Spatio-temporal inaccuracies of video-based ultrasound images of the tongue

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

[1]Articulate Instruments Ltd - Queen Margaret Campus, 36 Clerwood Terrace, Edinburgh EH12 8TS UK

[*]Speech Science Research Centre – Queen Margaret University College, 36 Clerwood Terrace, Edinburgh EH12 8TS UK

awrench@articulateinstruments.com , jscobbie@qmuc.ac.uk

Abstract. *This paper focuses on aspects of ultrasound technology that have an impact on the accuracy of this technique as an investigative tool for the study of displacement, timing and movement of the tongue during speech. The paper describes settings and elements in the design of ultrasound systems that can affect spatial and temporal resolution and provides recommendations for how to minimize distortion.*

1. Introduction

Speech scientists typically derive an ultrasound image signal from the video output of an ultrasound machine and synchronize it with the acoustic signal from a microphone by feeding both signals into a video recorder or a video capture card. It is important to consider how design of ultrasound technology affects both the spatial accuracy of such images of the tongue and the precision with which they can be aligned with the speech acoustics. There are of course other factors, which have a significant effect on the shape of the tongue image, such as the importance of maintaining the probe in a precise midsagittal alignment, the presence of artifacts and the properties of tissues. However, this paper focuses on technological factors. More specifically, by examining the images from several machines, the general implications of using images derived from the TV/video output are deduced and recommendations are provided for best practice.

2. Ultrasound image creation

B-Mode (Brightness mode) ultrasound imaging is used for speech research and is formed by displaying relative brightness information based on the strength of radio frequency (RF) signals in the 3-9 MHz range that are received by an ultrasound transducer during the pulse-echo operation. A full 2-D image is formed by processing

[*]Supported by SHEFC under SRIF 3.

and displaying the pulse-echo data acquired from individual scanlines fired along a set of specific look directions; usually 128 or more in number.

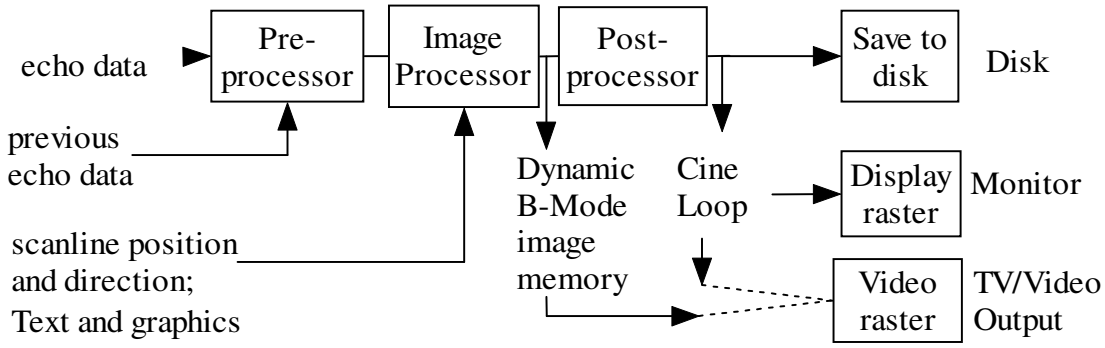

Figure 1 Block diagram of B-Mode Image creation storage and display

2.1 Pre-processing

The pre-processor provides options to enhance the raw echo data, such as edge enhancement and "persistence", where echo data is combined with data acquired from previous sweeps of the same location: a source of temporal blurring that should be avoided.

2.2 Image processing

During scan conversion, pre-processed echo signals are inserted into the B-Mode image memory (which may typically be 640x480, 800x600 pixels or larger) using information on position and direction of the scanline. The memory is continuously refreshed with new echo data. The converter also interpolates between scan lines to fill the whole memory with data. For a convex probe the interpolation can be observed where speckle appears as arcs of brightness that increase in sector size with depth rather than dots.

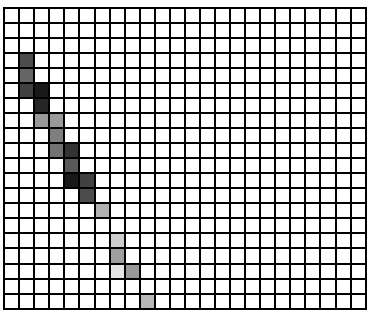

Figure 2 B-Mode Image memory showing one scanline built up by calculating the brightness from the echo strength, position and direction. Memory values for locations between scanlines are calculated by interpolation.

2.3 Post-processing

The post processor saves a complete B-Mode sweep in a buffer or in the cine loop (if this feature is implemented) and allows the greyscale mapping levels to be adjusted.

3. TV monitor/Video Format

European PAL comprises of 576 separate horizontal lines of image plus 49 blank lines (vertical blanking interval) making 625 lines in total with 50 interlaced fields per second (25 frames per sec). American NTSC comprises of 486 separate horizontal lines of image (525 including the vertical blanking interval) with 59.94 interlaced fields per second (29.97 frames per sec). NTSC operates at a faster frame rate than PAL at the cost of a lower vertical picture resolution. Note that although NTSC has 486 image lines video cards tend to discard 6 of these lines and save 480 lines, a number, divisible by 16, that makes mpeg coding easier.

Composite Video Format	Resolution (image lines)	Video card lines	Total rastered lines per sec	Frame rate (frames per second)
PAL	576	576	625	25
NTSC	486	480	525	29.97

Table 1 Composite video specification

In theory, setting an ultrasound video output to PAL should yield better resolution images. However, in practice, it would appear that in many, if not all machines the B-Mode memory images are rastered at a resolution of 480 lines and padded with extra blank lines if the final output format is PAL. As evidence of this, note that in figures 4 and 5 there is a black border around the PAL images to pad the 480 line image out to 576 lines. In such cases, there is therefore no advantage to operating ultrasound in PAL mode and it should be set to NTSC if this option is available.

In the example shown in Figure 3 the ultrasound frame (sweep) rate is about 1.5 times faster than the video frame rate and the resulting video output image is a composite of 2 or 3 ultrasound sweeps: the latter when the video scanning operates on the continually updated B-Mode image memory and the former when it operates on a buffered copy of the B-Mode memory which is updated only when a sweep is complete. In the unbuffered case (Figure3a), the resultant video frame, a mixture of parts of three B-mode sweeps, spans a time period of Tv=40ms in the case of PAL or Tv=33ms in the case of NTSC. On the left side of the output video image the interlaced scanlines consist of Sweep N and sweep N+1. In the center portion of the image both interlaced scanlines correspond to sweep N. On the right side the interlaced scanlines consist of Sweep N and sweep N-1. In the buffered case (Figure 3b) , the resultant video frame is a mixture of two whole

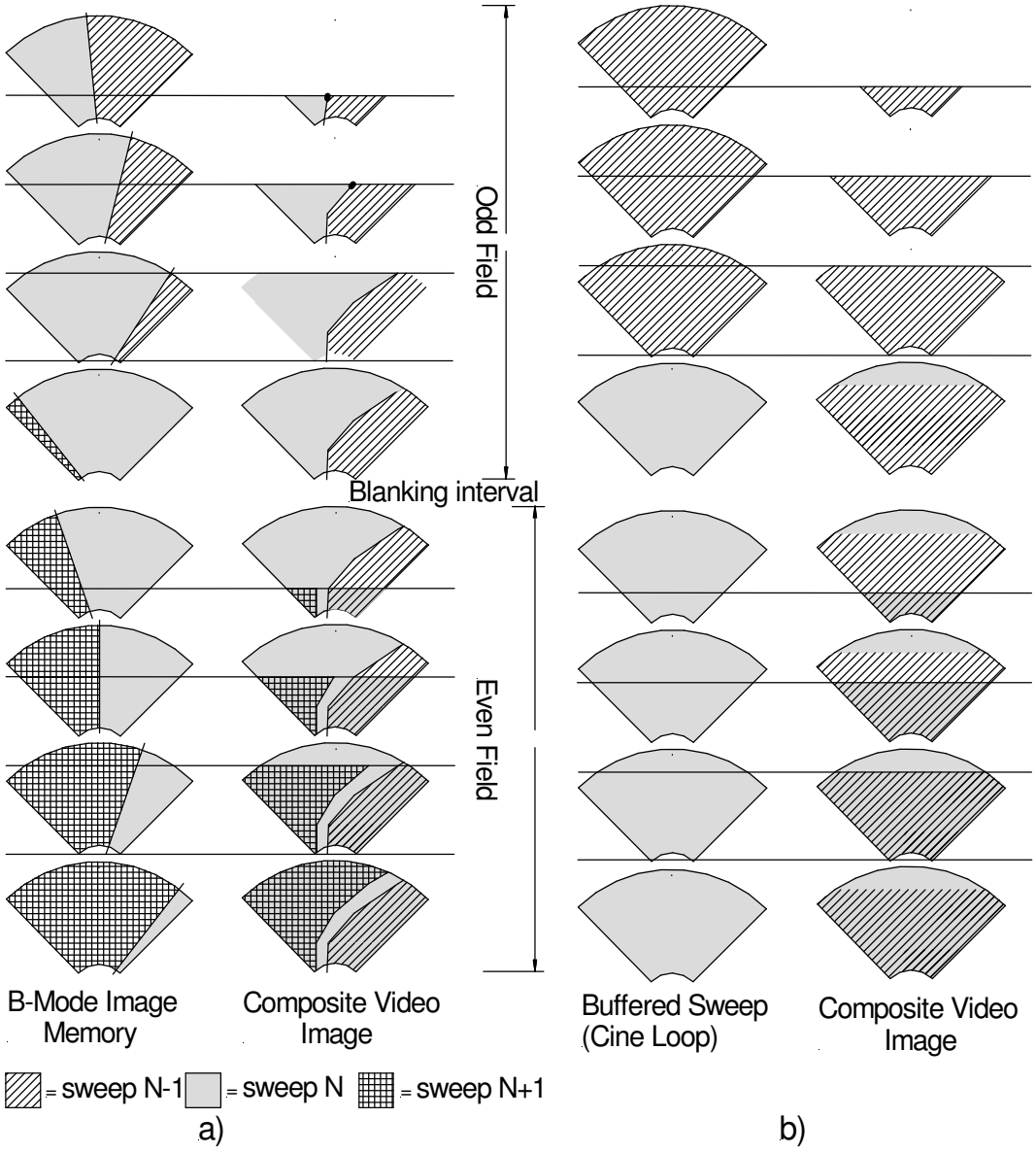

Figure 3 Two processes by which a single NTSC/PAL composite video image can be generated. a) directly from the B-mode image memory b) from buffered copy of the B-Mode image memory stored in the cine loop at the end of a complete sweep. The output image in each case is the bottom panel in the Composite Video Image column.

B-mode sweeps. Part of the output video frame consists of an interlaced combination of sweep N-1 and sweep N. This part is therefore made up from echo data gathered over a time period of twice the sweep period (Ts) where Ts is the time taken for a complete B-mode sweep (i.e. the period associated with the frame rate shown on the ultrasound screen). In this example, twice the sweep period is 53ms in the case of PAL or 44ms in the case of NTSC. The remainder of the output video in Figure 3b is comprised entirely from sweep N.

Figure 4 Example of video frame created as per figure 1a) sweep rate 44fps video frame rate 25Hz

In the buffered case there is also an inherent time lag. As the scanner is already part way through sweep N+1 at the time the composite video image is complete, the composite video image lags the B-Mode image memory by a variable amount of anywhere between 0 and Ts and this lag will change from frame to frame.

In general, regardless of which of the above processes is implemented, it is always true that different parts of the composite video image correspond to different instants in time and each image contains discontinuities. When the tongue is not stationary a striped double image appears. Speckle and finer features of the image become blurred.

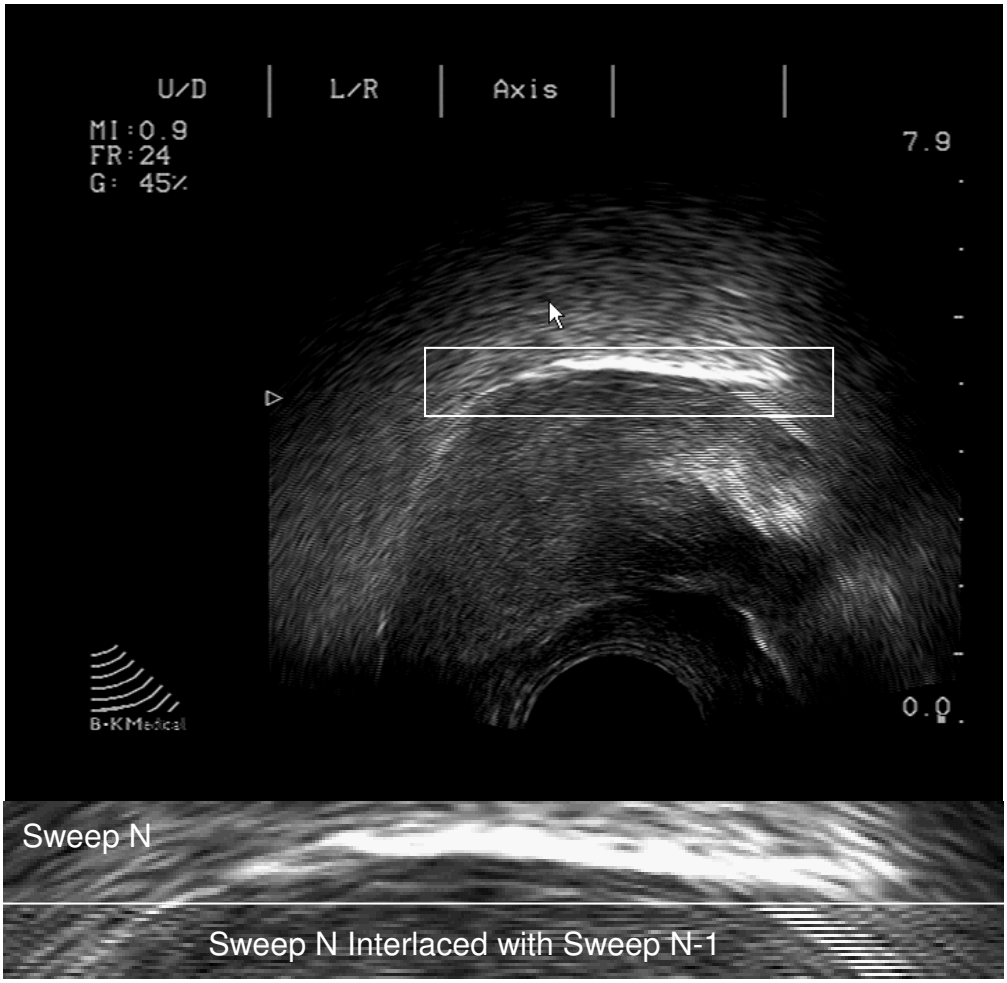

Figure 5 Example of video image created according to the process in figure 1b) sweep rate 24fps video rate 25fps

Although not observed in the systems examined for this paper, it would be possible for the manufacturer of an ultrasound system to add a further buffer stage that takes a copy of the image memory (or a copy of the cine loop frame) at the instant each video frame scan starts and hold it until the video frame has been generated. The resulting video output frames with reference to the scan sequence in Figure 3 would then look identical to the top panel in the B-Mode Image column and Cine Loop column respectively. This would reduce the time span in each frame but increase the lag by Tv ms.

If a video capture card is used to record the TV/video output from an ultrasound system then it should be noted that employing image compression may blur the image further making it difficult to see the distinct interlaced images and making it impossible to cleanly de-interlace the images. Furthermore, it is unclear whether any given soundcard synchronizes the audio with the start, middle or end of a video frame and what the tolerance is for doing so.

4. Cine Loop

Most modern ultrasound models provide a cine loop facility whereby each complete B-Mode sweep is stored in memory. Blocks of memory as large as 128Mbytes can be set aside for this purpose allowing around a sequence of up to 420 frames of 640x480x8 bit data to be stored at a frame rate determined by the ultrasound sweep rate rather than the NTSC video rate. Cine loop does however have two disadvantages: it does not currently provide a means for synchronizing with a microphone input; and it can take several minutes to save the cine loop to disk. Solving these two problems will be a key step in the development of ultrasound technology for the purpose of speech research.

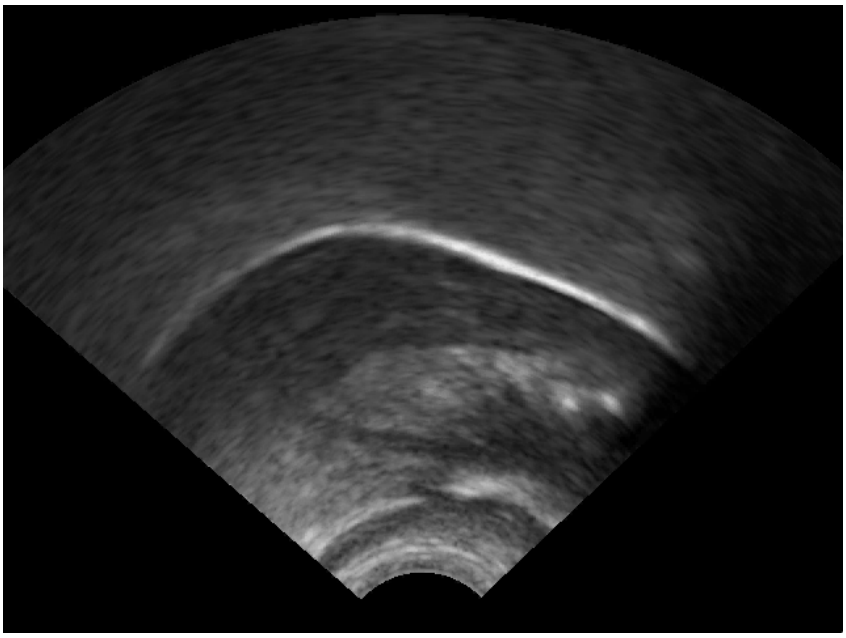

Figure 6 Frame from cine loop recorded at 101 frames per second provides a sharper image with no discontinuities

5. Conclusions

Images of the tongue derived from the video output of an ultrasound machine are subject to distortions in shape, temporal blurring and time delay. Over the 33ms period of time it takes to rasterize the NTSC video output, the tongue tip can move as much as 10mm and the tongue body 7mm. Even in steady states the tongue can move 1 or 2mm. In a video output frame where the tongue body is moving forward and the tip is raising, the image is likely to contain surface contours from the tongue body at its most advanced (sweep N+1) and from the tongue tip before it has raised (sweep N-1). An example of this can be seen in Figure 4. As a result, a distorted tongue surface contour is obtained if the bright edge closest to the probe is traced and recorded. In addition, the mixture of speckle patterns from different sweeps serves to reduce the contrast in the video output frames, making weak surface reflections more difficult to discern. Furthermore, temporal blurring, and a possible time lag mean that a given point in space can only be aligned at best to within 33ms of an acoustic speech event

Many modern ultrasound machines derive the video image according to the process in Figure 3a). In such machines the oldest part of an NTSC video frame will have been generated approximately 33ms prior to completion of the frame. Note however, if video output is derived from post-processed sweeps (as in Figure 3b) then oldest part of the NTSC video frame may be 33ms + the sweep period (i.e. for a machine set to 44 frames per second the oldest parts of the image may have been scanned 56ms prior to completion of the video frame).

Key recommendations from this paper are:

1. Set persistence to 0. This does not reduce the double images produced by the video raster process but prevents even older echo data from being mixed into the image.

2. Set the video output and video capture to NTSC format. This provides the better frame rate and in practice there is no reduction in the image resolution.

3. If possible, de-interlace the video output frames. That is to say, separate out the odd and even fields of the image and interpolate the missing lines in each field to create two images. This needs to be done prior to mpeg or jpeg image compression in order to get a clean separation. A de interlaced NTSC image consists of 240 lines out of a total of 525 lines per frame. The time period over which a de-interlaced image is generated is therefore (240/525)/29.97 = 15.25ms. This minimizes image distortion but discontinuities will still exist in a de-interlaced image. The time difference across the discontinuity will equate to the ultrasound sweep rate.

4. Set the frame rate to the highest practical value. This can be done by setting the minimum depth and minimum field of view that cover the full range of tongue movement that is of interest. Normally these values are around 8cm depth and between 90-120 degree field of view. It may be possible to set a dedicated frame rate parameter. Ideally it should be 60 frames per second or higher so consecutive de-interlaced frames contain distinct sweeps.

In the near future it should be possible to synchronize the acoustic speech signal with the cine loop output thus avoiding the problems associated with TV/video output but for speech scientists who only have access to ultrasound via video output, the preceeding guidelines should help to optimize the accuracy of measurements derived therefrom.

Further Reading

Hoskins, P. R., Thrush, A., Martin, K. and Whittingham, T. A. Diagnostic Ultrasound: Physics and Equipment, Greenwich Medical Media Limited, 2003.

Kremkau, F. W. Diagnostic Ultrasound: Principles and Instruments. W.B. Saunders Company, Sixth Edition, 2002.

Zagzebski, J. A. Essentials of Ultrasound Physics, Mosby, 1996.