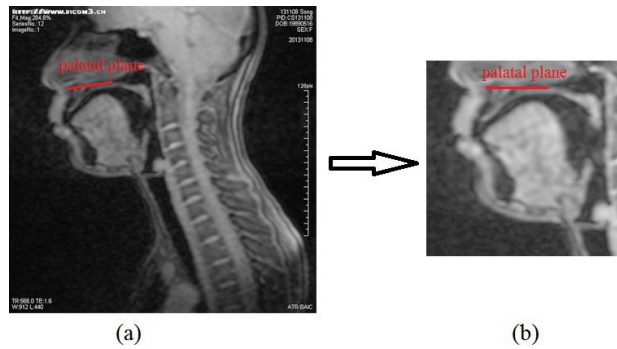


Figure 1: Images before and after standardization: (a) raw data; (b) standardized data.

2.4. Combining cine-MRI and tagged-MRI data

The purpose of the image processing is to synthesize a new set of images that shows marker points on the oral and pharyngeal surface of the tongue on cine-MRI data. To do so, the intersections of the tag lines with the surface contour of the tongue are manually marked by small circles as marker points on each tagged-MRI image. Then, the marker points are mapped onto the corresponding cine-MRI images. By repeating this procedure for all the frames, a file of motion images with the markers on the tongue surface is obtained.

2.5. Error correction procedure

Since those marker points were sampled manually, they include random errors of sampling. To minimize the sampling errors, the procedure used is simple temporal smoothing of the marker points among three consecutive frames. For this purpose, the coordinate values of the sampled marker points were extracted, and they were re-calculated based on three-point averaging using the formulas (1) and (2).

$$x_{i,j}' = (x_{i,j-1} + x_{i,j} + x_{i,j+1})/3, j=2,3,\dots,16 \quad (1)$$

$$y_{i,j}' = (y_{i,j-1} + y_{i,j} + y_{i,j+1})/3, j=2,3,\dots,16 \quad (2)$$

where x_{ij} and y_{ij} represent the horizontal and vertical coordinate of the i -th point in j -th frame.

2.6. Shortest distance calculation

To evaluate the reliability of the smoothed coordinates of the marker points, the shortest distance between each point's centroid and the tongue surface contour was calculated for all the markers of the motion image frames. The tongue contour line was digitally represented by the method using Fuzzy C-means (FCM) algorithm followed by Prewitt operator as described in Appendix. Then, the shortest distance between each point and the digital contour line was calculated to estimate the error of marker point mapping.

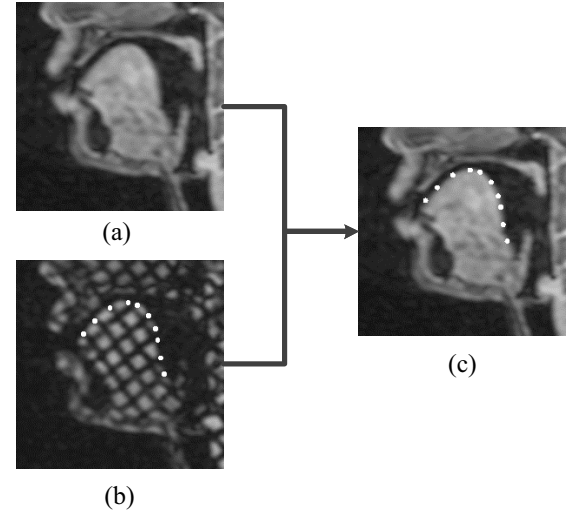


Figure 2: Outline of image processing procedure: (a) cine-MRI data; (b) tagged-MRI data with marker points; (c) result of combined image.

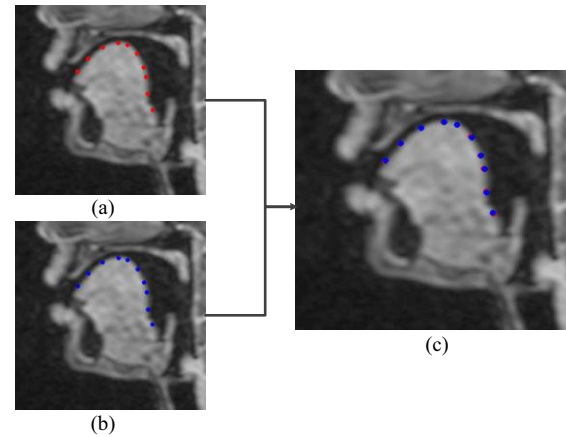


Figure 3: Images before and after smoothing: (a) original (manually sampled) result; (b) corrected result by smoothing; (c) overlay of original and corrected markers.

3. Results and discussions

The results obtained from the combined method are shown below for the utterance /midu/ obtained from the female subject. The same tendency was observed for the male speaker, which is not given in this paper.

3.1. Trajectories of the marker points

Figure 4 shows the marker positions mapped onto the four frames corresponding to each utterance segment in /midu/ with

the trajectories of all the marker data overlaid onto the corresponding frame images. This figure shows that each point follows tongue movement, suggesting that the marker tracking is successful judged by visual examination. The displacements of each marker points are mostly in the horizontal plane in this utterance, contrasting the gross tongue movement from /i/ to /u/.

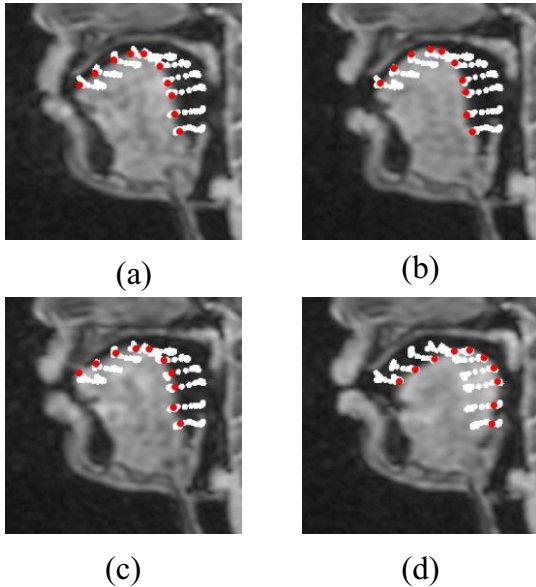


Figure 4: Positions of the marker points and cine-MRI data for the four frames corresponding to the segments: (a) /m/, (b) /i/, (c) /d/, and (d) /m/. All the trajectories of all the marker points are overlaid on each frame with the red markers indicating the corresponding marker positions for each frame.

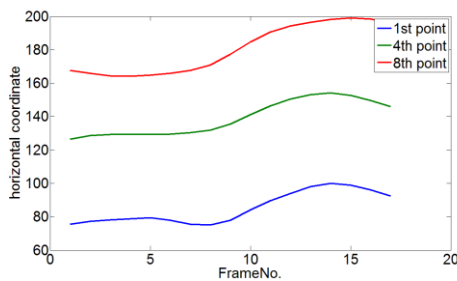


Figure 5: Horizontal movement of the 1st, 4th, 8th, point through the 17 frames.

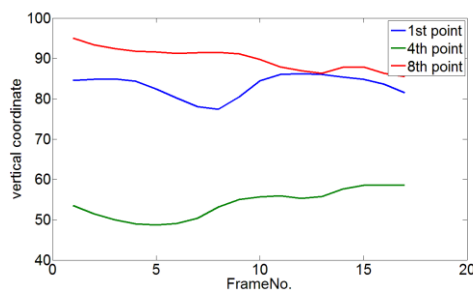


Figure 6: Vertical movement of the 1st, 4th, 8th, point through the 17 frames.

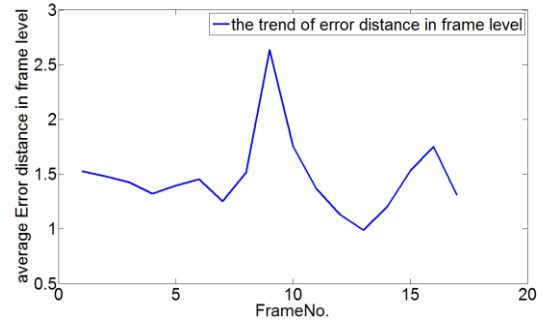


Figure 7: Time course of the frame-by-frame average distance errors between the smoothed marker points and the tongue contour line.

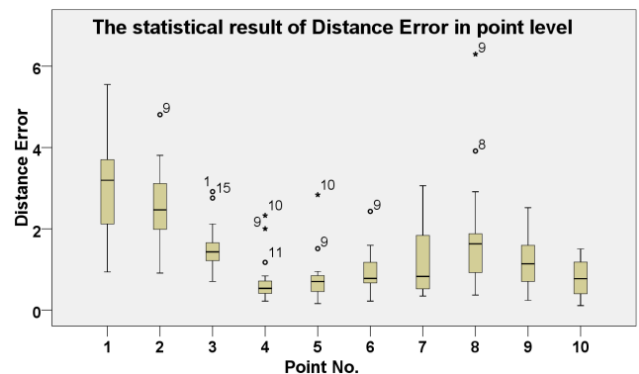


Figure 8: Box plots showing the distribution of the error distances for each marker point. The numerals for the outer points indicate the frame number.

3.2. Time functions of the selected marker points

Figure 5 and Figure 6 show the time functions of the selected three markers for all the frames during the utterance /midu/. Three markers were chosen from those on the tongue tip, tongue dorsum, and tongue base. Horizontal movements of the three markers are gradual from front positions to back positions as seen in Figure 5. The rapid movements of the markers are observed in the middle at about the frame number 10. In contrast, vertical movements of those points show certain irregularities during the time course as seen in Figure 6, and the most prominent is seen for the dorsum marker (4th point in the figure) at about the frame numbers 7 and 8. The deflection in the vertical displacement corresponds to consonant articulation for the word-medial /d/ because apical tongue tip articulation for stop consonants is commonly accompanied by the lowering of the tongue dorsum [10].

3.3. Result from evaluation of the distance errors

The result from the evaluation of the distance errors between the corrected marker points and the tongue surface contour is shown in the following two figures. Figure 7 shows the time course of the distance errors averaged over all the frames. From this figure, the largest errors are noticed at the 9th frame, which corresponds to the frame of maximum horizontal velocity of tongue movement as inferred from Figure 5. Figure 8 is the box plots for the distribution of distance errors for each marker point for all the frames. In this figure, the markers near the tongue tip (marker 1 and 2) have the larger distance

errors, and the markers near the tongue dorsum (marker 4 and 5) have the smallest errors. Those patters reflect the difference in velocity among the tongue regions: the tongue tip moves fast, while the tongue dorsum travels slow.

4. Conclusions

This study is a preliminary attempt to synthesize new articulatory data by combining two sets of MRI motion imaging data obtained from the techniques, cine-MRI and tagged MRI. The specific aim of the study is to supply numerical information for quantitative analysis of movements of the whole tongue surface. The limitation of the current method is the same as that for the tagged-MRI with synchronized imaging: The duration of the utterance to record is limited to the maximum duration of imaging using the tagged-MRI (about 1 sec). Despite the limitation, the synthesized motion images allow applications of motion analysis techniques to describe the characteristics of tongue articulation in speech. Numerical tracking of the markers on the pharyngeal surface of the tongue has been the issue for experimental studies of speech articulation, and the present study gives a solution to the inaccessibility to that region in the past. Several problems remain to further advance the technique, which include blurring of the tag lines at fast tongue movement, deformed tag-line patterns at large deformation of the tongue, and difficulty of automatic detection of the marker points. These problems are the topics for future studies.

5. Acknowledgements

This work was supported in part by the National Basic Research Program of China (No. 2013CB329305), and in part by grants from the National Natural Science Foundation of China (No. 61175016, No. 61304250 and Key Program No. 61233009).

6. Appendix

Region segmentation by edge extraction in the vocal tract is one of the tasks to describe the shape of articulators in the analysis of motion imaging [11]. In this study, the Fuzzy C-means (FCM) algorithm was used to segment the cine-MRI data based on the histogram of the gray level image. Then the Prewitt operator was employed to detect the edge. Then, the digital tongue surface was obtained.

Actually, the FCM algorithm is a variant version of the K-means algorithm. It is also used to cluster one dataset into several clusters. The divided groups are typically defined by a $c \times n$ -dimensional matrix named U . The main difference between the FCM algorithm and K-means algorithm is the range of the values from the matrix U . The elements are probabilities from 0 to 1 (they must satisfy that the sum of each column's element is equal to 1). Then the cost function of the FCM is:

$$J = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 \quad (3)$$

where $m \in [1, \infty)$ is a weighted parameter.

The procedure to determine the matrix U and cluster centroids c_i are:

The main steps for the FCM are listed below:

- 1) Use numbers from 0 to 1 to initialize the matrix U ;
- 2) Use the formula (5) to calculate the centroids of the clusters;
- 3) Calculate the cost function using the formula (3). When it is smaller than a given threshold or its changing value smaller than a given threshold, the algorithm terminates;
- 4) Use the formula (4) to update the matrix U , and return to step 2).

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (4)$$

$$c_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \quad (5)$$

After the segmentation, the Prewitt operator was used to detect the edge of the segmented cine-MRI data. To highlight the tongue surface, the regions apart from the tongue were eliminated. The results of each step in this section are shown in Figure 9.

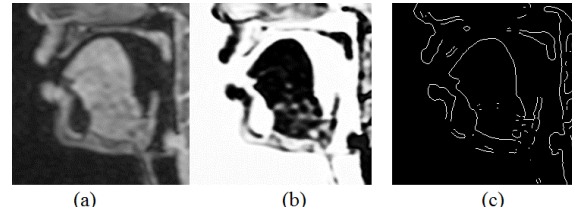


Figure 9: Procedure for edge extraction. (a) the pre-processed cine-MRI data; (b) the segmented image; (c) the data after edge detection;

7. References

- [1] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, pp. S56-S56, 1990.
- [2] A. A. Wrench, "A Multi-Channel/Multi-Speaker Articulatory Database for Continuous Speech Recognition Research," *Phonus.*, vol. 5, pp. 1-13, 2000.
- [3] D. H. Whalen, K. Iskarous, M. K. Tiede, D. J. Ostry, H. Lehnert-LeHouillier, E. Vatikiotis-Bateson, *et al.*, "The Haskins optically corrected ultrasound system (HOCUS)," *Journal of Speech Language and Hearing Research*, vol. 48, pp. 543-553, Jun 2005.
- [4] H. Takemoto, P. Mokhtari, and T. Kitamura, "Comparison of vocal tract transfer functions calculated using one-dimensional and three-dimensional acoustic simulation methods," in *15th Annual Conference of the International Speech Communication Association: Celebrating the Diversity of Spoken Languages, INTERSPEECH 2014, September 14, 2014 - September 18, 2014*, Singapore, Singapore, 2014, pp. 408-412.
- [5] X. Zhou, J. Woo, M. Stone, J. L. Prince, and C. Y. Espy-Wilson, "Improved vocal tract reconstruction and modeling using an image super-resolution technique," *Journal of the Acoustical Society of America*, vol. 133, pp. EL439-EL445, 2013.

- [6] K. Richmond and S. Renals, "Ultrax: An animated midsagittal vocal tract display for speech therapy," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012, September 9, 2012 - September 13, 2012*, Portland, OR, United states, 2012, pp. 74-77.
- [7] J. Lee, J. Woo, F. Xing, E. Z. Murano, M. Stone, and J. L. Prince, "Semi-automatic segmentation for 3D motion analysis of the tongue with dynamic MRI," *Computerized Medical Imaging and Graphics*, vol. 38, pp. 714-724, 2014.
- [8] H. Takemoto, K. Honda, S. Masaki, Y. Shimada, and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," *Journal of the Acoustical Society of America*, vol. 119, pp. 1037-49, 02/ 2006.
- [9] N. Kumar and S. S. Narayanan, "Hull detection based on largest empty sector angle with application to analysis of realtime MR images," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4-9 May 2014*, Piscataway, NJ, USA, 2014, pp. 6617-21.
- [10] W. S. Dodd, "J. C. CATFORD, A Practical Introduction to Phonetics (2nd edn.). Oxford: Oxford University Press, 2001. Pp xiii + 229. ISBN 0-19-924635-1," *Journal of the International Phonetic Association*, vol. 33, pp. 87-88, 2003.
- [11] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, pp. 323-38, 03/ 2009.