2013 IEEE 10th International Symposium on Biomedical Imaging:
From Nano to Macro
San Francisco, CA, USA, April 7-11, 2013

MRI ANALYSIS OF 3D NORMAL AND POST-GLOSSECTOMY TONGUE MOTION IN SPEECH

Fangxu Xing[a], Emi Z. Murano[b], Junghoon Lee[a,c], Jonghye Woo[a,d], Maureen Stone[d], Jerry L. Prince[a]

[a]Dept. Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, US 21218
[b]Dept. Otolaryngology–Head and Neck Surgery, JHU School of Medicine, Baltimore, MD, US 21205
[c]Dept. Radiation Oncology & Molecular Radiation Sciences, JHU School of Medicine, Baltimore, MD, US 21205
[d]Dept. Neural and Pain Sciences, University of Maryland School of Dentistry, Baltimore, MD, US 21201

ABSTRACT

Measuring the internal muscular motion and deformation of the tongue during natural human speech is of high interest to head and neck surgeons and speech language pathologists. A pipeline for calculating 3D tongue motion from dynamic cine and tagged Magnetic Resonance (MR) images during speech has been developed. This paper presents the result of a complete analysis of eleven subjects' (seven normal controls and four glossectomy patients) global tongue motion during speech obtained through MR imaging and processed through the tongue motion analysis pipeline. The data is regularized into the same framework for comparison. A generalized two-step principal component analysis is used to show the major difference between patients' and controls' tongue motions. A test is performed to demonstrate the ability of this process to distinguish patient data from control data and to show the potential power of quantitative analysis that the tongue motion pipeline can achieve.

Index Terms— Tongue, motion, glossectomy, MRI, tagged, HARP, IDEA algorithm, PCA

1. INTRODUCTION

The study of tongue muscle motion after surgical resection for cancer treatment or sleep apnea tongue reduction (glossectomy) is an important topic, because the tongue function may be seriously impeded for these patients. Studies of tongue motion differences in normal controls and post-glossectomy patients can be used to provide guidelines for surgery protocols. A pipeline of algorithms has been developed to extract and track the motion of internal tongue tissue points in 3D through a sequence of time frames during speech [1-2]. Principal component analysis (PCA) can be used to distinguish and elucidate motion patterns in the two subject groups. In the experiments carried out for this work, we analyze magnetic resonance (MR) images taken at 26 frames per second. The speech task is "a souk," a tightly controlled task that uses a forward tongue motion from /a/ to /s/ and an upward motion from /s/ to /k/.

The human tongue is a highly deformable object with the ability of performing fast and precise movements. The tongue is volume preserving and has a complex orthogonal muscle architecture due to its muscular hydrostat properties [3]. The measurement of tongue motion, although difficult, can be achieved by tagged MR imaging, which places magnetic "tags" in tissues that move and deform together with the tongue and record all motion information [4]. A sequence of cine MR images acquired at the same positions and same time frames during additional repetitions of the speaking cycle can provide high quality tongue edges for segmenting the tongue region. These data are processed using the tongue motion analysis pipeline [2], which uses the harmonic phase (HARP) algorithm to extract the 2D in-plane motion [5], the random walker algorithm to segment cine MR images to provide appropriate 2D tongue masks [6], a topology-preserving geometric deformable model (TGDM) to shrink-wrap a 3D tongue volume mask [7] and finally an incompressible deformation estimation algorithm (IDEA) to interpolate a dense 3D motion field for each time frame [8]. Such a field as the output of the tongue motion analysis pipeline possesses a few desired properties: (1) The motion field is 3D and its voxel resolution is optimized to the in-plane pixel resolution in all three directions. (2) The motion field preserves incompressibility which, as mentioned above, is one of the most important physical properties of the tongue. As a result, it can be considered as a proper representation of the real tongue motion during speech (Figure 1).

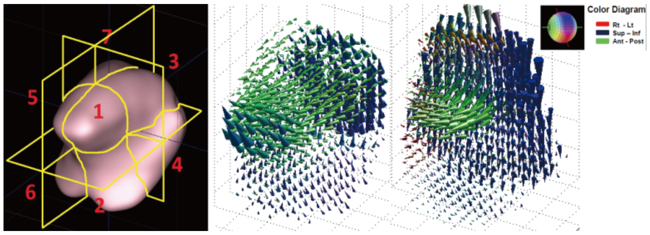

Figure 1. (a) 3D tongue mask and its division of eight VOIs. (b) 3D displacement field of a control at maximum /s/. The colormap follows conventional DTI scheme. (c) 3D displacement field of a patient at maximum /s/.

978-1-4673-6455-3/13/$31.00 ©2013 IEEE

Although we are able to obtain and process each subject's speech motion data, it is still challenging to compare them quantitatively. The first major obstacle is the inconsistent speaking rates across subjects, and the default computation of all displacements relative to the first time frame. Displacement fields relative to maximum /a/ need to be computed for each subject. And maximum /s/ and /k/ as two critical frames need to happen at the same time frames for each subject. Achieving this goal involves the inversion of discrete 3D vector fields and is not as obvious as one would think intuitively (flipping the sign). The second obstacle is that, although the patients have impeded speech function, they still manage to finish the task of speaking "a souk". Thus the PCA, which extracts the general motion pattern and major variance, has difficulty revealing the patients' unique, but subtle, pattern variations.

In this paper, we present the solutions we use to address these two problems and achieve a final statistical motion analysis and subject comparison. We evaluate this process on the dataset of seven controls and four patients. In the current stage, we focus our attention only on the average motion of the tongue at every time frame. It demonstrates the efficacy of the process and opens up the possibilities of future work on more local tongue motion quantity analysis.

2. METHODS

2.1. Computing the motion fields from a new reference frame

For each subject, the pipeline produces a sequence of 26 3D volumetric vector fields $\{\boldsymbol{D}_{1,1}(\boldsymbol{X}), \boldsymbol{D}_{1,2}(\boldsymbol{X}), \dots, \boldsymbol{D}_{1,26}(\boldsymbol{X})\}$. Each vector field $\boldsymbol{D}_{1,t}(\boldsymbol{X})$, as visualized in Figure 1(b) and 1(c), shows the displacement from time frame 1 (default reference frame) to the current time frame t. $\boldsymbol{X}$ is the 3D grid located at time frame 1. As a result, if we consider the vector field $\boldsymbol{D}_{1t}(\boldsymbol{X})$ as arrows, they grow from grid $\boldsymbol{X}$ and end up pointing at the non-grid positions (the tissue point locations) in the current frame.

The first time frame is normally a pre-speech relaxed position of the tongue. For speech motion, it is useful to observe the motion from /a/ forward into /s/ and then upward into /k/. More importantly, since every person's tongue relaxed position is different and unpredictable, the mid-central vowel /a/ has to be used as the common reference frame to compare motion across subjects [9]. Therefore we are forced to switch the reference frame to the maximum /a/.

Suppose maximum /a/ happens at time frame r and the current frame is t. Related motion fields are $\boldsymbol{D}_{1r}(\boldsymbol{X})$ and $\boldsymbol{D}_{1t}(\boldsymbol{X})$. Straightforwardly, if we are able to find the inverse field of $\boldsymbol{D}_{1r}(\boldsymbol{X})$, namely $\boldsymbol{D}_{r1}(\boldsymbol{X}_r)$, the following field is what we want:

$$\boldsymbol{D}_{rt}(\boldsymbol{X}_r) = \boldsymbol{D}_{1t}\big(\boldsymbol{X}_r + \boldsymbol{D}_{r1}(\boldsymbol{X}_r)\big) \tag{1}$$

Note that we have changed the symbol $\boldsymbol{X}$ to $\boldsymbol{X}_r$ because it is now the grid on the new reference r, instead of 1.

Because the field $\boldsymbol{D}_{1r}(\boldsymbol{X})$ is discrete, we only know that at time frame r, $\boldsymbol{D}_{r1}\big(\boldsymbol{X} + \boldsymbol{D}_{1r}(\boldsymbol{X})\big) = -\boldsymbol{D}_{1r}(\boldsymbol{X})$. To find $\boldsymbol{D}_{r1}$'s value at $\boldsymbol{X}_r$, we apply a fixed-point method [10] by iteratively solving the following equation:

$$\boldsymbol{D}_{r1}^{(n)}(\boldsymbol{X}_r) = -\boldsymbol{D}_{1r}\left(\boldsymbol{X}_r + \boldsymbol{D}_{r1}^{(n-1)}(\boldsymbol{X}_r)\right) \tag{2}$$

Substituting the converged result of Equation (2) into Equation (1) and repeating for every time frame, we get a new sequence of displacement fields $\{\boldsymbol{D}_{r,1}(\boldsymbol{X}), \boldsymbol{D}_{r,2}(\boldsymbol{X}), \dots, \boldsymbol{D}_{r,26}(\boldsymbol{X})\}$ for every subject starting at time frame /a/.

2.2. Two-step principal component analysis for tongue motion evaluation

Multi-subject PCA of the tongue motion requires a certain quantity of the motion to be in the same frame of reference. We have made the displacement field with respect to the same reference frame. Still, for different subjects, their tongue shapes (reflected by 3D tongue masks) vary extensively. In order to compare them, we need a common mask region (or a common tongue atlas space). For current development of this work, we consider the simplest approach – we divide the tongue into eight volumes of interests (VOIs) (see Figure 1(a)). Inside each VOI, we average the motion field to get one vector which represents its general motion, denoted by $\{\boldsymbol{d}_{r,1}, \boldsymbol{d}_{r,2}, \dots, \boldsymbol{d}_{r,26}\}_v$, where v is the VOI number from 1 to 8.

Furthermore, since we are only interested in the motion from /a/ to /s/ to /k/, we create a common time interval by taking the average motion between these two periods, and using cubic spline (denoted as "interp" in Equation (3)) to interpolate them into 17 time frames for all subjects, where /a/ is time-frame 1, /s/ is 7, and /k/ is 17. Denoting the time frame number of maximum /a/, /s/ and /k/ as a, s and k, we have

$$\{\widehat{\boldsymbol{d}}_{1,1}, \dots, \widehat{\boldsymbol{d}}_{1,7}, \dots, \widehat{\boldsymbol{d}}_{1,17}\}_v = \text{interp}\{\boldsymbol{d}_{a,a}, \dots, \boldsymbol{d}_{a,s}, \dots, \boldsymbol{d}_{a,k}\}_v \tag{3}$$

For any VOI, $\widehat{\boldsymbol{d}}_{1,t}$ is the interpolated mean motion we are interested in, which puts all subjects' motions in the same framework and ready for PCA (Figure 2). Labeling the subject number by m, we stack the mean motion of all 17 frames as one vector

$$\widehat{\boldsymbol{d}}^m = [\widehat{\boldsymbol{d}}_{1,1}^m; \dots; \widehat{\boldsymbol{d}}_{1,7}^m; \dots; \widehat{\boldsymbol{d}}_{1,17}^m] \tag{4}$$

which lies in a $3 \times 17 = 51$ dimensional space. $\widehat{\boldsymbol{d}}^m$ is the representation of the general motion in this VOI of subject m when performing the entire speech task of "a souk". Note that by doing so we have avoided treating each time frame independently. Instead, we consider the entire task as an evaluation of the subject's speech function.

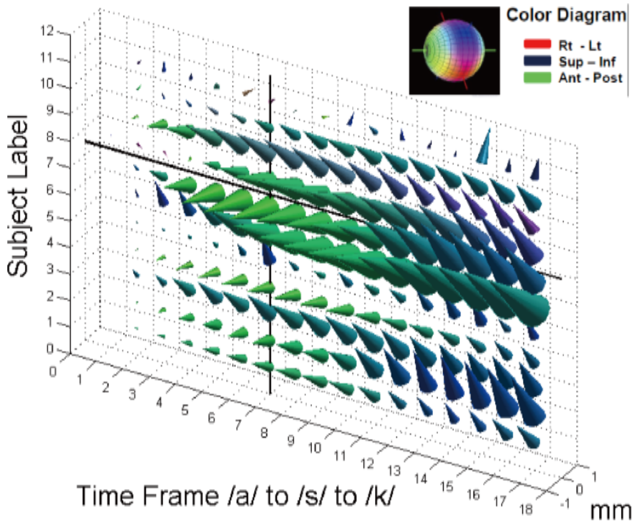

Figure 2. Average motion across all time frames from /a/ to /k/ for all subjects in VOI-1. Vertical line is at /s/. Horizontal line separates patients (top) from controls.

Suppose the number of controls is C and the number of patients is P. PCA of controls requires (1) subtract the mean of control motions $\hat{\boldsymbol{s}}^i = \widehat{\boldsymbol{d}}^i - \text{mean}\{\widehat{\boldsymbol{d}}^1, \dots, \widehat{\boldsymbol{d}}^i, \dots, \widehat{\boldsymbol{d}}^C\}$, $i = 1 \dots C$, (2) compute the covariance matrix $COV = [\hat{\boldsymbol{s}}^1, \dots, \hat{\boldsymbol{s}}^i, \dots, \hat{\boldsymbol{s}}^C][\hat{\boldsymbol{s}}^1, \dots, \hat{\boldsymbol{s}}^i, \dots, \hat{\boldsymbol{s}}^C]^T$ and (3) find the eigen-decomposition of COV to get $C-1$ principal directions $\{\boldsymbol{e}^1, \dots, \boldsymbol{e}^{C-1}\}$ and principal values (PCs) $\{\lambda^1, \dots, \lambda^{C-1}\}$. After projecting the subjects' motion onto these directions, we have observed the patients are hardly distinguished from test controls in their PC scores.

Therefore, we introduce another PCA step after performing the first PCA on controls. Observing the fact that the PC space has a rank of $C-1$, the remaining $51-(C-1)$ "principal directions" are only vectors generated by any feasible orthogonalization method (e.g., the Gram-Schmidt process). And this remaining $51-(C-1)$ dimensional space contains only the motion information of the patients, because the controls project a zero PC score in this space. As a result, we take the patient motion labeled by $j, j = 1 \dots P$, subtract the control mean and compute the patients' "normal" motion component, namely

$$\hat{\boldsymbol{s}}^j = \widehat{\boldsymbol{d}}^j - \text{mean}\{\widehat{\boldsymbol{d}}^1, \dots, \widehat{\boldsymbol{d}}^i, \dots, \widehat{\boldsymbol{d}}^C\} \tag{5}$$

$$\hat{\boldsymbol{s}}^j_{normal} = (\hat{\boldsymbol{s}}^j)^T \boldsymbol{e}^1 \lambda^1 + \cdots + (\hat{\boldsymbol{s}}^j)^T \boldsymbol{e}^{C-1} \lambda^{C-1} \tag{6}$$

The remaining motion is considered abnormal and is given by

$$\hat{\boldsymbol{s}}^j_{abnormal} = \hat{\boldsymbol{s}}^j - \hat{\boldsymbol{s}}^j_{normal} \tag{7}$$

We compute the covariance matrix of $\hat{\boldsymbol{s}}^j_{abnormal}$ and find its eigen-decomposition to get $P-1$ more vectors as the PC directions for abnormal motion $\{\boldsymbol{u}^1, \dots, \boldsymbol{u}^{P-1}\}$. Taken together, $\{\boldsymbol{e}^1, \dots, \boldsymbol{e}^{C-1}, \boldsymbol{u}^1, \dots, \boldsymbol{u}^{P-1}\}$ form a two-step PCA to represent the general normal vs. abnormal motions.

3. RESULTS

We evaluated the process on seven controls and four patients. Due to the limited amount of data, we doubled the number of subjects by only using the left half of the tongue and mirroring the right half to the left. This is reasonable because normal tongue motion is generally symmetric on the left and right side. Since the patients have only one side of the tongue that has received glossectomy, we distinguish this situation by "patient glossectomy side (PGS)" and "patient normal side (PNS)", which yields 14 controls, 4 PGSs and 4 PNSs on VOIs 1-4 (Figure 1(a)).

We took 10 of the controls to build the normal PC space and the 4 PGSs (re-labeled from 9 to 12) to build the abnormal PC space. Then we tested the remaining 4 controls (re-labeled from 1 to 4) and 4 PNSs (re-labeled from 5 to 8) for evaluation.

The principal directions for VOI-1 are shown in Figure 3. The first few normal PCs (top three) show mostly front/back and up/down motion. The first abnormal PC (topmost) is mostly left/right. The PC weights are also listed.

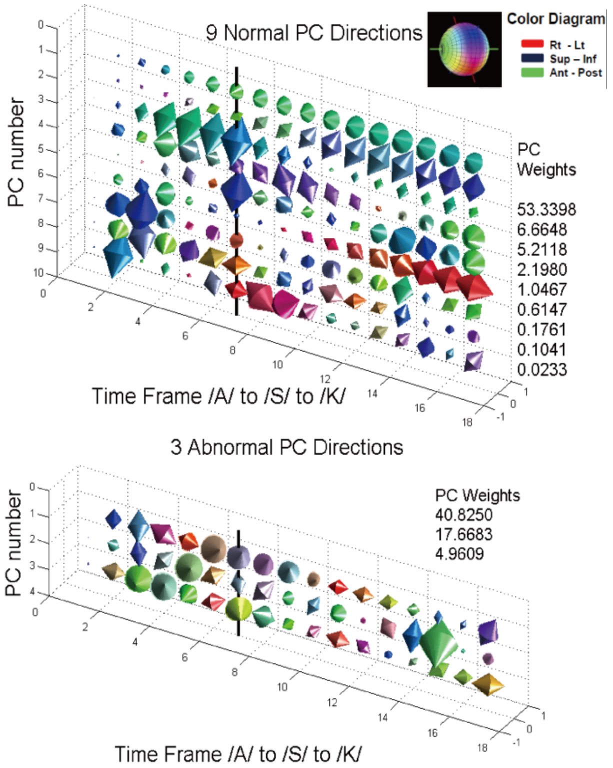

Figure 3. All PC directions (9 normal and 3 abnormal) of VOI-1. Vertical line identifies the position of /s/.

Then we projected each subject's motion onto these 12 directions to get its PC energy (projection score) in which we were interested in the abnormal ones (10th, 11th and 12th score). We plot these three scores for all subjects and all

VOIs in a 3D space in Figure 4. While the controls are all at origin having zero energy on the abnormal PCs, PGSs and PNSs have a wider spread than control tests. Lastly, we repeated the entire experiment with all possible combinations of training and testing data ($\binom{14}{10} = 1001$ cases for each VOI), obtained the three abnormal PC scores and averaged them in each case. The results of all subjects and all VOIs in all cases are shown in Figure 5. Control test data has lower and more consistent abnormal energy when comparing to PNSs, and they both are lower than PGSs in general. Especially, in all VOIs, the mean of the control test abnormal energy is lower than both PGS and PNS in 3829 out of 4004 cases. We conclude that despite the small amount of training data, this analysis is capable of distinguishing normal motion from patient motion ($p < 0.05$).

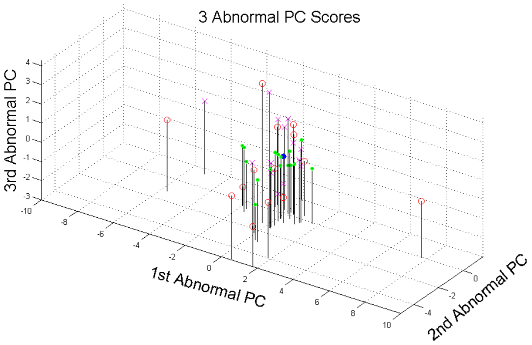

Figure 4. Abnormal PC energy space plot for all subjects in all four VOIs with origin as control, dot as control test, circle as PGS and cross as PNS.

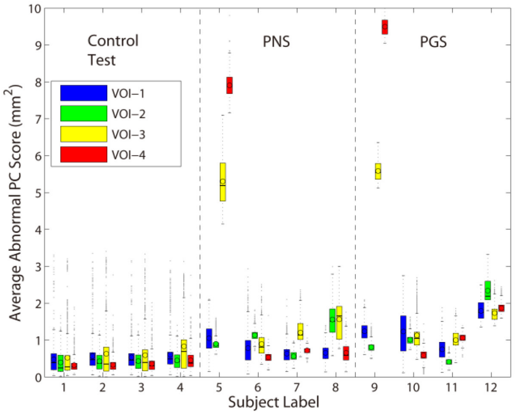

Figure 5. Boxplot of average abnormal PC scores of all subjects and all four VOIs in 1001 experiments. The center bar in a box indicates median and the circle indicates mean.

4. CONCLUSION

In this work, we described the process of acquiring and estimating 3D motion of the human tongue during speech. We provided the details for achieving consensus statistical analysis by using PCA, and showed that the analysis is capable of distinguishing control motion from patient motion. Although a number of limitations such as insufficient subject number and simple volume averaging may provide obstacles to the accuracy of the method, it shows much potential the tongue motion estimation pipeline can achieve for motion quantity analysis.

Acknowledgements. This work was supported by NIH/NCI grant 5R01CA133015. The authors also thank the anonymous reviewers for suggestions that improved both equation clarity and Figure 5.

5. REFERENCES

[1] V. Parthasarathy, J. L. Prince, M. Stone, E. Murano, and M. Nessaiver, “Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing,” *J. Acoust. Soc. Am.*, vol. 121(1), pp. 491–504, 2007.

[2] F. Xing, J. Lee, E. Z. Murano, J. Woo, M. Stone, and J. L. Prince, “Estimating 3D tongue motion with MR images,” *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, 2012.

[3] W.M., Kier, and K.K. Smith, “Tongues, tentacles and trunks: the biomechanics of movement in muscular-hydrostats,” *Zool. J. Linnean Soc*. vol. 83, pp. 307–324, 1985.

[4] E. A. Zerhouni, D. M. Parish, W. J. Rogers, A. Yang, and E. P. Shapiro, “Human heart: tagging with MR imaging – a method for noninvasive assessment of myocardial motion,” *Radiology*, vol. 169, pp. 59–63, 1988.

[5] N.F. Osman, E.R. McVeigh, and J.L. Prince, "Imaging heart motion using harmonic phase MRI", *IEEE Trans. on Med. Imaging*, vol. 19(3), pp. 186–202, 2000.

[6] L. Grady, “Random walks for image segmentation”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[7] X. Han, C. Xu, and J. L. Prince, “A topology preserving level set method for geometric deformable models”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 755–768, 2003.

[8] X. Liu, K. Abd-Elmoniem, M. Stone, E. Murano, J. Zhuo, R. Gullapalli, and J. L. Prince, “Incompressible deformation estimation algorithm (IDEA) from tagged MR images”, *IEEE Trans. on Med. Imaging*, vol. 31(2), pp. 326–340, 2012.

[9] R. D. Kent, and C. Read, *Acoustic Analysis of Speech*, Singular, San Diego, pp. 211, 1992.

[10] M. Chen, W. Lu, Q. Chen, K.J. Ruchala, and G. H. Olivera, “A simple fixed-point approach to invert a deformation field”, *Med. Phys.*, vol. 35(1), pp. 81–88, 2008.