Interpolation of Tongue Fleshpoint Kinematics from Combined EMA Position and Orientation Data

Andrew J. Kolb [1], *Michael T. Johnson* [1], *Jeffrey Berry* [2]

[1] Marquette University Electrical and Computer Engineering, Milwaukee, WI, USA
[2] Marquette University Speech Pathology and Audiology, Milwaukee, WI, USA

{andrew.kolb, michael.johnson, jeffrey.berry} @marquette.edu

Abstract

Articulatory data such as that collected via electromagnetic articulography (EMA) are valuable for many speech applications, including speech modeling, recognition, and synthesis. Nearly all current EMA applications and methods focus on the use of positional sensor data, even though modern 3D-EMA systems also capture sensor orientation, which provides significant additional information about articulator posture and vocal tract shape. To address this problem, this paper introduces a new method for interpolating untracked tongue fleshpoint positions from the combined position and orientation data of three EMA sensors on the tongue, using additional reference sensors for evaluating interpolation error. Comparison of interpolated and measured data illustrated effectiveness of the new method in providing additional tongue shape and position features. The results suggest that analytic methods that combine sensor position and orientation data are able to improve the characterization of tongue kinematics even using a small number of EMA sensors.

Index Terms: EMA, quaternions

1. Introduction

Electromagnetic Articulography (EMA) data are often used to obtain articulatory kinematic measures such as sensor speed, acceleration, and range of motion [1, 2], typically from sensors restricted to the midsagittal plane (c.f., [3]). These analyses take advantage of the temporal resolution provided by EMA, but do not require concomitant analysis of sensor orientation data. Orientation data have proven useful in applications such as head stabilization [4], and jaw angle measurement [5]. Orientation data have also been suggested for reconstructing tongue shape in the coronal plane [6], and have aided in the control of a skeletal animation of a tongue surface that was extracted from MRI data [7]. Yet no general method currently exists for interpolating tongue fleshpoint positions between EMA sensors using combined position and orientation data.

Other modalities for representation of tongue shape and movement include ultrasound [8, 9] at rates as high as 90Hz, and cine-MRI and real-time MRI [10, 11] at rates sample rates as high as 33 Hz [12]. EMA offers superior temporal resolution (up to 400 Hz), but somewhat lower spatial resolution because of the limited number of EMA sensors attached to the tongue surface. While EMA can be combined with MRI data to more completely characterize the moving tongue [13], such multi-modality approaches are cost-prohibitive and less practical for clinical assessment of tongue movement. Since increasing the number of EMA sensors is almost certain to influence movement patterns, reducing the validity of data for kinematic assessment, a method for interpolating tongue fleshpoint positions from relatively scant EMA sensor data would be very useful for both research and clinical applications. The goal of this work is to develop and evaluate such an interpolation method based on data from a three sensor EMA configuration.

2. Methods

The three points on the tongue used for interpolation included: a sensor placed on the tongue blade, about 1 cm back from the tongue tip along the midsagittal line (TB); a sensor placed on the tongue dorsum, as far back on the tongue as possible along the midsagittal line (TD); and a sensor placed on the left side of the tongue, one centimeter from the tongue's edge at the midpoint between TB and TD (TL). Each of these was a 5 Degree of Freedom sensor that included 3D position information and 2D orientation information that represented the plane of the sensor. With sensors oriented roughly parallel to the tongue, the data provided thus includes both position and gradient information at the vertices of a triangle on one half of the tongue.

Given data that has both position and derivative information, several different existing interpolation schemes are possible. The simplest is inverse-distance weighting, which expresses each interpolated point as a combination of the other points, weighted by how close the interpolated point is to each of the known points [14]. This is computationally simple and intuitive, but does not make use of the gradient information provided by the sensors. Other methods for surface fitting utilizing derivative information include Bezier patches, Hermite surface interpolation, and Clough-Tocher interpolation. A thorough summary of each of these techniques can be found in [15].

To interpolate inside the triangle formed by the sensors, a hybrid technique was chosen, drawing from techniques used in both Clough-Tocher interpolation and inverse-distance weighting. This approach makes the resulting algorithm computationally easy, while making adequate use of the derivative information at each of the known sensor points.

2.1 Quaternion data and baseline orientation

Since sensors are initially only roughly positioned, the first step in obtaining accurate derivative information and calculating the tongue mesh is to establish a baseline orientation, without which it is not possible to infer a gradient from the sensor orientation data.

There are several possible approaches to defining the tongue baseline, but in this initial work a simple averaging method that used the average sensor position across the data record to represent a baseline flat position was used. For each individual record, the average orientation of the sensors was calculated and assigned a value that represents the orientation pointing in the superior direction. Because orientation data are provided in quaternion format [16], the averaging technique employed the eigenanalysis method described in [17]. Defining this average quaternion value as the superior direction, an adjustment quaternion to be applied as a correction to all other data can be calculated as follows:

$$q_{adj} = q^{*}_{avg} q_{superior} \tag{1}$$

where q^{*}_{avg} is the conjugate of the average quaternion, and $q_{superior}$ is the quaternion whose orientation represents the superior direction. Right-multiplying the adjustment quaternion with all quaternions in a given record results in adjusted orientations that point in a direction roughly normal to the tongue surface:

$$q_{corrected} = q_{raw} q_{adj} \tag{2}$$

2.2 Tongue interpolation algorithm

With quaternions that more accurately provide gradient information about the tongue surface, virtual positions along the tongue can be calculated from the position and derivatives.

2.2.1 Clough-Tocher split

The method begins by breaking up the macro-triangle formed by TD, TL, and TB into sub-triangles, in a Clough-Tocher split [18]. This split breaks a single triangle into three new triangles, by connecting the three original vertices to the triangle's centroid. This process can be repeated with the sub-triangles as many times as needed to obtain the desired number of points inside the macro-triangle. For the purposes of this description, the triangle points will be described as points in the X-Z plane, with the Y-value at each XZ position representing the tongue height that needs to be found.

Using this split, once the Y-position and gradient at the centroid of the macro-triangle are found as described in the following sections, the process can be repeated to determine the position and gradient values at the centroids of each of the sub-triangles.

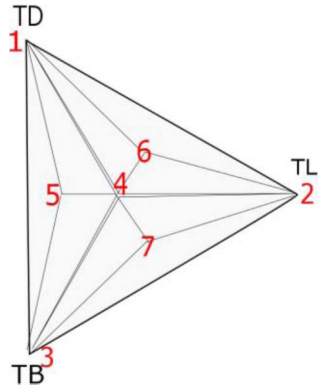

Figure 2.1: *First seven points found using a Clough-Tocher split. Point 4 is the centroid of the macro-triangle, with points 5, 6, and 7 found from the centroids of the sub-triangles.*

2.2.2 Interpolation using sensor orientation planes

Given the point, (b_x, b_y, b_z) and normal vector, (B'_x, B'_y, B'_z) represented by sensor TB, an equation for a plane can be written, representing a tangent surface on the tongue:

$$B'_x(x-b_x) + B'_y(y-b_y) + B'_z(z-b_z) = 0 \tag{3}$$

This plane can be used to project the tongue height y at some point in the X-Z plane by rearranging and solving for y:

$$y = P_B(x,z) = b_y + \frac{B'_x(x-b_x) + B'_z(z-b_z)}{-B'_y} \tag{4}$$

This process can be repeated for each of the sensors, resulting in three planes, $P_B(x,z), P_L(x,z), P_D(x,z)$ which are capable of providing projections at some point in the X-Z plane (Figure 2.2).

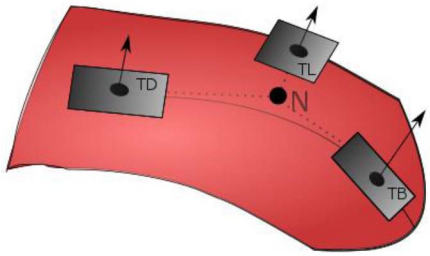

Figure 2.2: *Planes created from sensor position and orientation, used to project the value at a new point N.*

2.2.3 Inverse weighted centroid estimation

Inverse-distance-weighting was used to combine projections for the y-value at the centroid from each sensor into a single y estimate, with the weight given to a sensor's projection assigned to be the inverse of the distance between the sensor and the desired point. For example, the weight given to a projection from TB is:

$$w_B(x,z) = \frac{1}{\sqrt{(b_x - x)^2 + (b_z - z)^2}} \tag{5}$$

Others weights are determined similarly.

2.2.4 Calculate new point and normal vector

Combining equations (4) and (5) for each sensor yields the y-value for a new point (n_x, n_y, n_z):

$$n_y = \frac{w_B(n_x,n_z)P_B(n_x,n_z) + w_L(n_x,n_z)P_L(n_x,n_z) + w_D(n_x,n_z)P_D(n_x,n_z)}{w_B(n_x,n_z) + w_L(n_x,n_z) + w_D(n_x,n_z)} \tag{6}$$

The normal vector for the new point can be calculated analogously, as the inverse-distance-weighted average of the normal vectors at TB, TL, and TD.

2.2.5 Summary

With the macro-triangle's centroid position and derivative information calculated, this process can be repeated for other sub-triangle centroids on the tongue. It should be noted that to complete the interpolation, the calculated tongue heights are reflected across the midsagittal plane, which assumes that the tongue behaves approximately symmetrically (generally the

case for normal speakers). To allow for visualization, the points were connected to form a triangular mesh.

2.3 Experimental setup

To assess interpolation accuracy, six sensors were placed on the tongue, with the original three (TD, TL, TB) used to create the projected tongue values, and three additional sensors (labeled S1, S2, S3) served as ground truth against which the projected tongue values could be compared, as shown in Figure 2.3.

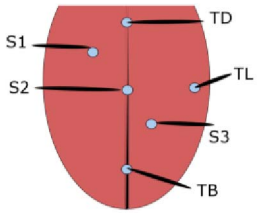

Figure 2.3: *Tongue sensor placement for evaluation*

The single subject used to conduct the experiment was a 22-year-old female, whose native language was American English. The data were collected at 400 Hz using the NDI Wave Speech Research System. The data were corrected for head movement using a reference sensor attached to glasses worn by the subject throughout all recordings. Bite-plate correction was also used to place the data in a physiologically defined coordinate space, with the X-Z plane corresponding to the maxillary occlusal plane and the X-Y plane corresponding to the midsagittal plane following bite-plate correction. The origin of the space was set at the tip of the central maxillary incisors.

2.4 Speech recordings

Six speech records were taken from the participant. The caterpillar script [19] was read twice, both as an acclimation record and a typical connected speech record. The remaining four records consisted of consonant-vowel-consonant (CVC) words or pseudo-words with consonants \k\, \t\, and \p\ flanking vowels \i\, \e\, \æ\,\ə\, \ʌ\,\u\,\o\, and \ɑ\ (or \IY\, \EY\, \AE\, \ER\, \AH\, \UW\, and \OW\ in ARPABET). Three of the records consisted of five repetitions of each word or pseudo-word, embedded inside a carrier phrase (for 40 repetitions in each record). The carrier phrase used was "It'll say (\kik\) again." This allowed for more natural use of the words in context, including more natural variability. The other single record was simply all 24 pseudo-words repeated without the carrier phrase. For all records, excluding the reading of the caterpillar script, the participant was provided a verbal prompt from demonstrating each target utterance.

2.5 Analysis method

For each of the six records, a basic error analysis was done to evaluate the predictions made by the interpolation method. This analysis included calculations of mean absolute error, standard deviation of absolute error, maximum overestimate, and maximum underestimate for each sensor. The errors were calculated by measuring the difference between the projected value and actual value within a sub-triangle. Additionally, these same metrics were calculated within only the vowels in each of the CVC records. Combined, these metrics give a picture of overall errors for the algorithm, as well as errors for a variety of vowels in certain consonant contexts.

3. Results and discussion

The results of the evaluation are shown in Table 1. Generally, S3 was the most accurately predicted from the three other points, but all sensors showed mean absolute errors below 3 mm, with standard deviations near 1.5 mm. The maximum errors averaged across all records for sensors 1, 2 and 3 were 10.09 mm, 12.46 mm, and 8.81 mm respectively. The vowel segments in the CVC records did not differ significantly in accuracy from all of the records as a whole. The total accuracy of the EMA sensors is 0.5 mm (with an unknown orientation resolution) [20], which places an upper limit on the accuracy achievable through interpolation.

Error Table	*Mean Error (mm)*	*Dynamic Range (mm)*
S1	2.56 ± 1.72	23.82
S2	2.28 ± 1.45	28.23
S3	0.90 ± 0.93	36.29

Table 1. *Mean absolute error and standard deviation across recordings, compared to sensor dynamic range.*

In a typical case, the interpolation method overestimates the value of S1, while underestimating the value of S2 and S3. The larger, overestimated error in S1 is reasonable, as by using only linear, first derivative components to interpolate the points, changes in tongue concavity near TD on either side of the medial sulcus will be smoothed over. S2 errors arise from only two sensors being used along the medial sulcus, which are inadequate for fully capturing midsagittal curvature. S3 was consistently the most accurate. Taken together, the accuracy of the sensors suggests that areas with less change in curvature can be more accurately interpolated.

The predictions were accurate enough that estimated gross changes in tongue position and shape (e.g. changes in concavity) matched the actual changes when viewed as a mesh. However, because the three data points contain only first derivative information, no changes in tongue concavity could be predicted within the interpolated area. A coronal cross-section of the interpolated tongue will always appear as a single concave or convex curve (u-or n-shaped), when in reality the tongue surface can contain multiple changes in concavity (allowing for an m-shape).

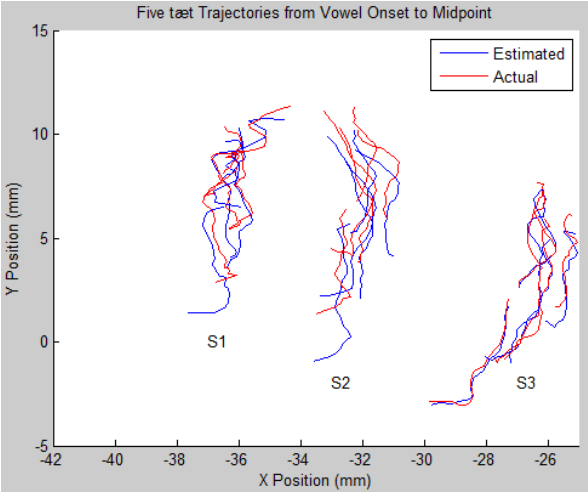

Figure 3.1: *Estimated and actual vowel trajectories for an example word (tæt).*

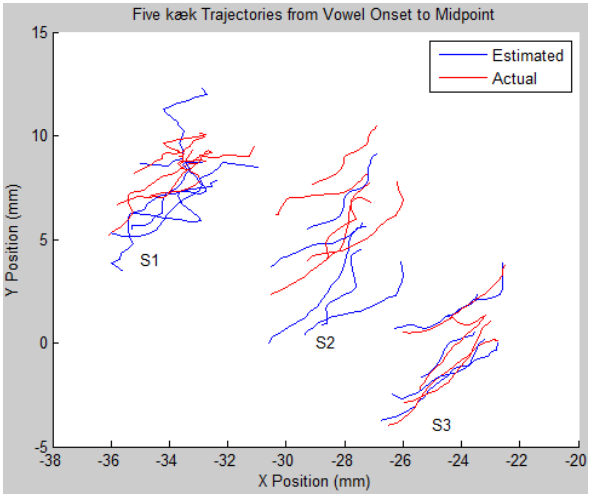

Figure 3.2: *Estimated and actual vowel trajectories for an example word (kæk).*

An example of actual versus estimated trajectories for the interpolated tongue points are shown in Figures 3.1 and 3.2, for two different consonant contexts. These trajectories illustrate the scale and nature of the interpolation accuracy, showing that the proposed interpolation method is able to characterize tongue motion throughout an articulatory movement. While the absolute range and positions show noticeable errors, the paths traced by the virtual and actual sensors appear very similar, suggesting measures like speed and distance travelled could be obtained via interpolation. Similarly, the variance in the actual and estimated trajectories is close to the natural variance in the real trajectories alone.

Future work includes a number of improvements to increase the accuracy of the interpolation method and extend it beyond its current limitations. The development of a more reliable baseline-orientation protocol would provide more accurate gradient information.

Additionally, it is possible to add more sensors to the tongue that can be used in the interpolation, which would add additional information and improve the overall accuracy. Adding a sensor outside of the macro-triangle used in the current work would allow for some extrapolation outside of the main portion of the tongue body. The current focus was on simple tongue configurations used in vowel articulation, but additional sensors could provide information about more complicated changes in concavity and curvature seen during liquids and consonant clusters. However, since adding sensors may affect sensor adhesion reliability and decrease the naturalness of articulation, increased accuracy must be weighed against these considerations.

By incorporating both orientation and position data from the EMA sensors, this approach substantially increases the overall amount of information and representative capability of EMA. In particular, the method provides a mechanism to estimate tongue position at target locations, both laterally and midsagittally. Combined with additional reference information such as palate height, this approach enables the creation of much more accurate articulatory features that directly correspond to vocal tract cross-section and shape. Applications of this work include clinical assessment of articulation, studies of pronunciation variation and accent, as well as speech processing technologies such as acoustic-to-articulatory inversion and articulatory audio and audio-video synthesis.

4. Conclusions

This work has presented a new approach for combining orientation and position data from a small number of EMA sensors to interpolate additional data points. Results show that the approach is able to accurately estimate tongue movement at the new interpolated locations, and thus can provide a more complete set of articulatory features than simple position-based analyses.

5. Acknowledgements

This paper is based upon work supported by the National Science Foundation under Grant No. IIS-1142826 and IIS-1320892.

6. References

[1] J. Berry, A. Kolb, C. North and M. T. Johnson, "Acoustic and kinematic characteristics of vowel production through a virtual vocal tract in dysarthria," in *Fifteenth Annual Conference of the International Speech Communication Association,* 2014, .

[2] J. R. Green and Y. Wang, "Tongue-surface movement patterns during speech and swallowing," *J. Acoust. Soc. Am.,* vol. 113, pp. 2820-2833, 2003.

[3] W. F. Katz, J. Wang and S. Mehta, "Using lateral sensors in flesh-point tracking of/l/and/ɹ/in American English," *J. Acoust. Soc. Am.,* vol. 135, pp. 2198-2198, 2014.

[4] C. Kroos, "Using sensor orientation information for computational head stabilisation in 3D electromagnetic articulography (EMA)," in *Tenth Annual Conference of the International Speech Communication Association,* 2009, .

[5] R. N. Henriques and P. van Lieshout, "A comparison of methods for decoupling tongue and lower lip from jaw movements in 3D articulography," *Journal of Speech, Language, and Hearing Research,* vol. 56, pp. 1503-1516, 2013.

[6] P. Hoole and A. Zierdt, "Five-dimensional articulography," *Speech Motor Control: New Developments in Basic and Applied Research,* pp. 331-349, 2010.

[7] I. Steiner and S. Ouni, "Progress in animation of an EMA-controlled tongue model for acoustic-visual speech synthesis," *arXiv Preprint arXiv:1201.4080,* 2012.

[8] Y. C. Chiang, F. P. Lee, C. L. Peng and C. T. Lin, "Measurement of tongue movement during vowels production with computer-assisted B-mode and M-mode ultrasonography," *Otolaryngol. Head. Neck. Surg.,* vol. 128, pp. 805-814, Jun, 2003.

[9] M. Stone and A. Lundberg, "Three- dimensional tongue surface shapes of English consonants and vowels," *J. Acoust. Soc. Am.,* vol. 99, pp. 3728-3737, 1996.

[10] M. Hasegawa-Johnson, S. Pizza, A. Alwan, J. S. Alwan and K. Haker, "Vowel category dependence of the relationship between palate height, tongue height, and oral area," *Journal of Speech, Language, and Hearing Research,* vol. 46, pp. 738-753, 2003.

[11] S. Narayanan, K. Nayak, S. Lee, A. Sethy and D. Byrd, "An approach ot real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America,* vol. 115, pp. 1771-1776, 2004.

[12] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker and J. Frahm, "Real- time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction," *Magnetic Resonance in Medicine,* vol. 69, pp. 477-485, 2013.

[13] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Commun.,* vol. 41, pp. 303-329, 2003.

[14] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proceedings of the 1968 23rd ACM National Conference,* 1968, pp. 517-524.

[15] G. Farin, "Triangular bernstein-bézier patches," *Comput. Aided Geom. Des.,* vol. 3, pp. 83-127, 1986.

[16] J. C. Hart, G. K. Francis and L. H. Kauffman, "Visualizing Quaternion Rotation," *ACM Transactions on Graphics,* vol. 13, pp. 256-276, 1994.

[17] F. L. Markley, Y. Cheng, J. L. Crassidis and Y. Oshman, "Averaging quaternions," *Journal of Guidance, Control, and Dynamics,* vol. 30, pp. 1193-1197, 2007.

[18] R. W. Clough and J. L. Tocher, "Finite element stiffness matrices for analysis of plates in bending," in *Proceedings of Conference on Matrix Methods in Structural Analysis,* 1965, pp. 515-545.

[19] R. Patel, K. Connaghan, D. Franco, E. Edsall, D. Forgit, L. Olsen, L. Ramage, E. Tyler and S. Russell, ""The Caterpillar": A Novel Reading Passage for Assessment of Motor Speech Disorders," *American Journal of Speech-Language Pathology,* vol. 22, pp. 1-9, 2013.

[20] J. J. Berry, "Accuracy of the NDI Wave speech research system," *Journal of Speech, Language, and Hearing Research,* vol. 54, pp. 1295-1301, 2011.