Automatic Generation of Movie with Sound during Speech Production for Assessing Velopharyngeal Insufficiency

A. J. van der Kouwe[1], P. Sagar[2], A. L. Silver[3], S. Maturo[3], K. Nimkin[2], and C. J. Hartnick[3]
[1]Athinoula A. Martinos Center, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, United States, [2]Pediatric Radiology, Department of Radiology, Massachusetts General Hospital, Boston, MA, United States, [3]Department of Otolaryngology, Massachusetts Eye and Ear Infirmary, Boston, MA, United States

Introduction

Velopharyngeal Insufficiency (VPI) is a condition in which a structural problem with the velum (soft palate) results in inadequate closure of the velopharyngeal port during speech production. Most English phonemes require that the velopharynx (VP) be closed, therefore VPI results in inappropriate nasal resonance during speech production. The condition can be corrected with surgery if the appropriate structures are identified. Imaging of the VP during speech is currently done with radiographic multiplanar videofluoroscopy and/or nasendoscopy [1]. Since these procedures are often performed in young children, it is particularly desirable to minimize discomfort and radiation exposure. To test whether MRI provides a reasonable non-invasive alternative, we developed a protocol and procedure running on the MR scanner that produces a movie with audio of the throat during speech production. The movie is generated automatically on the scanner shortly after the scan completes.

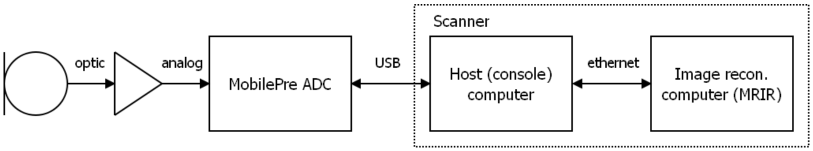

Figure 1: Optical microphone, preamplifier and analog-to-digital converter (ADC) connected to scanner. Image reconstruction computer provides timing to host.

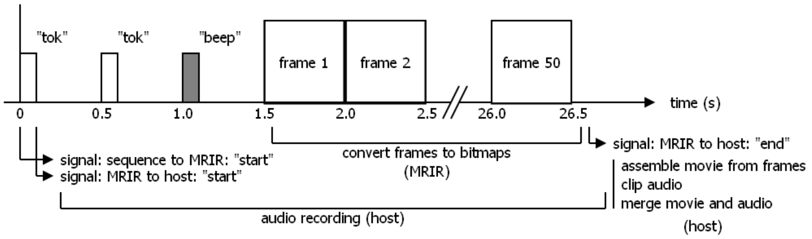
"tok"
"tok"
"beep"
frame 1
frame 2
frame 50
time (s)
0
0.5
1.0
1.5
2.0
2.5
26.0
26.5
signal: sequence to MRIR: "start"
signal: MRIR to host: "start"
convert frames to bitmaps (MRIR)
signal: MRIR to host: "end"
assemble movie from frames
clip audio
merge movie and audio
(host)
audio recording (host)

Figure 2: Timing and communications between devices and processes.

Methods

Experiments were performed on a 3 T Siemens (Erlangen, Germany) Tim Trio scanner with adult volunteers using a 12-element head and 4-element neck matrix coil. It was found that T1 contrast better elucidated the velopharyngeal structures than T2 contrast, therefore a simple single-slice 2D FLASH sequence was adopted rather than HASTE or FISP that might otherwise be favored for rapid cine-type imaging. A protocol generating exactly two frames per second was selected (FA 8°, TR 5 ms, TE 1.94 ms, BW 430 Hz/px, matrix 192x192, FoV 192 mm, slice thickness 6 mm, 2x GRAPPA with 24 reference lines, 50 repetitions, T_{acq} 25 s).

Sound was recorded using a FOM1-MR fiber optic microphone and FOM1-DRz preamplifier (Micro Optics Technologies, Middleton, WI) coupled to a MobilePre USB analog-to-digital converter (M-Audio, Irwindale, CA) connected by USB to the scanner console computer (host) (Figure 1). The recording was made using the open-source audio processing software *SoX* (Sound eXchange v. 14.3.1) [2] running on the host.

A TCP socket server was written to run on the host. The server waits for trigger messages from a TCP client on the image reconstruction computer (MRIR) (Figure 2). When a dummy image line denoting the beginning of the scan is received by the MRIR, it sends a message to the host that triggers the start of audio recording with *SoX*. During the scan, the MRIR generates the usual DICOM series but also stores each image in bitmap format on an area of the file system shared with the host. Bitmaps are generated using the open-source library EasyBMP (v. 1.06) [3]. When the scan completes, the MRIR triggers the host to assemble the bitmaps into a movie in mpeg format using the open-source software *FFmpeg* (rev. 24068) [4]. *SoX* clips the audio file so that the start time and length exactly match the frames of the movie. Finally, *FFmpeg* merges the audio and movie files to create a movie with synchronized sound in a single file in mpeg format. This process is completed within 3 s of the end of the scan. The final movie is time-stamped and stored in a temporary directory on the host for later retrieval by study staff.

The "tok-tok-tok" (introduction) sound at the beginning of the scan was modified in the sequence to a "tok-tok-beep" sound. The "beep" is a 2 kHz, 50 ms sinusoidal audio tone (generated by the gradients) that can be used to exactly synchronize the audio with the MR frames. This is not critical, as the delay between the start of the scan and the start of audio recording is predictable (approximately 120 ms). The synchronization tone can also be used to determine the exact start of the interfering scanner noise in the audio recording for the purposes of noise cancellation. If the optical microphone is carefully placed close to the subject's mouth, the voice is substantially louder than the scanner noise and no cancellation is necessary. Otherwise, cancellation can be achieved by subtracting the ensemble average across all TR-length epochs of the audio recording from each individual epoch. Note that the phase needs to be carefully aligned between epochs either by cross-correlation between the first epoch and all later epochs and/or by calculating the exact relationship between the scanner gradient clock rate and the sample rate of the M-Audio system. This was done offline using Matlab 7.8 (The MathWorks, Natick, MA).

Results and Discussion

Figure 3 shows several frames of a movie along with the audio waveform generated while an adult volunteer said the phrase "Pick up the puppy". Note the complete closure in frames 2 to 4. The next step is to validate whether such movies can provide suitable clinical information to replace videofluoroscopy/nasendoscopy. Frame rate, resolution and contrast may be adjusted if required. Noise cancellation was tested as an offline procedure but may be implemented online if necessary. Since the DICOM format supports mpeg movies, these movies could be submitted to the hospital PACS system as part of a clinical imaging session, but DICOM support has not been implemented.

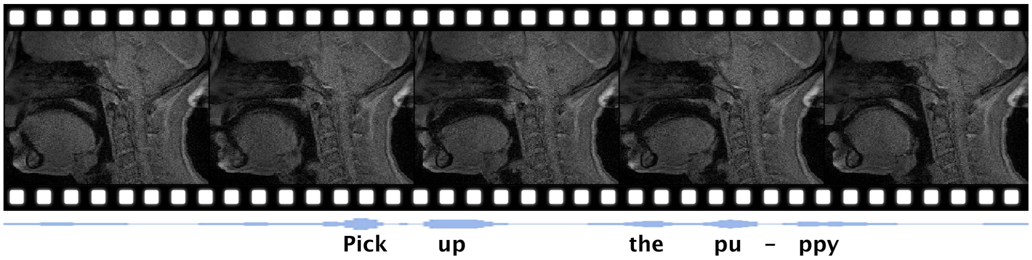

Figure 3: Five frames (2.5 s) from the middle of a movie with audio waveform showing velopharyngeal closure in sagittal view during production of the phrase "Pick up the puppy" in a healthy adult volunteer.

Acknowledgments

We thank Satrajit Ghosh and Ravi Seethamraju for recommendations on audio recording/protocols. Work partially funded by NIH P41RR014075.

References

[1] Rudnick et al., Curr Opin Otolaryngol Head Neck Surg 16(6):530-535, 2008.
[2] Bagwell et al., Sound eXchange (SoX), http://sox.sourceforge.net, 2010.
[3] Macklin, EasyBMP, http://easybmp.sourceforge.net, 2006.
[4] FFmpeg, http://www.ffmpeg.org, 2010.