Corrected high-speed anchored ultrasound with software alignment

Amanda L. Miller

Department of Linguistics

Totem Field Studios

University of British Columbia

Vancouver, British Columbia V6T 1Z4

Canada

and

Kenneth B. Finch

LoTense Inc.Ithaca, NY 14850

Running Title: Corrected high-speed anchored ultrasound
Abstract

The CHAUSA (Corrected High-speed Anchored Ultrasound with Software Alignment) computer system architecture for collection of high-speed ultrasound (US) used in the investigation of speech articulation, integrates existing hardware and software components. A method for using this architecture for speech data processing is presented. The architecture and method increase the frame-rate for data analysis from the standard NTSC video rate (29.97) to the US internal machine rate, in this case 124 FPS, by using DICOM data transfer. CHAUSA data are presented with alignment of the acoustic and articulatory signals to the correct high-speed frame (8 ms at 124 FPS). The method controls and reduces head position uncertainty by using a combined head stabilization and head movement correction paradigm. Techniques that export the US video through the VGA or S-Video port introduce spatio-temporal inaccuracies that are avoided with CHAUSA. Preliminary US data of the production of one speaker’s alveolar click in IsiXhosa reveal tongue dorsum retraction during the posterior release, and tongue tip recoil following the anterior release. These effects were invisible at lower frame-rates. The CHAUSA architecture and method enable the study of dynamics of rapid lingual speech events as they unfold in time, with incremental resolutions of 8 ms.
I. INTRODUCTION

By making affordable, safe and portable imaging of the tongue possible in real-time, ultrasound imaging has the potential to do for articulatory phonetics what the spectrogram has done for acoustic phonetics. The availability of portable ultrasound machines makes dynamic articulatory studies possible in linguistic fieldwork situations (Gick 2002) and clinical speech science environments. Both the tongue and palate can be imaged with ultrasound.

A Corrected High-Speed Anchored Ultrasound with Software Alignment (CHAUSA) computer system architecture and associated method, which uses DICOM (Digital Imaging and Communications in Medicine, NEMA 2008) file transfer protocol to transfer high frame-rate data, is presented. Video editing tools are used to undertake post-data collection software mixing of higher frame-rate ultrasound images, with the audio signal, and with the head position video that is used for head movement correction. Probe-to-head stabilization (anchoring) is achieved using the Ultrasound Stabilization Headset (Articulate Instruments 2008). Head movement correction is performed with the Palatoglossatron technique (Mielke et al 2005). The CHAUSA computer system architecture integrates existing hardware and software components, with a detailed proven method on how to effectively use these integrated components.

Lingual ultrasound imaging of speech has, in the past, been limited on several fronts. First, since the ultrasound rays are reflected at the interface with air, only the edge of the tongue is imaged, and not the hard bony structures of the palate, jaw and spine. This makes the interpretation of tongue contact with the palate difficult to gauge without concurrent imaging of the palate that is seen during swallowing (Epstein and Stone 2005). Second, alignment of ultrasound video images
of the tongue with the acoustic signal, which is paramount in speech studies, has been lacking. Dominant methodology in the fields of linguistics and speech science consistently uses analog VGA external monitor outputs or s-video ports of ultrasound machines, which limit the sampling rate to 29.97 frames per second. Both of these protocol conversions introduce unnecessary artifacts (Wrench and Scobbie 2006), and the conversions limit the speed of the events which can be observed.

The new integrated system presented here enables the field collection of head corrected, high-speed linguistic data. The high-speed characteristics of the CHAUSA (Corrected High-speed Anchored Ultrasound with Software Alignment) method (Miller 2007, 2008) were specifically developed for the study of dynamic consonants in fieldwork situations. A CHAUSA study is presented on alveolar click production in the Bantu language IsiXhosa, which demonstrates the merits of the architecture and method, and reveals new insights into the dynamics of the alveolar click release. This study also illustrates CHAUSA’s ability to align the acoustic signal to the exact high-speed frame, which is in this case 8 ms. Indications in the study presented here are that the combination of anchoring and head movement correction makes potentially possible palate and tongue position accuracy to within 1 mm., although this has yet to be definitively proven. The equipment used is portable and applicable to linguistic fieldwork and clinical settings.

CHAUSA uses high quality, high-speed DICOM images, that are software mixed to the exact frame with the acoustics, integrated with several prior head stabilization and head correction
techniques to increase field accuracy at high speeds. Therefore, prior techniques are reviewed in more detail next, with the most detail on the methods used, and why they were used.

II. BACKGROUND

Prior ultrasound architectures (Stone and Davis 1995, Gick et al. 2005, Mielke et al. 2005) are limited to the standard commercial video rate of 29.97 frames per second. This is too slow to document many aspects of speech production accurately, and previous studies have thus focused on vowels and sonorants that have stable articulatory gestures. Further, the use of VGA or S-Video outputs of US machines, with the accompanying acoustic-to-articulatory alignment accomplished via hardware mixing, can have significant unfixable mixing errors.

Both head stabilization and head movement correction techniques have been developed with high accuracy for research lab settings. Head and probe stabilization accuracy were kept to 1 mm. using the Head and Transducer Stabilization technique (Stone and Davis 1995), and head and probe movement correction has been achieved using an Optical tracking system for the Haskins Optically Corrected Ultrasound System (HOCUS) method developed by Whalen et al. (2005). Portable head stabilization methodology for linguistic fieldwork has also been developed by Gick et al. (2005). Mielke et al. (2005) developed the Palatoglossatron technique, whereby experimenters videotape subtle changes in head and probe position, which makes head movement correction techniques adaptable to fieldwork. Portable head and probe stabilization and correction techniques have yet to achieve the level of accuracy of methods developed for lab settings.
Prior Head stabilization techniques include “The Head and Transducer Support System (HATS)” (Stone and Davis 1995) which is a robust stationary system that has been shown to stabilize the head to 1 mm. Gick et al. (2005) showed that using a simple experimental field-work setup with minimal head stabilization by using a headrest, and a fixed transducer (held in place by either an arm holder in the lab or a portable microphone stand in the field).

Articulate instruments (2008) developed an Ultrasound stabilization headset, which anchors the ultrasound probe to the head, assuring that the probe maintains an optimal and constant position throughout a recording session. The headset achieves probe stabilization while allowing the head to move freely, and thus avoids the discomfort implicit in head stabilization and associated recording time limits. McLeod and Wrench (2008) measured probe slippage using the headset at about 5 mm. by overlaying palates from different recordings over time.

Prior Head Movement correction techniques include HOCUS (Haskins Optically Corrected Ultrasound System, Whalen et al. 2005) an optical head movement tracking system developed by HASKINS Laboratories and a less accurate and cheaper fieldwork technique Palatoglossatron developed by Mielke et al. (2005). HOCUS is accurate and reliable, but not portable and hence not viable for linguistic fieldwork or clinical settings. Palatoglossatron, developed at the University of Arizona, uses a video camera focused on two sticks each containing two dots. One stick is attached to the probe and the other stick is attached to the head, to track visually the movement of the head and the probe. The Palatoglossatron technique mixes the head video with the ultrasound signal using an audio-video mixer containing a hardware “blue screen” removal tool. Palatoglossatron uses a film industry clacker-board (at the beginning of the video) and a
bell (at the end) to re-synchronize the head video with the ultrasound signal post hoc. Mielke et al. (2005) developed the mathematical physics for the stick-movement correction method for head movement in the images of the tongue and the palate. While the laser approach of Optotrak used by HOCUS is more accurate, the Palatoglossatron head movement correction technique was integrated into the CHAUSA method because of its applicability to linguistic fieldwork.

III. HIGH-SPEED ARCHITECTURE

Ultrasound equipment has been able to display internally frame-rates greater than 29.97 FPS for many years. However, most speech labs choose to export those videos from the US machine using either the standard monitor video port or an s-video port in order to achieve synchronization with a hardware mixed audio signal. From a computer architecture point of view, this export process is a very defective process, introducing many artifacts and distortions. For instance, Wrench and Scobbie (2006) discuss some artifacts that are introduced by digital conversion to a lower frame-rate interface. Using DICOM to achieve lossless transmission of the internal images from the machine resolves these inherent difficulties.

A. The system architecture

The CHAUSA method avoids limitations found in video-based ultrasound by having the system architecture export the high frame-rate cine loops using DICOM technology, the same technology used in commercial medicine to move images from the hospital to the doctor’s office. The DICOM standard (NEMA 2008) was developed to optionally transmit perfect video with zero information loss over distances using networking technology. While it is a non-real time
transmission, the high quality greatly outweighs the time spent in the post data-collection step of synchronizing the US video with the real time audio. This system architecture is shown in Figure I. The US data are acquired by the notebook via two separate paths: a high-speed path, transferred via DICOM, which is referred to in Figure I and throughout this paper as the DICOM PATH, and the low-speed path transferred via the Canopus audio-video mixer, referred to here and throughout as the CANOPUS PATH. The hardware mixed CANOPUS PATH is used as a known starting alignment basis for the acoustic-video alignment process of the high-speed video from the DICOM PATH.

The CHAUSA method uses non-real time software mixing to achieve to-the-frame articulatory to acoustic alignment to avoid hardware mixing errors. US machines have several significant image generation or image translation times, which delay the US image output from the actual event, and hence from the audio signal as it enters the audio-video hardware mixer. There is a processing delay in formation of the reflected ultrasound data to the image on the machine screen, and another delay reformatting that image for output through the video port to the mixer. Meanwhile, the actual concurrent audio event has long since been mixed with prior US images. There may also be a mixing delay at the mixer itself. Brugman (2005) undertook a small pilot study to determine the total delay found with a GE Logiq book machine at 29.97 FPS and a somewhat different architecture. Brugman labeled 6 acoustic [k] bursts within the audio portion of each ultrasound recording (2 [k] bursts in each of 3 repetitions recorded in one take) using Praat (Boersma 2001), and then calculated the expected video frame. She independently marked the frame in the video signal that corresponded to the [k] release. Offsets were then calculated and determined to be 4-5 frames in each of the 50 tokens measured. In some cases, the delay may
just be a frame or two, but at 33 ms per frame, the delay is significant for high-speed, fast speech sound research. The actual ultrasound machine's processing delay times are unknown and vary from machine to machine as well as the state of the machine during US capture. For example, the amount of space available on the hard drive and other tasks the machine is undertaking contribute to delays. Delays have been measured from 33 ms to 150 ms. Even if the actual degree of asynchronization could be calculated for a specific recording environment, in current methodology there is little ability to re-synch the audio with the video post hoc.

A solution to both the image quality issue and the articulatory to acoustic alignment issues is to add to the standard hardware mixing path (the CANOPUS PATH in Figure I), the second DICOM PATH, which records the same US events, but with higher-speed and superior image quality. Using the CANOPUS PATH data (see Figure I) as a guide, the high-speed, high-quality DICOM PATH video is software-mixed with the audio in a post data collection stage. Data collection requires three researchers, as one person captures each take on the notebook within Adobe Premiere Pro via the CANOPUS PATH, one person saves the individual takes on the US machine hard drive in DICOM format for later transfer via the DICOM PATH, and a third person operates the clacker board. The clacker board is moved in front of the speaker and released after the beginning of the CANOPUS PATH data collection. A bell is rung at the end of the utterance, after the end of the US data collection on the US machine, but before the end of the CANOPUS PATH recording within Premiere Pro. This timing is necessary in order to keep the US recording within the 8-10 second window that can be recorded by the GE LogiqE US machine at this frame rate. The maximum window length varies with recording frame rate.
With the guide of the imperfectly mixed data, linguistic expertise is used to align the US and audio clips at multiple clearly identifiable linguistic events along the entire 8-second DICOM video clip, to the individual high-speed frame. The 29.97 FPS data transferred via the DICOM path is then discarded. Aligning multiple events (6-8) along the entire 8 second clip largely removes the concern that one researcher might interpret an isolated event differently than another, since all events can be checked, and an error in one event or one type of event would be self corrected by the need to align different types of events. All events must be correct as the entire timeline is moved as a whole for the final fine alignment, and hence the data self checks itself. Further we have additional electronic tools to prove and verify accurate alignment.

The Ethernet port is capable of perfectly transmitting the high quality, high frame-rate, DICOM video. The transfer of data can be done after the end of a study, as the hard drive can store DICOM data from an entire study. Some compression is still essentially lossless to the analysis tools, and this reduces transmission time. A lossless transmission protocol (RLE) is used. The US machine, the GE Logiq E, also stores the DICOM format cine loops on its hard drive. While these high quality cine loops can also be stored to an external hard drive or CD, the networking solution allows data to be transferred at higher non-standard frame-rates, such as the 124 FPS used here. The GE implementation of writing to an external drive uses a standard disk write (not DICOM). That is why the robust quality options of DICOM are not available when writing to external storage.

The notebook computer receives both the high quality DICOM images via the Ethernet port and the real-time Canopus mixed images via the fire-wire port, but at different times. The Canopus
Twin 100 is a good quality audio and video hardware mixer that has minimal mixing distortion. The delay implicit in the CANOPUS PATH comes from the GE LogiqE machine itself.

The third tier of the architecture uses a video camera, which records video of the head movements (using the Palatoglossatron sticks) concurrently with the ultrasound video recording. In the IsiXhosa data presented here, a standard 30 FPS video camera was used, whose output was brought in post-hoc through a synchronous fire-wire port of the notebook. Miller, Scott, Sands and Shah (2009) have replaced the standard video camera with a Prosilica GE 680C camera that has an adjustable frame rate. The GigE camera is set to match the frame rate of the ultrasound data, and can capture any frame rate up to 200 FPS. The GigE camera stream is captured digitally on a notebook computer.

Figure II pictures an IsiXhosa speaker wearing the Ultrasound Stabilization Headset (Articulate Instruments 2008). The headset anchors the probe to the head. The lower Palatoglossatron stick is attached to the probe to measure probe movement. A pair of glasses anchors the upper Palatoglossatron stick to the head, and does not touch the headset to allow independent recording of head movement.

The use of the Palatoglossatron method incorporated into CHAUSA differs from the standard method described in Mielke et al. (2005), which is tied to the standard video frame rate. Instead of using hardware mixing, software mixing is used to remove the blue screen background in the head video images, and to mix the then partially transparent head video with the high-speed ultrasound video that was transferred via the DICOM path. Figure III shows the speaker’s
tongue, along with the four dots on the Palatoglossatron sticks, as well as the headset and the high-speed ultrasound image.

**B. Software tools**

Commercial movie making software was adaptable to the software mixing used in the CHAUSA approach. Adobe Premier Pro was chosen for video mixing and this robust commercial software was coaxed into solving the unique challenges of this speech research technique. This software is surprisingly compliant to the non-standard frame-rates used in this research. Although nominally the maximum commercial high definition frame-rate available in the software was 60 FPS, the software package includes tools that allow the handling of arbitrarily fast frame-rates with high levels of reliability. There were no issues with 124 FPS US data used in the pilot experiment described in Section IV.

Figure III provides a single frame of the software-mixed video containing the speaker’s head with the stabilization headset, the head correction dots introduced via the Palatoglossatron sticks, and the DICOM US image of the tongue. The speaker and the stabilization headset are to the left, the pink colored dots are more towards the right, and the same frame of the US video is in the center. The stabilization headset is colored due to the “blue screen” removal process that software mixed the US video with the stabilization headset video containing the Palatoglossatron sticks. There is an interaction between the head correction and the accuracy of the audio-video alignment, such that the mis-alignment within the standard hardware mixing process (even if just one to four 29.97 frames), will degrade the head correction process. During the alveolar click
release imaged in the pilot experiment in Section V, the head moves very fast. If head correction were to be undertaken with the wrong video frame, it would introduce error.

Figure IV provides a data-analysis software tools diagram. The complete system architecture integrates these software tools with the hardware architecture of Figure I. The pyramid shape is used to indicate how the tools build on each other and the direction of flow of data processing, in this case from the top of the pyramid to the bottom. “Digital Jacket” software, made by Desacc, allows the user to acquire, view and manage DICOM data locally and remotely. In the CHAUSA method, the Digital Jacket DICOM software server (near the top in Figure IV) on the laptop receives the DICOM (top of Figure IV) transmissions from the US machine (via the DICOM PATH in Figure IV). Digital Jacket also accomplishes the export of the DICOM cine loop to an AVI file that preserves the non-standard frame-rate. Adobe Premiere Pro captures the low-speed (29.97 FPS) video from the CANOPUS PATH, and receives the high-speed AVI file from Digital Jacket. Alignment and software mixing of the head video, the Canopus audio, and the high-speed ultrasound video is undertaken in Adobe Premiere Pro.

A Matlab script is used to convert the mixed AVI file (Figure III) to a series of JPEGS required by Palatoglossatron. Palatoglossatron is used to trace the tongue and palate in the images of interest, as well as to undertake the head movement correction. Independent plotting software is used to graph the tongue and palate traces.
C. Articulatory – to-acoustic alignment

More detail on the articulatory to acoustic alignment process is provided, the process of aligning the DICOM high-speed ultrasound frame, to within 8 ms. Near perfect alignment, defined as alignment to the correct individual video frame (8 ms at a sampling rate of 124 FPS), is achieved using this method. At a frame rate of 165 FPS, at which the GE Logiq E machine has been shown to produce decent images of the entire tongue, the alignment would be to 6 ms. Since the US frame image is a static pictorial representing a finite amount of time, 8 ms in this data, perfect alignment would be articulatory–to-acoustic alignment that is to the correct portion of that 8 ms frame. For example, a 1-ms tone would appear in one eighth of the beginning, middle, or end of the 8-ms US frame. Figure IX illustrate this for a 3-ms tone used in the proof of alignment. CHAUSA results are near perfect, as opposed to perfect, in the sense that a key stop burst is shown anywhere within the correct US frame image.

Figure V depicts the workspace used within Adobe Premiere Pro. The timeline in the bottom panel of the workspace shows (a) the high-speed 124 FPS ultrasound video that was transferred via the DICOM PATH in the top row, (b) the low speed 30 FPS ultrasound video collected on the laptop via the CANOPUS PATH in the middle row, and (c) the audio channel recorded via the CANOPUS PATH in the bottom row. In Figure V, the entire 8.8 seconds of audio is visible, while the researcher zoomed in on a click burst to align the high-speed DICOM PATH video in (a) with the audio in (c). Zooming in and out of multiple independent events is undertaken to check the alignment of each of these events. Canonical oral stops preceding open vowels are recommended, as such events have both a clear acoustic burst and a large opening gesture on release that will be clearly identifiable. Multiple events are necessary to assure fine alignment.
The alveolar click burst shown here is an ideal candidate for alignment because of its fast release and abrupt acoustic burst.

Once rough alignment of the audio from (c) and the DICOM PATH US video (a) is accomplished for one event, the CANOPUS PATH video frames in (b) can be hidden, and attention focused on precise alignment of the DICOM PATH US video with the audio. The CANOPUS PATH ultrasound and audio are hardware aligned, and this is normally correct to within four to five 29.97 frames. Therefore, the precise alignment is a matter of moving the acoustic and ultrasound signals these few low speed frames or approximately four times that number of high-speed frames. This vastly simplifies the pattern recognition task described next.

Figure VI shows the Adobe Premiere Pro workspace that is used for the precise alignment. The timeline in the lower portion of the Premiere Pro workspace shows large scale audio and video. The program sequence in the top of Figure VI enables frame by frame movement through each event in (a), allowing the researcher to pick the precise frame that aligns with the audio seen in the timeline. The whole video track is then slid until the audio and video are aligned to within a single 8 ms video frame. Once the precise alignment of a single frame is completed, the researcher can zoom out and check that each of the stop bursts is aligned properly. If the alignment is off a bit in one of the frames, the other bursts will similarly not align properly. The long sequence is critical for this approach to work, as multiple independent bursts help to assure precise alignment. For the alignment to work, three repetitions of a sentence in each US video token should be recorded. This is possible within the 8-10 second recording window of the
LogiqE machine at the 100-125 FPS frame rate. Thus, for the CHAUSA method there are always six independent linguistic events (stop bursts) to be aligned.

The process of going back and forth to identify the precise alignment is vastly strengthened by the dynamic nature of the Adobe Premiere Pro Tool. Pattern recognition properties of the human mind make the temporal change recognition in the US images easier than one might imagine. The process produces a high level of confidence of the very best 8 ms frame. Changes jump out after one has looked at the sequence of frames a couple of times. By doing this with multiple burst/release pairs, an effective confidence level of 100% can be achieved by the researcher. This fine alignment, since it is along the entire timeline and includes multiple token types, is not overly dependent on varying researchers’ views of a particular token event, since all events can be checked to be aligned, after the stop releases are confirmed to be aligned. All these events check and balance any differing views of the articulatory-to-acoustic alignment of any one particular event.

**IV. PROOF OF ALIGNMENT**

First we provide figures demonstrating alignment of the various signals, and second we provide proof of alignment using a Tri-modal pulse generator, which is a recent development. Figure VII provides a close-up of the three aligned videos in a recording of production of a Mangetti Dune !Xung alveolar [!] click. The videos have been aligned using the procedure described in Section IIIC. The head video in the Video 3 track was collected using the Prosilica GE 680c camera at 114 FPS. This matches the 114 FPS DICOM PATH US data that is seen in the Video 2 track. The CANOPUS PATH video is in the Video 1 track, and the corresponding audio is in the Audio
1 track. The 114:30 frame rate ratio simplifies to a ratio of 3.8:1 DICOM PATH US video frames to CANOPUS PATH video frames. The anterior constriction release of the click occurs between frames 3 and 4 of the DICOM PATH video frames seen in the Video 2 track. Below, It can be seen simply by the frame boundaries lining up that the head video and the DICOM PATH US video are aligned. Also, the audio click burst, which is very sharp in onset, can be seen to be aligned to the click release in the high-speed DICOM PATH video, by viewing the DICOM frames closely.

Figure VIII provides two adjacent frames showing the anterior release of the alveolar click [!] in Mangetti Dune !Xung. The two frames are only 8 ms apart. The tongue tip is raised in the earlier frame in the left panel of Figure VIII, and lowered in the later frame in the right panel of Figure VIII. The anterior release of this click is abrupt (Miller and Shah 2009). Note the clarity of the data showing the click release in these two high-speed DICOM frames. The dramatic increase in the clarity of the data is difficult to communicate with static pictures. CHAUSA’s ability to single step forward and back in adjacent frames 8 ms in time means if one frame has an artifact, the next may not, and since the position of the tongues has hardly moved, it is similar to having also doubled the spatial density. Gradual tongue movements are very clear, and fast movements clearer than ever. This clarity for images of the whole tongue has never been seen before in phonetics. The clarity greatly enables and assists new knowledge acquisition, as well as increases certainty of this alignment method. Recall that alignment is along the entire 8 second timeline, over 900 frames, and all frames are as clear or clearer than these two representational frames. Past experience with less clear data, such as that seen in the CANOPUS PATH data in Figure VII, does not equip the easy understanding of the alignment to the high-speed frame that the CHAUSA method makes possible.
Since this alignment method is novel, we provide additional electronic proof of alignment. To do this alignment proof, we designed a Tri-modal pulse generator that simultaneously produces a 3 ms burst of ultrasound, 3 ms of audio from a buzzer, and 3 ms of light from a high-brightness LED (Light emitting diode) for use in daylight, which would be picked up by the various audio and video recordings. While the ordinary video camera used in the IsiXhosa study contains an audio signal (which can itself be misaligned with the video recording), the high-speed camera does not contain a simultaneous audio signal. Thus, the LED signal is picked up by the head video, the synchronizing ultrasound signal is picked up by the LogiqE ultrasound recording (and transmitted through both the DICOM and CANOPUS PATHS), and the buzzer is picked up by the audio recording mixed in the CANOPUS PATH. The Ultrasound probe used was designed for therapeutic ultrasound, and is called “Medical Products Online - Professional Ultrasound System.”

Originally the circuit generated a time programmable burst (a pulse), from 60 ms down to 200 usecs to control the ultrasound firing of the commercial ultrasound generator. At the identical time of this burst (to within a few millionths of a second), voltage pulses of the same duration were also sent to a bright LED, and to an electronic buzzer. The literature is sparse concerning the perceptibility of a few milliseconds of sound or light, but the authors determined by sequentially reducing the pulse width, that the 3ms stimulations produced by the synchronization circuitry were perceptible in both modalities.

The Tri-modal pulse generator is added to the CHAUSA computer architecture. The 3 ms burst is aimed at our standard 8C-RS GE Medical ultrasound probe, which marks the ultrasound frame that it occurs in with a bright flash in part of one frame. Likewise, the electronic buzzer is aimed at the standard audio microphone. The LED is aimed at the high-speed video camera and similarly marks the one video frame
simultaneously with the US frame and the audio track. An acoustic stand-off was used to couple the two ultrasound transducers without having the probe heads physically touch. The marked data is then synchronized as above. However, when the process is completed, the signals are examined to see if the various synchronizing marks line up.

Figure IX provides the head video recorded by the Prosilica GE 680C camera in (a), the DICOM PATH US video in (b), and the audio signal recorded in the CANOPUS PATH in (c) of the first author's production of the [!] click. We first aligned the recordings using the above procedure, and found that the synchronizing signals marking each of the signals were also aligned. Since we know that the US pulse, the lit LED, and the buzzer occur within a few millionths of a second of each other, and these frames line up, we know that the click audio and the US release image are aligned to the correct high-speed frame. In fact, note that the US pulse is located later in the time scan position of the left frame, indicating that the next frame (the right one) is about to happen in approximately the length of time between the buzzer burst and the click burst, a fraction of an 8-ms frame. This alignment precision has never before been possible.

These data demonstrate that the alignment process provided here is accurate to a frame, or 8 ms, given that the synchronizing marks are all aligned to within a single 8 ms DICOM US video frame. The synchronizing signals produced by the new circuitry may in fact prove to be a useful tool for future alignment assistance.
V. PILOT STUDY ON ALVEOLAR CLICK PRODUCTION USING CHAUSA

A. Introduction

A study to investigate the articulation of the alveolar click release in IsiXhosa was designed using the CHAUSA method in order to view the dynamics of the anterior and posterior releases of this click. Previous ultrasound investigations of clicks in Khoekhoe (Miller, Namaseb and Iskarous 2007) and N|uu (Miller et al. 2008), as well as X-ray recordings of !Xóõ clicks (Traill 1985) have all been sampled at 29.97 FPS. These previous studies all show large gaps in the very fast releases of alveolar clicks, and aliasing effects. Average durations of the alveolar click in this study were approximately 110 ms for the closure, and 15 ms for the release, which are comparable to the durations for this click in Mangetti Dune !Xung reported by Miller and Shah (2009) and N|uu by Miller, Brugman and Sands (2007). Thus, only one frame could be consistently imaged during the release of these clicks with the standard 29.97 FPS frame-rate (Miller et al. 2008). The abrupt release of the alveolar [!] click motivates an articulatory study of the dynamics of the release. Such a study was undertaken in IsiXhosa with the CHAUSA method. Miller (2008) provides preliminary results of this study.

B. Methods

High-speed 124 FPS ultrasound data were collected with the GE LogiqE portable ultrasound machine to image the alveolar click production in the utterance Ndi qaba isonga. ["di !aba isɔŋa] ‘I spread something on the bread.’ The sentence was repeated 3 times in each take, and 5 takes were recorded, yielding 15 repetitions of the target sound in the same phonetic context. The CHAUSA method was used to collect the data. 25 clearly visible frames were obtained during the production of the alveolar click, which showed a remarkably consistent pattern with no
effects of aliasing seen in previous studies. A single frame of the palate imaged during a swallow in the same headset seating was also traced. There are 6 pixels per mm. found in the 640 x 480 pixel images acquired with the Sony video camera used in this study and it is straightforward to trace to within a few pixels or a small fraction of 1 mm. This supports the claim that 1 mm. accuracy could be possible with this camera. The methodology described in Epstein and Stone (2005) was followed for tracing the palate, except that the IsiXhosa speaker held the water in his mouth for 1-2 seconds prior to swallowing.

The Palatoglossatron head-movement correction algorithm transforms tongue and palate traces to the same frame of reference. Therefore, the absolute values in millimeters in the axes are different for the uncorrected and corrected graphs, since the corrected graph is in this new frame of reference. However, the relative differences in mm. between tongue and palate traces are the same in the uncorrected and corrected graphs.

The [!] click anterior releases and the [ɡ] releases were used to align the DICOM PATH video and the audio of the CANOPUS PATH. Since three repetitions of the sentence were recorded in each take, these yielded 6 independent events for alignment in each take. The head video captured in the third tier of the CHAUSA architecture was aligned with the mixed ultrasound and dot-tracked video with the use of the clacker board and bell adopted from the standard Palatoglossatron method (Mielke et al. 2005).
C. Results

Seven non-consecutive tongue traces exhibiting the major stages of click production and a single palate trace are provided in Figure X. The results before head movement correction are provided in Figure Xa, while the head movement corrected version of these traces is provided in Figure Xb. The tongue and palate data in Figure Xa are the output of the ultrasound machine prior to Palatoglossatron head correction. That is, they exhibit head-to-probe anchoring using the Ultrasound stabilization headset (Articulate Instruments 2008), but no head movement correction. Figure Xb provides the same frames of tongue in the click production, and the same palate frame, traced in Palatoglossatron post-head movement correction. In both graphs, the portion of the palate that can be traced during a swallow is the thick, solid, black line that can be seen towards the top of these graphs. The palate traced from this speaker’s swallow is surmised to be in a neutral position.

To-the-correct US frame articulatory – acoustic alignment (+/- 8 ms) allows us to match the articulatory events seen in ultrasound traces with acoustic representations. Figure XI provides a waveform and spectrogram of the alveolar click with the acoustic events that correspond to the tongue traces of particular US video frames shown in Figure X labeled. Each of the numbered acoustic markers corresponds to the articulatory tongue traces provided in Figure X.

Traces 1-7 in Figure X show a complete cycle of click production. In trace 1 of the corrected graph, the tongue dorsum is touching the palate, but the front of the tongue is in a slightly raised position in the mouth. In Trace 2, which is imaged 2 frames or 16 ms later, the tongue body has lowered, resulting in two swells in the tongue. In trace 3, which is 5 frames or 40 ms from the
first trace, the tongue tip is more perpendicular, with the tip of the tongue forming a constriction just in front of the alveolar ridge. We surmise that the tongue body is touching the soft palate, which has lowered by this point. In trace 4, the tongue tip has retracted slightly and is touching the alveolar ridge. The tongue body has lowered, and the tongue dorsum has retracted, as part of the process of cavity expansion described in Traill (1985) and Thomas-Vilakati (1999). Trace 5 corresponds to the point on the waveform just before the anterior click burst seen in Figure XI. The tongue dorsum has retracted even more at this point, and shows the maximal cavity expansion just prior to the anterior release. Trace 6 shows the tongue root in the pharyngeal region for the vowel [a] following this click in the word qaba [ˈlabə]. The front of the tongue has released completely, and is low down in the mouth. Surprisingly, Trace 7 (121 ms) shows that the tongue tip has risen up again, and the tongue body has achieved the same shape as it held just prior to the anterior click release. These last two frames display a dynamic recoil effect of the tongue tip after its extremely rapid release.

As noted above, quantitative measurements of head movement using the probe stabilization headset alone showed from 2-3 mm. of head movement in the mid-sagittal plane (Scobbie et al. 2008) during speech. This makes it critical to perform head movement correction. The level of inaccuracy can be seen by the fact that the tongue tip is shown as moving through the palate in traces 3-5 of the uncorrected graph, while these same traces in the head-corrected graph show the tongue neatly touching the palate. These inaccuracies show clearly the need for head movement correction in addition to probe-to-head anchoring for lingual ultrasound imaging.
D. Discussion

Holistically, the dynamics of click production are seen clearly in the tongue traces in Figure Xb. The tongue dorsum retracts and the tongue body lowers to enlarge the click cavity prior to the anterior click release (trace 5). These dynamics can also be seen clearly in Traill’s (1985) alveolar click x-ray movies, particularly when viewed frame-by-frame with modern video tools. Tongue Traces 1 through 5 display the tongue dorsum and tongue root retraction and tongue body lowering. The intermediate un-shown frames move continuously and smoothly along approximately 1-mm. increments. Inducing from all these twenty-five consistent traces as they change over time, gives some weight to the induction that the entire body of data is accurately head corrected over time, and placed accurately within the oral cavity to a great deal of accuracy. Conversely, the uncorrected traces in Figure Xa are inconsistent with the x-ray data, and with the description of click cavity formation described in Traill (1985). The data is internally consistent to within approximately 1 mm. throughout the cavity and throughout the duration of the click. Induction of the 1 mm. hypothesis from the smooth movements of a large number of corrected tongue movements, while less firm than an analytical proof, has a great deal of precedence in other inductions from large amounts of data, and is a common technique in experimental physics.

The use of the palate as an articulatory landmark for ultrasound studies is encouraging. The movement of the soft palate during speech must lead to caution in interpreting tongue positions relative to the palate in ultrasound data. X-ray data, which images both the palate and the tongue simultaneously, is easier to interpret. However, collection of X-ray data is not safe, which inhibits its use in speech studies.
The tongue traces in Figure X show much more detail than has been seen previously in click production studies due to the low speed of previous ultrasound and X-ray methodologies (29.97 FPS). The tip of the tongue is seen to go completely down, and then rise back up again. This is interpreted as a recoil effect, which is seen in every instance of IsiXhosa alveolar click imaged (15 tokens). Every major stage of the anterior and posterior releases in the alveolar click can be seen in these traces, resulting in a complete picture of the process of cavity expansion used for rarefaction in clicks. As with the N|uu and Khoekhoe alveolar clicks studied previously, there is visible tongue dorsum and root retraction. This could not be seen in earlier X-ray studies (Traill 1985) or ultrasound studies with lower frame-rates (Miller, Namseb and Iskarous 2007; Miller et al. 2008).

VI. BENEFITS OF THE CHAUSA APPROACH

The new ultrasound architecture has provided high-speed data on click articulation (Miller 2008). CHAUSA data has more than quadrupled the frame-rate from the standard 29.97 FPS to 124 FPS, reduced distortion and artifacts, and combined probe-to-head anchoring and head movement correction in order to achieve accurate tongue to palate positioning. Benefits in terms of increased speed, improved spatial clarity, to-the-frame Articulatory-to-Acoustic Alignment, Probe-to-head anchoring and Head Movement Correction are found. The speed benefit can be seen by the capture of the tongue tip recoil effect seen here following Miller (2008) for IsiXhosa and Miller, Scott, Sands and Shah (2009) for Mangetti Dune !Xung.

Complex and significant spatial distortions caused by using analog ports of ultrasound machines have been previously documented (Wrench and Scobbie 2006). DICOM transfer avoids these
distortions and only has artifacts that are in the original cine loop. Figure XII provides two images taken at the same exact moment in time (the anterior release of the click), allowing us to clearly see the differences in terms of spatial quality between the CANOPUS PATH image in Figure XIIa and the non-real time DICOM PATH used in the CHAUSA method in Figure XIIb.

This pair of frames was chosen since the distortion differences are great. The frame in Figure XIIa shows two spatially distinct images of the same tongue in the same frame (center middle right), which makes it difficult to trace the tongue edge. Other frames have less distortion; however, this distortion is a serious issue that can lead to confounds in the data. The DICOM solution is an excellent one for spatial clarity. The posterior part of the tongue is much more difficult to see in the 30 FPS video CANOPUS PATH image in Figure XIIa than in the image transferred with the DICOM PATH at 124 FPS in Figure XIIb.

The Ultrasound stabilization headset (Articulate Instruments 2008) anchors the ultrasound probe to the head, assuring that the probe maintains an optimal and constant position throughout a recording session. Maintaining the same position is critical for allowing the comparison of different tongue motions implicit in different speech sounds. Since ultrasound does not image bony structures, there are few physical landmarks in the images. Anchoring assures that the tongue motions being compared are viewed from the same perspective throughout a single placement of the headset.
Anchoring also allows the researcher to determine the optimal image quality for each speaker, and lock this in over the duration of the recording session. Image quality is determined both by anchoring, and by a wide variety of ultrasound machine settings.

Previous studies have shown that perturbation of the jaw can lead to compensatory lingual movements (Kelso et al. 1984, Lindblom et al. 1979). Whalen et al. (2005) note that it is not known whether using a stationary probe setup reflects the same patterns as speech uttered in a less restrictive setting. In order to allow jaw movement that is central to the alveolar click release, the probe was made certain to only minimally be touching the submental area (and that the tongue was not compressed). The subject, who was a candidate for an advanced engineering degree and hence has expertise in this area, helped adjust the stabilization headset until he felt that his jaw was not being impeded by the stable probe during the production of the frame sentence. In addition, he confirmed that his production of the sentence sounded to him like his normal speech. The improvements in the Logiq E US machine imaging are a major reason why a light touch still produced excellent images. Thus, minimal perturbation was achieved for this study.

The Palatoglossatron head-movement correction method (Mielke et al. 2005) is portable and inexpensive, and suitable for use in a fieldwork setting. Although the exact accuracy of head-movement correction using this method cannot yet be quantified, a review of the corrected and uncorrected tongue and palate positions in this paper strongly suggests that head correction is required to make sense of articulatory movements. The goal of 1-millimeter accuracy appears
within reach as suggested by the closeness of the tongue to the palate in the corrected image in Figure X discussed in the results section of the pilot study.

Head movement correction is clearly needed with the use of the Ultrasound stabilization headset, since that accomplishes probe-to-head anchoring, but not head stabilization. If head and probe stabilization techniques described by Gick et al. (2005) are used instead, head movement correction is less necessary.

The CHAUSA method brings together different pieces of hardware and software for a unified whole. Each piece of the method described here is needed in order to achieve accurate high-speed ultrasound results. The issues of probe anchoring and head position accuracy are somewhat variable with respect to each other, and the choice of appropriate methodology for these two aspects of the method may be situation specific and interchangeable. What is clear, is that the lack of absolute spatial reference in ultrasound imaging makes it necessary to achieve both probe-to-head anchoring and either head stabilization or head movement correction.

VII. OTHER HIGH-SPEED APPROACHES

Other high-speed ultrasound approaches are becoming available. Hueber et al (2008) use the Terason T3000 ultrasound machine, which allows synchronization of two video streams (US and optical) with the audio stream. According to Hueber, the machine has some articulatory to acoustic alignment error, but is effectively aligned to the frame up to 71 FPS. This has been verified by a clever test protocol. The Hueber approach uses temporally tagged real-time software mixing, which adds an additional real-time burden to the same processor assisting
image formation. Additionally, much of the misalignment found in prior architectures comes from the delay from the US image formation prior to acoustic mixing. All of these issues may produce a significantly lower upper frame rate than CHAUSA’s 114 to 165 FPS. The Hueber approach is a superb accomplishment. However, Hueber has not published results faster than 71 FPS, slightly over half the frame rate of this study. The results shown in this paper would not have been seen clearly at this lower frame rate. Additionally, much of the misalignment found in prior architectures comes from the delay from the US image prior to acoustic mixing. A programmable mixing method such as that used by Hueber et al., which does not incorporate an intelligent human check with linguistic expertise, as CHAUSA does, could mean mixing errors go undetected. Additionally, the GE Medical LogiqE Ultrasound machine appears to have a more robust feature set than the Terason T3000 for complex high-speed linguistic studies. The GE medical machine has shown good images above 165 FPS, and the CHAUSA methods itself is only bound by this current machine rate.

Noiray et al. (2008) have developed a different high-speed approach for a lab setting using the HOCUS architecture with optical tracking of head movement. Given the reliance on a large and expensive additional Optical Tracking system, this architecture is not suitable for fieldwork, such as the CHAUSA approach. Nor has the high-speed acoustic to articulatory method been fully documented. In addition, it is not clear if this architecture has integrated a methodology to anchor the probe to the head. Anchoring is critical to assure that sounds being compared are imaged from the same perspective throughout a recording session. Further, locking in the optimal imaging position assures high quality images throughout the duration of the study.
Wrench and Scobbie (2008) compare video based and high-speed cineloop ultrasound tongue imaging approaches. A video based system based on a Mindray DP-6600 ultrasound machine with a frame-grabber card allows the capture of US video. The internal frame rate of the machine is 98 Hz, but the output is only 30 Hz. A post-hoc stage of deinterlacing results in a quasi 60 frames per second, however some of the distortions inherent in an NTSC analog port are still present. A high-speed system based on an Ultrasonix RP research ultrasound machine is controlled via the Ethernet port of a host computer. The Ultrasonix RP system, however, has unusually poor spatial clarity. The Ultrasonix machine records simultaneous synchronization pulses that can be used to align the audio and video in a post-hoc alignment stage. The Ultrasonix RP system may be promising, but it is not portable enough for linguistic fieldwork.

VIII. CONCLUSIONS

Results provided here show that the CHAUSA method is capable of capturing high-speed data for the investigation of fast speech sounds in linguistic fieldwork. An alignment procedure was described which achieves to-the-frame alignment (8 ms at 124 FPS, and higher) of the acoustic and articulatory signals. Results also show good positioning of the tongue relative to the palate. A hypothesis has been stated that this is close to 1 mm., based on the closeness of the tongue to the palate in the tongue traces in Figure X above, and tight spatial and temporal corrected tongue movements. Quantification of the degree of tongue to palate closeness achieved through the combination of probe anchoring and head movement correction used in CHAUSA is planned for future research. A pilot study has shown both tongue dorsum retraction and tongue tip recoil in the IsiXhosa alveolar click. The CHAUSA method opens the possibility for studies of dynamic consonants and vowels, and co-articulation that have not been possible with current ultrasound
methodology. Ultrasound studies of speech allow us to view the entire tongue, which contrasts with other high-speed articulatory methods such as EMA (Electromagnetic articulography) that track only a finite number of flesh points. CHAUSA makes possible the discovery of an important new body of knowledge about how the tongue works.

Acknowledgements

The development of the CHAUSA method was supported by a National Science Foundation Grant, NSF #BCS-0726200 (PI Amanda Miller) and BCS-0726198 (PI Bonny Sands): "Collaborative Research: Phonetic and Phonological Structures of Post-velar Constrictions in Clicks and Laterals" to Cornell University and Northern Arizona University. Any opinions, findings, and conclusions or recommendations expressed in this material are ours and do not necessarily reflect the views of the National Science Foundation. We would like to acknowledge the support of Abigail Scott, who assisted with the Tri-modal pulse generator proof of alignment data collection. We would also like to thank our IsiXhosa speaker, Luxolo Lengs, and our Mangetti Dune !Xung speakers, Jenggu Rooi Fransisko, Martin ||oshe Aromo, Sikunda |u’i Fly, Bingo Kanaho Costa, Caroline Tumbo Kaleyi, and Sabine Towe Riem.
References


Figure I. Ultrasound system architecture (Hardware)
Figure II. IsiXhosa Speaker wearing an Ultrasound Stabilization headset and Palatoglossatron movement tracking sticks
Figure III. Software-mixed image of an IsiXhosa speaker’s tongue edge in US (right), his head with stabilization headset in video camera image (left), and small pink dots showing position of head (superimposed on US image)
Figure IV. Data analysis software tools diagram

(DICOM, Digital Jacket, Adobe Premier Pro, Palatoglossatron)
Figure V. Example of the Adobe Premiere Pro workspace with the entire 8-second audio signal showing, as well as zoomed in traces of the low speed and high-speed video. Vertical line in timeline indicates position of video in upper right window.
Figure VI. Adobe Premiere Pro Workspace showing method of frame by frame movement while simultaneously viewing larger video and audio sequence. Thin line in timeline indicates exact point of image displayed in top window.
Figure VII. Adobe Premiere Pro editing workspace showing the alignment of the 114 FPS head video (Video 1), the 114 FPS DICOM PATH ultrasound video (Video 2), the 29.97 FPS CANOPUS PATH ultrasound video (Video 3), and the CANOPUS PATH 48,000 Hz audio signal of the initial part of the word !gif ‘to carry’ produced by Mangetti Dune !Xung speaker Jenggu Rooi Fransisko
Figure VIII. Close-up of three adjacent frames of the DICOM PATH ultrasound video illustrating the anterior release of the alveolar click in the Mangetti Dune !Xung word ǂlô! ‘to carry’ (Speaker Jenggu Rooi Fransisko)
Figure IX. Adjacent frames during the release of an alveolar click in the Ju’hoansi phrase Ha !áí. ‘He died.’ produced by the first author. Top panel shows head video with LED flashing in the left frame, middle panel shows DICOM PATH ultrasound video with therapeutic ultrasound signal flashing in upper right corner of left-most frame, a faint but distinctive narrow V shape, and bottom panel shows audio signal, with the 3 ms buzzer signal on the left, and the click burst on the right. Audio signal looks longer than its 16 ms duration because it has been stretched to align with the high-speed video signals.
Figure Xa. Ultrasound frames traced from a IsiXhosa alveolar click collected with an ultrasound stabilization headset with no head movement correction in the utterance *Ndī qaba isonka* [‘Ndī ḳaba isonga’] ‘I spread something on the bread.’
Figure Xb. Ultrasound frames traced from a IsiXhosa alveolar click collected with an ultrasound stabilization headset with head movement correction in the utterance *Ndi qaba isonka*. [ⁿ̃di laha išongə] ‘I spread something on the bread.’
Figure XI. Waveform and Spectrogram of a IsiXhosa alveolar click in the word *qaba* [!aba]

‘spread’ with labeled trace numbers corresponding to articulatory ultrasound traces in

Figure 8 (Color online)
Figure XII. A Single US Frame of Mixed video transferred through the CANOPUS PATH at 29.97 FPS, and the same single US Frame of Mixed video transferred through the DICOM PATH at 124 FPS