High-speed Cineloop Ultrasound vs. Video Ultrasound Tongue Imaging: Comparison of Front and Back Lingual Gesture Location and Relative Timing

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

1. Articulate Instruments Ltd 2. Queen Margaret University, UK
E-mail: awrench@articulateinstruments.com

Abstract

We compare two methods of acquiring ultrasound tongue images. A new system capable of recording directly from the cineloop image buffer at a high frame rate and which is more accurately synchronized with audio is compared with an optimised method of recording images via the NTSC video output of an ultrasound machine. As a focus for this comparison we gathered representative data on English /l from a single speaker, using a headset restraint system. Both systems performed well, but while the video system is at its limits, the cineloop system is inherently more accurate and offers greater opportunity for development.

1 Introduction

For many years ultrasound has been used to image the tongue [2]. More recently, midsagittal tongue contours have been extracted from image sequences and used in coarticulatory and other kinematic studies where timing and magnitude of the changes in tongue position are central to the study outcome. Despite the increasing application of ultrasound for this purpose, it is difficult to establish the accuracy of such measures both in general and in different specific laboratory settings. Not only are there variables in how the curve-fitting is achieved but also in the method by which the basic ultrasound images themselves are obtained and on the inner workings of the specific ultrasound system used.

The majority of current laboratory work is based on image sequences derived from the video output of medical ultrasound systems, and some of the sources of error in these images were presented by Wrench and Scobbie [4]. Many ultrasound machines also provide the option of saving "cineloop" image data. With cineloop sequences, each image is formed from a single sweep of ultrasound from one end of the transducer array to the other [1]. These images retain predictable temporal properties. In contrast, the image sequences acquired from the video output are often rasterised as the source ultrasound image is being updated, resulting in each image displaying discontinuities and uncertainty in the timing of different parts of the image.

2 High speed system

The high speed system is based on an Ultrasonix RP, a research ultrasound machine. This system provides two key features that are not found in most other systems. Firstly it permits the end user to control the operation of the system via software development kits, allowing access to every level operation. Secondly, it provides synchronisation pulses that allow the precise timing of each frame to be tracked with reference to the audio signal.

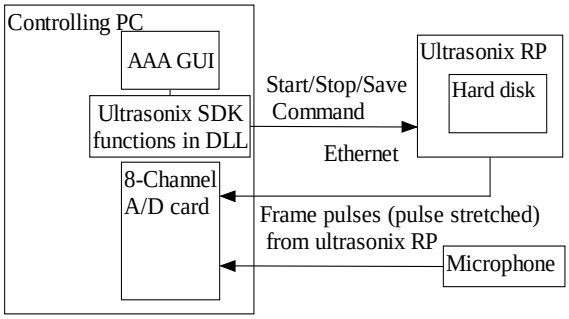

Figure 1: *High Speed ultrasound sync setup*

The probe is a microconvex type capable of 150 degrees but set to 112.5. The number of beam formed pulses is also simple to control and is set to a line density of 100, which means that there are 76 pulse firings for every sweep in the 112.5 degree sector. The depth is set to 8cm and the probe frequency to 5MHz for maximum penetration. The resulting frame rate is 100Hz. Higher rates could be achieved by reducing the line density still further but this would result in poorer special resolution. A typical 6 second recording takes about a second to save and be ready for the next recording, making the recording process efficient. At the end of a session

the ultrasound data is processed and synchronized with the recording using the pulse sequence recorded onto a secondary audio track to identify the exact instant of each ultrasound frame relative to the acoustics. The system is currently capable of recording 51 seconds of data in a single continuous sample but this could be increased with the addition of more system RAM.

To save time and space the data is kept in its raw pre scan converted form consisting of 76 lines and 412 16-bit samples per line. Further saving could be made by recoding the raw data to 8 bits per sample and only encoding samples from the depth range in which the tongue surface is visible e.g. 2-8cm. As well as saving space, this raw data format could provide a better basis for edge detection than processed images as it is free of processing artifacts.

The analysis software then converts the raw data to an image sequence to make it easy to review and hand correct tongue surface contour markers.

3 Video system

A different ultrasound machine was used for the recordings based on video output because the video output from the ultrasonix RP is surprisingly poor and unrepresentative of what can be achieved by most systems. A Mindray DP-6600 was used with NTSC video output at ~30fps. The settings were: Depth 7.55cm; Sector 160 degrees; Frequency 5MHz; frame rate 98fps. These settings provided the closest conditions to what was achievable with the ultrasonix system. In particular, the internal frame rate was very similar to the high speed system, and generally higher than video systems reported in the literature.

4 Data capture procedure

The high speed system is triggered remotely from the controlling PC via Ethernet. It starts and stops recording and saves the cineloop data to the local hard drive of the ultrasonix PC. As soon as the system starts recording, it generates pulses aligned with the start of each frame. These 25ns pulses are stretched to 11ms using dedicated hardware and fed into an A/D card on the controlling PC. A pre-amplified microphone signal is recorded on a second channel of the same A/D card. After the recording session is completed, the data on the ultrasonix hard disk is moved to the controlling PC, processed and imported into Articulate Assistant Advanced (AAA) software where each frame is tagged with a time code derived from the analogue frame pulse sequence.

The video system uses a frame grabber card to capture the uncompressed video output from the DP-6600. A “Brightup” unit is used to superimpose a flash onto the corner of the video image in response to a pulse that occurs immediately before recording commences. This pulse is also sent to the A/D card. A second A/D channel is used to record the pre-amplified microphone signal. The video and audio are then post-processed to ensure that the flash and the pulse are aligned. This innovation is also an improvement on standard video systems.

A probe stabilization headset was used to mount the probe. As far as possible the same probe position was maintained for both systems. An EPG recording was carried out simultaneously at 200Hz to provide confirmation of the timing of tongue/palate contact.

5 Materials

As a focus for this comparison we have undertaken a small study of intergestural timing between tongue root retraction and tongue tip raising in the production of /l/ consonants in syllable onset and coda [3]. The details of this phenomenon are well-understood, and the materials used are comparable to those used in previous studies.

The following two sentences were repeated 10 times each and recorded using each system.

1.Can we say “Pale Eva” retained a fiddlier status?
2.Can we say “Pay Laver” attained a fiddlier status?

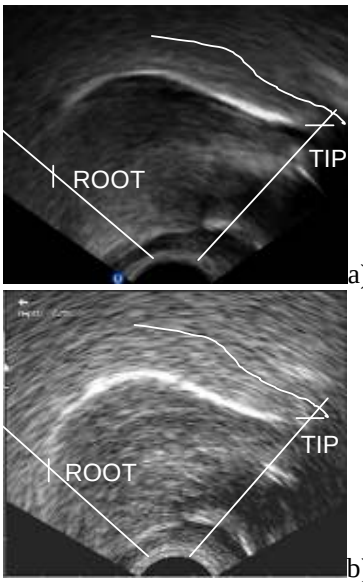

Figure 2: *“Laver” /l/ in a) cineloop and b) video*

6 Analysis

The NTSC video data was de-interlaced to provide ~60fps, then the image sequences from both systems were analysed in the same way using AAA software.

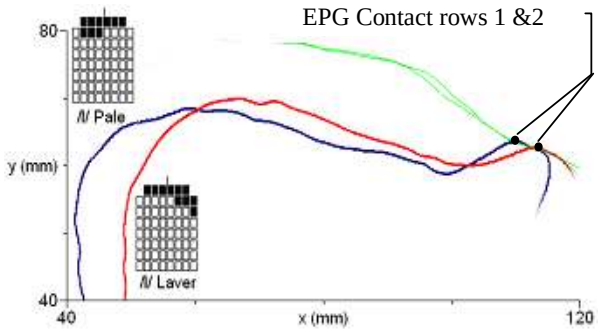

Figure 3: *Tongue contours and EPG palate patterns at the instant of maximum root retraction for /l/ in "Laver" and "Pale"*

A fan grid with 42 radial axes was laid over the image sequence, then a region of interest around the /ele/ transition was automatically edge detected, with hand corrections performed particularly where the edge fades due to the surface lying almost parallel to the pulse direction. Where the ultrasound image becomes indistinct towards the tip, the palate trace in combination with the EPG contact pattern were used to provide more confidence in the estmated tip position. The validity of the palate trace throughout a recording was supported by the fact that it matched a swallow at the start and end of each recording. There was however slight rotational movement of the probe relative to the head over the course of the session of 20 recordings.

Two radial distance measures were made from a point of origin of the ultrasound pulse scanline towards the tongue root and towards the tongue tip (see Figure 4, lower and upper traces respectively). For the tip measurement the palate boundary occurred at a dstance of 60mm and this is the reason for the plateau in the tip measurement plots. There was no such boundary to impede the tongue root movement and consequently the root measurement follows a generally smoother transition over time.

7 Results and Discussion

Figure 4 shows /ele/ transition data plotted as measured from the two systems. The general trend for a greater root retraction in word-final position is apparent in both sets of data. The measurement traces from the high speed system show an underlying bell shaped curve for the tongue root measure with a superimposed noise which is the result of error in the estimation of the tongue contour. Furthermore, you can make out from the highspeed data that there is a slight variation in the duration of the tongue root gesture and when it is shorter in duration it is also less extended.

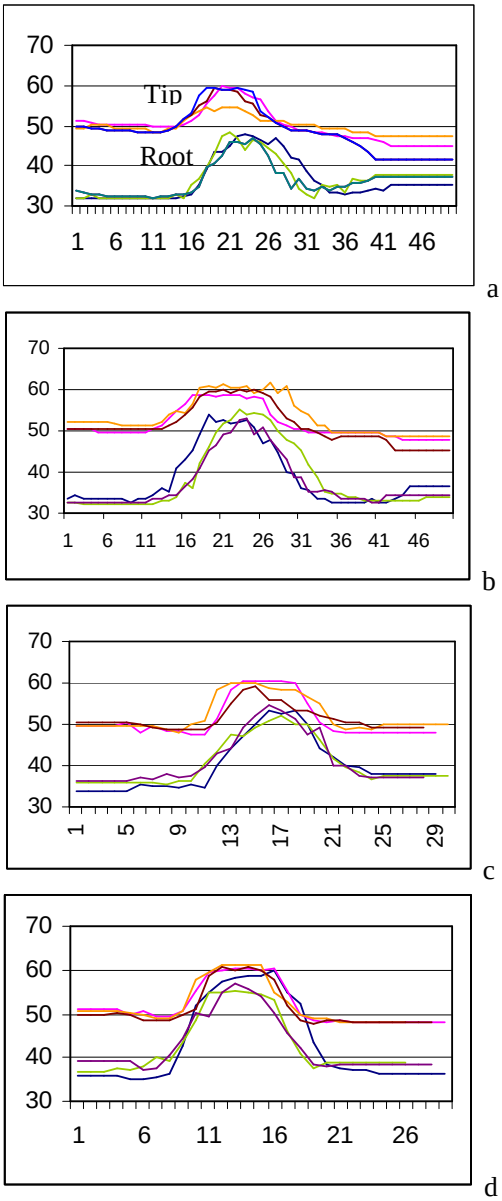

Figure 4 *Tokens of* a. *Cineloop "Pay Laver", b. Cineleep "Pale Eva", c. Video "Pay Laver", d. Video "Pale Eva". Vertical axis mm from probe, horizontal axis, sample frame.*

Great care was taken to optimize the video port ultrasound system for this comparison. Three key factors in getting the most from the video setup

were as follows. A) we set the internal ultrasound frame rate to much more than twice the video rate i.e. 98fps. This reduced the chance that successive de-interlaced frames contain duplicated parts of the same ultrasound sweep. B) the video data is in an uncompressed format that allows the images to be cleanly de-interlaced. Without these first two factors the video images would have been more discontinuous, with double images and half the frame rate. C) We measured the output video rate (not always exactly broadcast standard 29.97fps and in this case 30.52fps) and imposed a sync flash to enable synchronization with a pulse on the audio signal.

Positive aspects of the high speed cineloop system.

i. Efficient data storage to save disk space
ii. Precise acoustic synchronization
iii. No discontinuities in the images
iv. Higher temporal resolution in charted values

Negative aspects of the high speed cineloop system.

i. If tongue contours for every frame are drawn manually then it is more time consuming to analyse this data. Automatic edge detection helped provide a good starting point to speed up this process but automatic tongue contour prediction (as opposed to image edge detection) taking in various sources of information including image artifacts and EPG contact patterns is an area for future research.

Positive aspects of the video system.

i. A lot can be achieved with a relatively cheap basic ultrasound system.

Negative aspects of the video system.

i. Even with the high internal frame rate, it was the case that one frame in ten had part of the image unchanged between successive frames, forcing a compromise in how the tongue surface spline is interpreted from the discontinuous image.
ii. Even though the "brightup" flash allows the video frames to be aligned with the audio, there may still be a small variable delay due to the internal processing architecture of the ultrasound system.

There are some properties shared by the two setups. At 100fps and 98fps there is reduced distortion due to the difference in time of the first and last pulse return in each sweep. The high output frame rates of 100fps and ~60fps also mean that the tongue surface only moves a small distance between frames. This means that neighbouring frames can be used to help locate indistinct tongue surface contours with greater confidence than a standard ~30fps video system.

8 Conclusion

There are a large number of factors that contribute to confidence in the accuracy of ultrasound derived measures.

A. A cooperative participant who images well.
B. Good positioning of the microconvex probe to ensure US beams image the tip and root
C. Optimum settings for penetration, focus, clarity, line density
D. High ultrasound sweep rate
E. High output frame rate (e.g. ~60 de-interlaced)
F. Accurate frame rate estimation and alignment with the acoustic signal.
G. Reliable palate trace and probe-head stability
H. Synchronous EPG for tongue-palate contact

In this experiment we have pushed the video system to give the best possible performance, so that it generates similar data to the high speed system, in ideal conditions. However, it is more difficult to be sure of the position of the tongue tip due to uncertainty in timing of the image relative to the EPG, due to occasional discontinuities in the image and because there are fewer frames to provide continuity constraints. If some of the key elements B-G mentioned above are missing from a video based system then the disparity in the confidence and accuracy of the two methods is likely to increase.

References

[1] P.R. Hoskins, A. Thrush, K. Martin, T.A. Whittingham, *Diagnostic Ultrasound: Physics and Equipment*, Greenwich Medical Media Linited, London, 2003.

[2] K. Morrish, M. Stone, B. Sonies, D. Kurtz & T. Shawker. Characterization of tongue shape. *Ultrasonic Imaging*, 6(1): 37–47, 1984.

[3] A.A. Wrench, and J.M. Scobbie. Categorising vocalisation of English /l/ using EPG, EMA and Ultrasound. *Proceedings of ISSP 03*, 314-319, 2003.

[4] 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.

Acknowledgements

Infrastructure and development funded by SHEFC SRIF.