ACCURACY ASSESSMENTS OF HAND AND AUTOMATIC MEASUREMENTS OF ULTRASOUND IMAGES OF THE TONGUE

D. H. Whalen[1,2,3], Jaekoo Kang[1,2], Rion Iwasaki[1,2], Ghada Shejaeya[1,2], Boram Kim[1,2], Kevin D. Roon[1,2], Mark K. Tiede[2], Jonathan L. Preston[2,4], Emily Phillips[1,2], Tara McAllister[5] and Suzanne E. Boyce[2,6]

[1]City University of New York; [2]Haskins Laboratories; [3]Yale University; [4] Syracuse University; [5]New York University; [6]University of Cincinnati

dwhalen@gc.cuny.edu; jkang@gradcenter.cuny.edu; riwasaki@gradcenter.cuny.edu; gshejaeya@gradcenter.cuny.edu; bkim@gradcenter.cuny.edu; kroon@gc.cuny.edu; tiede@haskins.yale.edu; jopresto@syr.edu; emily.phillips@yale.edu; tkm214@nyu.edu; BOYCESE@ucmail.uc.edu

ABSTRACT

Ultrasound measurements of the tongue have become increasingly useful in phonetic research, but the consistency of hand measurements and the accuracy of automatic measures have yet to be extensively evaluated. Here, we examine three automatic methods (EdgeTrak, active contours (snakes) and SLURP) against hand measurements of two datasets made by three researchers. Discrepancies (errors) were calculated as the shortest distance between two contours. Results for adult speakers indicate that the three researchers were consistent with themselves and each other (mean errors < 1 mm) but higher for child data (errors < 5 mm). Automatic methods were initially consistent with hand measurements, but snakes and EdgeTrak were increasingly less consistent at later points. The results indicate that the SLURP method can be used for automatic extraction of full tongue shapes from ultrasound images. Intermittent hand correction of automatic procedures is also recommended. Future work will include more systems and more speakers.

Keywords: ultrasound, tongue shape, automatic measurements, reliability, accuracy.

1. INTRODUCTION

The tongue is a major speech articulator, but its measurement is challenging due to its location within the mouth. One of the various methods that have been devised for quantification of tongue movement is ultrasound [13, 23, 28]. This method has the advantages of being relatively inexpensive, well-tolerated by a wide range of participants [14, 27], and providing extensive coverage of the tongue. Disadvantages include difficulty of stabilization of the probe relative to the fixed vocal tract hard structure, relatively low sampling rate, and synchronizing with the audio signal.

Perhaps the major disadvantage of ultrasound is the difficulty of extracting the tongue surface from the recorded signal [1, 6, 15]. Many studies can be performed without extracting the entire shape [17, 26] or by selecting critical frames [9, 19]. However, some analyses, such as the exploration of pivots [7] or gemination processes [24], require dynamical data. In these cases, measurements of most if not all frames in an utterance are required. If measuring the required frames by hand is the ideal, it is prohibitively labor-intensive, and therefore (semi-)automated procedures for defining tongue contours are often used. Knowing the accuracy and reliability of automated measurements of the entire tongue surface is therefore crucial to many phonetic investigations, but previous systematic comparison of the existing automated procedures [e.g., 11] did not include validation of the single hand measurement that was taken as the ground truth.

We compared three semi-automated procedures—EdgeTrak [12], an edge-constrained active contour (snake) model (EPCS) [22] and SLURP [10, 11]—against the hand measurements of three phoneticians. Note that EPCS was designed for real-world visual parsing, and our novel application to ultrasound images was not a specific concern of its creators.

2. METHOD

Ultrasound images previously collected for research purposes from two different studies were re-examined for this test. One was a study of palatalization of adults' Russian consonants [20], and the other was of English-speaking children, either saying a variety of words or short sentences [18].

2.1. Stimuli

The Russian stimuli were produced by four native speakers from Moscow (aged 27-32, two female). Utterances were C_1VC_2 syllables including both actual and nonce words produced in the carrier phrase [a ɛtə _] 'and this is a _'. Stimuli included 6 word- and utterance-final consonants [t, j, s, s^j, l^j, r] where the preceding vowel was always [a] and C_1 was always [m], and 4 vowels [i, u, ɛ, a] where both flanking consonants were [p], e.g., [pap]. The first

repetition of each utterance (out of five produced by each speaker) was used for analysis. Midsagittal plane images of the lingual articulation were recorded using an Ultrasonix SonixTouch machine (BK Ultrasound, www.bkultrasound.com) with a C9-5/10 micro-convex transducer at a frame rate of 60 Hz.

The English-speaking child utterances were individual words (4 speakers, aged 4-6 years) or sentences (3 speakers, aged 10-12 years). Two of the younger children and all 3 of the older children had been diagnosed with a speech sound disorder. A Siemens Acuson X300 ultrasound with C8-5 or C6-8 transducer was used with a frame rate of 36 Hz. Stimuli analyzed for the younger speakers were 10 words with varying tongue shapes (2 repetitions per word). Stimuli for the older speakers were 5 sentences, split into 2 halves to make 10 files. Each file for both sets was measured twice (without the measurer knowing when a file was repeated).

The end of each utterance to be analysed was identified from the acoustics using annotation in Praat [2]. The 42 ultrasound video frames preceding the end of the target utterance (corresponding to ~700 ms) were extracted into individual video files. This amount of time ensured that all extracted frames were of the tongue during speech production for all but the single word productions; thus some frames were during rest. The video files were named such that the identity of the stimulus was not indicated. Each file was then duplicated so that it would be measured twice. For the Russian data, 10 stimuli x 2 copies x 4 speakers yielded 80 utterances to be analyzed. For the English data, 10 stimuli x 2 copies x 4 speakers x 2 ages yielded 160 utterances to be analyzed.

2.2. Hand measurements

The three hand measurers were phoneticians with previous experience in tracing ultrasound images of the tongue. Each measurer traced 9 contours by hand for each file: the first frame of the file, then frames 5, 10, 15, 20, 25, 30, 35 and 40 (every 83 ms for the adult data, 139 ms for the child data). Hand-measured contours were made using GetContours [25]. Measurers selected 16 "anchor points" along the underside of the white curve corresponding to the tongue surface. Each contour was traced starting at the most anterior visible point of the tongue surface behind the jaw shadow, and continuing until either the posterior end of the image or the hyoid shadow was reached. GetContours used those 16 anchor points to constrain a cubic spline of 100 equally spaced *xy* coordinates, which defined each contour. Although not all of the data points were defined by hand, the matching to the anchor points was highly constrained; therefore, we label these as hand measurements.

2.3. Automatic systems

Automatically fitted contours were created with EPCS, an active contours (snake) [22] implemented as a GetContours plugin using the hand-measured contour of the first frame of each file as the starting point ("seed"). In this implementation, the snake endpoints were constrained to lie on a line orthogonal to the last two anchor points at each end of the seed. The algorithm tracked contours for all of the frames in each file, using anchor points equally spaced along each fit for the current frame as the seed for the next.

EdgeTrak also defines contours as 100 *xy* coordinates. In order to ensure the same starting point for the automatic systems, the *xy* coordinates of the hand-measured contour for the first frame of each file were exported from GetContours and imported into EdgeTrak. The automated contour detection in EdgeTrak, also an active contour model, was then used to generate contours.

The third system tested was SLURP [10], which refines the snake algorithm by optimizing across multiple fits using a particle filtering algorithm [11]. Although still under active development, it has already shown promise, especially given that it does not need an extensive training set.

For logistical and technical reasons, it was not possible to test the automatic systems on the child data. Given the larger within-measurer error, we can predict that the automatic system would have a more difficult time than with the adult data.

2.4. Comparisons made

Hand measurements were compared to every other hand measurement, giving 1 within-measurer value and 4 across-measurer values for each frame. Automatic measures were compared to all six hand measured shapes at each of the 9 frames that were done. This allowed us to compare the change in accuracy of the automatic measures as the seed frame became more distant.

2.5. Discrepancy (error) measurements

Errors were calculated as the smallest distance between measurement points on the tongue surface. Because all systems produce 100 x-y points, each point of one surface was compared with each on the other. The average shortest distance between paired points was taken as the error. Because the length of the measured tongue surface was not always the same, it was sometimes the case that two points on one surface will be closest to the same point on the other. However, at the ends of the surface, if multiple points were closest to the endpoint of the other

surface, only the innermost point is used; extra points are excluded (see Fig. 1).

A further assessment was made using *canonical correlation* between matched contour point pairs. Similar to PCA, this method remaps the data onto two "canonical" variables such that the first maximizes correlation overall, and the second finds the remainder. The second value, reported here, represents the *lack* of correlation between given contours. As in any correlation, values range from 1 (perfect correlation) to -1 (perfect anti-correlation). It is complementary to our minimum distance measure because it reflects how well two contours track one another in *shape*.

Figure 1: Examples of error measures between two measurements of one artificially created tongue shape (nominal anterior to the right). The blue curve has 10 measurement points, the orange, 7. Arrows with solid lines indicate error magnitude. Dashed lines indicate measures at the extremes of the blue curve that are ignored. Note that two points in the middle of one curve can map onto a single point in the other curve (orange arrows).

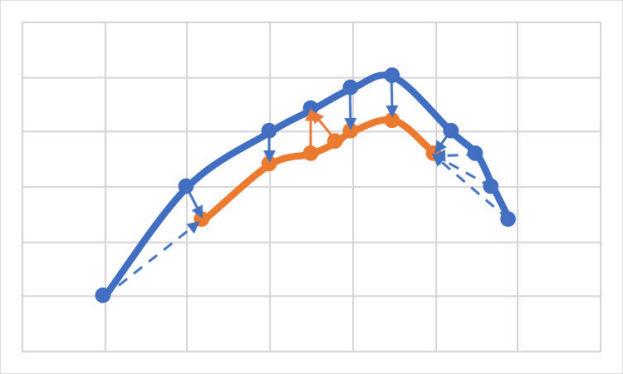


Table 1: Consistency (mean distance between tongue measurements, in mm). "Across" is mean of 8 comparisons.

Measurer	Within	Across
M1 (Russian)	0.50	0.50
M2 (Russian)	0.43	0.52
M3 (Russian)	0.41	0.55
M1 (child)	4.12	7.05
M2 (child)	5.10	----
M1 (young/old)	5.3/4.6	7.7/5.8
M2 (young/old)	4.0/4.4	----

3. RESULTS

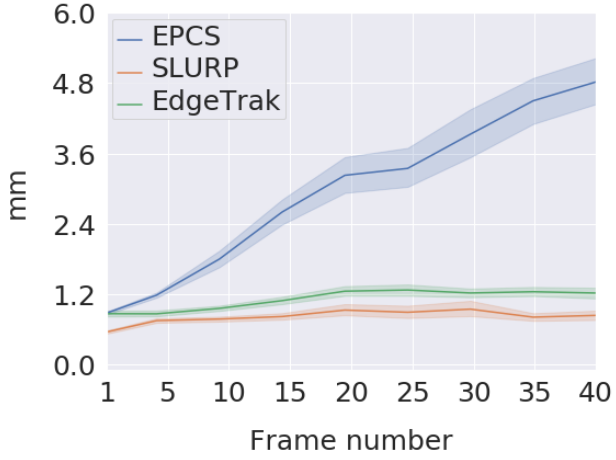
Results were considered for three tests: consistency of each hand measurer across the two tokens (within-M); consistency across hand measurers (across-M), and consistency of the automatic measurement (separately for each of the 3 systems) compared with the hand measurers for the adult data (Auto). Auto included a time series factor (levels 1-9, corresponding to frames hand-measured frames). The dependent measurement was the average error for each of the points on the tongue surface (see Fig. 1).

For the within-M, errors were at most 0.5 mm for the Russian data and 5.1 mm for the child data (see Table1). Across-M errors were, as can be expected, larger, but still less than 0.6 mm for the Russian data and 7.1 mm for the child data (Table 1). The adult differences correspond to an average of 2-3 pixels' worth of distance on the image.

Canonical correlations for adult data ranged from .93 to .97, with no further pattern apparent. Such high correlations confirm that the measurers and the automatic methods were in good agreement. For the child data, the values ranged from .83 to .87 for within- and across measurer; this is good agreement, but less so than for the adult data. Canonical correlations make fewer assumptions about the relevant portion of the surface than our first method, but still shows substantial agreement between measures and between the hand and automatic results.

Figure 2 shows the average distance between the shared portion (see Fig. 1) of the tongue edges found by our algorithms for the Russian (adult) data. The EPCS algorithm grew increasingly divergent from hand measurements in later frames. EdgeTrak had a somewhat similar trend but with smaller divergences. SLURP had little dropoff in consistency through the 40 frames measured. The magnitude of the differences was fairly small (about 5 mm for EPCS, about 1 mm for the others.)

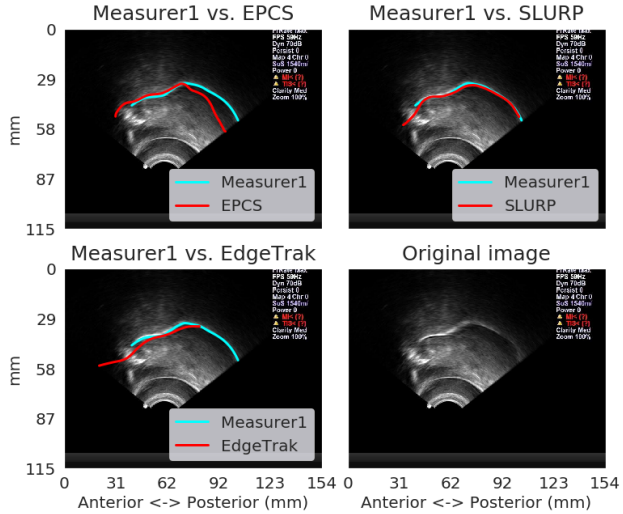
Figure 2: Change in discrepancies between automatic and hand measurements over time.



Our algorithm needs to be supplemented by an assessment of the endpoints. As can be seen in Figure 3, large discrepancies can be masked by ignoring end lack of agreement on the endpoints. SLURP was less

conservative than the hand measurement in the anterior region, while the hand measurement was less conservative in the posterior region. However, the near overlap extends over most of the tongue surface. By contrast, the EPCS algorithm was relatively close in length but differed by 10 mm for a long stretch. EdgeTrak was close in the common region, but had a sizable (ca. 2 cm) surface in the anterior region that was lacking in the hand version, and lacked almost as much in the posterior region. Thus the small deviation reported for this token is misleading.

Figure 3: Examples of discrepancies at a single frame for a single token for the three methods.



4. DISCUSSION

Our results show that hand measurements by trained phoneticians were quite reliable when the quality of the ultrasound images was good, and they support the standard in the field of using hand-measurements as the "gold standard" for defining tongue contours. However, when the image is less clear, human measurers disagree more, due to the fact that the "ground truth" is harder to establish in images with poor quality. Here, the consistency was less for the images from the child speakers, as can be expected due to the lack of head restraints and consequent movement of the speaker during a trial. Current fixed-probe systems are not usable with young children [21], and the clinical setting of the current stimuli further benefitted from the freedom from constraints. While comparable results have been found for hand-held and fixed systems for adolescents [30], the extent of movement artifacts is currently unknown, even when trials with obvious head movements are excluded [16].

The (semi-)automatic methods tested were fairly accurate for the adult data. Our current algorithm, as pointed out above, does not adequately address discrepancies in the length of the extracted surface. Revisions of the algorithm will address this issue.

In practice, researchers do not allow the semi-automatic measures to operate for 750 ms without correction, so the last measurement points in Figure 2 are simply confirmation that once a measurement diverges, it will continue to do so. The longer the method can continue without correction, however, the more useful it will be. The 40 frames that are within 1 mm for SLURP would allow a relatively quick process for ultrasound data.

There are other factors that would potentially affect the results of the semi-automated procedures, which we did not directly manipulate in this study: image quality, frame rate of the ultrasound video, frequency of adjustments to the automated contouring, etc. The degree of accuracy required for a specific purpose is also a consideration; if more approximate measures are sufficient to make a point, then added accuracy is not essential.

5. CONCLUSION AND FUTURE DIRECTIONS

We conclude that hand measurements are quite consistent for adult data (discrepancies of < 0.5 mm in our data) but less so for child data (discrepancies of < 8 mm). Automatic measures by the SLURP method, which takes a hand measurement as a seed, are quite accurate for many frames after the seed frame. For our adult data, this would often be 20-40 frames (333-666 ms).

Future work will expand the number of speakers and tokens analyzed. Comparisons are planned for the Articulate Assistant Advanced system [29], which will require changes for comparing procedures.

The length of the surface needs to be compared, with the discrepancies at both ends assessed. It is likely that a large discrepancy at one end of the surface will be considered more of an error than that same length split into the beginning and end of the surface. The algorithm for assessing the discrepancy remains to be developed, as it is not clear how much a measurement should be penalized for missing a relevant portion or, indeed, whether it is possible to create a penalty for finding a surface that is not present to the eye of the researcher.

Automatic extraction of tongue surfaces allows for greater use of ultrasound images. Although other methods, generally using PCA, that do not require edge extraction have been proposed [3, 4, 5, 8], edges provide the input to many applications. The current results indicate that (semi-)automatic methods can be accurate enough for many purposes. Our results also indicate that, for our current datasets, hand measurers can be quite consistent for adult data but less so for child data. Monitoring of automatic edges is therefore to be recommended, as is checking the consistency of hand measurements.

6. ACKNOWLEDGMENTS

Work supported by (US) NIH grants DC-002717 (Haskins Laboratories) and DC-013668 (City University of New York). We thank Chen Zhou for additional help, and Aude Noiray for helpful comments.

6. REFERENCES

[1] Akgul, Y. S., Kambhamettu, C., Stone, M. L. 1999. Automatic extraction and tracking of the tongue contours. *IEEE Trans. Med. Imaging* 18, 1035-1045.

[2] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version: 6.0.37. http://www.praat.org/.

[3] Carignan, C. 2014. TRACTUS (Temporally Resolved Articulatory Configuration Tracking of UltraSound) software suite. Version: 2.0. http://christophercarignan.github.io/TRACTUS.

[4] Hoole, P., Pouplier, M. 2017. Öhman returns: New horizons in the collection and analysis of imaging data in speech production research. *Computer Speech and Language* 45, 253-277.

[5] Hueber, T., Aversano, G., Cholle, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M. L. 2007. Eigentongue feature extraction for an ultrasound-based silent speech interface. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07.* I-1245-I-1248.

[6] Iskarous, K. 2005. Detecting the edge of the tongue: A tutorial. *Clin. Ling. Phon.* 19, 555-565.

[7] Iskarous, K. 2005. Patterns of tongue movement. *J. Phonetics* 33, 363-381.

[8] Kochetov, A., Faytak, M., Nara, K. 2018. The retroflex-dental contrast in Punjabi stops and nasals: A principal component analysis of ultrasound images. In: Yegnanarayana, *et al.* (eds), *Proceedings of Interspeech 2018.* Hyderabad: ISCA. 202-206.

[9] Kochetov, A., Sreedevi, N., Midula, K. 2012. Analysis of tongue shapes during the production of Kannada consonants. *Canadian Acoustics* 40, 30-31.

[10] Laporte, C. 2018. SLURP. Version: 0.9. https://github.com/cathylaporte/SLURP.

[11] Laporte, C., Ménard, L. 2018. Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. *Medical Image Analysis* 44, 98-114.

[12] Li, M., Kambhamettu, C., Stone, M. L. 2003. EdgeTrak, a program for band-edge extraction and its applications. In: (eds), *Sixth IASTED International Conference on Computers, Graphics and Imaging.* Honolulu, HI. 82–102.

[13] Lundberg, A. J., Stone, M. L. 1999. Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data. *J. Acoust. Soc. Am.* 106, 2858-2867.

[14] Miller, A. L., Finch, K. B. 2011. Corrected high-frame rate anchored ultrasound with software alignment. *J. Speech, Lg., Hear. Res* 54, 471-486.

[15] Noble, J. A., Boukerroui, D. 2006. Ultrasound image segmentation: a survey. *IEEE Trans. Med. Imaging* 25, 987-1010.

[16] Noiray, A., Abakarova, D., Rubertus, E., Krüger, S., Tiede, M. K. 2018 in press. How do children organize their speech in the first years of life? Insight from ultrasound imaging. *J. Speech, Lg., Hear. Res*, 1-14.

[17] Noiray, A., Iskarous, K., Whalen, D. H. 2014. Variability in English vowels is comparable in articulation and acoustics. *Lab. Phon.* 5, 271–288.

[18] Preston, J. L., McAllister, T., Phillips, E., Boyce, S. E., Tiede, M. K., Kim, J. S., Whalen, D. H. 2018. Treatment for residual rhotic errors with high and low frequency ultrasound visual feedback: A single case experimental design. *J. Speech, Lg., Hear. Res* 61, 1875-1892.

[19] Rodríguez, C., Recasens, D. 2017. An evaluation of several methods for computing lingual coarticulatory resistance using ultrasound. *J. Acoust. Soc. America* 142, 378-388.

[20] Roon, K. D., Kang, J., Whalen, D. H. submitted. Effects of ultrasound familiarization on production and perception of non-native contrasts.

[21] Rubertus, E., Noiray, A. 2018. On the development of gestural organization: A cross-sectional study of vowel-to-vowel anticipatory coarticulation. *PloS ONE* 13, e0203562.

[22] Shemesh, M., Ben-Shahar, O. 2011. Free boundary conditions active contours with applications for vision. In: Bebis, Boyle, Parvin *et al.* (eds), *Proceedings of the 7th International Symposium on Visual Computing.* Berlin: Springer. 180-191.

[23] Stone, M. L., Sonies, B. C., Shawker, T. H., Weiss, G., Nadel, L. 1983. Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. *J. Phonetics* 11, 207-218.

[24] Strycharczuk, P., Sebrechts, K. 2018. Articulatory dynamics of (de)gemination in Dutch. *J. Phonetics* 68, 138–149.

[25] Tiede, M. K. 2018. GetContours. Version: 1.3. https://github.com/mktiede/GetContours.

[26] Toda, M. 2011. Extracting kinematic properties from large-scale ultrasound corpora. In: Lee,Zee (eds), *International Conference on Phonetic Sciences.* Hong Kong: City University of Hong Kong. 1994-1997.

[27] Whalen, D. H., Shaw, P. A., Noiray, A., Antony, R. 2011. Analogs of Tahltan consonant harmony in English CVC syllables. In: Lee, Zee (eds), *Proceedings of the 17th International Congress of Phonetic Sciences.* Hong Kong: City University of Hong Kong. 2129-2132.

[28] Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E. S., Hailey, D. S. 2005. HOCUS, the Haskins Optically-Corrected Ultrasound System. *J. Speech, Lg., Hear. Res* 48, 543-553.

[29] Wrench, A. 2008. *Articulate Assistant Advanced User Guide v.2.05.*

[30] Zharkova, N., Gibbon, F. E., Hardcastle, W. J. 2015. Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. *Clin. Ling. Phon.* 29, 249-265.