Indexing head movement during speech production using optical markers

Kevin D. Roon,[a] Katherine M. Dawson,[a] Mark K. Tiede, and D. H. Whalen[a]

Haskins Laboratories, 300 George Street, New Haven, Connecticut 06511, USA
kroon@gc.cuny.edu, kdawson2@gradcenter.cuny.edu, tiede@haskins.yale.edu, dwhalen@gc.cuny.edu

Abstract: Optical marker tracking integrated with electromagnetic articulometry was used to assess the movement extent of various points on (a) forehead skin and (b) points on a head-mounted apparatus, relative to a fixed point just above the upper incisors, and to compare the accuracy of the two different approaches to indexing head position during speech production. Both methods can provide a satisfactory index of head position. If skin-affixed markers are used, a minimum of 4 is recommended. Locations for optimal marker placement are identified.

[DDO]
Date Received: December 14, 2015 **Date Accepted:** May 3, 2016

1. Introduction

It is often necessary in speech production research to index head position to investigate the relationship between head movements and other properties of speech, to "head-correct" movement data of vocal-tract articulators, or to calculate movement of other points (e.g., eyebrow, jaw) relative to the head. While electromagnetic articulometry (EMA) is frequently used to track head movement (Hoole and Zierdt, 2010), EMA is often not feasible due to, e.g., expense of equipment, difficulty of use with certain populations such as very young or old participants.

Some studies have indexed head position by attaching a single visual marker to a point at the nasion (e.g., Scarborough *et al.*, 2009). This method relies on the assumption that the skin at the chosen point does not move relative to the skull enough to qualitatively affect analyses. Skin on the forehead can move relative to the skull during speech, though the extent of such movement has not been clearly established. A more serious limitation of this approach is that a single optical marker cannot provide any information about rotational movements, making this method unsuitable for, e.g., head-correcting vocal tract movement data. Other researchers have indexed head position by using visual markers affixed to a tightly fitted, head-mounted apparatus. For example, Dohen *et al.* (2004) used points on glasses to measure jaw aperture relative to the head. Whalen *et al.* (2005) proposed using visual markers affixed to either head-mounted goggles or some other head-mounted device during the collection of ultrasound data of the tongue to place tongue contours into a coordinate system relative to the palate during post-processing. By using multiple markers, this approach allows tracking of both the positional and rotational movements of the head, but it is also unknown whether and how much such an apparatus moves relative to the skull.

An alternative, simpler approach is to affix multiple visual markers to the forehead skin, but it is not known whether apparatus-affixed or skin-affixed markers provide more accurate data. Several factors influence how well a given set of optical markers indexes head position. The most obvious factor is the degree to which the markers move relative to the skull. Other factors include the number of markers used, the distance from the markers to the skull, the distance between the markers, and the independence of individual marker movements. The combined interaction of these factors will determine which method provides a more accurate index of head position. A head-mounted apparatus needs to be custom-made, adds complexity to the experimental design, and may cause participant discomfort. If the two methods provide an equally good estimation of head position, then affixing markers directly to the skin is to be preferred as being simpler, conceptually and practically. The present study evaluated whether skin- or apparatus-affixed markers provide a more accurate index of head position.

[a]Also at: Program in Speech-Language-Hearing Sciences, CUNY Graduate Center, 365 Fifth Avenue, Room 7107, New York, NY 10016, USA.

2. Experiment

Our study had eight participants (aged 22–50 years, 4 female). Six were native speakers of English (Southern British English P1F or American English P2M, P3F, P5F, P6M, P8M). Two had non-native but professional level English (one native Dutch speaker P4F, one native Hindi speaker P7M). All participants reported no linguistic or neuromuscular disorders, signed informed consent forms approved by the City University of New York (CUNY) Institutional Review Board, and were paid for their time.

All data were collected at the CUNY Graduate Center. Two types of movement data were collected from participants simultaneously at a sampling rate of 100 Hz using an optical camera (Optotrak Certus) system integrated with an EMA (WAVE) system (both made by NDI, Northern Digital Inc.). The Optotrak system was used to track infrared emitting diodes (IREDs) that were affixed to various points on each participant's skin. Participants sat in a chair with ten IREDs affixed to their skin using double-sided adhesive tape [Fig. 1(A)]: one at the nasion, and then three rows of three across the forehead. IREDs were placed as far apart as possible while still remaining in the line of sight of the Optotrak cameras. Participants later wore a custom-made device (the "tiara") fitted with six IREDs [Fig. 1(B)], four on a plastic plate and one on each of two side-mounted posts. Note that IREDs could move independently in the skin condition but were fixed relative to each other on the tiara.

An EMA sensor was affixed at the upper incisors ("UI") to index skull position. The sensor at the UI was affixed to the gingiva just above the UI [see Figs. 1(A) and 1(B)] either with dental glue (four participants) or embedded in a plastic dental mold (cf. Neto Henriques and van Lieshout, 2013) that was cast over the UI. The Optotrak system collects three-dimensional positional data, having a reported spatial precision of $\leq$0.2 mm (Richards, 1999). The WAVE system collects five-dimensional data (Cartesian x, y, and z positions, plus sensor azimuth and elevation orientation angles) with a static precision of 0.5 mm (Berry, 2011). Movement data recorded from both systems were registered within a common coordinate system, made possible with NDI-provided data-capture software referencing four IREDs mounted to the field-generating unit of the WAVE system.

Movement data were recorded in several conditions. In the utterance conditions, participants read a series of short utterances having the form "Do read *key* before *coup*," with 12 different italicized word pairs, at self-paced, normal and fast speaking rates. The duration of data capture on each trial was fixed at a certain number of seconds (either 7 or 10 s depending upon participant), so in the normal and fast conditions, participants repeated the short sentence a few times (2–5), with the number of times varying by participant and speaking rate. Next, to establish the likely maximum possible amount of skin movement for each participant, in the deliberate movement condition they were asked to try to move their skin deliberately and as much as possible without speaking by crinkling their forehead, raising eyebrows, etc. In the conversation condition they engaged in brief, spontaneous conversations with the experimenter. In the passage condition they read a short passage (the "Grandfather Passage," Reilly and Fisher, 2012) from a computer screen. The nine IREDs affixed to the forehead were removed and the participant then donned the tiara. The IRED at the nasion and UI EMA sensor were left in place, and the participants repeated all of the conditions outlined above.

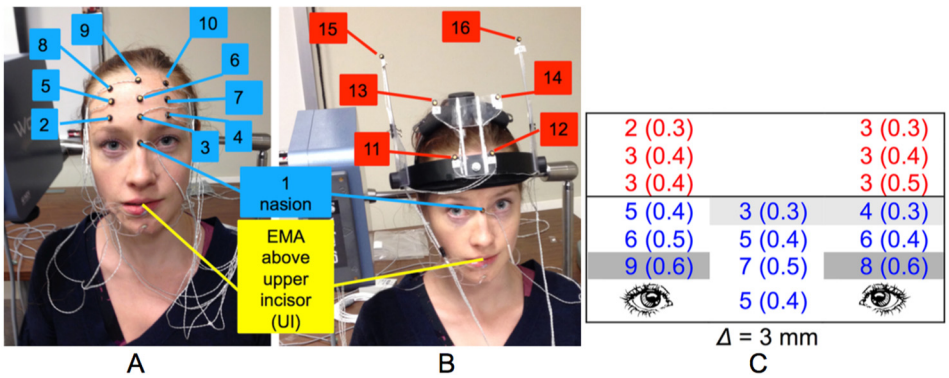

Fig. 1. (Color online) IRED and EMA sensor placement in the (A) forehead skin and (B) "tiara" methods. (C) IRED movement ranges (with SD) in mm for participant P5F in the utterance condition, normal rate.

Head movement data from EMA sensors were low-pass Butterworth filtered at 5 Hz to compensate for noise in the signal-capture process. This low cutoff for the filter was chosen because the WAVE was used to track relatively slow head movement. Optotrak data were not filtered due to the low level of noise during signal capture. Only frames when the participant was speaking were analyzed, except in the deliberate movement condition. For P3F and P8M, only skin method data were analyzed due to problems with the UI EMA sensor with the tiara. For P6M, the IRED affixed to the right tiara post [IRED 15 in Fig. 1(B)] was excluded because that IRED was out of the camera line of sight for many trials (five IREDs are sufficient for analysis).

3. Results

The first analysis determined how much the IREDs moved relative to a single reference point on the skull (at the UI) for the two methods (skin and tiara) and how much difference there was in those movements between the speaking conditions. First, the 3D Euclidian distance from each of the IREDs to the UI EMA sensor was calculated at each time frame of every trial. Distance ranges were calculated as the maximum distance between each IRED and the UI sensor minus the minimum distance. Figure 1(C) shows the distance ranges in mm (with the SD of the distance) for the IREDs in the utterance condition, normal speaking rate, for P5F (who showed the largest ranges for the skin method and typical ranges for the tiara). Ranges for tiara-affixed IREDs are in the top three rows. All other ranges are for the skin-affixed IREDs. The two skin-affixed IREDs that moved the most were those just above each eyebrow (dark shading). Two of the three most superior IREDs moved least (light shading). The difference in the IRED distances between the methods (Δ) was calculated as the mean of the distance ranges for the skin method minus the mean of the distance ranges for the tiara method. Δ for P5F was 3 mm, meaning that the skin IRED distance ranges were on average 3 mm larger than those of the tiara IREDs.

Some generalizations are possible. P2M and P4F did not show much movement ($\leq$1 mm) within or across methods in the speech conditions. IREDs placed over one (P7M) or both (P1F, P3F, P5F, P6M) eyebrows had the largest distance ranges (1–9 mm). The two IREDs that moved the least had movement ranges of 1–4 mm and always included either the nasion [IRED 1 in Fig. 1(A)] or superior-most [IREDs 8, 9, 10 in Fig. 1(A)] locations. The tiara-affixed IREDs also moved relative to the UI for all participants, with distance ranges from 1–5 mm. Δ ranged from -1 to 3 mm across participants. Distance ranges were the largest in the deliberate movement condition for all participants, as expected, ranging from the smallest single IRED range of 5 mm (P4F, nasion) to the largest of 23 mm (P2M and P5F, both above left eyebrow). There was no obvious pattern of material differences in distance ranges of skin-affixed IREDs in the other speech conditions compared to the utterances at normal rate, where a "material" difference was defined as four or more of the ten IREDs differing by more than 0.5 mm (the static precision of the WAVE system).

Since the IREDs move in both methods, it was useful to see how well the 3D position of the EMA UI sensor was predicted by the 3D positions of the IREDs in both methods. The trials were paired for each participant by content (e.g., repetitions of the same utterance at a normal speaking rate) and method. For the skin-affixed trials, a series of prediction models was created in which the position of the UI EMA sensor was predicted by the position of the IREDs. Three linear regression models were generated for each trial: one with IRED x-axis positions predicting EMA sensor x-axis positions, one predicting y-axis positions, and one predicting z-axis positions. The predicted positions of the UI EMA sensor were then calculated by subtracting the residuals of these three models from the actual x, y, and z positions of the EMA sensor. The Euclidean distance (in mm) between the predicted positions and the actual positions was calculated at each sample across the entire trial, of which a median distance was calculated. Thus, smaller distances between the predicted and actual EMA positions indicate a better estimation.

For each participant, one such model was created for each possible combination of skin-affixed IRED predictors. Figure 2(A) shows the results of these prediction models for all of the possible combinations of the ten IREDs used to predict UI position across all trials in all conditions, for participant P4F (whose pattern of results was indicative of all of the other participants while having the largest difference between predicted and actual positions). Each boxplot shows the range of median distances when all possible combinations of x number of skin-affixed IREDs were used to predict the UI position. The value above "10" on the x axis indicates the single median distance across all trials using all ten IREDs, whereas the boxplot above "1" indicates

the range of median distances across the same trials where each individual IRED was the sole predictor. The boxplot above "2" indicates the range of median distances across the same trials for all 45 possible combinations of two IREDs, etc. For all participants, the median distance between the positions predicted by the skin-affixed IREDs and actual UI positions was very small (from 0.15 to 1.27 mm) and decreased as more IREDs were used to predict UI position. For P4F and most other participants, the median distance of individual IRED predictors (boxplot above "1" on the x axis) was rather poor compared to multiple predictors. Similar models were then generated for the matched trials using the tiara predictor IREDs. In Fig. 2(A), the solid horizontal line indicates the median distance between the predicted and actual UI positions across all trials using the tiara IREDs as predictors, and the dashed lines indicate the maximum and minimum distance ranges. Figure 2(A) shows that the difference between the median range of the tiara predictors and the worst single skin predictor was <0.7 mm. When three skin IRED predictors were used, it was <0.1 mm.

These results indicate that using four markers provides a good balance between the convenience of fewer markers and the increased accuracy of using more markers. For all participants, the worst combination of any four IREDs was a better predictor of UI position than the best single IRED predictor. The median distance for four skin-affixed predictors was either less than for the tiara-affixed predictors, or no more than 0.2 mm greater, which is below the 0.5 mm level of precision of the WAVE system. Therefore, the calculated differences in the distances between the two methods are so small that they are immaterial. Another significant advantage of using three or more markers is that three are the necessary minimum for tracking the head as a rigid body, and a fourth marker provides redundancy should one marker fail, become unreliable, or become obscured (see, e.g., Hoole and Zierdt, 2010).

The question then is where those four markers should be affixed. Gauging how well each IRED predicted UI position on its own [boxplot above "1" in Fig. 2(A)], the IRED affixed at the nasion was the best single predictor for all participants. Thus one marker should be placed at the nasion, but what three other locations should be chosen remains unanswered. Figure 2(B) shows the results of a "round-robin" scoring of how many times a given predictor location was part of the best-performing combination for a given number of IREDs. A given IRED location got one point for each time that IRED location was part of the predictor set with the smallest distance between the predicted and actual UI position, i.e., it belonged to the predictor set representing the bottom of every boxplot in Fig. 2(A). This resulted in a score range from 10 points, indicating both that that IRED was the best individual predictor and part of every best combination of predictors, to 1 point, indicating that that IRED was part of the best combination only when all 10 IREDs were used. Higher numbers indicate better predictors. The four best IRED locations for each participant are shaded, with the best in dark gray. For all participants, the nasion is one of the four best-scoring predictors, but there is no clear pattern for determining the other three locations, given idiosyncratic physiological differences across participants.

To determine a recommended set of IRED locations, a count was made of the number of times each location was one of the highest-scoring locations in Fig. 2(B) (shaded). Note that references to specific IRED locations are participant-specific and qualitatively defined, similar to referring to the location of EMA sensors across participants all as being affixed to "the tongue blade." The four locations that did the best across participants by that measure (score following) were: the nasion (8), just above

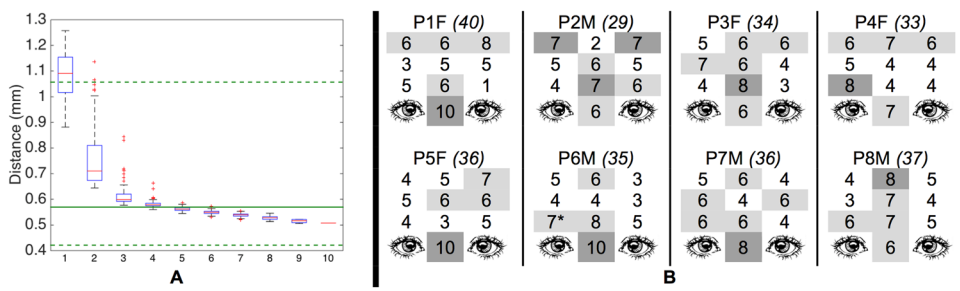

Fig. 2. (Color online) (A) Range of median distances between predicted and actual UI EMA sensor position across all trials for each combination of skin-affixed IREDs compared with tiara affixed IREDs (whose median = solid horizontal line, range indicated by horizontal dashed lines), for participant P4F. (B) Round-robin scoring of each IRED predictor (greater score = better predictors). Numbers in parentheses indicate the number of matched-trial pairs for each participant.

the nasion between the eyebrows (6), the superior midline location (6), and superior above the left eyebrow (5)—IREDs 1, 3, 9, and 10 in Fig. 1(A). For all participants, the difference in the distances between the UI position calculated using the tiara-affixed IREDs and those of these four skin-affixed IREDs were quantitatively very small (<0.25 mm) and qualitatively the same since the difference in distances were well below the static precision of the WAVE.

4. Discussion

These results have particular relevance to the collection of ultrasound images of the tongue during speech. Movement of the head relative to the ultrasound probe introduces error in the data. Two types of approaches have been used in the literature to address this problem. The first approach involves stabilizing the head relative to the ultrasound probe, either by immobilizing the participant's head (e.g., Davidson and De Decker, 2005), or by attaching the probe to a helmet that the participant wears (e.g., Scobbie *et al.*, 2008). Immobilization reduces natural head movements that accompany speech. While the helmet-attached probe avoids the issue of immobilization, it can cause discomfort and has still been found to shift up to 10 mm (Scobbie *et al.*, 2008). The second approach is head-correction (Whalen *et al.*, 2005; Aron *et al.*, 2006), which leaves the head and jaw more free to move by tracking the relative position of the head and probe using optical markers, but use of a tiara shares the potential discomfort of helmet-mounted probes. Since there was no material difference in accuracy between using forehead-affixed optical markers and markers affixed to a head-mounted tiara, the head-correction approach can be improved by using forehead-affixed optical markers, which simplifies the experimental setup and avoids potential participant discomfort.

In summary, all points on the forehead and at the nasion moved relative to the skull during speech, as did markers attached to a custom head-mounted apparatus. Four skin-affixed optical markers and markers affixed to a head-mounted apparatus predicted head position equally well. Given the additional complexity that a head-mounted apparatus introduces to the design of an experiment and potential discomfort that it can cause, our results show that the method using skin-affixed optical markers is superior.

Acknowledgments

The authors gratefully acknowledge support from NIH Grant DC-002717 to Haskins Laboratories and City University of New York. The authors thank our participants, Susan Alimonti for assistance with data collection, and Hosung Nam for assistance with analyses.

References and links

Aron, M., Kerrien, E., Berger, M.-O., and Laprie, Y. (**2006**). "Coupling electromagnetic sensors and ultrasound images for tongue tracking: Acquisition setup and preliminary results," in *Proceedings of the 7th International Seminar on Speech Production*, pp. 435–450.

Berry, J. J. (**2011**). "Accuracy of the NDI Wave speech research system," J. Speech Lang. Hear. Res. **54**, 1295–1301.

Davidson, L., and De Decker, P. M. (**2005**). "Stabilization techniques for ultrasound imaging of speech articulations," J. Acoust. Soc. Am. **117**, 2544.

Dohen, M., Lœvenbruck, H., Cathiard, M.-A., and Schwartz, J.-L. (**2004**). "Visual perception of contrastive focus in reiterant French speech," Speech Commun. **44**, 155–172.

Hoole, P., and Zierdt, A. (**2010**). "Five-dimensional articulography," in *Speech Motor Control: New developments in basic and applied research*, edited by B. Maasen and P. H. H. M. van Lieshout (Oxford University Press, Oxford), pp. 331–349.

Neto Henriques, R., and van Lieshout, P. H. H. M. (**2013**). "A comparison of methods for decoupling tongue and lower lip from jaw movements in 3D articulography," J. Speech Lang. Hear. Res. **56**, 1503–1516.

Reilly, J., and Fisher, J. L. (**2012**). "Sherlock Holmes and the strange case of the missing attribution: A historical note on 'The Grandfather Passage,'" J. Speech Lang. Hear. Res. **55**, 84–88.

Richards, J. G. (**1999**). "The measurement of human motion: A comparison of commercially available systems," Human Move. Sci. **18**, 589–602.

Scarborough, R., Keating, P. A., Mattys, S. L., Cho, T., and Alwan, A. (**2009**). "Optical phonetics and visual perception of lexical and phrasal stress in English," Lang. Speech **52**, 135–175.

Scobbie, J. M., Wrench, A. A., and van der Linden, M. (**2008**). "Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement," in *Proceedings of the 8th International Seminar on Speech Production*, pp. 373–376.

Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., and Hailey, D. S. (**2005**). "The Haskins Optically Corrected Ultrasound System (HOCUS)," J. Speech Lang. Hearing Res. **48**, 543–553.