DeepLabCut tongue model V4 evaluation on child data

A. A. Wrench[1,2]

[1]Queen Margaret University, UK, UK, [2]Articulate Instruments Ltd, UK
[1,2] awrench@articulateinstruments.com
ORCHID ID 0000-0003-2547-9671

Keywords: Ultrasound, keypoints, DeepLabCut, accuracy, tongue, speech, articulation

Abstract

This study aims to assess how accurately DeepLabCut [1], when applied to ultrasound tongue images, can estimate child ultrasound tongue contours. Keypoint positions of hand labelled child images were compared with the corresponding estimated keypoint positions. Mean RMSE between hand labelled and estimated positions was 0.7-1.2mm along the tongue contour and 0.5-0.9mm perpendicular to the tongue contour. Pearson correlation scores showed very high correlation. X co-ordinates along the tongue contour were in the range 0.87-0.98 and Y coordinates perpendicular to the tongue contour were in the range 0.94-0.99.

Introduction

In previous work [2] we trained and tested a DeepLabCut pose estimation model using hand labelled data. We subsequently relabelled the training set twice. In version 3 we included water swallowing data but in V4 we removed swallowing data and made a separate swallowing model. It is rather difficult to hand label tongue tip position when it extends into the mandible shadow and to discern the true surface from artifacts. The DLC tongue model V4 is based on hand labelled data from 53 speakers including 17 children. 16 with 10mm radius 5MHz microconvex probe and one with a 20mm radius 2MHz convex probe. Different Ultrasound machines and probe geometries were used. This model was hand labelled after experience observing the correlation between EMA sensors and estimated keypoints in a co-registered dataset which provided a "gold standard" reference. In this paper we ask the question: How accurate is the V4 model in estimating child ultrasound images.

Method

Hand labelled data was taken from 10 child speakers in the ultrasuite dataset who were not included in the training set. Approximately 47 frames from each speaker were hand selected to be as different in tongue shape as possible. 468 frames were labelled with 14 points amounting to 6552 labelled keypoints. 6 typically developing (TD) children form the Ultrax project and 4 children from the Ultraphonix project with speech sound disorders (SSD) were selected on the basis that the probe did not lose contact and the hyoid and mandible were in the field of view. In general, the disordered speakers showed less tongue shape change so a 6:4 rather than 5:5 ratio of TD to SSD was selected.

UltraSuite speakers

ultraphonix 10	ultraphonix 11	ultraphonix 12	ultraphonix 14	Ultrax 09	Ultrax 10	Ultrax 20	Ultrax 25	Ultrax 26	Ultrax 31

Results

Pearson correlation scores (Table 1) show very high agreement between estimated and hand labelled keypoints of 0.87 to 0.99.

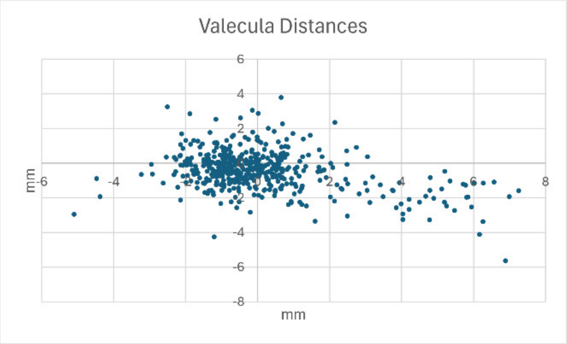
Table 1: Pearson correlations between hand labelled keypoint positions and corresponding keypoint estimates.

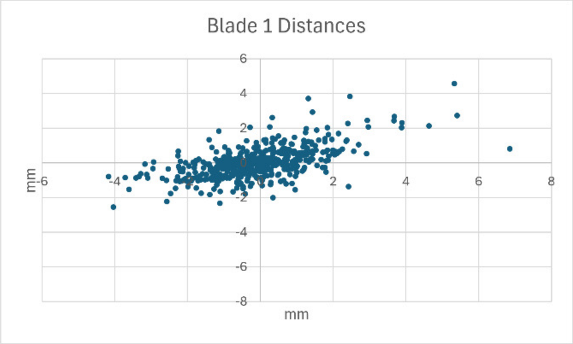
vallecula x/y 0.87/0.95	Root1 x/y 0.96/0.94	Root2 x/y 0.98/0.94	Body1 x/y 0.98/0.96	Body2 x/y 0.96/0.98	Dorsum1 x/y 0.93/0.99	Dorsum2 x/y 0.93/0.99	Blade1 x/y 0.95/0.99	Blade2 x/y 0.95/0.99	Tip1 x/y 0.96/0.99
Tip2 x/y 0.95/0.98	hyoid x/y 0.97/0.97	mandible x/y 0.96/0.97	shortTendon x/y 0.94/0.96						

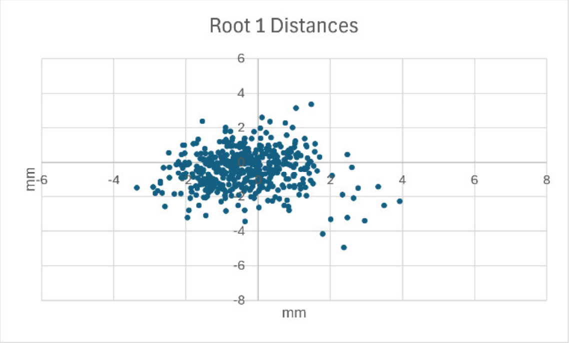
Average root mean squared error (RMSE) distances are in mm (table 2) and vary from 0.64mm to 1.21mm with standard deviations (sample) of 0.43mm to 1.03mm

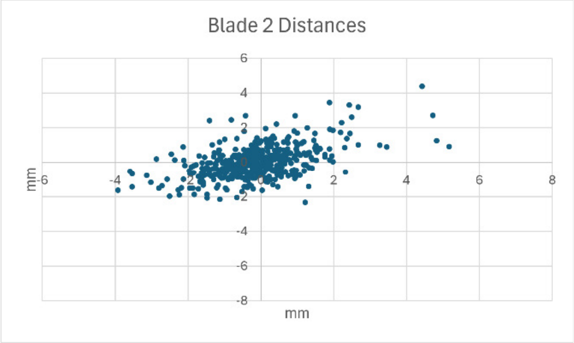
Table 2: Mean RMSE distance between hand labelled keypoint positions and corresponding keypoint estimates with standard deviations.

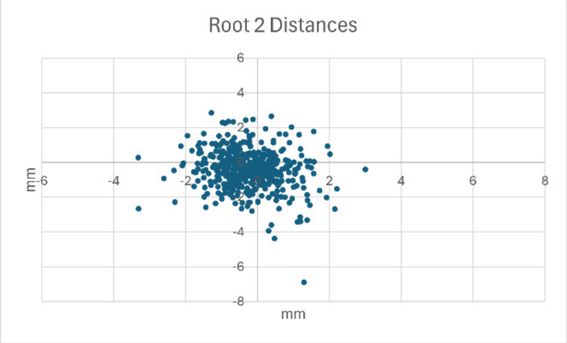
	Vallecula x/y	Root1 x/y	Root2 x/y	Body1 x/y	Body2 x/y	Dorsum1 x/y	Dorsum2 x/y	Blade1 x/y	Blade2 x/y	Tip1 x/y
mean RMSE	1.05/0.86	0.92/0.86	0.66/0.85	0.68/0.74	1.00/0.63	1.21/0.52	1.20/0.59	0.9/0.62	0.82/0.64	0.76/0.71
STDEV	0.95/0.77	0.67/0.70	0.51/0.74	0.55/0.64	0.86/0.56	0.99/0.43	1.03/0.47	0.87/0.53	0.76/0.53	0.67/0.58
	Tip2 x/y	Hyoid x/y	Mandible x/y	ShortTendon x/y						
mean RMSE	0.91/0.92	0.78/0.79	0.68/0.57	0.64/0.69						
STDEV	0.80/0.77	0.59/0.69	0.58/0.45	0.48/0.56						

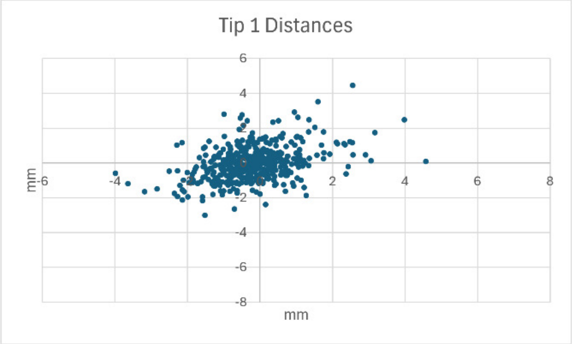

Valecula Distances


Blade 1 Distances


Root 1 Distances


Blade 2 Distances


Root 2 Distances


Tip 1 Distances

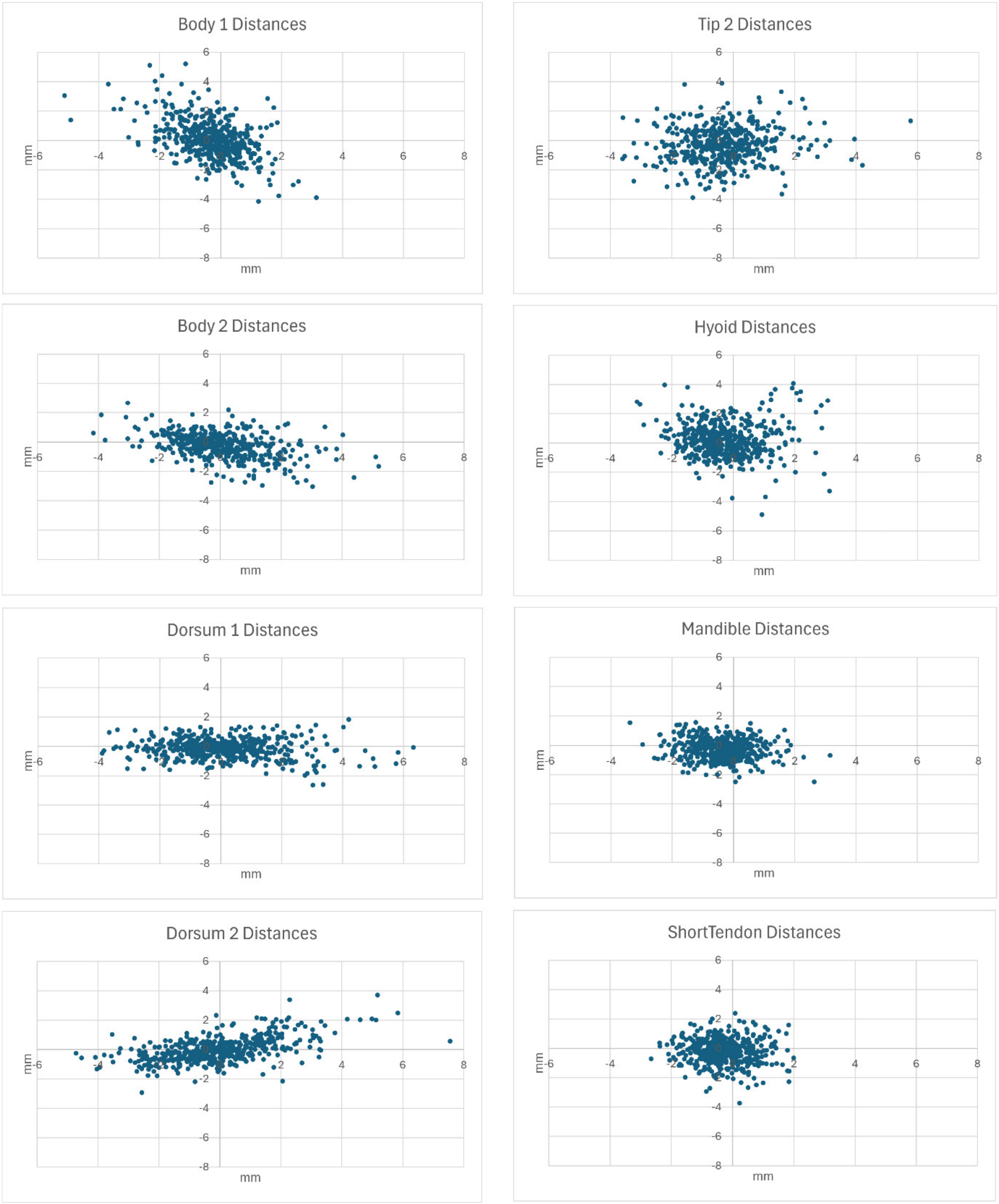


Figure 1: Shows differences in mm of 468 keypoint 2D position estimates for 14 keypoints compared to hand labelled positions

Discussion

The results show that child keypoints can be estimated from ultrasound images using the V4 model with acceptable accuracy. Figure 1 gives a visual report of the spread which, consistent with previous analysis, shows generally more spread along the contour (x) as opposed to below or above the contour (y).

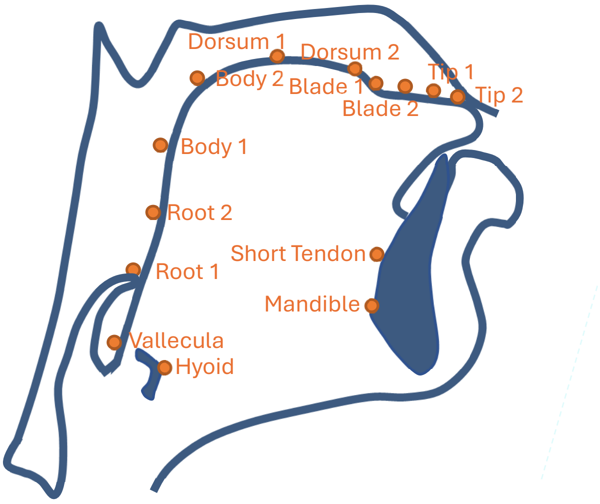


Figure 2 Location of keypoints discussed in this paper

It should be noted that the test data was recorded using an Ultrasonix microconvex probe with 10mm radius. No child data recorded with a 20mm radius 2-4MHz convex probe as provided by the MicrUs system now more widely in use has been tested in this study. The MicrUS system 10mm 5MHz microconvex probe is largely equivalent though, and these results should apply to child data recorder with it. The training data for the V4 model only includes one child speaker using a 20mm convex probe geometry. Additional training and test data of sufficient quality and a new test will be required to evaluate this probe geometry performance. The microconvex images MUST be recorded with full probe contact at all times and with the mandible and vallecula within the field of view. Also a clearly defined tongue contour. Otherwise the accuracy will be significantly less than reported here.

References

[1] Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9), 1281-1289.

[2] Wrench, A., & Balch-Tomes, J. 2022. Beyond the edge: markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut. *Sensors*, 22(3), 1133.