

Technical Note

Recording High Quality Speech During Tagged Cine-MRI Studies Using a Fiber Optic Microphone

Moriel S. NessAiver, PhD, Maureen Stone, PhD,* Vijay Parthasarathy, MS, Yuvi Kahana, MS, and Alex Paritsky, PhD 

Purpose: To investigate the feasibility of obtaining high quality speech recordings during cine imaging of tongue movement using a fiber optic microphone.

Materials and Methods: A Complementary Spatial Modulation of Magnetization (C-SPAMM) tagged cine sequence triggered by an electrocardiogram (ECG) simulator was used to image a volunteer while speaking the syllable pairs /a/-/u/, /i/-/u/, and the words "golly" and "Tamil" in sync with the imaging sequence. A noise-canceling, optical microphone was fastened approximately 1-2 inches above the mouth of the volunteer. The microphone was attached via optical fiber to a laptop computer, where the speech was sampled at 44.1 kHz. A reference recording of gradient activity with no speech was subtracted from target recordings.

Results: Good quality speech was discernible above the background gradient sound using the fiber optic microphone without reference subtraction. The audio waveform of gradient activity was extremely stable and reproducible. Subtraction of the reference gradient recording further reduced gradient noise by roughly 21 dB, resulting in exceptionally high quality speech waveforms.

Conclusion: It is possible to obtain high quality speech recordings using an optical microphone even during exceptionally loud cine imaging sequences. This opens up the possibility of more elaborate MRI studies of speech including spectral analysis of the speech signal in all types of MRI.

Key Words: MRI; tongue; speech; audio; audibility

J. Magn. Reson. Imaging 2006;23:92-97.

© 2005 Wiley-Liss, Inc.

THE USE OF MRI for measurements of the vocal tract during speech is becoming increasingly popular as MRI frame rates become faster (1,2). In ideal circumstances the acoustic wave would be recorded simultaneously with the MRI recording, to allow vocal tract shape data to be aligned with its acoustic output. A difficulty for researchers, however, is that the speech wave cannot be collected simultaneously with the MRI data due to the highly noisy environment of the MRI. In these cases speech is collected in the scanner after the MRI sequence is completed, or at a separate recording session. Humans, however, do not repeat items identically. This is true even when repetitions are sequential. Thus, speech data collected separately cannot be perfectly aligned and compared to the MRI image sequence.

A good quality speech wave is the fundamental tool of speech research because the speech wave is what the speaker means to produce and is what the listener hears. Furthermore, in acoustics, it is known what frequencies and what size spectral changes are important. The physiological equivalent is not known. Therefore, comparison of acoustic and physiological features helps validate the importance of the observed physiological changes. As MRI is applied to speech, the speech wave will be used to corroborate physiological events in the vocal tract, or to predict the events using mechanical and physical inverse models of the vocal tract. The ability to collect a high quality speech wave during MRI recording will significantly enhance the use of MRI in modeling and understanding the control of the vocal tract during speech. Although this application is the main goal of the present work, the use of the speech signal for retrospective gating will also be mentioned. Since speech is a time-varying event, like heart motion, cine-MRI and tagged cine-MRI (tMRI) require multiple repetitions and ensemble summation to create a composite image sequence. Multiple repetitions of the speech material are not identical, however, so simple summation can cause reduced image quality and blurred tags. Therefore, the speech wave collected, simultaneously with each MRI repetition, could be used to synchronize those repetitions.

To record the speech wave, a microphone needs to be introduced into the MRI scanner. There are several problems with this idea. First, the metal in most microphones would cause poor performance in the center of

Department of Diagnostic Radiology, University of Maryland Medical School, Baltimore, Maryland, USA.

Contract grant sponsor: National Institutes of Deafness and Other Communication disorders, National Institutes of Health (NIH); Contract grant number: R01-DC01758.

*Address reprint requests to: M.S., University of Maryland Dental School, Dept of Biomedical Sciences, 666 W. Baltimore St., Baltimore, MD 21201.

E-mail: mstone@umaryland.edu

Supplementary material referred to in this article can be accessed at <http://www.interscience.wiley.com/jpages/1053-1807/suppmat/index.html>.

Received May 30, 2003; Accepted September 6, 2005.

DOI 10.1002/jmri.20463

Published online 5 December 2005 in Wiley InterScience (www.interscience.wiley.com).

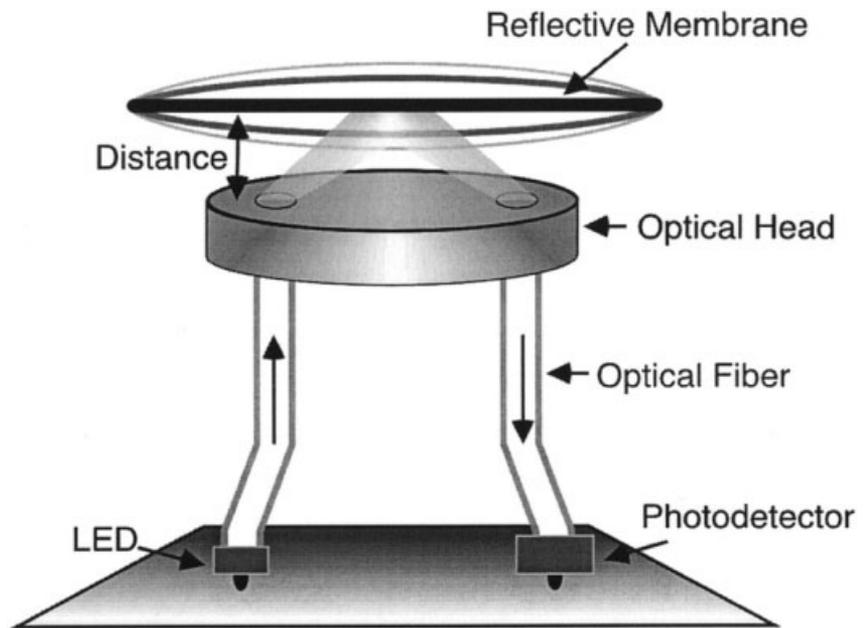


Figure 1. Principle of operation of the fiber optic microphone (used with permission of Phone-Or Corp, www.phone-or.com/temp/technology.htm).

the magnet and would interfere with the imaging sequence, distorting the final images. Second, the noise from the imaging sequence itself is quite loud and it would be difficult to isolate the volunteer's voice.

Recently, a fiber optic microphone (FOM) (Phone-or corporation, Or-Yehuda, Israel) with no metallic components has become available. This article presents early results using a version of the FOM with built-in noise-cancellation, combined with software we developed that removes background gradient noise to obtain high quality recordings during tMRI cine imaging.

MATERIALS AND METHODS

The operation of the FOM is centered on an optical sensor that measures the distance to a membrane by detecting the intensity of light that is being reflected by that membrane (3). As shown in Fig. 1, light emitted by a light-emitting diode (LED) travels over an optical fiber to the optical head, which in turn beams the light onto a sound-sensitive membrane. Sound causes the membrane to vibrate, thereby changing the intensity of the light reflected off the membrane. This light is then transmitted back to the photodetector over a second optical fiber, which transforms the intensity-modulated light into an audio signal through simple electronic processing.

The intensity of the light reflected off the membrane is determined by the angle of the optical fibers, as well as the location of the membrane in relation to the optical head. As the membrane moves away from the optical head, a greater amount of light transferred by the first optical fiber is collected by the second fiber. The FOM uses a figure-eight microphone configuration in which sound may enter from both sides of the microphone. Ambient sound entering from both sides, such as gradient noise, is significantly cancelled while sound entering from only one side, such as speech, produces a strong signal.

The FOM weighs 0.5 g, has a signal-to-noise ratio (SNR) (at 1 kHz) of >63 dB, and has a frequency response of ± 2 dB over the range of 10 to 10,000 Hz.

All scanning was performed using a Complementary Spatial Modulation of Magnetization (C-SPAMM) (4) cine sequence implemented on a Marconi 1.5-T Eclipse system using a bilateral temporal mandibular joint (TMJ) array coil. The FOM was attached to a flexible tube taped between the two TMJ coils and positioned less than 2 inches above the volunteer's mouth, with the microphone affixed slightly to the right. Informed consent was obtained as per the Institutional Review Board (IRB) protocol. The fiberoptic cable was passed out of the scan room through an radiofrequency (RF) shield penetration tube and connected to an Apple iBook laptop. Recordings were made using a 44.1 kHz sampling rate and were saved as ".wav" files.

The scan parameters were as follows: TR/TE = 4.17 msec/2.1 msec; flip angle = 10° Radiofrequency spoiled-Fourier Acquired Steady State (RF-FAST); bandwidth (BW) = 25 kHz; matrix size = 16×64 ; field of view (FOV) = 14×28 (sagittal orientation); tag pulse = 1-1 (90° - 90°); tag spacing = 7 mm; phase encode group (PEG) size = 16; RR interval (from ECG simulator) = 1007 msec; number of images = 12, temporal resolution = 66.7 msec (15 fps), Number of Signal Averages (NSA) = 1. CSPAMM tagging requires that four sets of intermediate cine images be acquired per slice. The final image is the combination of all four and has a 64×64 resolution. Although this resolution seems low, it meets the requirements for Harmonic Phase Analysis (HARP) analysis (5). The tags were applied perpendicular to the frequency encode direction; twice in the Anterior-Posterior (AP) direction and twice in the Head-Foot (HF) direction. By using a PEG size of 16 and only 16 phase-encode (PE) views, each cine can be acquired in a single cycle or utterance comparable to a single heartbeat. However, due to limitations of the scanner acquisition software, a single "throw-away" cycle (with-

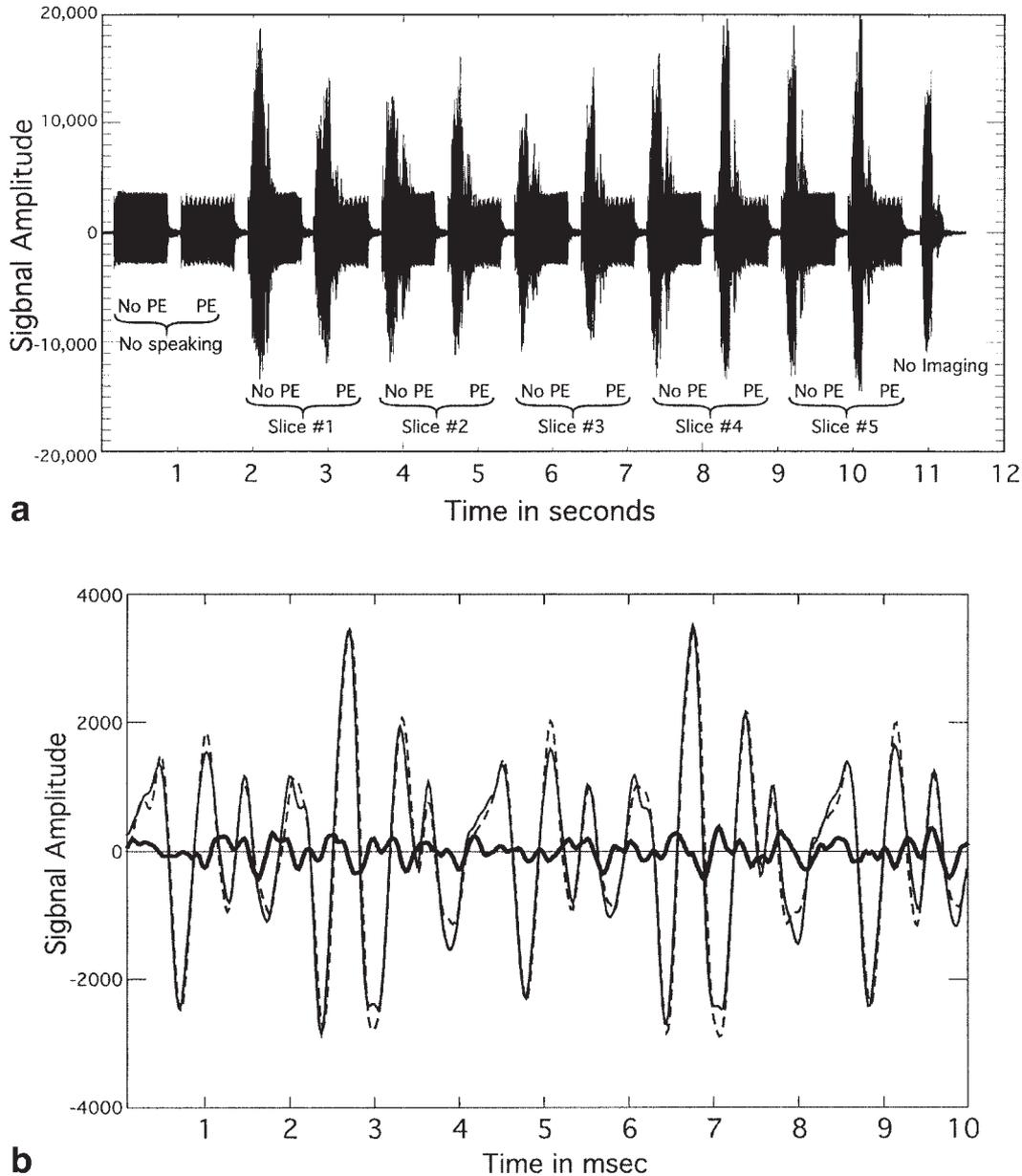


Figure 2. a: Recording of “golly.” During the first two “bursts” the volunteer was not speaking. The next 10 bursts were acquired while imaging five sagittal slices. The last burst was the word alone, recorded after the imaging was completed. Variations in the peak amplitude of each burst is due to the volunteer speaking more or less loudly. **b:** A 10 msec sample of speech. The dashed line is the recording of gradient noise without speech (e.g., rep1,PE). The thin line is recording during speech (e.g., rep2,PE). The solid black line is the subtraction.

out PE) must precede the data cycle to drive the z-magnetization to a consistent starting point (see Fig. 2, top) with each new slice. Raw k-space data was saved for use with temporal registration (described below). The final images were reconstructed using the magnitude image C-SPAMM reconstruction (MICSR) algorithm (6). Defining the two phase alternated images of CSPAMM imaging as A and B, then the MICSR reconstruction is defined as $|A|^2 - |B|^2$.

The speaker was a 25-year-old, male, native speaker of Tamil-accented English without dental work. Data were collected for four utterances: /a/-/u/, /i/-/u/, “Tamil,” and “golly.” Data from a second subject saying

/i/-/u/ are also displayed. Spectral analysis was done to track the vocal tract resonant frequencies (formants) for the speech utterances to determine the improvement in spectral quality after postprocessing and the usefulness of the resulting data for acoustic research. The three extreme vowels in English are /i/ (beat), /a/ (father), and /u/ (tune). They are spectrally extreme and physiologically quite different. Two words, “Tamil” and “golly,” contain the five consonants, /t/, /m/, / [symbol:SILDoulosIPA/241][font“SILDoulosIPA”] / [font“SILDoulosIPA”]/, /g/, and /l/. These consonants were studied because they also have a wide variety of spectral properties and physiological properties. The two

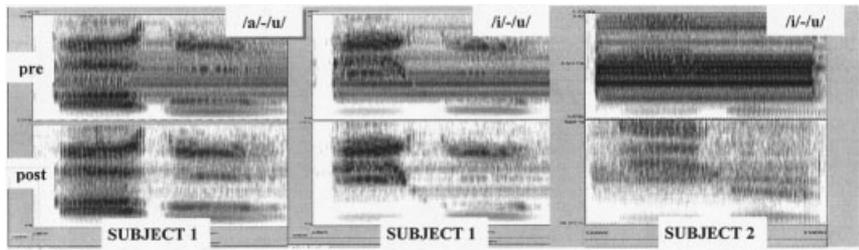


Figure 3. Pre- and postprocessed spectrograms for /a/-/u/ and /i/-/u/ for two subjects.

stops /t/ and /g/ are of short duration, but spectrally distinct. They are also physiologically quite different from each other with vocal tract closure being made with the tongue tip for /t/ and the tongue body for /g/. The /m/ is labial, /l/ has 2 points of tongue elevation and the /[symbol:SILDoulosIPA/241][font“SILDoulosIPA”] [/font“SILDoulosIPA”]/ is a non-English sound found at the end of the word “Tamil” (spelled as “l” in English). The speech utterances had to be less than one second in duration to be recorded in the MRI environment.

A total of six sagittal slices were proscribed for each utterance, requiring a total of 12 cycles for each tag and matrix orientation combination. The speaker was instructed to remain quiet for the first two MRI bursts to allow collection of a reference recording of the gradient sounds. Figure 2 (top) shows the gradient bursts and speech repetitions for “golly.” The first two bursts are the throw-away series (no PE gradient activity) and the imaging series (with PE gradients) for the first slice when the speaker was not speaking. The last burst (13th) was the speaker speaking one last time after the imaging sequence had ended.

Postprocessing of audio recordings were performed on a PowerMac G4 using code written in IDL (Research Systems Inc., Boulder, CO, USA). One method that was considered for removing the contribution from the gradients was to apply a notch filter to the spectra. However, all three utterances have one or both syllables with a significant frequency component that overlaps one or more of the gradient peaks, particularly near 960, 1200, and 1920 Hz. A notch filter to remove those gradient spectral peaks would have adversely affected the desired spectra. Subtraction in the temporal domain, described below, reduced the unwanted spectral peaks to insignificant levels without distorting the vocal spectra.

The postprocessing of the audio data for each cine series consisted of the following steps:

Gradient Noise Subtraction

1. Determine approximate starting point of each burst by looking for upswing in signal power.
2. Isolate the first two bursts as references #1 and #2 (without and with PE gradient activity, no voice).
3. Determine data shifts necessary to align the “throw-away” bursts with reference #1 and the imaging bursts with reference #2 using cross-correlations with the first 16,000 points of each burst. Subdata-point resolution was accom-

plished by zero-filling the cross-correlation vector to 256,000 points.

4. Shift each of the target data bursts, applying the appropriate phase ramp in Fourier space.
5. Subtract the appropriate reference waveforms.

Figure 2, bottom, shows an example of a reference and target waveform after alignment and their difference.

Retrospective Alignment of k-Space Data

The use of CSPAMM imaging requires multiplying together the subtraction of two pairs of images. Small misalignments caused by variations in how the volunteer repeats each utterance can result in significant degradation of the final image. We attempted to reduce this degradation by temporally aligning the k-space data with the vocal recordings as a reference using the following method:

1. Using the vocal recordings after gradient noise subtraction, manually identify the start and end time of each vowel sound.
2. Manually identify the single repetition, out of four for each slice, with start and end times closest to the mean values and define this as the reference series.
3. Use linear interpolation over time of the k-space data for the other three series.
4. Reconstruct the temporally aligned images and combine using the $|A|^2 - |B|^2$ MICS method.

RESULTS

Improving Speech Quality

The original and postprocessed spectrograms appear in Fig. 3 for a single repetition of /a/-/u/ and /i/-/u/; the latter for two subjects. The original waveforms (top) come from the subtraction microphone, which used passive noise cancellation to produce a fairly good spectral representation of the speech wave; the speech was in fact audible. However, the signal was too noisy for successful waveform analysis. Postprocessing (subtraction) was able to reduce the gradient component by 21 dB when averaged over the complete burst. The postprocessed waveform was substantially improved (bottom) and spectral analysis could be performed.

The important acoustic features of vowels are their resonant frequencies or formants, which appear as dark horizontal bands. The first formant is low for /i/ and /u/ and high for /a/. The second formant is high,

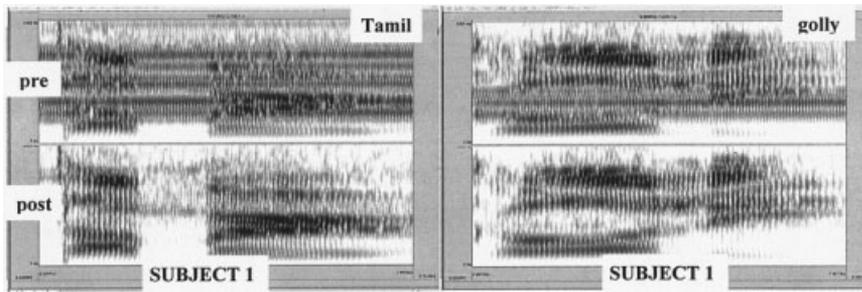


Figure 4. Pre- and postprocessed spectrograms for “Tamil” (left) and “golly” (right).

middle, and low for /i/, /a/, and /u/, respectively. These features are very well known and are the primary distinctions between vowels (7). The postprocessed spectrograms are noticeably clearer than the originals. The /a/-/u/ dataset is the better of the two with gradient noise almost gone and the formants very distinct. The /i/-/u/ dataset has some visible gradient noise, but the formants are still visible.

Spectrograms for the two words, Tamil and golly, are seen in Fig. 4. The upper spectrograms are original and the lower are postprocessed. The postprocessed spectra shown in Figs. 3 and 4 display significantly reduced residual contribution from the gradients, particularly between 900 and 2200 Hz, where the gradient noise is most prominent. Both temporal and spectral features of the postprocessed words are enhanced. The vowel formants are measurable and the consonants are also improved. The short bursts of noise following /t/ and /g/ are more apparent, and the formants for /l/ are more accurately displayed in the postprocessed signal. The improvement is particularly noticeable on quiet sounds like /i/ and /m/. These data are of sufficient quality for acoustic analysis.

Time-Alignment

A second use for the simultaneously collected speech wave is time-alignment of the multiple MRI repetitions. An important assumption in gated cine imaging is that the motion being imaged is reproducible. Figure 5 shows the signal power, smoothed over a sliding 9-msec window, of 10 repetitions of “golly.” One can clearly see that there are substantial variations in the start and stop times of various features. The first two vertical dashed lines, labeled A and B, show the range of times when the power associated with the first syllable reaches a plateau. The earliest and latest values are circled. These start points differ by roughly 95 msec. The second two lines show the range in the start of the rise to the second syllable. The time between the points labeled C and D is roughly 60 msec. Figure 5 clearly displays a lack of reproducibility at the start, the end, and during the speech, when repeating verbal utterances. As a result, aligning these data sets using the start points of the speech waves had limited success; most images showed little or no improvement while other images were reduced in quality. Acoustic analysis of the four waveforms themselves are useful, however, to compare the multiple repetitions for variability. This can be used to determine the trustworthiness of the MRI tags. Moreover, when one is laid on the composite

movie, the sound facilitates interpretation of the motion. For example, in /a-u/ motion, the tongue moves up for /u/ and then forward for the inhalation. Without the speech wave, the second gesture might appear to be part of the /u/ sound.

To improve the tMRI alignment results a better method may be to improve the subject’s precision. Increased precision through training would reduce variability in onset and throughout the utterance. In that case, prospective triggering of the speech might be adequate. In the present data set subjects spoke in rhythm with the MRI noise, but were only given brief exposure to the noise prior to data collection. One subsequent subject was trained to speak on the beat in three 10-minute sessions. Two others were trained using evenly-timed metronome-like beats (white noise bursts) to pace their syllable onsets (8). Once the subjects were trained on the MRI noise or beats, they were able to hold the rhythm well enough to produce good reproducibility. Good reproducibility means that tissue points can be tracked from the composite tMRI images. By combining methods to improve MRI alignment and

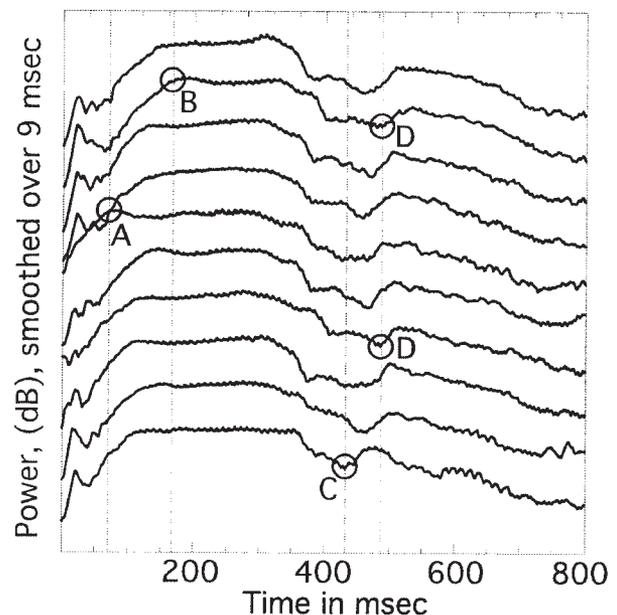


Figure 5. Signal power of 10 repetitions of “golly.” The circles labeled A and B denote the earliest and latest times when the power reaches a plateau. The circles labeled C and D denote the earliest and latest times when the power starts to sweep up for the second syllable.

speech quality, the speech acoustic and tongue physiological features can be better compared.

DISCUSSION

We have found that an optical microphone with MRI specific noise cancellation can be used for recording high quality speech during MRI scanning. Due to the high reproducibility of the MR gradient pattern and the low noise of the optical microphone, it is possible to reduce the acoustic noise of the gradients to inconsequential levels. The resulting recordings are suitable for spectral analysis, allowing simultaneous MRI and speech data collection for comparative analysis.

We have also developed some initial postprocessing methods to temporally synchronize the images being reconstructed using the acoustic waveform. However, the results were disappointing due to two sources of temporal variability: speech onset time relative to the external trigger and inconsistent duration of the sounds. The effect of these two temporal variabilities is to cause similar speech events to occur at different times, in multiple repetitions. The result is spatial misalignment of the anatomical structures and improperly matched tag intensities, both of which adversely affect the MRI data summation. Temporal features of the speech wave, such as voice onset, could be used as a trigger for the tagging sequence, just as features of the ECG wave are a trigger for the heart. In addition, training the subject to speak to the noise or to beats improves subject repeatability and precision during data collection. Prospective gating with a good quality microphone and appropriate interface to the scanner would significantly reduce the first source of variability. Subject training and rhythmic cueing is also an idea for subjects to improve onset and internal timing errors.

The ability to collect high quality audio data during an MRI scan is remarkable and will greatly enhance the applicability of MRI to speech research. In functional

MRI (fMRI) and tMRI, language and speech are heavily studied. The fields of linguistics and speech pathology are ready to compare acoustic features of speech to MRI measures of vocal tract geometry, muscle compressions, and structural kinematics. Simultaneous audio and MRI collection will be of invaluable assistance to speech studies. With well-timed subject repetitions, tissue-point measurements can be compared to the speech wave with greater precision and confidence to answer questions about vocal tract behavior and its resulting speech signal.

Supplemental Material

Close-up images of the FOM, the scan setup as well as sound and movie clips are available at <http://www.interscience.wiley.com/jpages/1053-1807/Suppmat/index.html>.

REFERENCES

1. Stone M, Davis EP, Douglas AS, et al. Modeling the motion of the internal tongue from tagged cine-MRI images. *J Acoust Soc Am* 2001;109:2974–2982.
2. Stone M, Davis EP, Douglas AS, et al. Modeling tongue surface contours from Cine-MRI images. *J Speech Lang Hear Res* 2001;44:1026–1040.
3. Jost BM, Stec JP. Refractive fiber optic microphones with ambient acoustic noise-canceling capabilities. *J Acoust Soc Am* 1995;98:1612–1617.
4. NessAiver M, Stone M, Parthasarathy V, Prince JL. Using tagged cine-MRI to extract muscle activity within the tongue. Proceedings of the Radiological Society of North America, Chicago, IL.; 2002 (Abstract 552).
5. Osman NF., McVeigh ER, and Prince JL. Imaging heart motion using harmonic phase MRI. *IEEE Trans Med Imaging* 2000;19:186–202.
6. NessAiver M, Prince JL. Magnitude image CSPAMM reconstruction (MCSR). *Magn Reson Med* 2003;50:331–342.
7. Hillenbrand J, Getty L, Clark M, Wheeler K. Acoustic characteristics of American English vowels. *J Acoust Soc Am* 1995;97:3099–3111.
8. Masaki S, Tiede M, Honda K, Shimada Y, Fujimoto I, et al. MRI-based speech production study using a synchronized sampling method. *J Acoust Soc Jpn* 1999;20:375–379.