Head-Probe Stabilisation in Ultrasound Tongue Imaging Using a Headset to Permit Natural Head Movement

James M Scobbie*, Alan A Wrench† and Marietta L van der Linden*

*Queen Margaret University Edinburgh, †Articulate Instruments Ltd
E-mail: jscobbie@qmu.ac.uk, awrench@articulateinstruments.com

Abstract

Translation and rotation movements of an ultrasound probe relative to the speaker's head induce error. We examine one means of reducing such errors, the headset stabilising system made by Articulate Instruments Ltd., using a Vicon 612 3D motion analysis system during and between episodes of speech. Probe movements relative to the head of the speaker were derived. The headset restricted unwanted movement to the midsagittal plane even though a speaker was free to move the head naturally during utterances. Stressed low vowels moved the probe dynamically within this plane by as much as 10mm in extreme cases before it returned to near its original position. Long-term slippage ranged from 1.4mm to 2.9mm and is within acceptable limits.

1. Introduction

A major challenge for articulatory research using Ultrasound Tongue Imaging (UTI) is ensuring that the probe's 2D plane of analysis, once obtained, is consistent. The probe's orientation and position relative to the head must be maintained across sequences of frames. If the speaker moves their head relative to the probe, or if the probe moves, then two problems arise. One may be fatal for the research; the other of which is correctable if resources permit.

The worst-case scenario is that the probe translates or rotates so that different 2D slices of the tongue (i.e. outside the single targeted plane) are obtained. Such frames can be discarded if these distortions are detected, but 3D data must be gathered from the probe and speaker's head for this to be possible. A far less problematic error is rotation or translation of the image *within* the targeted plane.

The extra demands that error-detection and/or correction make on data collection are onerous for laboratory work and are impractical or impossible in many of the areas in which UTI might usefully be employed, such as field work, work with vulnerable subjects (such as children or the infirm), or where the ecological validity of the speech is important.

Various techniques have been developed to

a. immobilise the head and/or the probe
b. measure any change in probe-head alignment
c. discard or correct frames of data with movements beyond certain thresholds

Mostly, laboratories employ head immobilisation with a chair or speaker restraint [6]. When the head is immobilised, 2D video camera-based systems can be used to correct residual probe motion in the mid-sagittal plane (e.g. [5]). Hau et al. [2] are careful to note that subjects need to be compliant to ensure that they stay still in the immobilising set-up being evaluated.

The HOCUS system [7], on the other hand, does not immobilise the head, but measures the relative orientation of head and probe in order to discard or correct frames too far from the norm. They suggest thresholds for permissible error in measurements of mid-sagittal data: 5º of yaw, or rotation in the vertical axis (0.7mm estimated movement), 5º-7º of pitch (rotation of the probe-head forward in the mid-sagittal plane) and 5º-7º of roll. Lateral translation of the probe of 2-4mm is said to be acceptable. HOCUS is a motion-capture-based system which is able to correct for mid-sagittal rotation and translation (backing-fronting and lowering-raising).

A third way is to allow natural head movement, while restricting head-probe movement as much as possible (to avoid the need for systems to identify frames as requiring correction or disposal). This approach requires a headset to hold the probe stable while the head moves in a natural manner.

The specific headset used here was developed (and sold commercially) by Articulate Instruments Ltd. [1]. It clamps the probe under the chin into a holster attached to a highly-adjustable mechanical restraint system. McLeod and Wrench [4] analysed the long-term slippage of this system. They overlaid palate traces gathered during a data collection session and found slippage to be about 5mm. Here we use motion-capture to analyse the 3D movement of probe relative to head to assess both long-term slippage and short-term dynamic movement.

3. Method

Three naïve compliant non-linguist participants (s1, s2, s3) undertook tasks such as paragraph reading, nonsense word reading and spontaneous speech for about 20 minutes. The headset (Figure 1) was individually fitted with the main goals of obtaining a. an image appropriate for phonetic analysis and b. stability. A VICON 612 motion-capture system with 8 MX cameras and synchronised acoustics captured markers locations every 10ms.

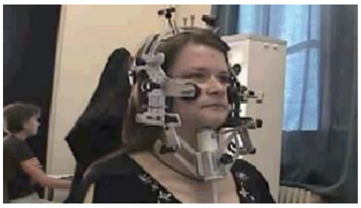

Figure 1: *Speaker wearing headset and probe with VICON markers attached.*

Head markers were on the nose and temples (widely-separated minimally moving locations). Probe markers included two on an anterior stalk, providing better 3D data on the probe (Figure 2). Markers were placed on upper and lower lips to help align speech activity to probe movement, though phonetic analysis was not intended. Articulate Assistant Advanced™ was used to capture UTI and acoustic data.

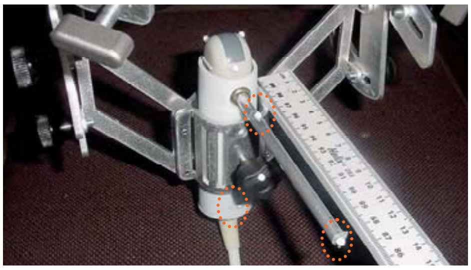

Figure 2: *Probe housed in plastic sleeve to support horizontal rod. The three probe markers (c. 5mm in size) are circled. Other markers were not analysed.*

Probe movement was transformed from a room or global coordinate system to local head coordinate system based on the temple and nose markers using a custom written MATLAB script The apparent distance between two fixed points on the probe varied due to inherent Vicon measurement error, measures as approximately 0.5mm (s1), 0.7mm (s2) and 1mm (s3) during speech.

To calculate long term slippage of the headset, marker locations were taken from a single sample point between materials, when the speaker was silent (Figure 3). The Euclidean distance from the probe marker nearest the chin to the nose marker will be presented here. Other Euclidian distances as well as movements in each of the three independent dimensions were also examined.

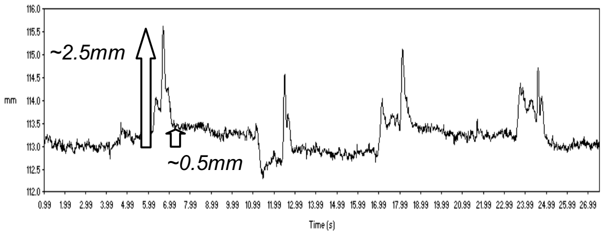

Figure 3: *Distance from nose to probe (mm) against time (s) Five rest periods flank four speech events. Approximate slippage between the first and second rest periods of 0.5mm and dynamic error of 2.5mm during the first utterance are indicated.*

Results are presented as the maximum range per speaker, and as 4 standard deviations, i.e. as 95.45% of the variance (assuming it to be normally distributed) around the mean. The 4sd figure is, we feel, more representative of error: 95% of the subject's slippage is predicted to be smaller than the value. For short-term dynamic movement during speech, *peak* error was recorded for each utterance, (The *average* error during each speech event is far less.) Peak error was calculated for each of the three possible translations and rotations, by treating both the probe and head as rigid 3D objects.

4. Results

Euclidean distance between the probe and various markers and distance in each of the three planes were negligible. Error greater than background noise was only evident in the mid-sagittal plane (Table 1).

Table 1. *Long term slippage of probe to head (Euclidean)*

	s1	s2	s3
n	11	9	9
range	1.6mm	2.0mm	1.1mm
4sd	2.2mm	2.9mm	1.4mm

Results for short term peak dynamic error are presented for backing and lowering translations (Table 2). The largest errors from individual tokens were 10mm backing and 7.5mm lowering. Due to an obscured marker, only two subjects' rotational errors are presented. In the worst tokens, pitch was maximally 1.4° (s2) and 4.8° (s3).

Table 2. *Mean peak backing (x), lowering (y) and pitch.*

	s1	s2	s3
n	8	7	9
mean trans x (mm)	5	3.5	6
mean trans y (mm)	3.4	-1.1	4.5
mean rot z (clockwise)	n.a.	1.0°	2.6°

Table 3. *Errors in lateral translation (z), roll and yaw.*

	s1	s2	s3
trans z (mm)	0.5 - 1	0.5 - 1	0.5 - 1
rot x (roll)	n.a.	< 0.5°	< 0.5°
rot y (yaw)	n.a.	< 0.5°	≅ 0.5°

5. Discussion

The long term slippage shows a smaller error than McLeod & Wrench's reported 5mm [4] – this could be due to extra error introduced by the drawing of tongue curves error in that study. Slippage in the mid-sagittal plane is within the HOCUS acceptable limits. There is no long term movement in any particular direction. Slippage in other planes is negligible.

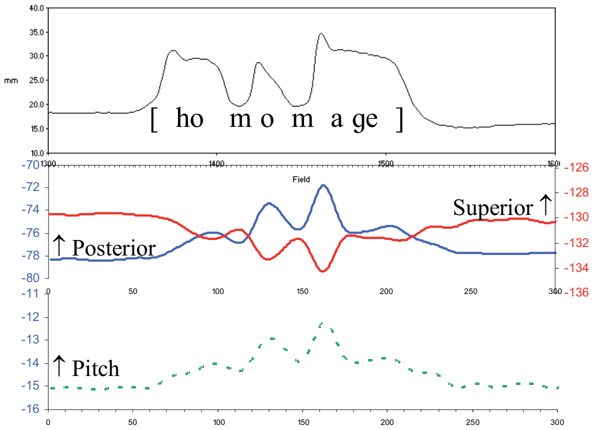

Figure 4: *Lip aperture (top); posterior (blue, left scale, mm) and superior (red, right scale, mm) translation (middle); pitch (degrees, bottom) in "ho-mo-**Ma**ggie"*

Dynamic error is also restricted to the mid-sagittal plane, so the headset removes any a priori need to discard frames for that kind of error. Peak error is quite high (Table 2), but fast-changing and short-lasting (Figure 3, Figure 4). Qualitatively it appears from close examination of lip aperture distance (and the acoustics) that the error arises because jaw lowering pushes the submental surface down on the probe, rotating the entire headset and probe slightly. In a phrase like "ho-mo-**Ma**ggie", the low stressed vowel in "Maggie" causes probe backing, lowering, and about 3° clockwise midsagittal rotation. Because the "vertical" orientation of the image is more anterior the image rotates anti-clockwise. This is what would be expected by a probe riding down on the jaw with or without a headset. In effect, during low vowels, tongue lowering will be *under-estimated* and backing exaggerated – but this is probably dependent on the nature of the materials and subject.

During continuous sentential speech, the whole system is in constant motion, and peak errors are less pronounced. Instead of a background to single point of peak error, the average location and orientation of the probe (relative to head) during 2.5 seconds of a random sentence was calculated. Error is, as informally observed, far less than the extremes reported for single stressed nonsense words. 95% of variance in the nose-probe distance during continuous speech falls between 1.3mm & 2.2mm, and Table 4 shows the detailed results.

Table 4. *Sentential mid-sagittal translation and pitch.*

	s1	s2	s3
4sd trans x (mm)	2.9	2.1	3.8
4sd trans y (mm)	1.9	0.5	2.8
4sd rot z (pitch)	n.a.	2.7 °	1.8 °

6. Conclusions

Like temporal synchronisation and spatial resolution [8], stabilisation is a crucial aspect of ultrasound data collection and should be quantified by each laboratory. This headset is capable of restricting UTI head-probe movement error to within acceptable limits [7] without any post-hoc correction. Most crucially, error occurs only in the mid-sagittal plane.

Correction within this plane is in principle possible, given 3D data on head and probe position. Such correction would be highly desirable for the high peak error, observed during stressed low vowels in isolated words. However, peak error is short-lived. This does not mean it is unimportant, just that, with the sampling rate of UTI at 30Hz, error-correction is itself inherently inaccurate. Even if the entire system is accurately synchronized, the rate of change in error is much faster than the sample rate of the video-based systems on which UTI is based (~30Hz). If correction systems are video-based, the difficulties are compounded. High speed UTI with correction also based on high sample-rate data is preferable, but in that case the demands on accurate synchronization are even greater. In our data, a single UTI frame could be matched against three VICON frames, and during 30ms, the probe can move a few millimetres. "Correction" cannot be more precise, and may be making the errors worse, because generally it is not known when the raw data underlying the ultrasound image was captured [8].

People move their heads when talking, and so a headset which lets people move naturally is inherently attractive. However, it means that any head-correction would *have* to be based on 3D data. For laboratories using immobilization, 2D data from a single fixed video camera can be used for correction (taking into account the problems mentioned above) but this assumes head and probe movement only occur in the target plane. This might be possible if speakers are compliant, but cannot be guaranteed. Speech behaviour may be affected by immobilisation. With the set-up described here, vernacular speech patterns do not appear to be affected [3].

Two of the benefits of UTI are its simplicity and portability. Headsets are a workable solution to stabilisation in a range of circumstances, from the laboratory setting with 3D correction, to situations in which head immobilisation is undesirable, or simultaneous motion capture and post-processing is impractical. Further developments in high-speed ultrasound and error correction are required for detailed quantitative research.

References

[1] Articulate Instruments Ltd. *Ultrasound Stabilisation Headset Users Manual: Revision 1.13*. Edinburgh, UK: Articulate Instruments Ltd. 2008.

[2] C. Hau et al. 3D ultrasound on a budget: Reconstruction of 3D tongue shapes from multiple coronal planes. Oral paper at Ultrafest IV, NYU. 2007.

[3] E. Lawson, J. Stuart-Smith and J.M. Scobbie. Articulatory insights into language variation and change: preliminary findings from an ultrasound study of derhoticisation. *Proceedings of NWAV 36*. In press.

[4] S. McLeod and A.A. Wrench. Protocol for Restricting Head Movement when Recording Ultrasound Images of Speech. *Asia Pacific Journal of Speech, Language and Hearing*, **11**: 23-29, 2008.

[5] J. Mielke et al. Aligning tongues and palates with Palatron. Oral paper at Ultrafest 3, Tucson. 2005.

[6] M.L. Stone and E.P. Davis. A head and transducer support system for making ultrasound images of tongue/jaw movement. *JASA*, **98**: 3107–3112, 2995.

[7] D. Whalen et al. The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech Language and Hearing Research*, **48**: 543-553, 2005.

[8] A.A. Wrench and J.M. Scobbie. Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. *Proceedings of ISSP 06*, 451-458, 2006.

This research was partly funded by ESRC (R000-22-2032) and SHEFC SRIF. Many thanks to our subject participants.