Laboratory Phonology
Journal of the Association for Laboratory Phonology

Krause, P. A., et al. 2020 Automatic motion tracking of lips using digital video and OpenFace 2.0. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 9, pp. 1–16. DOI: https://doi.org/10.5334/labphon.232

JOURNAL ARTICLE

Automatic motion tracking of lips using digital video and OpenFace 2.0

Peter A. Krause[1,2], Christopher A. Kay[1] and Alan H. Kawamoto[1]

[1] Department of Psychology, University of California Santa Cruz, Santa Cruz, CA, US

[2] Department of Psychology, California State University Channel Islands, Camarillo, CA, US

Corresponding author: Peter A. Krause (peakraus@ucsc.edu)

Because the lips are external organs, they can be easily observed by means that are non-invasive, including simple video recording. The current paper introduces a free-of-charge, highly portable, automatic solution for extracting oral posture from digital video, based on an existing face-tracking utility. We describe how this solution might benefit various lines of laboratory phonology research, including analysis of articulatory coordination and audiovisual speech, as well as phonetic fieldwork. We then provide a tutorial, framed as a simple experiment in metronomic speech. The tutorial describes how to configure the software to work with one's digital camera of choice, how to extract relevant articulatory parameters from the face tracker, and how to approach statistical analysis. We have supplied pre-written Python scripts to aid the reader in following along with the experiment. The portability, ease, and affordability of the described solution have important implications for the accessibility of articulatory measurement, both in the lab and out in the field.

Keywords: articulatory measurement; digital video; lip aperture; motion tracking

1. Introduction

Articulatory tracking methods offer clear value to researchers in various branches of laboratory phonology. The benefits offered to phoneticians are obvious and well-established: Kinematics afford a detailed characterization of the motoric mechanisms of speech. However, as we have argued elsewhere (e.g., Kawamoto, Liu, Lee, & Grebe, 2014; Krause & Kawamoto, 2019) psycholinguists examining verbal reaction times also have much to gain from employing articulatory methods. For one thing, articulatory operationalizations of verbal reaction time often tell a very different story than do acoustic ones, in part because different speech-related movements are not obligatorily phase-locked (Kawamoto et al., 2014; Holbrook, Kawamoto, & Liu, 2019). Continuous articulatory trajectories can also be used to examine phonological expectations preceding an overt response (Krause & Kawamoto, 2019, 2020), similarly to how mouse-tracking provides insight into continuous decision mechanisms in manual reaching tasks (e.g., Spivey, Grosjean, & Knoblich, 2005).

The present paper describes a method for extracting articulatory variables such as vertical and horizontal lip apertures from digital video of participant speech, using a freely available face-tracking utility (OpenFace 2.0, Baltrušaitis, Zadeh, Lim, & Morency, 2018). In the example we will present, video data were collected using an affordable and widely available webcam, although no particular model is required. Our method does not require applying explicit reference markers to the face, nor does it require anchoring the camera relative to the face. As a result, participant preparation time is negligible. In addition, because any digital video camera can serve as the primary sensor, including those

integrated into laptop computers or cellular phones, our method could be taken outside the laboratory and into the field. In other words, this method presents a particularly low barrier to entry, and should be readily adoptable by researchers who have not yet incorporated articulatory methods into their work. In addition, it can easily complement those existing articulatory methods that permit video recording of the face.

1.1. Linguistic information accessible via lip tracking

Vertical lip aperture affords rich information. For example, while phonetic research typically associates changes in vertical lip aperture with labial consonants, non-labial consonants (and vowels) will necessarily involve target configurations in which the lips are open; the very fact of this contrast can be exploited by appropriately considered designs. We ourselves have done so when analyzing articulatory reaction times of bilabial- or alveolar-initial utterances following open- or closed-mouth initial oral configurations (Holbrook et al., 2019) and anticipatory oral postures produced when participants expect upcoming utterances to have either bilabial or nonbilabial initial consonants (Krause & Kawamoto, 2019). In some circumstances, vertical lip aperture may contrast even between non-labial consonants (such as between velars and coronals, which, given the different requirements they place on contact between tongue and palate, and the varying degree to which they receive functional contributions from the lower incisors, differently constrain jaw height, e.g., Lee, Beckman, & Jackson, 1994).

Other features of oral configuration are also informative. In similar phonetic contexts, horizontal lip aperture can be used to distinguish between rounded and unrounded vowels (due to the narrowing of the horizontal lip aperture associated with increased lip protrusion, Fromkin, 1964). Lip area has shown close correspondence to the acoustic envelope and to vocal tract resonances (Chandrasekan, Trubanova, Stillitano, Caplier, & Ghanzafar, 2009).

Furthermore, there is a long line of evidence that complex interdependencies may arise between articulatory variables that are externally visible, and ones hidden inside the vocal tract. In an early study comparing EMA to optical lip tracking, Yehia, Rubin, and Vatikiotis-Bateson (1998) found that most of the variability in vocal tract motion (~80%) was recoverable from orofacial movement alone. Similarly, we have reported preliminary evidence that the onset of labial movement reliably differs between voiced and voiceless initial consonants, suggesting possible reciprocal coordination between lips and glottis (Liu, Holbrook, Kawamoto, & Krause, under review). Other recent work has explored the implications for modern computer vision. For example, a recent computer vision study found that tongue movements could be partly predicted from dynamic observation of the external face (Kroos, Bungaard-Nielsen, Best, & Plumbley, 2017).

Additionally, visible articulatory movements contribute to face-to-face speech, in a manner that varies across contexts and populations. For example, listeners attend more closely to oral movements when there is background noise (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Speakers appear to leverage this fact, making larger articulatory motions in noisy, compared to quiet, environments (e.g., Hazan & Kim, 2013), even when silently 'mouthing' speech (Herff, Janke, Wand, & Schultz, 2011). Additionally, while mouthed speech leads to a reduction in tongue articulation, lip articulation remains distinct, presumably because of its visual salience (Bicevskis et al., 2016). Speakers striving for clear speech exaggerate their lip movements when the listener is sighted, but not when the listener is visually impaired (Ménard, Trudeau-Fisette, Côté, & Turgeon, 2016).

It is clear that measurements obtained from orofacial movement alone can be used to answer a number of relevant questions in the present literature, even without obtaining direct information from the inner vocal tract. As such, for many lines of research it is

advantageous to forgo measurements from the inner vocal tract to allow participants to speak in a more naturalistic environment. The method we propose further extends this logic, allowing participants to speak more naturally without the use of facial markers, and instead drawing measurements from digital video of ordinary speech.

It is worth noting that because this method draws from the visual information provided by the video, the participant must remain facing the camera for maximally effective measurement. Therefore, this method works best with study designs that force participants to reference an external speech cue during measurement (e.g., word or picture naming studies, in which the camera could be mounted on the computer screen used to display the words to be named). However, some leeway exists, and later portions of this paper discuss the method's robustness to the participant's distance from the camera.

1.2. Comparison with other approaches

Other techniques have previously been applied to measure speech articulation, most notably EMA and optical motion tracking with markers. Perhaps the most well-known example of the latter is the Optotrak family of systems (NDI, Waterloo, Ontario), the newest of which claim accuracy down to 0.1 mm and temporal resolutions of 4,600 Hz. Independent empirical testing of the Optotrak 3020, when calibrated to an operating range of 2–4 meters, found that in-plane motion was accurate to within .01 mm, and out-of-plane motion to within .05 mm (Schmidt, Berg, Ploeg, & Ploeg, 2009). Work concerning the accuracy of EMA found the AG-500 EMA tracking system to have a median error of 0.5 mm, and a maximum of 2 mm with a temporal resolution of 200 Hz (Yunusova, Green, & Mefferd, 2009). We provide these specifications for the sake of comparison with our present method.

Mik et al. (2018) recently explored the benefits of using conventional EMA analysis alongside a series of three high-speed optical cameras registering the movements of reflective markers at 200 fps. Their approach affords an extremely nuanced characterization of the data and allows researchers to cross-validate oral configuration estimates provided by both components of the system. While our proposed method is in some ways similar to the high-speed optical portion of the system they describe, it seeks to serve a different purpose, trading spatial and temporal resolution for affordability, portability, and the absence of facial markers.

We ourselves have elsewhere described a different video-based tracking system than the one in this paper (e.g., Holbrook et al., 2019; Kawamoto et al., 2014). That system differed in two primary respects. Firstly, explicit reference marks were applied at four key positions around participants' lips using nontoxic face paint; these paint marks were motion-tracked using a commercial video-editing product (Adobe AfterEffects CS4). Secondly, a lipstick camera was anchored to a fixed position, relative to the face, by mounting it on a steel arm attached to a batting helmet. (Because AfterEffects expresses motion-tracked positions in raw video pixels, this anchoring was used to ensure that a given distance in pixels retained the same meaning over the course of an experimental session.) The present system eliminates the need to apply explicit tracking references, by replacing Adobe AfterEffects with OpenFace (more details below). This also results in financial savings, in that OpenFace is free to academic researchers. However, by using a head-mounted camera, our previous system offers one advantage the present method cannot: the ability to track participants' lips while using tasks that do not limit their head movement. We see no reason that the head-mounted approach could not be combined with the use of OpenFace for tasks in which speech is not elicited by external stimulus cues, or in paradigms which otherwise require participants to look freely around their environment. However, when task demands can be used to orient participants' faces in

the preferred direction, the current approach offers clear benefits to participants' comfort, as well as the flexibility to take the system 'into the wild.'

1.3. OpenFace 2.0

The method here described is made possible by using the OpenFace face-tracking utility (Baltrušaitis et al., 2018). The essential operational details of that system are well-described by its authors; however, it is worth highlighting a few key points.

OpenFace is a deep-learning-based system built on top of the OpenCV computer vision framework. It estimates several values pertinent to facial behavior; the ones most relevant to our present purpose are facial landmark location and head pose. Its approach to detecting facial landmarks extends the technique of using constrained local models (CLMs) by adding a set of 'patch experts,' resulting in an approach the authors call a convolutional experts constrained local model (CE-CLM, Zadeh, Baltrušaitis, & Morency, 2017). These patch experts specialize in detecting features in a completely local fashion, irrespective of the estimated locations of other features, and in a manner robust to variability in landmark appearance. The global configuration of landmarks is then updated according to a three-dimensional point distribution model (PDM), which both controls landmark positions and penalizes misalignments. The PDM is computed using the mean values of each landmark in all three spatial dimensions, as well as rotation and translation matrices and a vector of non-rigid shape parameters (for full details, see Equation 3 in Zadeh, Baltrušaitis, & Morency, 2017).

If OpenFace is facilitated with accurate values for the intrinsic camera parameters, it can express the positions of facial landmarks in three-dimensional space (with respect to the camera origin). This is possible because OpenFace uses orthographic projection to establish correspondences between the CE-CLM's three-dimensional PDM and the two-dimensional points on the camera image. OpenFace then uses these correspondences, as well as the intrinsic camera parameters, as inputs to a direct least-squares solution to the perspective-*n*-point (P*n*P) problem (Hesch & Roumeliotis, 2011). This P*n*P solution determines the camera pose with respect to the facial landmarks. The camera pose not only permits real-world positional estimates of individual facial landmarks, but also allows the system to estimate head pose in six degrees of freedom.

1.4. Goals of the current paper

The current paper will introduce the reader to our method by applying it to a familiar paradigm in laboratory phonology: metronomic speech with alternating syllables.

We have three goals for this paper, which we will elaborate after introducing them:

1. Present a tutorial for using OpenFace effectively in laboratory phonology research.
2. Demonstrate this method's ability to detect linguistically meaningful contrasts in oral configurations produced during speech.
3. Demonstrate the robustness of lip aperture estimates obtained from OpenFace across different distances of the face from the camera.

1.4.1. Presenting a tutorial

Our description of the steps taken during our metronomic speech task will be more than usually detailed. Our description of the method in fact begins *before* articulatory measurement, with an explanation of the procedure for empirically determining the focal lengths and optical centers (along both the x and y axes) of the camera to be used. This step is necessary to maximize the accuracy of OpenFace's positional estimates. Several procedures we will describe, including camera calibration, segmentation of video into

trials, video analysis using OpenFace, extracting lip aperture trajectories from OpenFace's raw output, and trimming trials with extreme head orientation will be facilitated by custom scripts written in Python 3. Where possible, we have provided commented versions of these scripts as supplemental materials to appear with this article. (In one case, we have instead referred the reader to code appearing in an existing online tutorial, which served as the basis for our own script.)

1.4.2. Demonstrate the detection of linguistic contrasts

We will visualize and statistically analyze the data obtained from the metronomic speech task. The reader will see our method's ability to statistically distinguish the lip aperture trajectories produced for each of the four syllables in our inventory.

1.4.3. Demonstrate the robustness of positional estimates

The way OpenFace solves the P*n*P problem (see above) allows it to determine its positional estimates for facial landmarks on a world coordinate system, given with respect to the camera origin. We will demonstrate that, when defined in these coordinates, lip apertures of the same magnitude are assigned comparable estimates even when measured at different distances from the camera. Consequently, so long as participants remain generally facing the camera during a speech elicitation task, the camera need not be anchored in a fixed position relative to the head.

We have achieved this demonstration by using two identical webcams, sitting side-by-side, at the same height, and at a fixed distance of 12 inches from each other in the depth dimension. Participants were simultaneously recorded by both cameras, allowing us to directly compare measurements of the exact same utterances, measured from two different distances.

2. Method

2.1. Calibration

To make accurate estimations of the camera's position in space, OpenFace should be given parameters obtained through a camera calibration, specifically the focal length (fx and fy) and optical center (cx and cy) of the camera that was used for recording. These parameters are constant across recordings, and thus only need to be taken once unless the focus of the camera is changed. As such, any autofocus function must be turned off before calibration, and remain off as participants are run.

To obtain the parameters used for camera calibration, we employed a Python 3 script using the OpenCV computer vision library to perform a chessboard calibration. We have not included this script as a supplement to the current article, because the script we used was copied almost in its entirety from the tutorial freely available here: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_calib3d/py_calibration/py_calibration.html. We refer the reader to that tutorial for appropriate Python 3 code that can be copied, with minimal changes, into their editor of choice.

In the chessboard calibration method, camera parameters are calculated from a series of still images featuring an asymmetrical black-and-white chessboard with vertices of known dimensions, photographed at a variety of distances, angles, and locations in the camera frame. We found such an image online (specifically of a 9×6 chessboard), printed it onto standard printer paper, and attached it to a binder to ensure it stayed flat while being photographed from multiple angles. Then 25–28 pictures were taken using each camera and put through the Python script to obtain the camera parameters for use during data coding. Although there is no strict requirement for the number of pictures required to produce accurate parameters, the documentation for the OpenCV library recommends at least 10.

2.2. Participants

Six speakers from the University of California Santa Cruz participated in this study. The study was approved by the University of California Santa Cruz's Institutional Review Board, and participants gave informed consent before participating.

2.3. Design

The present design was chosen to show that this method can detect linguistically meaningful contrasts between phonemes using OpenFace's *estimated* mm as a dependent variable in a within-participants experimental design. These estimated mm should not necessarily be assumed to map directly to real-world mm. A great deal of work in the field of computer vision seeks to maximize the accuracy of computer estimates of distance despite the inherent difficulty of estimating three-dimensional distances using two-dimensional images. This challenge is known as the P*n*P problem (see Introduction). However, for our present purposes, it is extremely difficult to verify that OpenFace's estimated mm are accurate to the level required to replace more traditional physical measurements like calipers for a variety of reasons, including a) inability to match OpenFace's estimated digital landmarks with physical landmarks at mm-level precision, b) the number of participants required to prove that OpenFace's estimation works to the same level of precision across an extremely large number of potential skin tones, face shapes, and lighting conditions, and c) the fact that both measurements cannot be taken simultaneously, as using a tool like calipers would obstruct the face in video data, leading to a drop in accuracy. Therefore, we leave such validations to those doing empirical work in the area of the P*n*P problem, and restrict our own claims to the ability to detect statistically reliable distinctions between phonemes in within-participant designs, and that OpenFace's measure of estimated mm is internally consistent regardless of the distance of the speaker from the camera.

We used a 2×2 design of syllable x camera distance. Participants produced four syllables, /nu/, /mu/, /ni/, and /mi/, chosen to contrast in vertical lip separation during both consonant production (alveolar /n/ being expected to have greater vertical lip separation than bilabial /m/) and vowel production (unrounded /i/ being expected to have greater horizontal lip separation than rounded /u/). These productions were simultaneously recorded by two cameras at distances of 18 inches from participant (hereafter 'close camera') and 30 inches from participant (hereafter 'far camera'). Two dependent variables were drawn from the resulting camera footage, vertical lip separation in pixels, and vertical lip separation in OpenFace's estimated mm.

2.4. Apparatus

Speakers were recorded using two Razer Kiyo webcams shooting at 60 fps and 1280×720 resolution, each connected to a separate computer running OBS Studio 21.0. The particular model of webcam possesses an adjustable soft white light to better illuminate faces, and informal testing proved the dimmest setting to be the best compromise between improved lighting conditions and participant comfort.

The close camera was placed atop a computer monitor, and the far camera was placed atop an identical computer monitor twelve inches behind it. Due to the fact that the close camera would obstruct the view of the far camera if placed directly behind it, the cameras were placed six inches apart horizontally, the minimum required for the close camera to no longer obstruct the far camera.

In order to make the comparison across cameras as fair as possible, both were set to use the same focal length; the specific length was selected so that participants' faces would be in proper focus in the close camera. Recall that OpenFace's positional estimates rely

on the calibrated camera parameters, which themselves are partly a function of the focal length. The goal was to simulate how OpenFace would compensate for postural changes in front of a *single* camera. Therefore, using a common focal length was essential. This necessarily meant that video taken by the far camera was suboptimally focused. While this may appear to present an undesirable tradeoff, Baltrušaitis et al. (2018) have shown that OpenFace's tracking is reliable under a range of subobptimal conditions.

2.5. Procedure

Participants were seated approximately 18 inches from the closer camera, and thus approximately 30 inches from the far camera, given the 12-inch distance between the cameras. If their height made it necessary, participants were provided with an additional pillow for the chair to ensure that they were high enough to be in frame for both cameras. They were then given the instructions to produce the syllables /nu/, /mu/, /ni/, and /mi/ in time with the beat of an electronic metronome (set to 124 beats per minute) with one syllable being produced per beat. Participants were told to keep their heads still, although they were permitted to keep the beat with fingers or toes if so desired. They were permitted to practice producing syllables in time with the metronome until they were sure they understood the task and felt comfortable performing it. Participants were instructed to remain oriented toward the computer monitor during all productions.

Once participants had finished the practice, they produced ten sprints of the syllables. During each sprint, participants were instructed to remain silent for at least four beats, then begin speaking on the first beat of any subsequent measure. They then produced the sequence /nu/, /mu/, /ni/, /mi/ five times in a continuous cycle, such that each syllable was produced a total of five times. Following this, they were given a brief rest, then told to begin the next sprint of five cycles through the syllables.

2.6. Data preparation

A trained research assistant and one of the authors used HandBrake open source video transcoder and the Adobe Premiere Creative Cloud video editing software to prepare video segments of each of the participant's sprints. First, all raw video recordings were converted from variable frame rate to constant frame rate in order to correct for transient fluctuations in frame rate that would have made the videos difficult to align in the editing software. (This process was not automated via script due to the fact that HandBrake already has a batch processing feature.) Then, for each participant, the two video recordings (one from each camera) were imported into Adobe Premiere and brought into synchronous alignment using the 'synchronize by audio' feature. A small amount of post-hoc color correction was then applied to maximize the similarity of the recordings. Finally, the two videos were cut at the same ten time points, found by referencing key metronome beeps and then searching for visual evidence of the lingual closure for the first /n/, and separate videos were saved out for each of the participant's 10 sprints. Because fatigue and verbal stumbling were common, we retained each participant's best five sprints for analysis.

We used a custom Python script (Appendix A [Python Script 1, hereafter PS1]) to chop each sprint recording into separate segments for each constituent syllable. PS1 started at the beginning of each sprint-sized video and made a cut every 483 ms (i.e., 60s/124, given our 124 BPM metronome rate). PS1 then generated and executed a Unix bash file to pass each of these smaller video segments through OpenFace for processing (a Windows version that instead uses a batch file is included in the additional files). PS1 configured OpenFace with the camera lens parameters obtained earlier (see Method), in order to maximize the accuracy of positional estimates. OpenFace permits a variable level of detail in its outputs. In this case, PS1 generated a csv file containing estimates of head pose at

each frame (in the three translational and three rotational degrees of freedom), as well as facial landmark location estimates in two different coordinate systems: raw video pixels, and estimated mm. OpenFace2.0 gives pitch, yaw, and roll of the head in radians, and x-, y-, and z-translation of the head in estimated mm with respect to the camera origin. In one of the facial landmark location coordinate systems, facial landmarks are located with respect to the two-dimensional grid of video pixels. In the other, facial landmarks are in x-, y-, and z-translation coordinates given in estimated mm with respect to the camera origin.

A second custom Python script (Appendix B [Python Script 2, hereafter PS2]) iterated over all the OpenFace output files for a participant and constructed a new csv file summarizing the trial-by-trial data for that participant. For each trial (at each camera distance), PS2 extracted three head-pose parameters for every frame: pitch, yaw, and z-translation. Pitch and yaw were converted to absolute values. PS2 then computed the maximum absolute pitch, maximum absolute yaw, and mean z-translation produced over the trial, and wrote those values to the appropriate row of the summary data file. These values were used as criteria for retaining or excluding trials from statistical analysis (see below). PS2 also extracted OpenFace's tracking confidence estimates at each frame and wrote the minimum obtained value for the trial to the summary data file. This minimum value was also used as a retention/exclusion criterion.

Finally, PS2 extracted two facial landmark positions from each frame (each was extracted once from the pixel coordinates and once from the estimated mm coordinates). The OpenFace documentation gives a diagram of how each landmark is labelled, and where each appears on a parametric face: https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format. Because our analyses in this demonstration will consider vertical lip aperture, we extracted positions of the outer lips just above and below the midline; that is, landmarks 51 and 57. (If one wished to analyze horizontal lip aperture, one could extract the locations of landmarks 48 and 54, as was done by Krause & Kawamoto, 2020). For each coordinate system, PS2 computed Euclidean distance between landmarks 51 and 57 in the (x, y) plane, for each video frame. PS2 then smoothed the resulting vertical lip aperture trajectories by submitting them to a Savitzky-Golay filter, parameterized with a window size of seven samples and a third-degree polynomial. PS2 wrote the resulting smoothed trajectory values to the summary data file. **Figure 1** depicts OpenFace's facial landmarks imposed on a photograph of a face. Landmarks 48, 51, 54, and 57 have been labelled for reference.

2.7. Data trimming

As already mentioned, by using a task that oriented the face toward an external stimulus in the direction of the cameras, we helped to eliminate distortions that might arise due to suboptimal head orientation. Such distortions can be further minimized by trimming trials with extreme head orientations, as well as trials on which OpenFace's confidence in its own tracking is poor. In all the cases described below, the elimination of any trial entailed deleting both the close camera distance data and the far camera distance data. Although when originally preparing these results, we dropped trials by hand, we have since developed a Python script (Appendix C) that automates this process.

First, we dropped trials for which tracking confidence fell below 85%, for either camera distance. Second, we used the close camera data as the basis for determining trials with extreme head orientations, and dropped those. The general approach was to identify trials in which the face was angled very far off center (i.e., extreme absolute pitch and/or yaw) and trials in which the participant leaned very far back or forward, with respect to their normal range over the session (i.e., extreme z-translation). Specifically, we eliminated

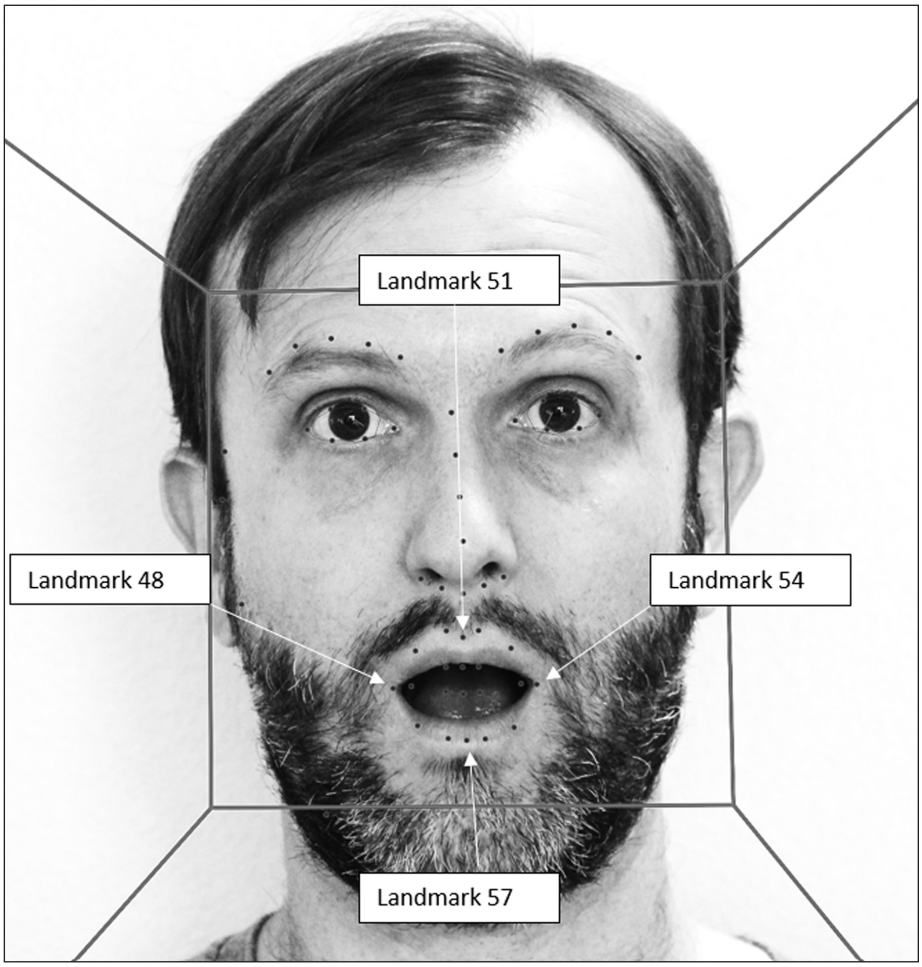

Figure 1: A depiction of OpenFace2.0's facial landmarks, as arrayed on a photograph of a real face. Landmarks critical to the determination of vertical and horizontal lip aperture have been labelled.

trials in which absolute pitch and/or yaw fell more than 2.5 standard deviations above the omnibus mean. We also eliminated trials in which mean z-translation for the trial fell more than 2.5 standard deviations above or below the overall mean z-translation for that participant. Taken together, these criteria led to the exclusion of 12 trials (2% of original data).

2.8. Statistical analysis

The goal of our statistical analyses was to illustrate two important points about this procedure: that it facilitates detection of linguistically-meaningful contrasts in lip aperture, and that its estimates in real-world coordinates are robust to changing distances of the face from the camera, eliminating the need for an apparatus that fixes this distance. To that end, our analyses will consider the contributions of camera distance and initial consonant to measurements of minimum vertical lip aperture, which should differ between bilabial and nonbilabial initial consonants.

In keeping with current best practices in phonetics and psycholinguistics, we analyzed data using linear mixed-effects (LME) models with random intercepts for participants and items (i.e., syllables), and random slopes (where possible) for fixed factors. There remains some debate about how best to specify the random effects structures of these models: While Barr, Levy, Scheepers, and Tily (2013) advocate the use of maximal models, there are reasons to believe that maximal models are at an increased risk of convergence failure

and overparameterization (Bates, Kliegl, Vasishth, & Baayen, 2018). We have therefore applied Bates et al.'s (2018) recommended simplification procedure, which begins with the maximal model and uses a combination of principal components analyses and model comparison to find a 'parsimonious' model justified by the underlying structure of the data. Experts also disagree on how to use LME models in hypothesis testing; however, Luke's (2017) recent simulations suggest that hypothesis testing using Satterthwaite- or Kenward-Roger-estimated degrees of freedom leads to lower Type I error rates than the 't-as-Z' approach or model comparison via likelihood ratio testing. We therefore performed t- and F-testing of fixed effects using Satterthwaite-estimated degrees of freedom.

3. Results

3.1. Visualization of vertical lip aperture trajectories

Two of the four syllables began with /m/ (a bilabial consonant, requiring closed lips to execute) and two with /n/ (an alveolar consonant, requiring open lips to execute). We therefore should expect minimum vertical lip apertures to be smaller for /m/- than for /n/-initial words. Our analysis will consider minimum vertical lip apertures as defined in two coordinate systems: raw camera pixels, and OpenFace's estimated mm, as computed after correcting for the z-translation component of the head pose. While we should expect camera distance to have a large impact on vertical lip aperture as expressed in raw pixels, the effect should be reduced or eliminated when expressed in estimated mm.

Figures 2 and **3** generally bear out these expectations. Both plots directly compare mean vertical lip aperture trajectories measured from the two camera distances. The graphs are paneled such that columns distinguish initial consonants and rows distinguish vowels. **Figure 2** depicts trajectories in raw video pixels, and **Figure 3** in estimated mm. The scaling of both graphs is comparable, in that the y-axes of each graph run from 90% of the minimum mean value depicted to 110% of the maximum mean value depicted.

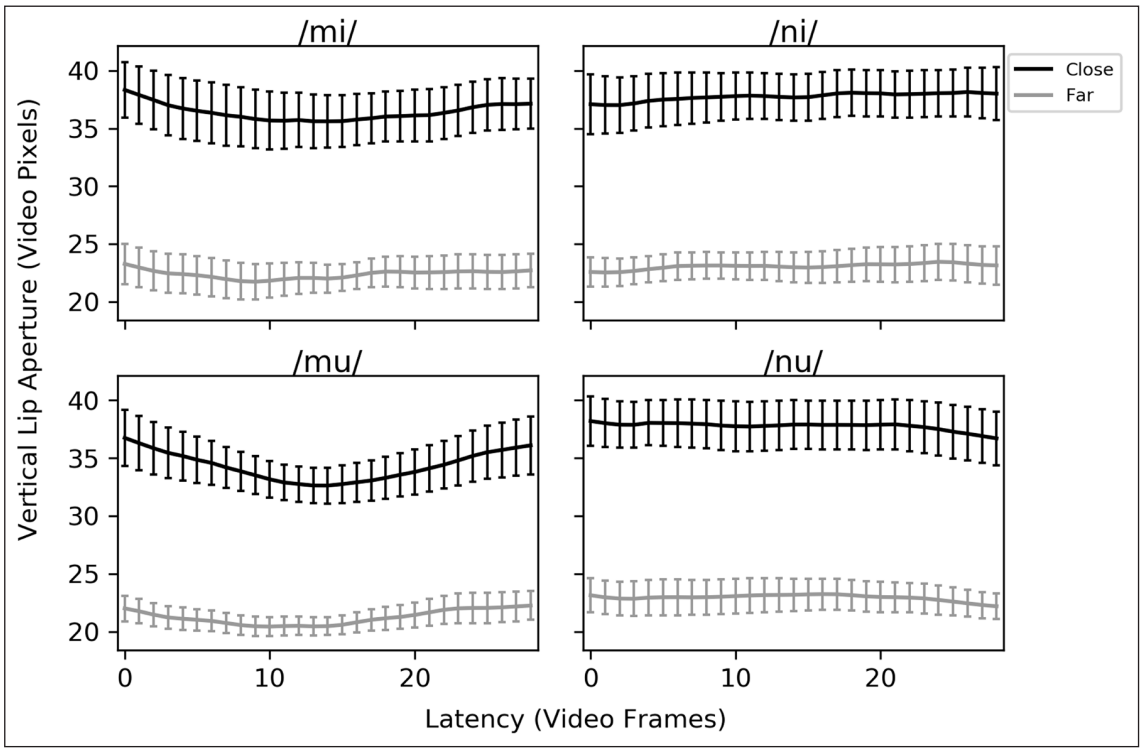

Figure 2: Mean vertical lip aperture trajectories, as measured at both distances from camera and given in raw video pixels. Each video frame spans ~16.67 ms. Error bars: ± 1 SEM.

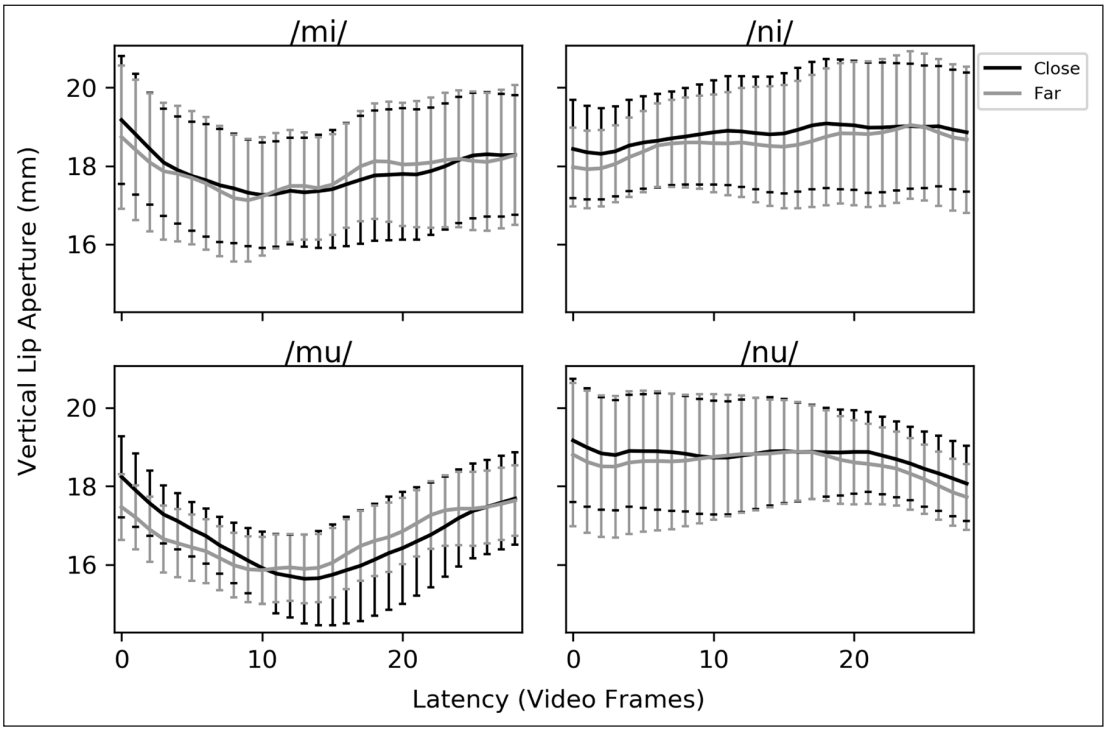

Figure 3: Mean vertical lip aperture trajectories, as measured at both distances from camera and given in OpenFace2.0's estimated mm. Each video frame spans ~16.67 ms. Error bars: ± 1 SEM.

Note that, even during bilabial closures, distances in both sets of coordinates are well above 0. This is a consequence of using outer (rather than inner) points on the lips as the basis for vertical lip aperture. Previous experience with OpenFace has led us to the conclusion that outer points are tracked more smoothly and consistently. Should a researcher wish to center complete closures of the lips on 0, they could do so by determining the mean minimum distance measured for a participant during bilabial-initial syllables, and subtracting that value from all the participant's data. However, further inner-lip tracking may be needed if the researcher wishes to estimate lip compression (which can serve as a basis for discriminating different types of closures, and possibly for detecting subglottal pressure), which may be an important consideration to weigh when deciding whether and how to apply this method.

In our case, we forewent re-centering based on bilabial-initial syllables, because either we would have had to perform the subtraction based on the data from one camera, which would have resulted in data from the other camera looking notably strange, or we would have had to re-center data from each camera separately, which would have distorted the cross-camera comparisons we were pursuing.

3.2. Tests of minimum vertical lip aperture

We constructed two LME models of the minimum vertical lip aperture produced during the trial: one with distances expressed in estimated mm and the other with distances expressed in video pixels. Both models included fixed factors for camera distance, initial consonant, and the two-way interaction. Maximal random effects structures for these models would include random intercepts for participants and items, participant-level random slopes for camera distance, initial consonant, and the two-way interaction, an item-level random slope for camera distance, and random-effect correlation terms within participants and items. **Table 1** shows the parsimonious random effects structures for each model, determined using Bates et al.'s (2018) procedure.

Table 1: Summaries of random effects structures for LME models of minimum vertical lip aperture.

	Random Effects	
Units	**Participants**	**Items**
mm	Intercept + Camera Dist. + Initial Consonant + Camera Dist. × Initial Consonant	Intercept
pixels	Intercept + Camera Dist. + Initial Consonant + Camera Dist. × Initial Consonant	Intercept

In the model constructed using estimated mm, neither the main effect of camera distance nor the camera distance × initial consonant interaction was statistically reliable, *p*s > .05. However, as expected, there was a reliable main effect of initial consonant, $t(6) = 2.89$, $p = .03$, 95% CI of the difference [0.25, 3.18].

While the failure to detect a main effect or interaction term involving camera distance is consistent with our expectation (and a desirable feature of the method), one might protest that such a null effect could be obtained by simply using an underpowered dataset. Our testing of the model constructed using raw video pixels helps put this concern to rest.

In the model constructed using pixels a reliable main effect emerged for camera distance, $t(5) = 13.88$, $p < .001$, 95% CI of the difference: [9.45, 13.62], though the effect of initial segment was only marginal, $t(6) = 2.40$, $p = .055$, 95% CI of the difference: [–0.07, 4.18]. The camera distance × initial consonant interaction was also reliable, $F(1, 5) = 9.30$, $p = .03$. Unpacking this interaction into its simple main effects found that the difference across initial consonants was not reliable at the close camera distance, $t(5) = 3.23$, $p = .07$, 95% CI of the difference: [–0.31, 5.73], or at the far camera distance, $t(7) = 1.52$, $p = .47$, 95% CI of the difference: [–1.65, 4.46]. (*P*-values and confidence intervals for the simple main effects were Tukey-adjusted.) Although the simple main effect at the close difference was not individually reliable, the general form of this interaction is exactly what one should expect given that overall dynamic range, in raw video pixels, should be reduced at the further distance. Our detection of positive results for both the main effect of camera distance and the camera distance × initial consonant interaction demonstrates that the null effects in the estimated mm model do not reflect a lack of statistical power.

4. Discussion

This paper has provided readers with a tutorial for using OpenFace, in conjunction with readily available digital camera hardware, to collect and analyze articulatory phonetic data. The empirical results we presented also provided support for two important claims: that the method is sensitive to linguistically meaningful distinctions in articulatory variables, and that the method is reasonably insensitive to incidental changes to displacement of the face from the camera. We will unpack the support for these claims consecutively.

We found that the places of articulation of the syllables' initial consonants affected the smallest vertical lip apertures produced during a syllable. As expected, vertical lip apertures were smallest when participants produced /m/-initial words, requiring a bilabial closure, than when they produced /n/-initial words, whose alveolar lingual closure left the lips free to separate. These findings suggest that lip aperture estimates derived from OpenFace's facial landmark tracking capture meaningful differences in articulatory constraints.

Our use of vertical lip aperture has been intended as a simple introduction to what is possible with this method. However, it should be readily apparent that more sophisticated articulatory measures could be easily derived from this approach. For example, by taking the first temporal derivative of instantaneous (vertical or horizontal) lip aperture, one could obtain a continuous lip aperture velocity curve. By determining a local maximum of the velocity curve's magnitude component, one could determine maximum speed. By

finding zero crossings in this velocity curve, one could determine changes in the direction of a lip aperture's motion. If a researcher is interested in questions of reaction time, they could use changes in lip aperture to determine an index of articulatory reaction time. Multiple solutions to this problem have already been proposed. For example, Holbrook et al. (2019) determined the first frame at which movement exceeded a fixed distance criterion. Mooshammer et al. (2012) is one of several papers that have used a variable criterion that defines articulatory onset as the moment at which movement speed first rises above a given percentage of the maximum speed for that trial.

We also compared identical vertical lip apertures recorded at two camera distances. We found that, when comparing the measurements of the further camera to the closer one, vertical lip apertures spanned a smaller number of video pixels, but *not* a smaller number of OpenFace's estimated mm.

We believe that adoption of the method outlined here could lead to a much broader deployment of articulatory tracking. For laboratory speech scientists who have to date relied solely on acoustic recordings, due either to cost or tractability, we hope the relative ease and affordability of this method will be all the incentive they need to supplement their approach.

There are other intriguing ways in which this method might increase the proliferation of articulatory recording. One of the most exciting possibilities is the potential to take it into the wild. Acoustic recordings have been *de rigueur* in phonetic fieldwork for several decades, but the use of articulatory recordings in fieldwork is a somewhat newer phenomenon, which until now has depended on portable ultrasound technology (Gick, 2002). Given that our modest spatial resolution was enough to reveal meaningful articulatory differences, there is nothing preventing enterprising phoneticians from taking laptops, cellular phones, and so forth outside the laboratory to use as portable recording devices. This approach offers potential benefits even to those phoneticians already using ultrasound in the field; while the primary benefit of ultrasound is in recording tongue movements, to our knowledge articulatory fieldwork has not yet involved lip tracking. We look forward to seeing what fieldwork of this kind might reveal about the phonetic richness of the world's languages.

Data Accessibility Statement

The articulatory trajectory data analyzed in this article has been made available using the Open Science Framework, and is hosted at the following DOI: 10.17605/OSF.IO/7VXT9.

Additional Files

The additional files for this article can be found as follows:

- **Appendix 1.** Python 3 script for segmenting video of a participant conducting a 'sprint' of metronomic speech into syllables, and feeding these video segments through OpenFace2.0 for facial tracking. DOI: https://doi.org/10.5334/labphon.232.s1
- **Appendix 2.** Python 3 script for extracting vertical lip aperture trajectories from OpenFace2.0 output data for each trial, and assembling the consolidated data sheet for a participant. DOI: https://doi.org/10.5334/labphon.232.s2
- **Appendix 3.** Python 3 script for eliminating trials with extreme head orientations. DOI: https://doi.org/10.5334/labphon.232.s3

Ethics and Consent

The UCSC Office of Research Compliance Administration approved this study. Human participants gave informed consent and were treated in accordance with the Declaration of Helsinki.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

The three authors made equal contributions to the current submission.

References

Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)(FG)* (pp. 59–66). Xi'an, China. DOI: https://doi.org/10.1109/FG.2018.00019

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory & Language*, *68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. Available from arXiv:1506.04967 (stat.ME).

Bicevskis, K., de Vries, J., Green, L., Heim, J., Božič, J., D'Aquisto, J., Fry, M., Sadlier-Brown, E., Tkachman, O., Yamane, N., & Gick, B. (2016). Effects of mouthing and interlocutor presence on movements of visible vs. non-visible articulators. *Canadian Acoustics*, *44*(1), 17–24.

Chandrasekan, C., Trubanova, T., Stillitano, S., Caplier, A., & Ghanzafar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. DOI: https://doi.org/10.1371/journal.pcbi.1000436

Fromkin, V. (1964). Lip positions in American English vowels. *Language and Speech*, *7*, 215–225. DOI: https://doi.org/10.1177/002383096400700402

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, *32*(2), 113–121. DOI: https://doi.org/10.1017/S0025100302001007

Hazan, V., & Kim, J. (2013). Acoustic and visual adaptations in speech to counter adverse listening conditions. *Audio-Visual Speech Processing (ASVP) 2013*. Annecy, August 29–September 1, 2013.

Herff, C., Janke, M., Wand, M., & Schultz, T. (2011). Impact of different feedback mechanisms in EMG-based speech recognition. *Interspeech, Volume 12*. Florence, August 27–31, 2011.

Hesch, J. A., & Roumeliotis, S. I. (2011). A direct least-squares (DLS) method for PnP. *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV 2011)* (pp. 383–390). Barcelona, Spain, 2011. DOI: https://doi.org/10.1109/ICCV.2011.6126266

Holbrook, B. B., Kawamoto, A. H., & Liu, Q. (2019). Task demands and segment priming effects in the naming task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *45*(5), 807–821. DOI: https://doi.org/10.1037/xlm0000631

Kawamoto, A. H., Liu, Q., Lee, R. J., & Grebe, P. R. (2014). The segment as the minimal planning unit in speech production: Evidence from absolute response latencies. *The Quarterly Journal of Experimental Psychology*, *67*(12), 2340–2359. DOI: https://doi.org/10.1080/17470218.2014.927892

Krause, P. A., & Kawamoto, A. H. (2019). Anticipatory mechanisms influence articulation in the form preparation task. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 319–335. DOI: https://doi.org/10.1037/xhp0000610

Krause, P. A., & Kawamoto, A. H. (2020). Nuclear vowel priming and anticipatory oral postures: Evidence for parallel phonological planning? *Language, Cognition, & Neuroscience*, *35*(1), 106–123. DOI: https://doi.org/10.1080/23273798.2019.1636104

Kroos, C., Bundgaard-Nielsen, R. L., Best, C. T., & Plumbley, M. (2017). Using deep neural networks to estimate tongue movements from speech face motion. *14th International Conference on Auditory-Visual Speech Processing (AVSP2017)*. Stockholm, Sweden, 2017. DOI: https://doi.org/10.21437/AVSP.2017-7

Lee, S.-H., Beckman, M. E., & Jackson, M. (1994). Jaw height and consonant place. *The Journal of the Acoustical Society of America*, *95*, 2820. DOI: https://doi.org/10.1121/1.409680

Liu, Q., Holbrook, B. B., Kawamoto, A. H., & Krause, P. A. (under review). Verbal reaction times based on tracking lip movement.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. DOI: https://doi.org/10.3758/s13428-016-0809-y

Ménard, L., Trudeau-Fisette, P., Côté, D., & Turgeon, C. (2016). Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind speakers. *PLoS One*, *11*(9), e0160088. DOI: https://doi.org/10.1371/journal.pone.0160088

Mik, L., Lorenc, A., Król, D., Wielgat, R., Święciński, R., & Jędryka, R. (2018). Fusing the electromagnetic articulograph, high-speed video cameras and a 16-channel microphone array for speech analysis. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, *66*(3), 257–266. DOI: https://doi.org/10.24425/122106

Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E., & Tiede, M. (2012). Bridging planning and execution: Temporal planning of syllables. *Journal of Phonetics*, *40*(3), 374–389. DOI: https://doi.org/10.1016/j.wocn.2012.02.002

Schmidt, J., Berg, D. R., Ploeg, H.-L., & Ploeg, L. (2009). Precision, repeatability, and accuracy of Optotrak optical motion tracking systems. *Journal of Experimental & Computational Biomechanics*, *1*(1), 114–127. DOI: https://doi.org/10.1504/IJECB.2009.022862

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398. DOI: https://doi.org/10.1073/pnas.0503903102

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Muhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception and Psychophysics*, *60*(6), 926–940. DOI: https://doi.org/10.3758/BF03211929

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1–2), 23–43. DOI: https://doi.org/10.1016/S0167-6393(98)00048-X

Yunusova, Y., Green, J. R., & Mefferd, A. (2009). Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech, Hearing, and Language Research*, *52*(2), 547–555. DOI: https://doi.org/10.1044/1092-4388(2008/07-0218)

Zadeh, A., Baltrušaitis, T., & Morency, L.-P. (2017). Convolutional experts network for facial landmark detection. Available from arXiv:1611.08657v5 [cs.CV]. DOI: https://doi.org/10.1109/CVPRW.2017.256

How to cite this article: Krause, P. A., Kay, C. A., & Kawamoto, A. H. 2020 Automatic motion tracking of lips using digital video and OpenFace 2.0. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 9, pp. 1–16. DOI: https://doi.org/10.5334/labphon.232

Submitted: 30 September 2019 **Accepted:** 06 June 2020 **Published:** 07 July 2020

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS