Improving ultrasound pose estimation accuracy

BY TRAINING ON CO-REGISTERED EMA DATA.

J. Balch-Tomes[1,3], A. A. Wrench[1,3], J. Scobbie[1], C, Macmartin[2] and A. Turk[2]

jbalchtomes@articulateinstruments.com

[1]Queen Margaret University, UK [2]Edinburgh University, UK [3]Articulate Instruments, UK

Aim

We assess how accurately DeepLabCut [1], when applied to ultrasound tongue images, can estimate Electromagnetic articulography (EMA) sensor positions.

Dataset

200 phonetically balanced sentences recorded by a male Scottish English speaker 50-60yrs. Synchronously recorded using Articulate Assistant Advanced (AAA). Instrumentation included 22kHz audio, Carstens AG501 EMA, 4 tongue sensors, upper/lower lip, jaw sensors. MicrUS 81Hz midsagittal Ultrasound. 60Hz head stabilised lip video .

Image quality and resolution

1. **Quality of the ultrasound image** Great care must be taken in configuring the ultrasound field of view, depth, frequency, dynamic range, power, contrast, probe alignment, amount of gel and firmness of probe contact. If the images are indistinct or incomplete, then the results will be adversely affected.
2. **Resolution of the ultrasound image**. The variance of the estimated keypoints is limited by the resolution of the image data provided to DeepLabCut. In a trade-off with speed of estimation, we use 320x240 pixel images which, for this dataset, corresponds to a **pixel resolution of 0.44mm.** We use 2MHz ultrasound pulses which, due to ultrasound physics corresponds to a **radial resolution of 1mm.** We use 64 scanlines to image a 101.2degree field of view. So at the tongue surface (approximately 70mm) the **angular resolution is 2mm.**

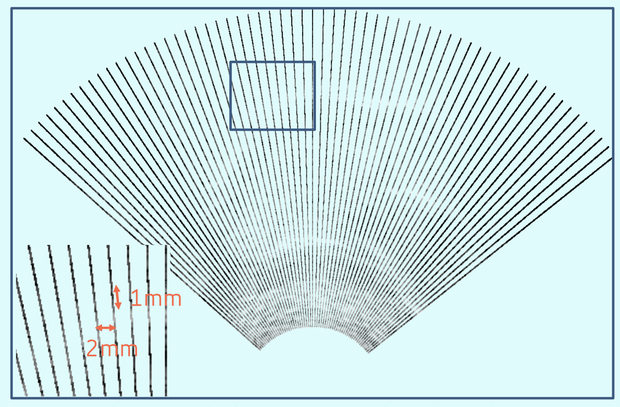

Figure 1 *Shows underlying raw scanline data that is interpolated to form the ultrasound image and the radial (1mm) and angular (2mm) resolution*

Co-registering EMA and ultrasound

Step 1. Run a **speaker independent (SI) DeepLabCut model** that is trained on hand-labelled data from lots of systems and speakers [2] to get an estimate of the keypoint positions.

Step 2. Measure the spacing in mm between the keypoints.

Step 3. Use AAA to record synchronous EMA and ultrasound, placing the sensors at the spacing of step 2.

Step 4. In AAA, create Analysis Values corresponding to the x and y positions of the keypoints and EMA sensors and x/y plot them in AAA (see figure 4).

Step 5. Create a fiducial (cyan) and interactively move it to shift the keypoints in the AAA x/y plot so they match the four corresponding EMA sensor positions (see figure 2).

Step 6. To correct for probe movement, pick a point in time and record the position of the probe origin EMA sensor. Then subtract any difference in position frame by frame from the keypoint co-ordinates.

Step 7. Check that the keypoints and EMA sensor still match and adjust the fiducial if necessary.

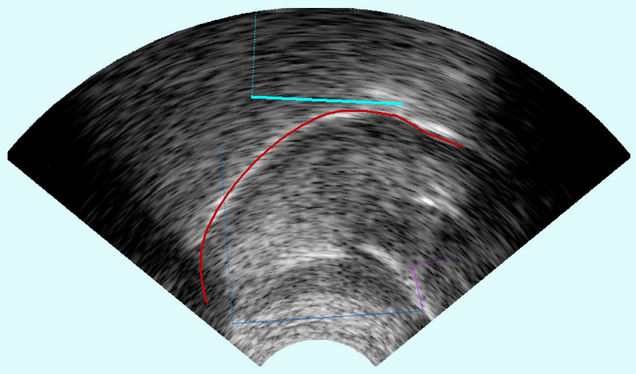

Figure 2 *The cyan fiducial is manually positioned and indicates the origin and orientation of the EMA Euclidean space*

Step 8. Calculate the reverse mapping of EMA sensors into ultrasound image pixel co-ordinates. Create DeepLabCut label files from these mapped EMA sensor positions.

Step 9. Hand label remaining training keypoints by extrapolating from the EMA data.

Step 10. Train the speaker/session dependent DeepLabCut model based on 560 frames (50ms apart from 6 sentences).

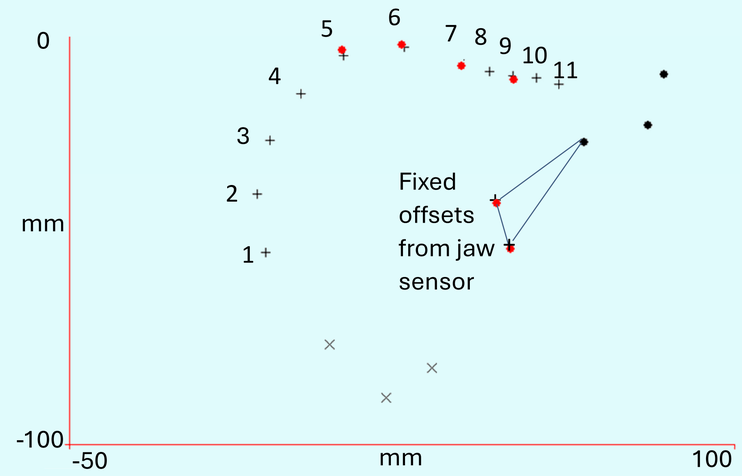

Figure 3 *EMA sensor positions and keypoints (from SI hand labelled model) for a single frame after aligning Euclidean spaces using fiducial. Short tendon and mandible keypoints assigned as fixed offsets from the EMA jaw sensor.*

Results

Comparison between the positions of the four EMA tongue sensors and the corresponding keypoints generated by the EMA trained DeepLabCut model (Step 10 above) was carried out on 1059 frames from 10 holdout sentences by the same speaker and session. Keypoint displacements were smoothed using a Savitzky-Golay filter (m=4, 100ms window).

Table 1 shows **Pearson correlation values were in the range 0.96-1.00** and the standard deviations of the keypoint errors were between 0.5 and 1.5mm.

Figures 4 and 5 show how similar EMA and keypoint kinematic data are.

By comparison, the Speaker/system Independent DeepLabCut model currently employed in AAA has Pearson correlations in the range 0.63-0.95 and standard deviations in the range 1.0 to 11.4mm largely due to poorer estimation of tip extension.

Train	Test	Test results (P) Pearson correlation scores compare EMA and keypoints (s.d.) S.D. in mm of keypoints w.r.t. EMA sensor positions				
			tip	blade	dorsum	body
560 frames	1060 holdout frames	P	x=0.96 y=0.97	x=0.97 y=0.99	x=0.96 y=1.00	x=0.96 y=0.99
		s.d.	x=1.5 y=0.9	x=1.2 y=0.6	x=1.3 y=0.5	x=1.2 y=0.6
SI model	As above	P	x=0.63 y=0.84	x=0.75 y=0.95	x=0.76 y-0.98	x=0.77 y=0.95
		s.d.	x=11.4 y=1.7	x=8.6 y=1.0	x=8.3 y=1.5	x=6.7 y-3.4

Table 1. *Pearson correlations and standard deviations*

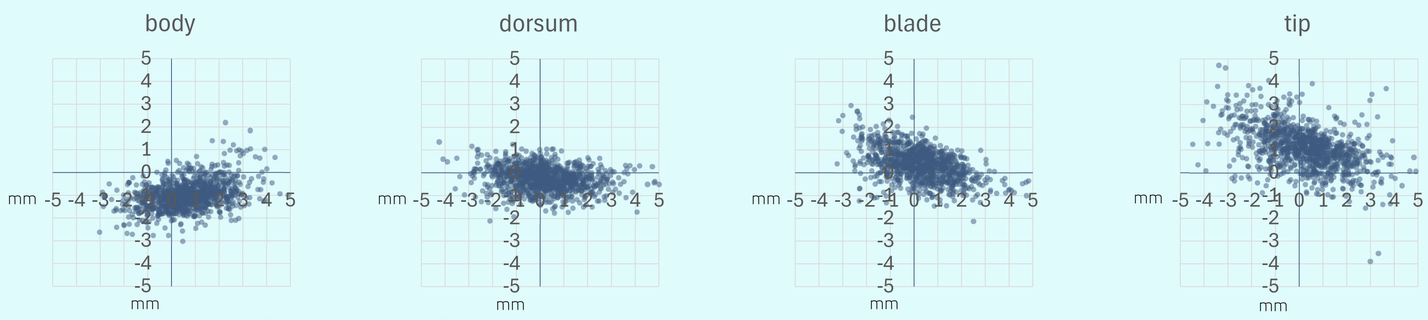

Figure 4 *Shows offsets in mm of 1059 keypoint 2D position estimates compared to each EMA sensor position*

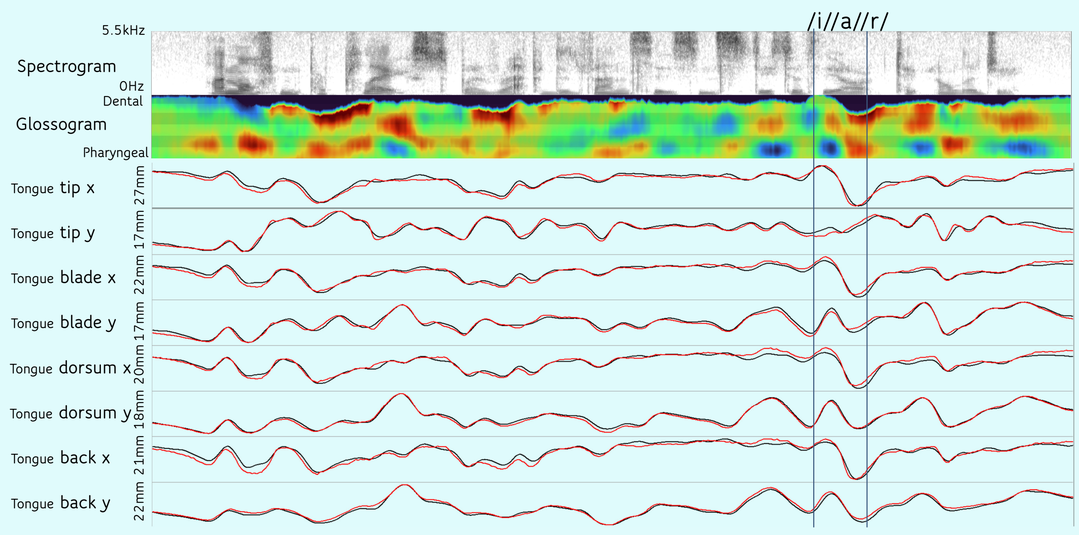

Figure 5 *Chart of x and y displacement over time of the 4 EMA sensors (red) compared to corresponding keypoints (black) for the holdout set sentence "I'm often perplexed by rapid advances in state-of-the-art technology" with probe movement correction*

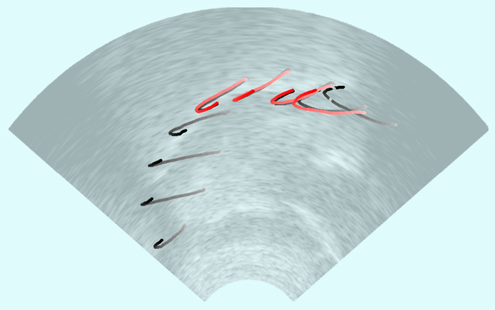

Figure 6 *Shows movement of the keypoints (black) and EMA sensors (red) over the time period of the phone sequence /i//a//r/ indicated in figure 5.*

Discussion

The results show that **kinematic data can be estimated from ultrasound images with remarkable accuracy** if probe movement is accounted for. The **standard deviations in the y axis are 0.5-0.9mm** based on an underlying radial resolution of +-0.5mm. The **standard deviations in the x axis are 1.2-1.5mm** based on an underlying angular resolution of +-1mm. Higher underlying image resolution may provide higher accuracy at a cost of more expensive ultrasound and longer processing times.

EMA is not a perfect standard. Ultrasound generally estimates the tongue surface at the base of the papillae whereas EMA sensors sit 1 or 2mm above the papillae. During a velar constriction, the velar EMA sensor can be seen to stop raising , while ultrasound shows the tongue continuing to raise to make firmer contact. During retroflexion, the tip sensor moves closer to the blade sensor than the corresponding point on the tongue surface (see Figure 7). The opposite is true for dorsiflexion.

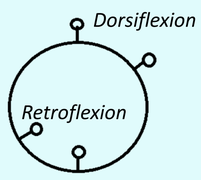

Figure 7. *Circle represents tongue surface and small circles on stalks represent EMA sensors sitting above the tongue surface*

Future

To take this forward and improve the accuracy of the SI model provided with AAA we can record more co-registered speakers, then use the resulting multispeaker model on ultrasound-only speaker recordings. Using knowledge gained from the co-registered datasets any errors can be hand corrected before using this data to broaden the training set. For best results it may be necessary to limit this model to data recorded with the MicrUS 20mm probe. It will also be necessary to monitor and correct for probe movement relative to the cranium.

Acknowledgements

Access to EMA and expertise of A. Turk and C. Macmartin funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; ERC Advanced Grant awarded to A. Turk; Grant agreement No. 101019847).

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.

articulateinstruments

Sketchfab articulateinstruments

@awrench