JSLHR

Article

Corrected High–Frame Rate Anchored Ultrasound With Software Alignment

Amanda L. Miller[a] and Kenneth B. Finch[b]

Purpose: To improve lingual ultrasound imaging with the Corrected High Frame Rate Anchored Ultrasound with Software Alignment (CHAUSA; Miller, 2008) method.
Method: A production study of the IsiXhosa alveolar click is presented. Articulatory-to-acoustic alignment is demonstrated using a Tri-Modal 3-ms pulse generator. Images from 2 simultaneous data collection paths, using dominant ultrasound technology and the CHAUSA method, are compared. The probe stabilization and head movement correction paradigm is demonstrated.
Results: The CHAUSA method increases the frame rate from the standard National Television System Committee (NTSC) video rate (29.97) to the ultrasound internal machine rate—in this case, 124 frames per second (fps)—by using Digital Imaging and Communications in Medicine (DICOM; National Electrical Manufacturers Association, 2008) data transfer. DICOM avoids spatiotemporal inaccuracies introduced by dominant ultrasound export techniques. The data display alignment of the acoustic and articulatory signals to the correct high–frame rate (FR) frame (±4 ms at 124 fps).
Conclusions: CHAUSA produces high-FR, high-spatial-quality ultrasound images, which are head corrected to 1 mm. The method reveals tongue dorsum retraction during the posterior release of the alveolar click and tongue tip recoil following the anterior release of the alveolar click, both of which were previously undetectable. CHAUSA visualizes most of the tongue in studies of dynamic consonants with a major reduction in field problems, opening up important areas of speech research.

Key Words: ultrasound, articulation, speech and language, data collection methods and instruments, tongue

By making affordable, safe, and portable imaging of the tongue possible in real time, ultrasound (US) imaging has the potential to do for articulatory phonetics what the spectrogram has done for acoustic phonetics. The availability of portable US machines makes dynamic articulatory studies possible in linguistic fieldwork situations (Gick, 2002) and clinical speech science environments (Bernhardt, Gick, Bacsfalvi, & Ashdown, 2003). A Corrected High Frame Rate Anchored Ultrasound With Software Alignment (CHAUSA) computer system architecture and associated method, which use the Digital Imaging and Communications in Medicine (DICOM; National Electrical Manufacturers Association, 2008) file transfer protocol to transfer high frame rate (FR) US data, is presented. Video editing tools are used to undertake post–data collection software mixing of higher FR US images with the audio signal and with the head position video that is used for head movement correction. Both probe-to-head stabilization (anchoring) and head movement correction are combined to reduce head position uncertainty. The result is high-FR (more than 124 fps), high-spatial-quality tongue image data that are acoustically aligned and head corrected (1 mm) to the high-FR frame (±4 ms at a FR of 124 fps).

Lingual US imaging of speech has, in the past, been limited on several fronts. First, because the US rays are reflected at the interface with air, only the upper edge of the tongue is imaged, and not the hard bony structures of the palate, jaw, and spine. Furthermore, the tongue tip and root are more difficult to image than the tongue body. This makes the interpretation of tongue contact with the palate during speech difficult to gauge without superimposition of the palate, which is imaged separately during a swallow (Epstein & Stone, 2005). Second, alignment of US video images of the tongue with the acoustic signal, which is paramount in speech studies, has not been previously precisely measured. Third, higher articulatory US FRs (>100 fps) are required to accurately

[a]The Ohio State University, Columbus
[b]Ithaca, New York
Correspondence to Amanda L. Miller: amiller@ling.osu.edu
Editor: Anne Smith
Associate Editor: Maureen Stone
Received May 29, 2009
Revision received December 22, 2009
Accepted September 9, 2010
DOI: 10.1044/1092-4388(2010/09-0103)

trace the dynamics of fast speech sounds, such as stop releases. CHAUSA resolves these alignment accuracy issues and FR issues by exporting the US data with the complex DICOM protocol. The high-FR characteristics of the CHAUSA method (Miller, 2008) were specifically developed to better visualize rapid speech phenomena, such as stop-release dynamics and coarticulation, and they are suitable for lab investigations as well as linguistic fieldwork situations.

This article is organized as follows. First, we provide background on previous US architectures and previous ways of coping with the problem of head position uncertainty in US data collection. Additionally, we review other currently developed high FR US approaches. We then present the CHAUSA computer system architecture, the data collection techniques, and the data analysis techniques that comprise the method. We provide validation of the system before presenting an application of the method in the form of a pilot study on alveolar click production in IsiXhosa. Finally, we discuss the limitations and benefits, and conclude the article.

Background

Standard Video US Systems

Most prior US architectures (Gick, Bird, & Wilson, 2005; Mielke, Baker, Archangeli, & Racy, 2005; Stone & Davis, 1995) are limited to the standard commercial video rate of 29.97 fps. These systems export the US signal out of the VGA or S-video ports of US machines to VCRs or computers. An audio/video (A/V) mixer is used to synchronize the audio and US video signals. Studies using these systems have made great gains in describing vowels (e.g., Stone & Lundberg, 1996), sonorants (Gick, Campbell, Oh, & Tamburri-Watt, 2006), and fricatives (Stone, Faber, Raphael, & Shawker, 1992) that have stable articulatory gestures. Distortions are present with these earlier technologies, but averaging tongue shapes has been successfully used to help identify and minimize distortions (Li, Kambhamettu, & Stone, 2005a). Using VGA or S-video outputs of US machines, with the accompanying acoustic-to-articulatory alignment accomplished via hardware mixing, can cause significant unfixed alignment mixing errors. The US machine image formation time delays the articulatory image compared with the talker's voice audio. In addition, the conversion of the image displayed on the US monitor to the image transmitted out of the VGA port introduces further delay and introduces scan-to-VGA mismatches. These timing mismatches produce spatial artifacts such as multiple scans in the same frame so that one frame contains half of two different scans of the same tongue (Wrench & Scobbie, 2006). The VGA FR limit reduces the speed of the events that can be observed. We have measured these mixing errors from one and a half to three low FR frames, about 45 ms to 99 ms, using the GE LogiqE machine and a Canopus TwinPact 100 mixer.

Head Position Uncertainty

US does not image the bony palate simultaneously with the tongue during speech (Stone, 2005). Therefore, head position uncertainty, and movement of the soft palate, can lead to errors in detecting relative tongue position (Stone, 2005). Both head stabilization and head movement correction techniques have been developed with high accuracy for research lab settings to overcome head position uncertainty. Prior head stabilization techniques include the Head and Transducer Support System (Stone & Davis, 1995), which is a robust stationary system that has been shown to stabilize the head to 1 mm. 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). HOCUS is accurate and reliable but not portable and, hence, not viable for linguistic fieldwork or clinical settings. Portable head stabilization methodology for linguistic fieldwork has also been established and tested by Gick et al. (2005). They use a simple experimental fieldwork 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 with a boom arm in the field). However, portable head and probe stabilization and correction techniques have yet to achieve the level of accuracy of methods developed for lab settings.

Mielke et al. (2005) developed the Palatoglossatron head movement correction technique at the University of Arizona, whereby experimenters videotape subtle changes in head and US probe position, which makes head movement correction techniques adaptable to fieldwork. Palatoglossatron uses a video camera focused on two rods, each containing two dots. One rod is attached to a pair of sunglasses worn by the subject with a strap, and the other rod is attached to the US probe to visually track the movement of the head and the probe. Mielke et al. developed the equations for the rod movement correction for head movement in the images of the tongue and the palate. The standard Palatoglossatron technique mixes the head video with the US signal using a video mixer containing a hardware blue screen mixing option. A modified version of the Palatoglossatron head movement correction technique was integrated into the CHAUSA method because of its applicability to linguistic fieldwork. Palatoglossatron hardware is discussed in the CHAUSA Data Collection Methodology section of this article.

Articulate Instruments (2008) developed an Ultrasound Stabilization Headset, which anchors the US 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 reduces the discomfort implicit in head stabilization. Scobbie, Wrench, and van der Linden (2008) showed from 2 mm to 3 mm of head movement relative to the probe in the mid-sagittal plane during speech. These measurement errors, greater than our 1-mm goal, make it critical to perform head movement correction as well as probe stabilization.

Other High-FR Approaches

Several high FR US approaches are under development. Hueber, Chollet, Denby, and Stone (2008) used the Terason T3000 US machine, which allows synchronization of two video streams (US and optical) with the audio stream. According to Hueber et al., the US and video streams are found always to be synchronized up to 71 fps, and the authors have verified that it is aligned to the correct US frame, with a subtle constraint on small interframe time gaps on this machine. The Hueber approach uses temporally tagged real-time software mixing of the audio with the US video and the optical video, which adds an additional real-time processing burden to the same processor receiving the US video, the optical video, and the audio, which have slightly variable programmable time states. Additionally, much of the misalignment found in prior architectures comes from the delay from the US image formation prior to acoustic mixing, which can be significant. There are also innate differences between the Terason and GE Medical US machines. Hueber et al. have more than doubled (to 71 fps) the FR of standard approaches. However, likely for the reasons stated above, the achieved FR is only slightly more than half the FR of the CHAUSA approach (more than 124 fps) demonstrated in Miller (2008) and in the Application section of this article.

Noiray, Iskarous, Bolanos, and Whalen (2008) have developed a different high-FR approach for a lab setting using the HOCUS architecture (Whalen et al., 2005) 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, as the CHAUSA approach is. Nor has the high FR articulatory-to-acoustic alignment method been fully documented yet. In addition, it is not clear if this architecture has integrated a methodology to anchor the probe to the head. Anchoring, which locks the physical position and direction of the probe with respect to the tongue over multiple takes, is critical to ensure that sounds being compared are imaged from the same perspective throughout a recording session. Furthermore, locking in the optimal imaging position ensures the best quality images for each subject throughout the duration of the study.

Wrench and Scobbie (2008) compared video-based and high-speed cineloop US tongue imaging approaches. A Mindray DP-6600 US machine transmits to an analysis computer with a frame-grabber card, which gives increased control of US video frames. This approach produces deinterlaced video. While the FR of the machine is 98 fps, the output FR is only the standard NTSC/VGA 29.97 fps. With the increased control, they deinterlace this video, which has approximately the same number of transmitted data bits, but the interlaced time layers are separate. The interpretive value of this is not well understood, and their upper FR limit (2 × 29.97 fps), with half the spatial clarity, is well below Hueber's system of 71 fps. To clearly see stop-release kinematics, or to clearly see other somewhat slower speech events such as [l] and [r] without statistical averaging, FRs must be greater than 100 fps. Distortions inherent in a low FR AVI analog video export are still present.

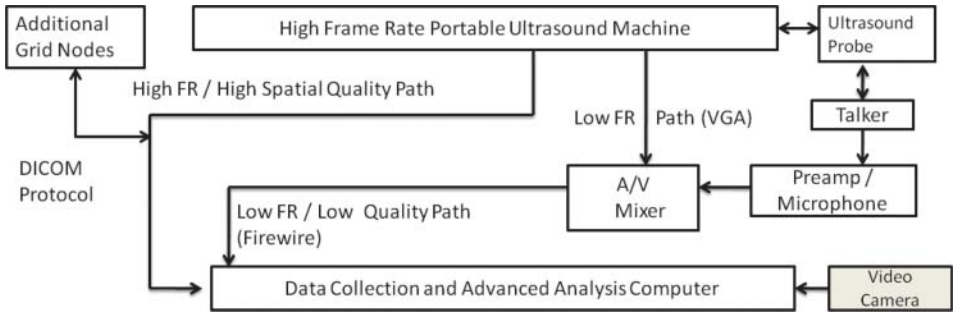
Another high-speed system described by Wrench and Scobbie (2008) is based on an Ultrasonix RP research US machine, which is controlled via the Ethernet port of a host computer. The Ultrasonix RP system, however, has unusually poor spatial clarity. The resolution—using data that is kept in its raw, prescan converted form—is 76 lines with 412 samples per line. The GE LogiqE uses 480 lines with 640 samples per line, as do most US machines. Currently, of all field approaches, CHAUSA has the highest FR (above 124 fps) with the best possible spatial quality, and with expected 1-mm head position accuracy.

High-FR, High-Transmitted-Spatial-Quality CHAUSA Grid System Architecture

Figures 1 and 2 provide the hardware and software that are involved in the CHAUSA architecture. Specific hardware components and software programs are all crucial for the data collection and data analysis phases of the method. Figure 1 provides a schematic of the CHAUSA grid system hardware. The main system components are the high-FR portable US machine, a gigabit Ethernet (GigE) video camera, and at least one data analysis computer.

In the current instantiation of CHAUSA, we use a GE LogiqE machine. The same US video, captured on the GE machine, travels through two distinct hardware paths shown in Figure 1 to the data analysis computer: the *high-FR/high-spatial-quality path* and the *low-FR/low-spatial-quality path*. The US image data are generated by the software on the US machine (first block of Figure 2) at the native high FR of the US machine. The second block of Figure 2 shows the translation of the high-FR US images for export out the low-FR/VGA path. The high-FR/high-spatial-quality path data are transferred directly to a data collection and advanced analysis computer using DICOM software, which is seen in the third block of the software

Figure 1. Schematic of Corrected High–Frame Rate (FR) Anchored Ultrasound With Software Alignment (CHAUSA) grid system hardware. DICOM = Digital Imaging and Communications in Medicine; A/V = audio/video.

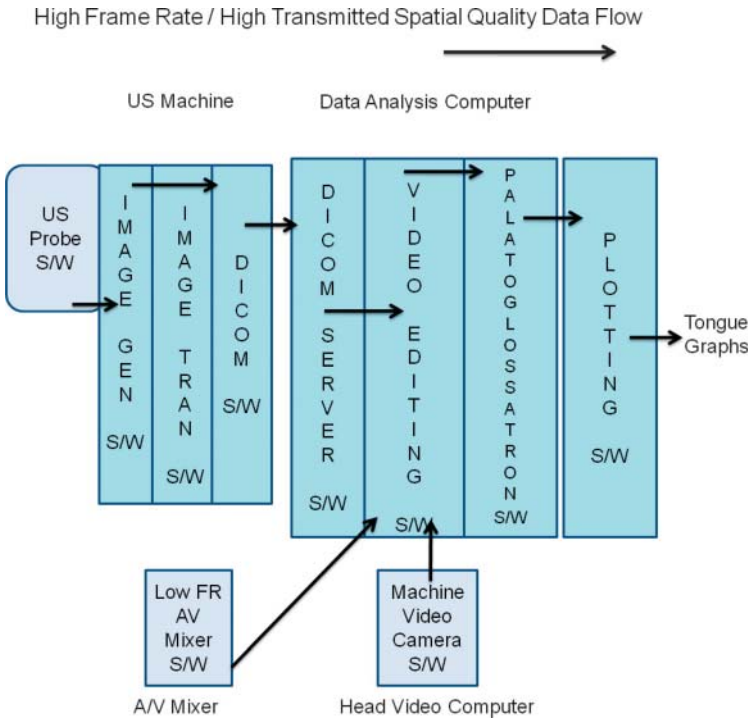


data tools diagram of Figure 2. The DICOM transmission protocol preserves both the native high FR of the US machine and the native image quality during transmission to the data analysis computer. Figure 1 shows "additional grid nodes," which are additional storage, additional data analysis computers, or additional US machines, which can be local or distributed over the Internet network.

A high-FR GigE camera is used to film video of the speaker's head from a side angle in order to capture head and probe movement using Palatoglossatron hardware. The high-FR video camera is captured in a data analysis computer through a GigE connection. The head video and the US video of the tongue are transmitted separately to the data analysis computer through two high-FR digital Ethernet paths.

The low-FR/low-spatial-quality path in Figure 1 uses older protocol conversions and is similar to what most researchers are using to transmit US data synchronized with an audio signal at 29.97 fps, as described in the CHAUSA Data Collection Methodology section. The US

Figure 2. Software data tools for the CHAUSA grid system. US = ultrasound; S/W = software; GEN = generation; TRAN = transmission.



machine video signal is transmitted through the VGA port of the US machine to an A/V mixer. The speaker's voice is captured by a preamp, which is mixed with the US signal in the A/V mixer. The mixed A/V signal is then transmitted through a firewire port to a data analysis computer and received with video editing software (in our case, Adobe Premiere Pro).

The software used in the data analysis phase of the CHAUSA system is shown on the right side of Figure 2. The DICOM server software receives the high-FR US video files from the US machine. The low-FR path audio is then delinked from the low-FR video/audio and mixed with the high-FR US video. The low-FR video is then discarded. The right arrows in the body of Figure 2 indicate that the high-FR, high-spatial-quality image data, originally transmitted by DICOM, move through each software layer to the next layer of software on its right. During this process of different software programs processing CHAUSA data, the FR and image quality remain unchanged. The high-FR head video is mixed with the high-FR US video using video editing software on the data analysis computer. The US video overlaid with the head video is analyzed using Palatoglossatron software. This involves two phases: tracing the tongue edge in the US video and correcting those traces for head and probe movement. Although there are several software packages available for tracing the tongue edge, Palatoglossatron is the only software that has a built-in algorithm for correcting the traced tongue positions for movement of the head and probe captured with the head video. Finally, plotting software is used to make dynamic tongue graphs, which are the output of the CHAUSA architecture. These are discussed in the CHAUSA Data Analysis Component section of this article.

CHAUSA Data Collection Methodology

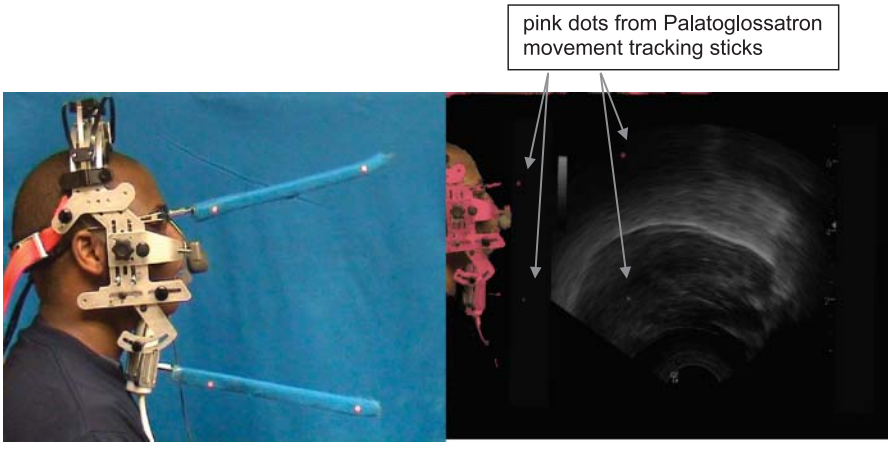
Palatoglossatron Hardware

CHAUSA uses the standard Palatoglossatron hardware designed by Mielke et al. (2005). The hardware is seen in the left panel of Figure 3. A pair of metal rods with the same dimensions used by Mielke et al. was covered with blue felt. 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 low-FR path recording within Premiere Pro. When using the GigE video camera, which has no audio recorder, a second clacker board drop instead of the bell is used to mark the end of the utterance. The lower Palatoglossatron rod is attached to the probe with a screw to measure probe movement. A pair of glasses anchors the upper Palatoglossatron rod to the head. It is crucial to note that the pair of glasses does not touch the headset. This allows independent recording of head movement. The glasses are firmly attached to the head using a strap. Thus, the upper rod moves with the head, and the lower rod moves with the probe. The right panel of Figure 3 is the resulting video image, and it is therefore discussed in the CHAUSA Data Analysis Component section.

Tri-Modal 3-ms Pulse Generator

The articulatory-to-acoustic alignment of the low-FR path audio, and the high-FR, high-spatial-quality US video is undertaken in a post–data collection stage using video editing software. The second author designed the Tri-Modal 3-ms Pulse Generator, which simultaneously produces a 3-ms pulse of US, 3 ms of audio from a buzzer, and 3 ms of light from a high-brightness light-emitting

Figure 3. IsiXhosa speaker wearing an Ultrasound (US) Stabilization Headset that cradles the ultrasound probe, and Palatoglossatron movement tracking sticks (left). The software-mixed US image shows the speaker's tongue surface, superimposed with his head wearing the stabilization headset in the video camera image. The image also shows the small pink Palataglossatron head correction dots (right).



diode (LED) for use in daylight, which is picked up by the various audio and video recordings. The US probe used in the Tri-Modal 3-ms Pulse Generator was designed for therapeutic use by Medical Products Online and is called the International Professional Ultrasound System (http://www.medicalproductsonline.org/mepron10ulsa.html). Standard digital logic design techniques determine the 3-ms pulse width. The three modes have a $\frac{1}{10}$ms synchronicity, which is determined by digital logic design techniques, as given in the LED, buzzer, and US crystal component data sheets. The Tri-Modal 3-ms Pulse Generator has been added to the CHAUSA computer architecture. The therapeutic US probe is placed next to the GE LogiqE probe. An acoustic standoff is used to couple the two US transducers without having the probe heads physically touch. The 3-ms therapeutic US pulse transfers to our standard 8C-RS GE Medical US probe, which marks the US frame in which it occurs with a bright flash in part of one frame. The buzzer is placed close to the microphone, and the LED is taped to the side of the US stabilization headset so that the light flashes are picked up by the GigE video of the head.

Anchoring

Articulate Instruments' Ultrasound Stabilization Headset (Articulate Instruments, 2008) was adopted for use with CHAUSA. The left panel of Figure 3 shows an IsiXhosa speaker wearing the headset. The headset anchors the probe position and direction to ensure optimal imaging across multiple takes. The quality of US imaging of many speakers' tongues is sensitive to small variations in the position of the probe for spatial clarity. A slight movement of the probe can degrade future takes in a multiple-take session.

A/V Mixer

An A/V mixer is a crucial component of the low-speed, low-FR path seen in Figure 1. We chose the Canopus TwinPact 100 device for two reasons. First, it has frame locking, which keeps the audio and video signals that enter the A/V mixer synchronized to within a single frame. Second, it has the appropriate connectors needed. We transmit the video signal from the US machine through the VGA port into the VGA port of the Canopus mixer. The mixer also converts from VGA analog video to the IEEE 1394 digital video standard for transmission out the firewire port. A firewire cable transmits the digital video signal from the A/V mixer to a laptop computer.

Researchers

Data collection requires three researchers, as one person captures each take on the notebook within Adobe Premiere Pro via the A/V mixer, low-FR path; another person saves the individual takes on the US machine hard drive in DICOM format for later transfer via the high-FR/high-spatial-quality path; and a third person saves the high-FR head video files onto a second laptop. One of these researchers also hits the start button on the Tri-Modal 3-ms Pulse Generator. If a standard digital video camera is used, the number of researchers is reduced to two. Educated laypeople with some computer experience can be trained to perform two of these saves (high-FR head video data collection and low-FR-path data collection).

Word List and Frame Sentence Creation

Each take is limited to the maximum 8- to 12-s window that can be recorded with the GE LogiqE machine. The window size varies with FR, and is about 8 s at rates of 114 fps–124 fps. Experiments are limited from 1/2 hr to 1 hr in length, based on the recommended durations of wearing the headset to avoid head and neck strain. To meet these conditions, we limit the word list for one experiment to six to eight words. If the experimenter does not have a Tri-Modal 3-ms Pulse Generator, a frame sentence is developed that contains additional stops that will be used in the manual alignment process. We recommend the use of dorsal stops, as they have distinctive noise bursts that serve as landmarks in the acoustic signal, and the posterior part of the tongue is easiest to image with US. Alveolar clicks are ideal for use in alignment because of their abrupt releases (Thomas-Vilakati, 2008) and clear, abrupt anterior release bursts (Ladefoged & Traill, 1994), but not all speakers of nonclick languages can produce clicks. The experiment name, word list number, and talker initials are written on the clacker board so that they are filmed in the video of the head.

Video Camera

A video camera records head movements (using the Palatoglossatron rods) concurrently with the US video recording. In the IsiXhosa data presented in the Application section, a standard 29.97-fps video camera was used, whose output was brought in post hoc through a synchronous firewire port of the notebook. Miller, Scott, Sands, and Shah (2009) replaced the standard video camera with a Prosilica GE 680C camera that has an adjustable high FR. The high-FR GigE camera is set to match the FR of the US data and can capture any FR up to 200 fps. The GigE camera stream is captured digitally on a notebook computer with a fast ($\geq$ 7,200 rpm) hard drive.

CHAUSA Data Analysis Component

We now turn to a description of the data analysis components of the CHAUSA method. We discuss the

components in order of application: (a) DICOM transmission of the files from the US machine to the data analysis computer; (b) articulatory-to-acoustic alignment of the high-FR, high-spatial-quality US AVI files to the audio track of the low-FR-path AVI files; (c) mixing of the head video with the high-FR US video and low-FR audio; (d) tracing the tongue edge and performing head movement correction; and (e) plotting the tongue edge and palate data, in concert with analysis of the acoustic data.

DICOM Transmission

The Ethernet port is capable of perfectly transmitting the high-spatial-quality, high-FR DICOM video. The transfer of data is done after the end of a study, as the US machine hard drive can store DICOM US data from an 8-hr recording session. The US machine, the GE LogiqE, also stores high-spatial-quality, high-FR cineloops on its hard drive with its own storage protocol. Although these high-quality cineloops can also be stored to an external hard drive or CD, the networking solution allows these data to be transferred at higher nonstandard frame rates, such as the 124 fps used in the pilot study presented below. The GE implementation of writing to an external drive uses a standard disk write (limited to 50 fps, not DICOM rates) that is different from the custom disk write to its own internal hard drive. The robust quality and high-FR options of DICOM are not available when writing to external storage if DICOM is not installed.

The notebook computer receives the high-quality, high-FR DICOM images via the Ethernet port and the real-time A/V mixed images via the firewire port but at different data collection times. The real-time A/V mixed video is recorded by streaming directly to the laptop computer during recording. The delay implicit in the low-FR path comes from processing times in the GE LogiqE machine itself.

Articulatory-to-Acoustic Alignment

After the DICOM transmission, the CHAUSA method uses non–real time software mixing to achieve "to-the-frame" articulatory-to-acoustic alignment precision and to avoid locked-in hardware mixing errors caused by delays in video export from the US machine and possibly A/V mixer errors. The best way to resynch the audio to video, if such delay is present and undesired, is to import the AVI file into video editing software, unlink the audio from the video, and manually realign and relink according to known articulatory landmarks. Not all A/V hardware mixers have mixing errors. The Canopus TwinPact 100 specification claims frame-locked or "perfect" mixing. However, more important is the delay in the scan image creation time in the software of Figure 2, and those file conversion times, to convert from internal video memory to the VGA standards for output to the VGA port, also shown in Figure 2. We measure total mixing error (image creation and image translation) in our low-FR path to be between one and one half and three 29.97-fps frames (with an average of 2.2 frames). This is determined by closely examining the low-FR alignment of the click, by comparing the to-the-frame high-FR audio-aligned video (accurate ± 4 ms) to the exact same articulatory and auditory click in the low-FR mixed A/V path. The measurement accuracy is limited by the accuracy of the high-FR alignment. The net result is that the normal VGA export mixing error varies from 45 ms to 90 ms, averaging 65 ms, with a measurement error of 4 ms. In a previous experimental setup, there was a For.A brand video-video mixer in the critical path between the VGA port and the recording computer, and an older US machine (GE LogiqBook) was used. Delays averaged four 29.97-fps frames, or 122 ms of error. CHAUSA removes this error.

In the data collection phase, the Tri-Modal 3-ms Pulse Generator marks each of the video and audio recordings. In the data analysis phase, the marked data of each mode is synchronized by sliding the video track until the three synchronizing marks line up. Without the Tri-Modal 3-ms Pulse Generator, a manual alignment process may be used. The hardware-mixed low-FR, low-spatial-quality video is used as a known starting alignment basis for the A/V alignment process of the high-FR video principally by having starting and stopping information embedded, including a video clacker board at the beginning and a bell (or second clacker board) at the end. The final fine alignment is accomplished using high-FR articulatory-to-acoustic speech events. The hardware mixed video is particularly useful for understanding field data collection issues such as identification of different starting and stopping times in the different modalities. This is particularly important because the recording window of the GE LogiqE machine is a moving window that continues recording new material at the end, while losing recorded material at the beginning, if the recording is not stopped prior to the end of the 8–10 s. The low-FR path video and audio are also available as a backup for manual alignment in case the Tri-Modal 3-ms Pulse Generator stops working in the field.

Mixing the Head Video With the High-FR/High-Spatial-Quality-Path US Video and Audio

The commercial video editing software used for the alignment of the data is also used to mix the Palatoglossatron video with the high FR US video. This time, we are mixing two videos rather than aligning video with audio. First, we use Chroma keying to remove the blue screen

background of the head video file, replacing it with a transparent background. Next, we mix the transparent head video with the mixed audio/high-FR US video file, being careful to position the lower arc of the US cone in the position expected by Palatoglossatron. The US cone relates to the origin of the US probe and the conical fan out of the US rays from this origin. The US cone will line up with the cone-shaped grid overlaid by the Palatoglossatron software to track the tongue edge. This exact positioning is needed by the Mielke head-correction equations, and precision depends on good video editing tools. Furthermore, the head video should show a clear and large view of the pink dots on the Palatoglossatron rods.

After mixing the files, we export the video-mixed AVI file out of Adobe Premiere Pro. The right panel of Figure 3 provides a single frame of the software-mixed US video containing the speaker's head with the stabilization headset, the head-correction dots introduced via the Palatoglossatron rods marked with arrows, and the DICOM-transmitted US image of the tongue. The speaker and the stabilization headset are to the left, the hard-to-see pink colored dots are to the right of these, and the synchronized frame of the US video is in the center. The stabilization headset is colored because of the blue-screen removal process that software-mixed the US video with the stabilization headset video containing the Palatoglossatron rod. A MATLAB script is used to convert the blue-screen mixed AVI file (right panel of Figure 3) to a series of JPEGs required by Palatoglossatron.

Tracing the Tongue Edge, Performing Head-Movement Correction, and Plotting the Data

Mielke et al. (2005) developed Palatoglossatron software, which allows the researcher to trace the tongue and palate in the images of interest as well as perform head-movement correction on these traces from the locations of the pink dots in the mixed video. For the tongue images to be head-corrected, the pink dots are precisely identified graphically so that the Mielke formulas can operate according to the different movements of the rods in each high-FR frame to adjust and correct the traced tongue edges. The software corrects for head position, as the dots change on each JPEG, independent of FR. This is one reason why the Palatoglossatron software, which was originally designed for 29.97 fps, still works at 124 fps. Although other tongue-edge-tracing software is available, such as EdgeTrak (Li, Khambamettu, & Stone, 2005b), we use Palatoglossatron because of its head-movement-correction capabilities. We also use a head- and probe-movement correction software, in addition to Palatoglossatron software, called Peterotron (Arizona Phonological Imaging Laboratory, 2009). The Peterotron method is described fully on the Arizona Phonological Imaging Laboratory wiki (Arizona Phonological Imaging Laboratory, 2009). This software ensures accurate head-movement correction.

Peterotron software exports a numerical description of each head-corrected tongue trace in a table. The Arizona Phonological Imaging Laboratory (2009) developed scripts to convert this format to the smoothing spline analysis of variance (SSANOVA) format. SSANOVA is a statistical package within the *R* statistics software that allows pairwise comparisons using SSANOVA models (Davidson, 2006). In the pilot study presented below, we plot the tongue surfaces using Microsoft Excel, as we are plotting more than two traces allowed by the SSANOVA script together. The head-corrected data, as opposed to the uncorrected data, is presented in a transformed spatial coordinate space by the Mielke equations.

The CHAUSA dynamic tongue graphs are the output of the CHAUSA method. They are a variation of waterfall tongue graphs, which have a third dimension of time, to show tongue movement over time (see, e.g., Li et al., 2005a). In clicks and other fast consonants, the tongue moves so fast that waterfall graphs can seem difficult to understand. The data also indicate that the tongue sometimes covers a good portion of the total movement range in a very short time, and later in the same click production, the same physical distance movement takes four or five times that amount of time. For clarity, we have chosen to space, more or less evenly, the tongue movement in physical space, regardless of the time between tongue positions, in our dynamic tongue graphs. This means, in the abstract, that the third time dimension is plotted neither linearly nor logarithmically but in arbitrary and varying units of time depending on the actual articulation movement details. This could make for a very confusing third plotted time axis. For this reason, our version of the waterfall type graphs actually have the third time axis essentially projected into two dimensions, with the actual timeline indicated by annotations of the tongue-trace numbers given in ms from the first tongue trace. They are, nevertheless, dynamic tongue graphs that had an implied time alignment with an acoustic waveform. The annotated times allow cross-references with audio in a timeline (e.g., in a waveform or spectrogram).

The tongue traces from different frames often end up being variable in length because of difficulties in tracing the tongue tip and root in particular images. Improved US imaging in newer high-performance machines helps but does not resolve this issue. The surfaces software package developed by Li et al. (2005a) contains an option to krig the tongue traces so that each of the tongue traces is the same length prior to averaging. If not using this software, an independent method for controlling for

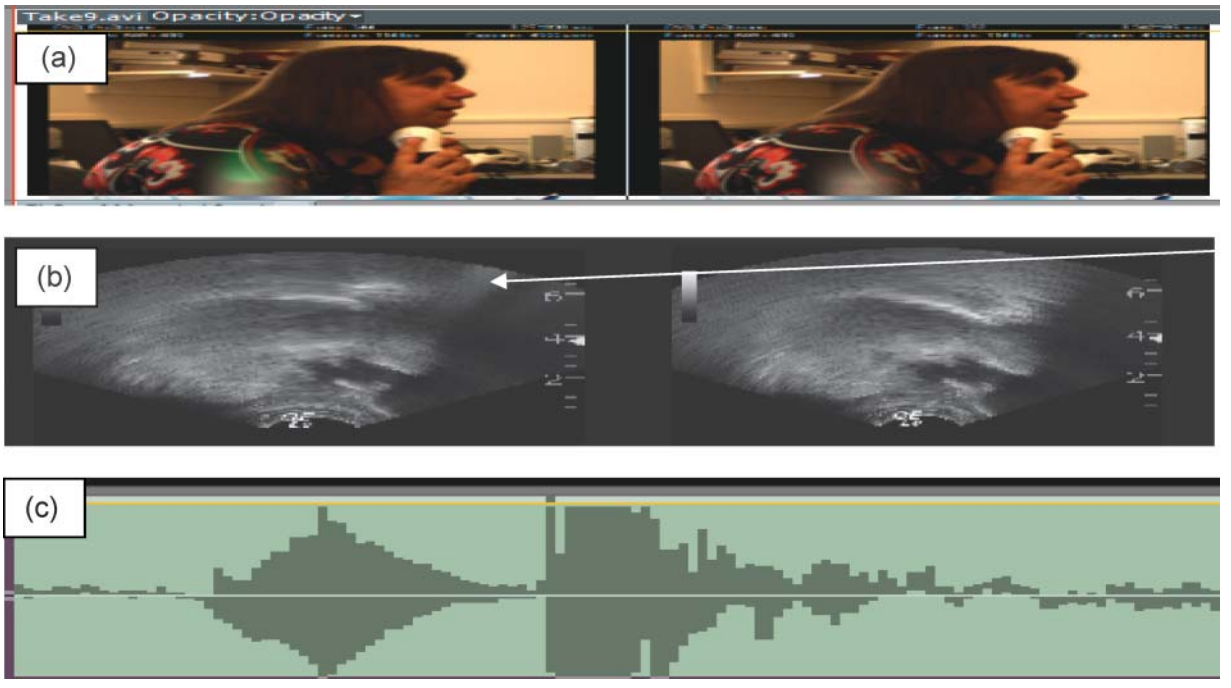
tongue length is necessary. We used a map-measuring wheel, available from a drafting supply store, to measure the length of the tongue in each of the traces in the graphs presented in the pilot study. We are assuming that the tongue does not substantially change length from one frame to another in eight thousandths of a second. In certain well-known situations, such as the tongue moving from mostly vertical to mostly horizontal in a single frame, we can credibly assert that the tongue length, while vertical, maintains its standard premeasured length. Therefore, dots were added to the tip of the tongue edge in Palatoglossatron software to attain tongue edges that were all the same length. At this stage of our knowledge, this was possible only because there were supporting US light spot markers when the tongue tip first contacts the hard palate. In prior technologies, these light spots were more difficult to interpret, but in high-FR data, they can more often be understood. By going back and forth in 8-ms frames, the exact moment of contact and release can be determined in many cases. The tongue length was determined both by the length to the contact spot from the US visible tongue and by the assumption that the length remained constant over an 8-ms interval. If these both supported each other, we would then show that length in the dynamic tongue graphs.

Validation

Proof of Alignment With the Tri-Modal 3-ms Pulse Generator

To provide proof of alignment, a manual alignment process is first completed using video editing software. Next, the signals are examined to see if the various tri-modal synchronizing marks line up. Figure 4 provides the head video recorded by the Prosilica GE 680C camera in (a), the high-FR-path US video in (b), and the audio signal recorded in the low-FR path in (c), of the first author's production of the [!] click. First, we aligned the recordings manually by aligning the three tokens of the anterior click releases in the US signal with the three acoustic release bursts in the audio track of the mixed video. Then, we checked to see if the synchronizing marks, the 3-ms US pulse, the 3-ms buzzer, and the 3 ms of light in the head video, were also aligned. Because we know that the US pulse, the lit LED, and the buzzer occur within $\frac{1}{10}$ of a millisecond 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-FR frame. The US crystals scan from left to right. The 114-fps US frame rate means that we are practically limited to increments of 8 ms in the US signal. However, we can, in some sense,

Figure 4. Adjacent frames during the release of an alveolar click in the Ju|'hoansi phrase *Ha !áí.* "S/he died." produced by the first author. Top panel (a) shows head video with the light-emitting diode flash visible in the left frame. Middle panel (b) shows high-FR ultrasound video with therapeutic ultrasound signal flashing in upper right corner of left-most frame, a faint but distinctive narrow *V* shape that is marked by an arrow. Bottom panel (c) shows audio signal with the 3-ms buzzer signal on the left and the click burst on the right.



see better resolution in the US video than the 114-fps FR allows because of the position of the 3-ms pulse within the 8-ms frame. Note that the US pulse is located later in the time scan position of the left frame. This indicates that the next frame (the right one) is about to happen in approximately the length of time between the buzzer burst and the acoustic click burst, a fraction of an 8-ms frame, because the length of time (several ms) between the buzzer and the click burst is the same as the distance between the therapeutic US pulse and the anterior click release. Such a fraction of frame-alignment precision has never before been possible.

Temporal Errors

With Tri-Modal 3-ms Pulse Generator alignment, the error is ± .05 ms. Using manual alignment, at 124 fps, an error band of 4 ms is achieved, with the understanding that the error bar starts at the center of the 8-ms frame.

Spatial Errors

Tracing error that would occur when the tongue and palate edge are traced within the Palatoglossatron software is one source of spatial error found in the CHAUSA method. Six pixels per mm are found in the 640 × 480 pixel frame size of the US images mixed with the Sony video camera used in this study. It is straightforward to trace errors to within one or two pixels, leading to a data analysis error of 16%–33% of 1 mm. This data collection error is small compared with our claim of the accuracy of the data and supports the claim that 1-mm accuracy could be possible with the US tongue traces. In the pilot study described next, the tongue moves a great deal within 8 ms. We know by Palatoglossatron dot changes that the head is also moving fast in relation to the probe in these frames. For best accuracy, the high-FR GigE video camera is required so that each high-FR US frame is corrected for head and probe movement.

At present, we can induce indirectly the limit to the inaccuracy of Palatoglossatron spatial results. We hypothesize, by induction, that the spatial inaccuracy of the entire analysis process is ≤ 1 mm. This induction is explained more fully in the next section. This 1-mm error band is the sum of tongue- and palate-tracing errors, temporal alignment errors between the head video and US video, and spatial dot–tracking errors.

Application: Pilot Study on Alveolar Click Production Using CHAUSA

The CHAUSA method was applied to the investigation of the articulation of the alveolar click release in IsiXhosa in order to view the dynamics of the anterior and posterior releases of this click. Previous US investigations of clicks in Khoekhoe (Miller, Namaseb, & Iskarous 2007) and N|uu (Miller, 2010; Miller et al., 2009), as well as X-ray recordings of !Xóõ clicks (Traill, 1985) have all been sampled at 29.97 fps. All of these previous studies show large gaps in the very fast releases of alveolar clicks and aliasing effects. 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) provided the first results of this study.

We collected CHAUSA data of one IsiXhosa speaker's production of the utterance *Ndi qaba isonka.* [ⁿdi !ɑbɑ isɔŋgɑ] "I spread something on the bread." The sentence was repeated three times in each take, and five takes were recorded, yielding 15 repetitions of the target sound in the same phonetic context. Twenty-five to 30 clearly visible frames were obtained during the production of the alveolar click, which showed a remarkably consistent pattern without the 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. 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 s prior to swallowing, rather than swallowing immediately. The [!] click anterior releases and the [g] releases were used to align the high-FR/high-spatial-quality path video and the audio of the low-FR path. Because three repetitions of the sentence were recorded in each take, these yielded six independent events for alignment in each take.

Seven ordered but nonconsecutive tongue traces exhibiting the major stages of click production and a single-palate trace are provided in Figure 5. The results before head-movement correction are provided in the upper panel of Figure 5, and the head-movement-corrected version of these traces is provided in the lower panel of Figure 5. The traces in the upper panel of Figure 5 show probe-to-head anchoring using the Ultrasound Stabilization Headset (Articulate Instruments, 2008) with no head-movement correction. The lower panel of Figure 5 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 hard palate that can be traced during a swallow is the thick, solid, black line that can be seen toward the top of these graphs. As noted by Wilson (2006), a swallow is not a rest position, and thus palate images traced from swallows do not necessarily correspond to a neutral vocal tract position, nor do they correspond to a speech position.

Articulatory-to-acoustic alignment, which is correct to the US frame (± 4 ms), allows us to match the articulatory events seen in the US traces with acoustic representations.

Figure 5. Ultrasound frames traced from an IsiXhosa alveolar click collected with an Ultrasound Stabilization Headset with no head-movement correction (upper panel) and with head-movement correction (lower panel) in the utterance *Ndi qaba isonka.* [ⁿdi !aba isɔŋga] "I spread something on the bread."

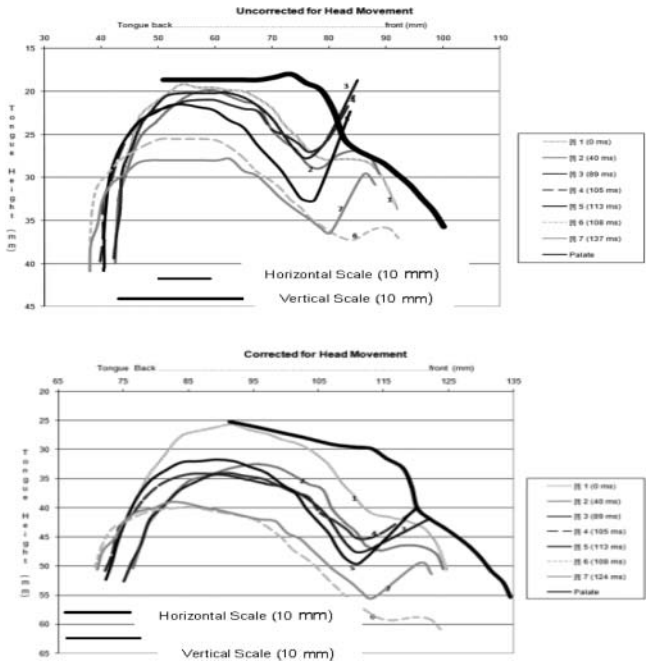
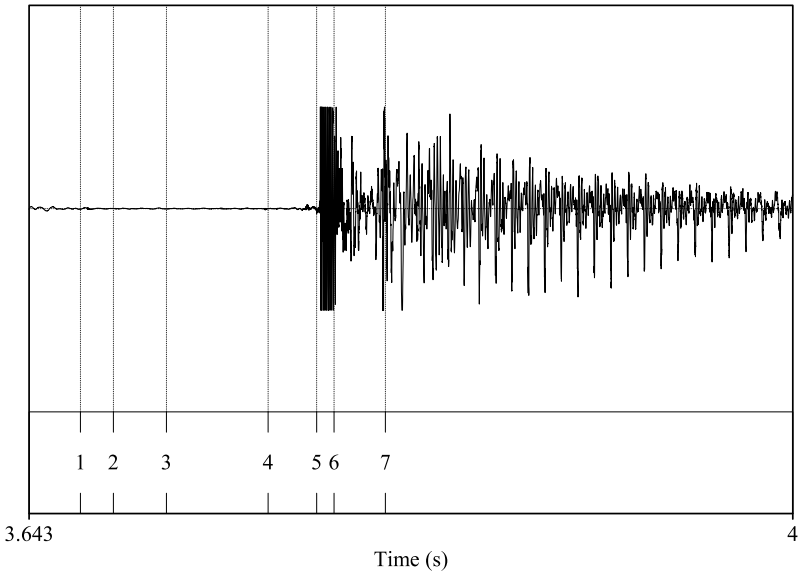


Figure 6 provides a waveform of the alveolar click with the acoustic events that correspond to the tongue traces of particular US video frames shown in Figure 5. Each of the numbered acoustic markers corresponds to the articulatory tongue traces provided in Figure 5. Notice that Trace 5, which shows the tongue shape just before the anterior release, occurs milliseconds before the click burst.

Traces 1–7 in Figure 5 show a complete cycle of alveolar click production. Low-FR and high-FR slideshows of this click are provided at http://www.chausa.info/qualityimages.html. 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 five frames (40 ms) later, the tongue body has lowered, resulting in two swells in the tongue. In Trace 3, which is six frames from the second trace and 11 frames (89 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, which is four frames from the third trace (13 frames, 105 ms from the first trace), the tongue tip has retracted slightly and is contacting 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 (2008). Trace 5, which is one frame from the fourth trace (14 frames, 113 ms from the first trace), corresponds to the point on the waveform just before the anterior click burst seen in Figure 6. The tongue dorsum has retracted even more at this point and shows the maximal cavity expansion just prior to the anterior release. Trace 6, which is only one frame later than Trace 5 (15 frames, 121 ms from the first trace), shows the tongue root in the pharyngeal region for the vowel [a] following

Figure 6. Waveform of an IsiXhosa alveolar click in the word *qaba* [!aba] *spread* with labeled tongue trace numbers corresponding to articulatory ultrasound traces in Figure 5.



this click in the word *qaba* [!aba]. The front of the tongue has released completely and is down low in the mouth. Surprisingly, Trace 7, which is two frames (16 ms) from Trace 6 (17 frames or 137 ms after the first), 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.

Holistically, the dynamics of click production are seen clearly in the tongue traces in the lower panel of Figure 5. The tongue dorsum retracts, and the tongue body lowers to enlarge the click cavity prior to the anterior click release (Trace 5). These overall dynamics can also be seen relatively clearly in Traill's (1985) alveolar click X-ray movies when viewed frame by frame with modern video tools, even if the exact instant of release is uncertain. Tongue Traces 1–5 display the tongue dorsum and tongue root retraction and the tongue body lowering. The intermediate unshown frames move continuously and smoothly along approximately 1-mm increments. We induce from all of these 25 consistent traces as they change over time that the entire body of data is accurately head corrected over time and is placed correctly within the oral cavity. Conversely, the uncorrected traces in the upper panel of Figure 5 are inconsistent with the X-ray data and with the description of click cavity formation described in Traill. The data are internally consistent to within approximately 1 mm throughout the cavity and throughout the duration of the click. Inducing the 1-mm hypothesis from the smooth movements of a large number of corrected tongue movements, although less firm than an analytical proof, has precedence in many other cases of scientific inductions from large amounts of data.

The tongue traces in Figure 5 show much more detail than has been seen previously in click production studies because of the low FRs of previous US and X-ray methodologies. 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 in US studies with lower FRs (Miller et al., 2009; Miller, Namaseb, & Iskarous, 2007).

There are anomalies in the plotting of the transformed space, and the vertical axis plot units are different in the before and after transformation plots. The horizontal axis units are also different; however, the percentage change of the vertical axis before and after transformation is much greater than that seen in the horizontal axis. This produces a plotting illusion, not in the underlying data, as the head correction rotates the tongue more vertically than the tongue or palate shape changes. Furthermore, Mielke (personal communication, 2008) suggested that the plotting software spline choices could produce similar distortions. We confirmed Mielke's prediction that Excel plotting software distorted shapes under plot rotation, such as by spline-rounding error, by an informal study with computer-generated shapes. In the future, the authors will be alert to such illusions and plotting anomalies.

The inaccuracies seen in the tongue traces on the upper panel of Figure 5 clearly show the need for head-movement correction in addition to probe-to-head anchoring for lingual US imaging. The use of the palate, hard or soft, as an articulatory landmark seen by US lightening the moment the tongue touches the palate for US studies is encouraging. The movement of the soft palate during speech must lead to caution in interpreting tongue positions relative to the soft palate in US data. X-ray data, which image the palate and the tongue simultaneously, are easier to interpret. However, collection of X-ray data is not safe, which inhibits its use in speech studies. Interpreting the moment of contact between tongue and palate by changes in lightness of the US images requires close examination of several high-FR, high-spatial-quality frames before and after the moment of presumed contact. These can be seen in the slideshow on the CHAUSA website.

Discussion: Limitations, Benefits, and Conclusions

Limitations and Costs of the CHAUSA Approach

1. Capturing directly on US machines is limited to 12-s takes.

   With FRs of 114–124 fps, the recording window length on the GE LogiqE machine is limited to 8–10 s per take. With three to four repetitions of the frame sentence in a single take, we can record large amounts of data. In the study presented here, we were able to record about two hundred fifty 9-s takes before DICOM transmission of the data to the data analysis computer. The method is not suitable for sociolinguistic experiments in which a large corpus of continuous data is needed.

2. Data collection is more time consuming per take.

   Recording forty 9-s takes and saving the data to the machine's hard drive each time took about 2 hr. Recording the same amount of data continuously streaming through a video mixer would take less than 15 min. If a qualitative description of the phenomenon

under study is sufficient, total data analysis time could be less than prior approaches because a single CHAUSA take is a more accurate representation of the phenomenon.

3. The CHAUSA architecture and method require two to three researchers.

   Two researchers are needed to start and stop the high-FR and low-FR-path recordings. To gain the full benefits of head-movement correction on every frame, a high-FR camera is needed, which requires an additional researcher. The Tri-Modal 3-ms Pulse Generator reduces the requirement to one PhD researcher and two educated lay assistants and reduces the data analysis effort.

4. The analysis of US data is time consuming.

   The addition of DICOM transfer of the high-FR data, file conversion from DICOM to high-FR AVI, the articulatory-to-acoustic alignment, and the Palatoglossatron head-movement correction procedure makes the entire data analysis phase more time consuming than standard low-FR US data analysis. The Tri-Modal 3-s Pulse Generator reduces the alignment time to about 2 min per take. Again, because of the increased precision, the data collection can be reduced to fewer repetitions to capture the phenomenon under study, and, therefore, total data analysis time could be less than prior approaches. Seven traces of the high-FR data can accurately capture the dynamic rarefaction gestures associated with clicks. It is not necessary to trace every frame.

5. The CHAUSA approach is somewhat more expensive than other approaches.

   The DICOM software, DICOM server software, video editing software, and a data analysis computer are needed. A used GE LogiqE US machine costs $18,000 or less. The US stabilization headset is an added expense; however, much less expensive approaches work adequately. Therefore, the added cost of DICOM, the DICOM server software, and video editing software is $7,850 if new software is purchased. As with other computer hardware and software, this cost may drop substantially in the future.

Benefits of the CHAUSA Approach

1. CHAUSA data have more than quadrupled the FR from 29.97 fps to 124 fps or higher.

   The results of this study would not have been seen clearly at the 71-fps FR achieved by Hueber et al. (2008). The GE medical machine has shown adequate images at FRs up to 165 fps, and the CHAUSA method is itself bound only by the scan rate of the US machine, which may improve in future models.

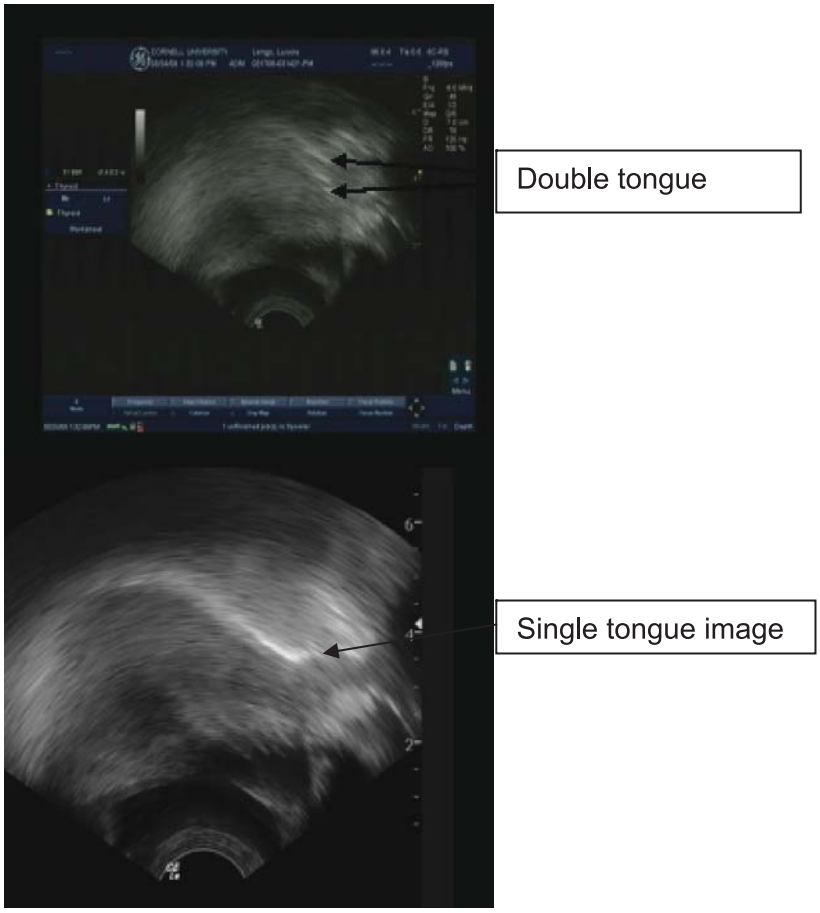
2. Reduced distortion and artifacts improved transmitted spatial clarity.

   Figure 7 provides two images taken at the same moment in time (the anterior release of the click), thus allowing us to clearly see the differences in terms of spatial quality between the two types of image transmission. The low-FR-path image is in the upper panel of Figure 7, and the non–real-time high-FR path used in the CHAUSA method is in the lower panel of Figure 7. Innate differences in image sizes are due to the inclusion of screen information in the streamed low-FR data (upper panel) that are not transmitted in the high-FR data (lower panel). This pair of frames was chosen because the distortion differences are great. The frame in the upper panel of Figure 7 shows two spatially distinct images of the same tongue in the same frame indicated by the arrows, 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. High spatial clarity, in addition to high FR, is a significant benefit for data analysis, and the two benefits are sometimes additive for the interpretation of fast articulatory events. Furthermore, the posterior part of the tongue is much more difficult to see in the 29.97-fps video low-FR-path image in the upper panel of Figure 7 than in the image transferred with the high-FR/high-spatial-quality path at 124 fps shown in the lower panel of Figure 7.

   Figure 8 provides two adjacent frames showing the anterior release of the palatal click [ǂ] in Mangetti Dune !Xung. The two frames are only about 8 ms apart. The tongue tip is raised in the first frame in the left panel of Figure 8 and lowered in the second frame in the right panel of Figure 8. The tip of the tongue is not fully visible in the left frame, a common US effect for a vertical tongue tip, but was seen visibly rising in several adjacent preceding frames (not shown), and is seen lowering in the following frames (not shown). Note the clarity of the data showing the click release in these frames. The change in tongue position between the left and right images (Δ8 ms) is abrupt (Miller et al., 2009). Constant exposure to this high-FR data produces the understanding of various fast and less fast speech events. Both high FR and high spatial clarity assist the correct interpretation of fast speech events.

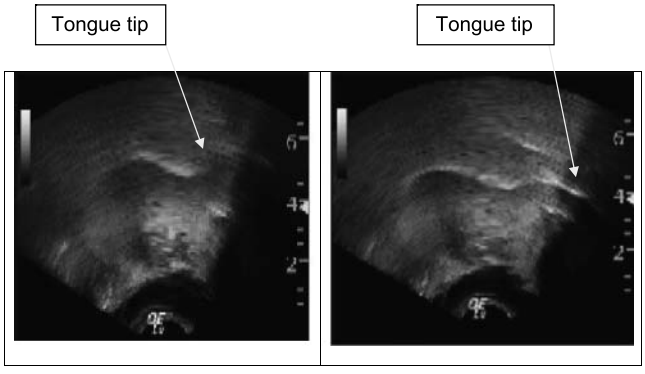
   Earlier images of the palatal click release in Khoekhoe collected at 29.97 fps with the GE LogiqBook machine had the tongue root obscured by the hyoid shadow, which moves with the hyoid during the release (Miller et al., 2007). Single stepping of high-FR, high-spatial-density images shows clearly gradual tongue movements becoming unambiguous and shows fast movements becoming visible. For example, the

Figure 7. A single US frame of mixed video transferred through the low-FR path at 29.97 fps (upper image), and the same US frame of mixed video transferred through the high-FR path at 124 fps (lower image).



left frame of Figure 8 only definitively shows a raised tongue tip. However, by single stepping between the before and after frames, the actual tongue movement and length can be known. The clarity of time-sequenced images of most of the tongue length has not been available with prior clinical and phonetic fieldwork techniques.

Figure 8. Close-up of two adjacent frames of the high-FR US video illustrating the anterior release of the palatal click in the Mangetti Dune !Xung word ǂii *malaria*.



3. High-FR Palatoglossatron head-movement correction allows us to compare tongue positions in different consonants recorded separately (Mielke et al., 2005).

The Palatoglossatron head movement correction method is portable and inexpensive as well as 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 article strongly suggests that head movement correction is required to make sense of articulatory movements. The goal of 1-mm accuracy appears within reach. CHAUSA's high-FR modification to Palatoglossatron is required for studies of fast articulatory events if the absolute position of the tongue with respect to the palate, or to other tongue positions, is necessary.

4. Anchoring using the Ultrasound Stabilization Headset allows us to lock in optimal image quality throughout a recording session (Articulate Instruments, 2008).

   Probe-to-head anchoring allows the researcher to determine the optimal image quality for each speaker and to lock this in over the duration of the recording session. The best probe position is specific to both the speaker and the phenomenon investigated.

5. The GE LogiqE machine has superior image quality compared with its predecessors.

   Improvements that occurred in the new design of the LogiqE—compared with its predecessor, the LogiqBook—are a major reason why a light touch produced excellent images and are one reason why high-FR US imaging is possible with the CHAUSA method. We expect that image quality will continue to improve with future generations of US machines.

   Minimal probe perturbation of the jaw is also a benefit of the GE Medical LogiqE machine used in CHAUSA; thus, anchoring does not likely cause compensatory lingual movements because of perturbation of the jaw, as found in earlier studies (Lindblom, Lubker, & Gay, 1979).

6. Precise articulatory-to-acoustic alignment (± 0.05 ms) allows for synchronized high-quality articulatory and acoustic analyses.

   The ultrasound data collection is relatively quiet compared with electromagnetic articulography (EMA) data collection, allowing better simultaneous acoustic recordings. The system is thus an excellent one for research into the articulatory-to-acoustic mapping of speech.

7. Dynamic tongue-to-palate graphs are accurate to ± 4 ms and 1 mm from a single take.

   CHAUSA dynamic tongue graphs, where the time axis is projected onto the spatial graphs, allow us to better visualize articulatory movement by evenly spacing tongue movements in physical space, regardless of the time between traces. These graphs capture kinematics because the time intervals can be very brief or very long.

Conclusions

The CHAUSA grid computer system architecture and method bring together distinct pieces of hardware and software as well as integrate them to produce a unified goal. This architecture produces high-FR, aligned, and accurately head-corrected dynamic US tongue graphs. Each piece of the architecture and method described here is needed in order to achieve these results. Results provided here show that the CHAUSA method is capable of capturing high-FR data, a fourfold improvement in FR, for the investigation of fast articulatory events (especially stop releases) in linguistic fieldwork. An alignment procedure was described, which achieves to-the-frame alignment (± 4 ms at 124 fps) of the acoustic and articulatory signals. Results also show good positioning of the tongue relative to the palate. A hypothesis has been stated that the accuracy is close to 1 mm, based on the closeness of the tongue to the palate in the tongue traces in Figure 5 and in the tight spatial and temporally 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 architecture and method open the possibility for studies of rapidly articulated consonants and vowels as well as studies of co-articulation that have not been possible with current US methodology. US studies of speech allow us to view most of the tongue, which contrasts with other high-FR articulatory methods such as EMA that track only a finite number of flesh points. The improved accuracy of CHAUSA means that multiple takes can be used for the quantitative investigation of linguistic variation, as opposed to averaging out error-induced variation. CHAUSA makes possible the discovery, in field and in clinical settings, of an important new body of knowledge about speech kinematics.

Acknowledgments

Development of the CHAUSA method was supported by National Science Foundation grants BCS-0726200 (Amanda Miller, principal investigator) and BCS-0726198 (Bonny Sands, principal investigator), titled "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 thank Abigail Scott, who assisted with the tri-modal proof-of-alignment data collection. We also thank our IsiXhosa speaker, Luxolo Lengs, and our Mangetti Dune !Xung speaker, Jenggu Rooi Fransisko.

References

Arizona Phonological Imaging Laboratory. (2009). *The APIL ultrasound manual*. Retrieved from http://apil.arizona.edu/labmanual/index.php?n=Main.HomePage.

Articulate Instruments Ltd. (2008). *Ultrasound Stabilization Headset user's manual, Revision 1.3*. Edinburgh, United Kingdom: Author.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Ashdown, J. (2003). Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated

by trained listeners. *Clinical Linguistics & Phonetics, 17,* 199–216. doi:10.1080/0269920031000071451.

Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America, 120,* 407–415. doi:10.1121/1.2205133.

Epstein, M. A., & Stone, M. (2005). The tongue stops here: Ultrasound imaging of the palate. *The Journal of the Acoustical Society of America, 118,* 2128–2131. doi:10.1121/1.2031977.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association, 32*(2), 113–122. doi:10.1017/S0025100302001007.

Gick, B., Bird, S., & Wilson, I. (2005). Techniques for field application of lingual ultrasound imaging. *Clinical Linguistics & Phonetics, 19*(6–7), 503–514. doi:10.1080/02699200500113590.

Gick, B., Campbell, F., Oh, S., & Tamburr-Watt, L. (2006). Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. *Journal of Phonetics, 34,* 49–72. doi:10.1016/j.wocn.2005.03.005.

Hueber, T., Chollet, G., Denby, B., & Stone, M. (2008). Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the Eighth International Seminar on Speech Production* (pp. 365–368). Strasbourg, France: INRIA.

Ladefoged, P., & Traill, A. (1994). Clicks and their accompaniments. *Journal of Phonetics, 22,* 33–64.

Li, M., Kambhamettu, C., & Stone, M. (2005a). Tongue motion averaging from contour sequences. *Clinical Linguistics & Phonetics, 19*(6–7), 515–528. doi:10.1080/02699200500113863.

Li, M., Kambhamettu, C., & Stone, M. (2005b). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics, 19*(6–7), 545–554. doi:10.1080/02699200500113616.

Lindblom, B. E., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics, 7,* 147–161.

Mielke, J., Baker, A., Archangeli, D., & Racy, S. (2005). Palatron: A technique for aligning ultrasound images of the tongue and palate. *Coyote Papers, 14,* 97–108.

Miller, A. (2008). Click cavity formation and dissolution in IsiXhosa: Viewing clicks with high-speed ultrasound. In R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the Eighth International Seminar on Speech Production* (pp. 137–140). Strasbourg, France: INRIA.

Miller, A. (2010). Tongue body and tongue root shape differences in N|uu clicks correlate with phonotactic patterns. In S. Fuchs, M. Toda, & M. Zygis (Eds.), *Turbulent sounds: An interdisciplinary guide* (pp. 245–280). Berlin, Germany: Mouton de Gruyter.

Miller, A., Brugman, J., Sands, B., Exter, M., Namaseb, L., & Collins, C. (2009). Differences in airstream and posterior places of articulation in N|uu clicks. *Journal of the International Phonetic Association, 39,* 129–161. doi:10.1017/S0025100309003867.

Miller, A., Namaseb, L., & Iskarous, K. (2007). Posterior tongue body constriction locations in clicks. In J. Cole & J. Hualde (Eds.), *Laboratory phonology 9* (pp. 643–656). Berlin, Germany: Mouton de Gruyter.

Miller, A., Scott, A., Sands, B., & Shah, S. (2009). Rarefaction gestures and coarticulation in Mangetti Dune !Xung clicks. In M. Uther, R. Moore, & S. Cox (Eds.), *Proceedings of the 10th Annual Conference of the International Speech Communication Association* (pp. 2279–2282). Brighton, England: Causal Productions.

National Electrical Manufacturers Association. (2008). *Digital imaging and communications in medicine PS 3.1-2008*. Retrieved from http://medical.nema.org.

Noiray, A., Iskarous, K., Bolanos, L., & Whalen, D. H. (2008). Tongue–jaw synergy in vowel height production: Evidence from American English. In R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the Eighth International Seminar on Speech Production* (pp. 81–84). Strasbourg, France: INRIA.

Scobbie, J., Wrench, A., & van der Linden, M. (2008). Head probe stabilization in ultrasound tongue imaging using a headset to permit natural head movement. In R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the Eighth International Seminar on Speech Production* (pp. 373–376). Strasbourg, France: INRIA.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics, 19*(6–7), 455–501. doi:10.1080/02699200500113558.

Stone, M., & Davis, E. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *The Journal of the Acoustical Society of America, 98,* 3107–3112.

Stone, M., Faber, A., Raphael, L., & Shawker, T. (1992). Cross-sectional tongue shape and lingua-palatal contact patterns in [s], [k] and [l]. *Journal of Phonetics, 20,* 253–270.

Stone, M., & Lundberg, A. (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *The Journal of the Acoustical Society of America, 99,* 3728–3737.

Thomas-Vilakati, K. D. (2010). *Coproduction and coarticulation in IsiZulu clicks*. Berkeley, CA: University of California Press.

Traill, A. (1985). *Phonetic and phonological studies of !Xóõ Bushman*. Hamburg, Germany: Helmut Buske Verlag.

Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., & Hailey, D. S. (2005). The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language, and Hearing Research, 48,* 543–553.

Wilson, I. (2006). *Articulatory settings of French and English monolingual and bilingual speakers* (Unpublished doctoral dissertation). University of British Columbia.

Wrench, A., & Scobbie, J. (2006). Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. In H. Yehia, D. Demolin, & R. Laboissière (Eds.), *Proceedings of the Seventh International Seminar on Speech Production* (pp. 451–458). Ubatuba, Brazil: CEFALA. Retrieved from www.cefala.org/issp2006/cdrom/main_index.html.

Wrench, A., & Scobbie, J. (2008). High-speed cineloop ultrasound vs. video ultrasound tongue imaging: Comparison of front and back lingual gesture location and relative timing. In R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the Eighth International Seminar on Speech Production* (pp. 57–60). Strasbourg, France: INRIA.