Very High Frame Rate Ultrasound Tongue Imaging

Alan A. Wrench[1], James M. Scobbie[2]

[1]Articulate Instruments Ltd – Queen Margaret University Campus,
Musselburgh, EH21 6UU, Scotland, UK

[2]CASL (Clinical Audiology, Speech and Language) Research Centre,
Queen Margaret University, QMU Drive, Musselburgh, EH21 6UU, Scotland, UK

awrench@articulateinstruments.com

Abstract. *This paper examines the trade-off between temporal and spatial resolution in ultrasound tongue images at fast frame rates. The fastest lingual speech movements are investigated using a variety of echo pulse densities. Benefits and drawbacks of using higher frame rates are considered. Faster frame rates reduce distortion of the shape of the tongue during highly dynamic segments but it becomes increasingly difficult to discern the detail of that shape. The best temporal and spatial resolution is achieved with shorter distances between the probe and the tongue surface.*

1. Introduction

The high-speed ultrasound system at Queen Margaret University is uniquely capable of capturing images at very high frame rates (>300Hz) and automatically aligning them with the acoustic speech signal and other speech production instrumentation. As with all ultrasound systems, an increased frame rate comes at a price: for a fixed depth setting, either the scan-line density must be reduced or the field of view (FoV) must be reduced, meaning images have a either a lower spatial resolution or a narrower field of view. The maximum achievable frame rate depends on how many ultrasound echo pulses (shown in figures 2a, 3a and 4a as visible scan-lines) are used to make up each image and how long it takes to process each received pulse.

The fundamental physical limit, based on the speed of sound in tissue, is 13 microseconds per 10mm penetration. With the system described here, there is in addition a processing overhead of 27µs per frame, while in other systems the overhead will be different. At an 80mm depth setting, each echo pulse takes 8x13µs+27µs (131µs total) to process. The QMU system can generate an image made up of 25 echo pulses at a frame rate of 306Hz; or an image made up of 76 echo pulses at a rate of 100Hz; or an image made up of 152 echo pulses at 50Hz, etc. The angular spatial resolution is then dependent on the selected field of view, which in this specific system may range from 50 to 150 degrees, and the distance of the tongue surface from the transducer (because scan-lines diverge as they extend outwards from the convex transducer). Comparable trade-offs exist in every laboratory's ultrasound system, though a small set of manufacturer's pre-set options may define the limits of flexibility.

This flexibility prompts the question: how fast is fast enough? What sample rate is best suited for ultrasound tongue imaging? Before moving to this question, we will briefly cite some findings on articulator speed from previous EMA and acoustic studies.

Trills: Shosted (2008) and Ladefoged (1977) report that apical trills are produced at a frequency of ~25Hz. Uvular trills can have a higher frequency (in the range 26-33Hz) but in this case it is the uvula which is the principal vibrating mass and the uvula cannot be observed by ultrasound except for a faint signal when it contacts the tongue.

High speaking rate: Jannedy et al. (2010) used EMA sampling at 200Hz to study a speaker who was officially recognized for speaking German tongue twisters at the fastest syllable rate. They recorded the tangential velocity of a sensor attached approximately 10mm from the tongue tip for typical and fast speaking rates. Speaking rate made no significant difference to peak tangential velocity. The segment /l/ had the highest speed (~900mm/s), closely followed by /t/ closure (~800mm/s).

Taps/Flaps: Jannedy et al. (2010) included taps as allophonic variants of /t/. Taps therefore may fall into the group with velocities in the range 400mm/s to 800mm/s.

Click releases: Sharf et al. (1995) used EMA sampling at 200Hz to observe (alveo-) palatal click movement in Xhosa. The Tongue tip coil was placed 5mm from the tip. They report a maximum tongue tip velocity of 1094 mm/s.

2. Method

Ultrasound data was acquired using an Ultrasonix SonixRP machine remotely controlled via Ethernet from a PC running Articulate Assistant Advanced software™ (Articulate Instruments, 2010). The transducer is a short-handled paediatric microconvex probe operating at 5, 6 or 8MHz. The system allows full control of frame rate (up to 386Hz), Field of View (up to 150°), echo pulse density and depth. The ultrasound machine generates a hardware pulse at the instant that each sweep of echo pulses is completed. This synchronization pulse sequence is recorded on a multichannel analogue acquisition system at 22,050Hz along with the acoustic speech signal. These synchronization pulses are then detected in a post processing operation allowing each ultrasound frame to be accurately time tagged relative to the audio stream. The system (Figure 1) captures other channels of data, such as an NTSC video stream and EPG.

Unlike most ultrasound scanners, the SonixRP allows us access to the set of echo pulse vectors before it is turned into an image. The ultrasound data is therefore stored as a set of M-mode type vectors (Figure 2a) rather than the conventional 2D interpolated image (Figure 2b). This interpolation stage is normally built into the ultrasound scanner and the actual number of echo pulses used to create the image is not apparent to the user. (The echo-pulse density, or equivalently, the scan-line density, in other systems can be estimated using the length of the speckle arcs resulting from interpolation.) Access to raw echo return vectors has three benefits. First, it provides access to the actual data before interpolatory smoothing providing the potential for efficient edge detection. Second, it minimizes the storage requirements that become onerous at very high frame rates if full 640x480 images are stored. Third, it reduces the time taken to process and transfer ultrasound data, speeding up the recording session.

The echo return data is transferred to the remote PC via Ethernet. On viewing data, a standard graphical interpolation is performed on the raw data to convert it to an image for analysis in AAA, similar to the image processing that is normally carried out

within the ultrasound scanner (see below). The ultrasound data is transferred piecemeal during a recording to provide real-time feedback, which is fast enough to provide a smooth live display. However, since frames can be dropped, a block transfer is made at the completion of each recording for permanent storage with complete reliability.

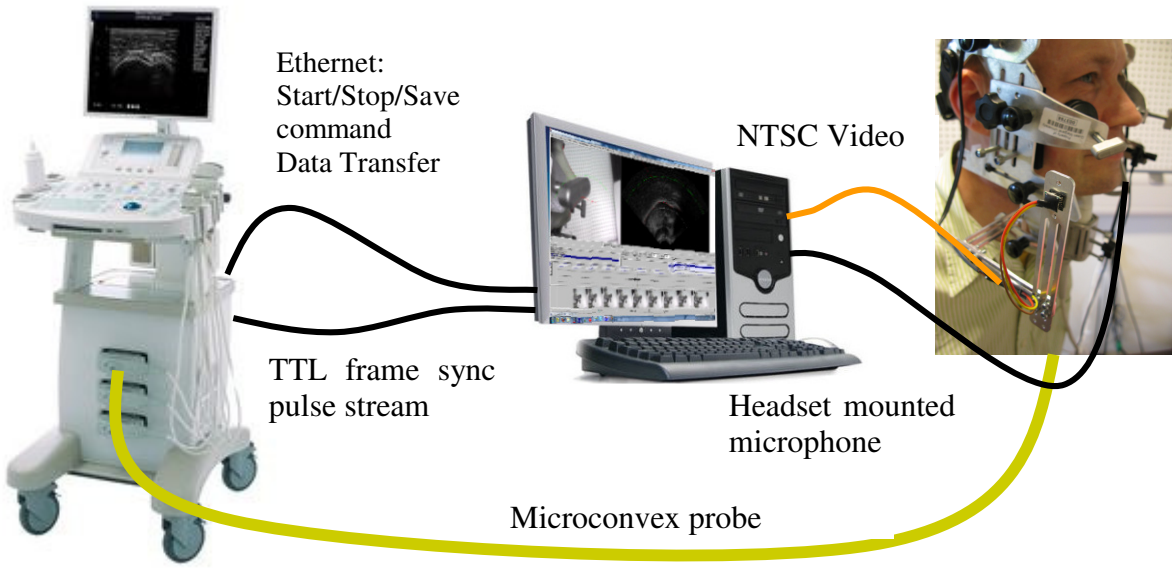


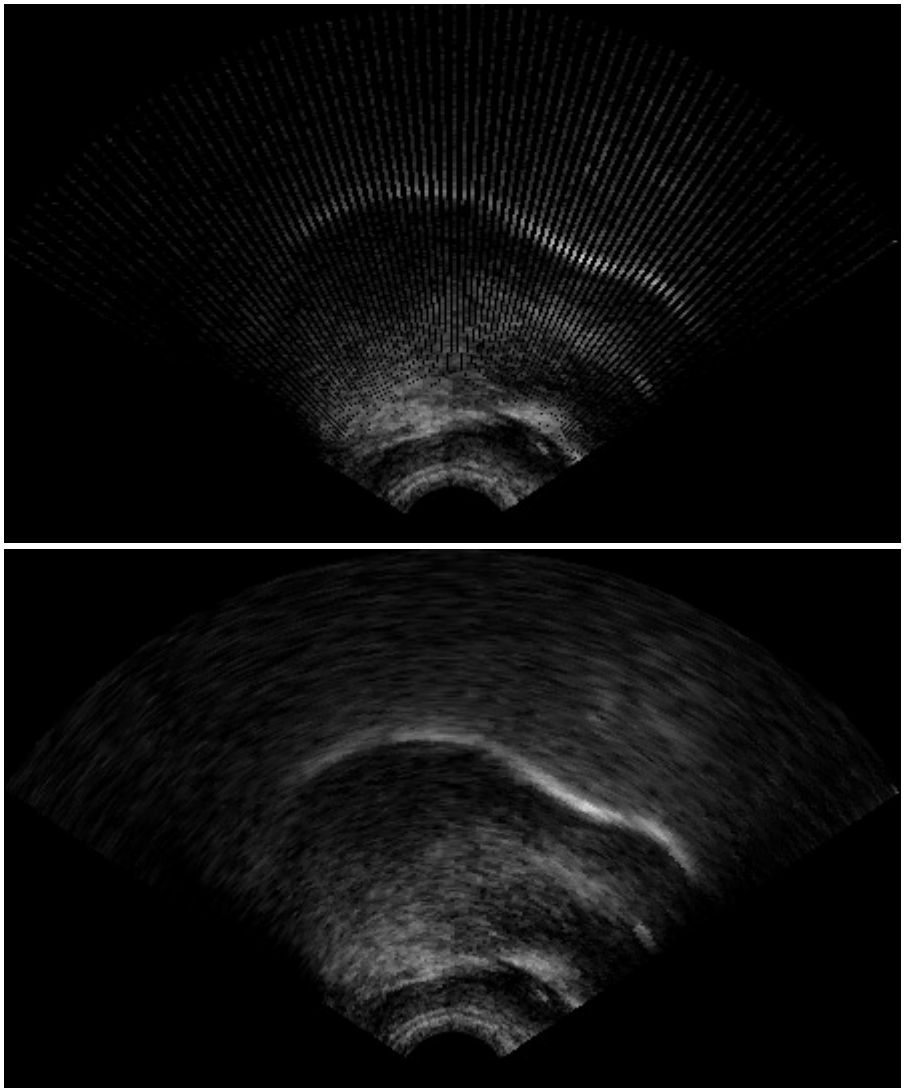
Figure 1. System configuration for recording ultrasound data, audio, sync pulse stream and NTSC video from a headset-mounted camera.

For this investigation the depth setting was set to 80mm and the echo return vectors had 412 discrete samples (approximately 5 pixels per mm). The transducer frequency was set at 5MHz providing an axial (i.e. radial) resolution of approximately 0.9mm. The probe frequency therefore determines the measurable axial resolution in this case.

Figures 2a, 3a and 4a below show the reduction in angular resolution as the frame rate is increased and the depth (80mm) and FoV (Field of View) (112°) settings are kept constant. Figures 2b), 3b) and 4b) show the interpolated images derived from this raw data. Despite their apparent continuity, they have exactly the same spatial resolution, and exactly the same reduction in spatial resolution at higher frame rates, as the uninterpolated images. Figure 2's scan at a 100Hz sample rate has 76 visible scan-lines, one from each echo pulse. Figure 3 shows a 196Hz scan with 39 scan-lines. Finally, Figure 4 shows a scan with a 306Hz sample rate and 25 scan-lines.

Since the scan covers a 112° FoV, the angle between scan-lines / echo pulses in Figure 4 is 112/24 = 4.67°. At a depth of 5cm, the distance measured along an arc between echo pulses is 2*π*(depth + distance from virtual origin to probe surface)*4.67/360 = 4mm. So this means that although the axial (i.e. radial) spatial resolution along each echo pulse remains high (0.9mm) the angular resolution is significantly lower (4mm at depth of 40mm from the probe surface and 5.6mm at 60mm from probe surface). Even at slow frame rates, the angular resolution may be lower than the axial resolution, and always worsens as distance from the probe increases. Table 1 shows a range of angular resolutions for different settings.

a)



b)

Figure 2. 76 scan-lines 100Hz frame rate. a) raw echo pulse data b) image constructed from echo pulse data, interpolated in arcs around the scan-lines to fill in the gaps in the image. FoV 112°.

Despite the low density of echo pulses, the image quality at 306Hz frame rate (Figure 4) is sufficient to allow the full contour of the tongue from root to tip to be extracted and a smooth interpolated image created. Up to 25 data points per contour could be directly measured from the tongue. Tongue contours can be extracted with high temporal accuracy, while the different axial and angular resolutions give rise to interesting effects. The spatiotemporal quality possible is exemplified using an isolated production of a click (Figure 5) (Automatic contour fitting was performed using AAA software). The rarefaction of the air trapped between the tongue and the palate can be seen to comprise both a lowering of the tongue blade and a sliding of the tongue tip in a posterior direction along the palate. Once the seal is broken, the tip springs forward reaching velocity around 1.2m/s (estimated by assuming the tip moves in an arc) and recoils from the floor of the mouth at the end of the articulation.

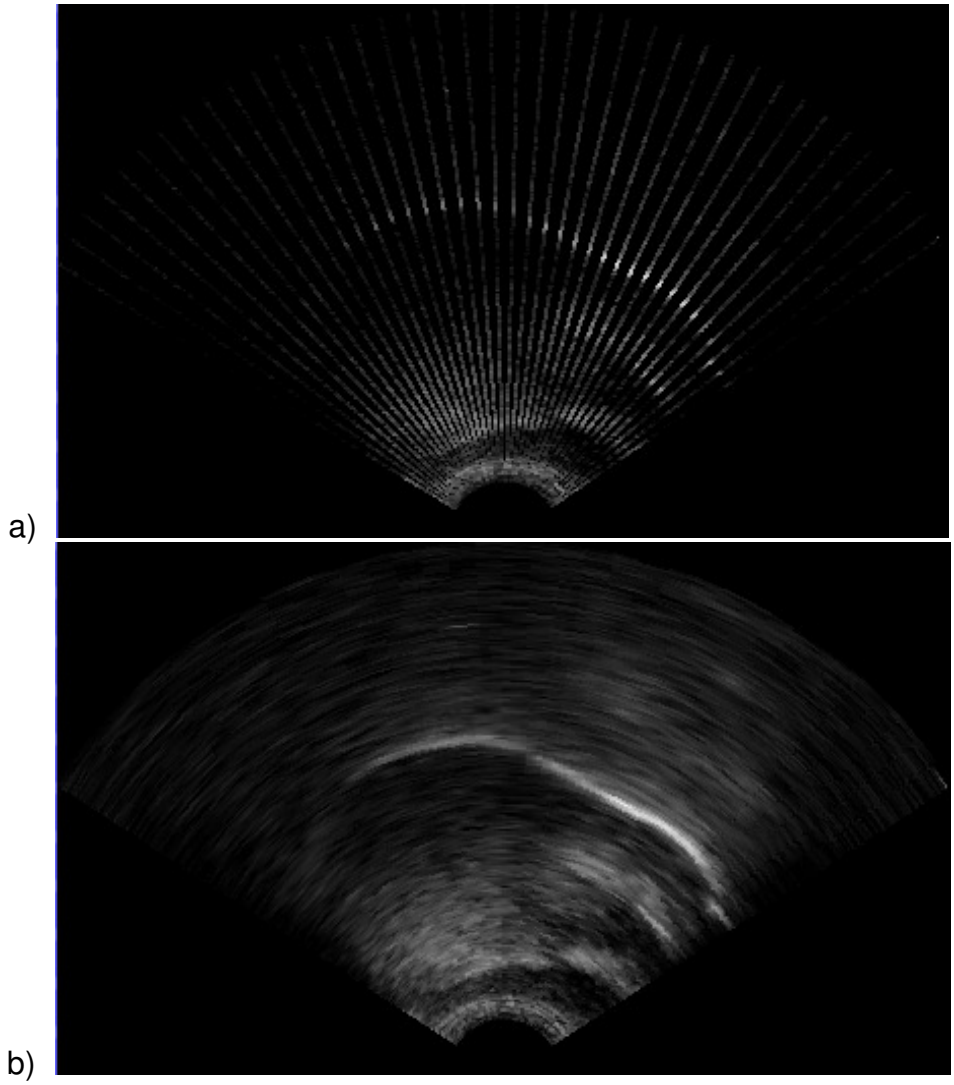


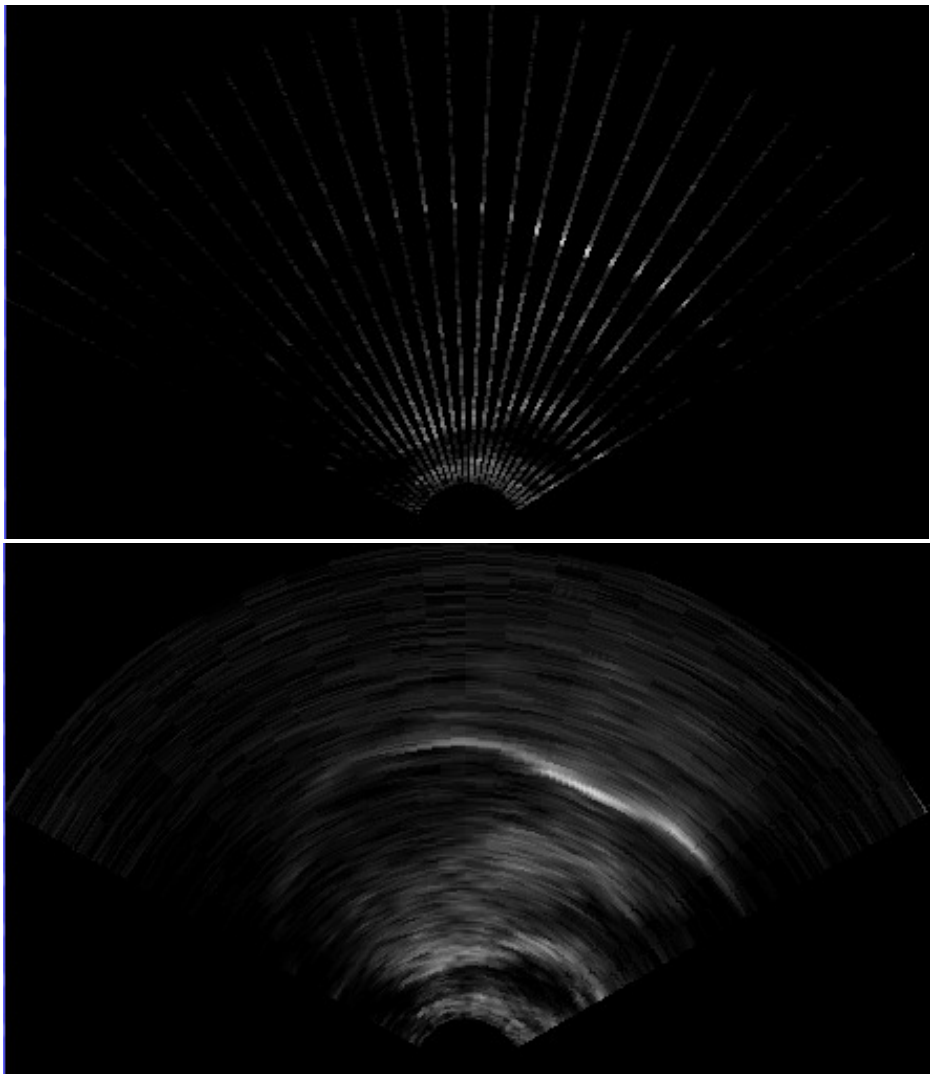
Figure 3. 39 scan-lines 196Hz frame rate. a) raw scan-line data b) image constructed from scan-line data interpolated in arcs around the scan-lines to fill in the gaps in the image. FoV 112°.

On inspection, the reduced angular spatial resolution in Figure 5 is rather apparent. During the rarefaction phase, in this example, the tongue body stays in a more-or-less fixed position while the blade lowers gradually. As there is a resolution of 0.9mm axially, the lowering can be observed in fine increments. In contrast, the posterior movement of the tongue in the pre-release phase appears to jump from one position to the next. This is largely due to the low angular resolution at the palatal surface of ~5mm (at a distance from the probe of ~60mm).

What is the minimum frame rate required for such a fast-moving articulation? Figure 6 shows how, in the exceptionally high velocity release phase of a click, intermediate samples provide more detail of the path that the tongue tip takes. However, de-interlaced video rates of 60Hz appear adequate for capturing most tongue dynamics (Wrench and Scobbie, 2008). If the frame rate is doubled from this basic de-interlaced video mode rate, to 120Hz, then one intermediate frame could be resolved. If tripled, the two intermediate frames could be resolved. Figure 6 displays the two extra frames possible at a rate of 182Hz. At 306Hz four extra frames could be resolved. (In Figure 6,

the lips and nose contours are derived from synchronous NTSC video fixed to the same headset holding the probe, demonstrating low head movement).

a)



b)

Figure 4. 25 scan-lines 306Hz frame rate a) raw scan-line data b) image constructed from scan-line data interpolated in arcs around the scan-lines to fill in the gaps in the image. FoV 112°.

3. Discussion

With the particular system described in this paper, storage is not an issue for the high data rates: the higher the frame rate, the fewer the number of echo pulses. Thus the amount of data describing each frame reduces in line with the increased number of frames. The storage requirement is 6.3Mbyte/s, regardless of frame rate. By contrast 640x480 8bit greyscale images at 120Hz would require. 36.9Mbyte/s storage.

The optimal frame rate for ultrasound analysis seems to lie somewhere in the range of 60-200Hz. Unless there is particular interest in the very fastest articulations, such as click releases, the very fastest frame rates possible should be avoided. There is always a tradeoff between decreased angular resolution at high frame rates and the distortion possible in tongue shape due to the time taken to complete a sweep from the

first echo pulse to the last (Wrench and Scobbie 2006, 2008). Angular resolution is important when the tongue surface is nearly parallel to the echo pulses. This often occurs with the posterior surface of the tongue body. An acute surface angle also substantially reduces the surface reflections. Thus the combination of fewer and weaker reflections can cause this part of the tongue to be very faint and blurred in an ultrasound image (Figure 5). Lower frame rates will reduce any blurring by providing more echo pulses and finer-grained interpolation between closer scan-lines. The other area where angular resolution is important is at the tongue tip during blade retroflexion (Figure 5).

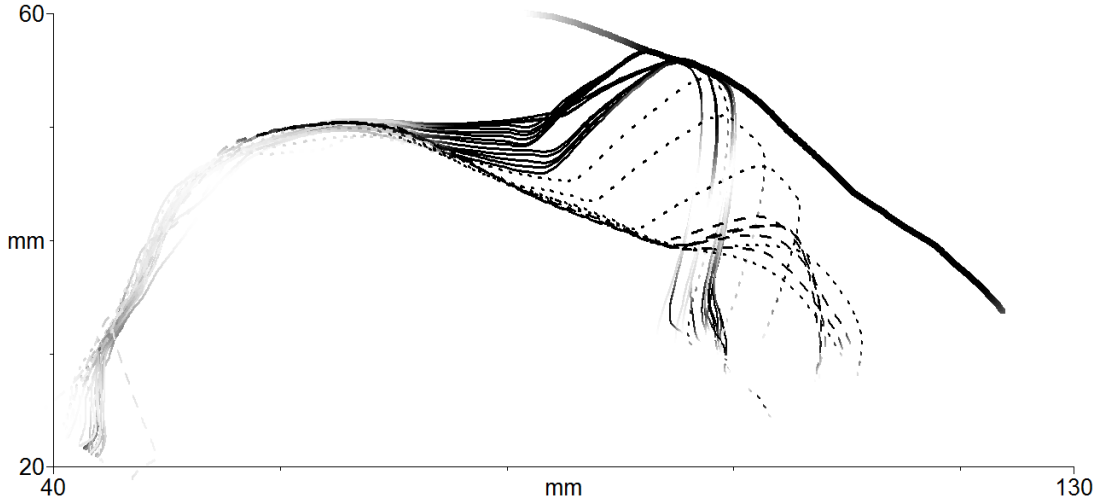


Figure 5. Tongue surface contours of successive frames throughout the production and release of a click. 306Hz sample rate. Solid lines= pre-release rarefaction; Dotted = release phase; Dashed lines = recoil from floor of mouth. Faded contours show low confidence of auto-fit.

With peak tongue tip velocities in speech of ~800mm/s, the displacement of the tip between frames at 60Hz is ~13mm. At 120Hz, it is ~7mm. This is therefore the distance the tip can move between the first echo pulse and the last in a sweep. Tongue movement during a sweep can lead to shapes which are distorted even with dense scan-lines, due to the time difference between the emission of pulses towards the root and tip (Wrench and Scobbie 2006, 2008). Field of View and depth settings should therefore be set to the smallest values that keep the vocal tract in view, and a balance struck between fast frame rate and high resolution. In sum, ultrasound tongue imaging requires careful consideration of these factors during experimental design and when reporting results.

Frame rate	D 70 FoV 112°	D 80 FoV 112°	D 70 FoV 150°	D 80 FoV 150°
60Hz	1.0mm	1.1mm	1.3mm	1.5mm
100Hz	1.6mm	1.8mm	2.2mm	2.4mm
120Hz	2.0mm	2.2mm	2.6mm	2.9mm
180Hz	3.0mm	3.3mm	3.9mm	4.4mm
300Hz	4.9mm	5.5mm	6.6mm	7.3mm

Table 1. Angular resolution at 60mm from probe surface (70mm radius) for different frame rates, depth settings (D, in mm) and FoV.

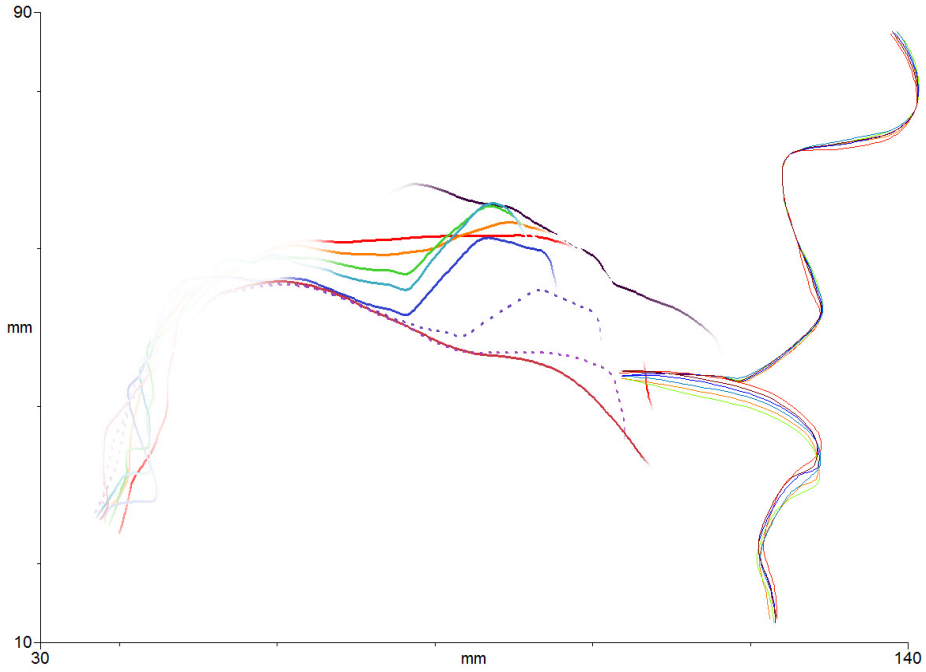


Figure 6. 42 scan-lines at 182Hz (~triple the interlaced NTSC video rate of 60Hz). Successive solid traces are shown at 60Hz. Dotted traces are at 182Hz. Colour indicates the same time point in the camera profile and tongue contour. Advancing time order = red-orange-green-blue-purple.

References

Articulate Instruments. *Articulate Assistant Advanced Ultrasound Module User Manual, Revision 2.13*, Articulate Instruments Ltd, 2010.

Jannedy, S., Fuchs, S. & Weirich, M. Articulation beyond the usual: Evaluating the fastest German speaker under laboratory conditions. In S. Fuchs, P. Hoole, C. Mooshammer & M. Zygis (eds.), *Between the regular and the particular in speech and language*, 205-234. Frankfurt/M.: Peter Lang, 2010.

Ladefoged, P., Cochran, A., & Disner, S. F. Laterals and trills. *Journal of the International Phonetic Association*, 7:46–54, 1977.

Scharf, G., Hertrich, I., Roux, J., & Dogil, G. An articulatory description of clicks by means of electromagnetic articulography. *Proceedings of ICPhS-13, Vol. 1 378-379.* Stockholm, 1995.

Shosted, R. An aerodynamic explanation for the uvularization of trills? *Proceedings of ISSP-8*, 421-424. Strasbourg, 2008.

Wrench, A.A. & Scobbie, J.M. Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. *Proceedings of ISSP-7, 451-458.* Ubatuba, 2006.

Wrench, A.A. & Scobbie, J.M. High-speed cineloop ultrasound vs. video ultrasound tongue imaging: comparison of front and back lingual gesture location and relative timing. *Proceedings of ISSP-8,* 57-60. Strasbourg, 2008.