Received: 27 November 2018 | Revised: 11 February 2019 | Accepted: 28 February 2019

DOI: 10.1002/mrm.27745

FULL PAPER

Magnetic Resonance in Medicine

Intermittently tagged real-time MRI reveals internal tongue motion during speech production

Weiyi Chen[1] | **Dani Byrd**[2] | **Shrikanth Narayanan**[1,2] | **Krishna S. Nayak**[1]

[1]Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, California

[2]Department of Linguistics, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, California

Correspondence
Weiyi Chen, University of Southern California, Los Angeles, 3740 McClintock Avenue, EEB 400, CA 90089-2564.
Email: weiyic@usc.edu

Funding information
This work was supported by NIH Grant R01DC007124 and NSF Grant 1514544.

Purpose: To demonstrate a tagging method compatible with RT-MRI for the study of speech production.

Methods: Tagging is applied as a brief interruption to a continuous real-time spiral acquisition. Tagging can be initiated manually by the operator, cued to the speech stimulus, or be automatically applied with a fixed frequency. We use a standard 2D 1-3-3-1 binomial SPAtial Modulation of Magnetization (SPAMM) sequence with 1 cm spacing in both in-plane directions. Tag persistence in tongue muscle is simulated and validated in vivo. The ability to capture internal tongue deformations is tested during speech production of American English diphthongs in native speakers.

Results: We achieved an imaging window of 650-800 ms at 1.5T, with imaging signal to noise ratio $\geq$ 17 and tag contrast to noise ratio $\geq$ 5 in human tongue, providing 36 frames/s temporal resolution and 2 mm in-plane spatial resolution with real-time interactive acquisition and view-sharing reconstruction. The proposed method was able to capture tongue motion patterns and their relative timing with adequate spatiotemporal resolution during the production of American English diphthongs and consonants.

Conclusion: Intermittent tagging during real-time MRI of speech production is able to reveal the internal deformations of the tongue. This capability will allow new investigations of valuable spatiotemporal information on the biomechanics of the lingual subsystems during speech without reliance on binning speech utterance repetition.

KEYWORDS
real-time MRI, speech production, spiral, tagging, tongue

1 | INTRODUCTION

The vocal tract is a complex system that consists of both movable and immovable structures. Speech production involves complex spatiotemporal coordination of multiple vocal organs in the upper (oral) and lower (pharyngeal) airways. Visualization of the movements of the organs can provide important information about the spatiotemporal properties of speech actions, or "gestures." Several modalities have been used to visualize speech, including X-ray,[1] computer tomography (CT),[2] electromagnetic articulography (EMA),[3] ultrasound,[4] and MRI.[5-10] MRI can uniquely provide both static images with excellent soft tissue contrast and dynamic images with high frame rate, without the use of ionizing radiation,

making it a promising tool. Real-time MRI (RT-MRI) now plays an important role in interpreting dynamics of vocal tract shaping during speech production, swallowing, and other human functions such as vocal performance.[5,11-13]

In speech production, the upper respiratory tract forms a series of connected resonance cavities that can be modified in size and shape using coordinated movements of the velum, jaw, pharyngeal tongue root, tongue body, tongue tip, and lips.[6] Among these articulators, the human tongue is the most powerful enabler of the remarkably complex shaping occurring in speech. The tongue is a muscular hydrostat comprised of numerous intrinsic and extrinsic muscles.[14] The internal deformation of tongue muscles cannot be easily interpreted by the contours of the tongue surface, and the relationship between muscle activity and tongue shaping is the subject of scientific investigation as an important component in understanding how healthy speech is controlled and how it is disrupted in disease.[15,16] However, scientists remained reliant on inverse modeling of surface contours heavily contingent on modeling assumptions.[16-19]

RT-MRI techniques have been extensively used in the last decade to study speech production, specifically the dynamics of vocal tract shaping with a focus on tracking the air-tissue interface at articulator and vocal tract surfaces.[13] Recent RT-MRI advances include improvements in spatiotemporal resolution,[12,20,21] increasing spatial coverage,[22-25] reducing reconstruction latency,[26-28] mitigating off-resonance artifacts,[21,29] and combinations of the above. However, these techniques all lack the ability to measure internal muscle activity and to image and quantify the deformations of local regions within the human tongue, arguably the most important articulator, during natural speech.

Tagged MRI has been used to capture internal tongue deformation since early 1990s. Static MRI was used as snap shot at designated points in a tongue movement to visualize the deformation.[30-32] Later, tagged CINE-MRI was used to analyze the motion of the internal tongue during speech.[33-35] Recently, it has been used to provide images for measurement of 4D tongue motion and to generate an atlas of the human tongue during articulation,[36,37] Such CINE methods rely on repetition with perfect synchronization, thus allowing tagged MRI to be used to analyze cardiac motion,[38,39] as heart beats in sinus rhythm are highly repeatable, independent of rate of contraction,[40] and can be easily synchronized with electrocardiography. However, the heart differs from the tongue in important ways, most notably in that speech production possesses great token and type variability due to its voluntary, information encoding, and highly context-sensitive nature. Real-time tagged MRI with Cartesian sampling was explored for cardiac applications.[41,42] However, these methods only provide 1D deformation in real-time, as they implement fast imaging by either compromising resolution on phase encoding direction[41] or by only acquiring a small island of harmonic peak in k-space.[42] They need at least 2 heartbeats to resolve motion on both directions. Real-time Strain ENCoding (SENC) techniques,[43,44] although able to provide quantitative strain for cardiac applications, nevertheless measure on a plane that is perpendicular to the imaging plane and are not compatible with speech applications.

In this work, we demonstrate a tagging method compatible with RT-MRI for the study of natural human speech production. We apply tagging as a brief interruption of continuous RT-MRI data acquisition. We explore the selection of imaging parameters for such speech studies to optimize image quality and tag persistence. We evaluate this method using simulations and in vivo studies of American English diphthong and consonant production. We show that the proposed method can capture tongue motion patterns and their relative timing through internal tongue deformation, and, therefore, provide a potential tool for studying muscle function in speech production and similar scientific and clinical applications.

2 | METHODS

2.1 | Tagged RT-MRI implementation

Experiments were performed on a Signa Excite HD 1.5T scanner with a custom 8-channel upper-airway coil.[12] The pulse sequence was implemented within a real-time imaging platform (RTHawk Research v2.3.4 , HeartVista, Inc., Los Altos, CA).[45]

Figure 1 illustrates the acquisition timing and the pulse sequence diagrams for tagging and imaging. As shown in Figure 1A, tagging is applied as a brief interruption to continuous real-time spiral acquisition. A button was added to the RTHawk graphical user interface to allow operator control of intermittent tagging. Manually pushing the button initiates the tagging module to be applied right after the current imaging repetition time (TR) and before the next imaging TR. Real-time spiral data acquisition experiences only a brief interruption of less than 6 ms (comparable to 1 imaging TR). The persistence of the tag grid depends on longitudinal relaxation (T_1) of the tongue muscle and the effect of imaging RF excitations.[46]

Figure 1B illustrates the tagging sequence, which is a standard 2D 1-3-3-1 binomial SPAtial Modulation of Magnetization (SPAMM) sequence,[47,48] with a 1-cm spacing in both in-plane directions. Two SPAMM pulses were sequentially applied along the x and y axes, followed by crushers to eliminate any remaining transverse magnetization.[49] The second composite SPAMM sequence had its phase shifted by 90° relative to the first one[48] and used a different crusher area to avoid stimulated echoes. The overall duration was 5.66 ms.

Figure 1C illustrates the imaging sequence, which is a standard spiral spoiled gradient echo (GRE), and is designed to make the maximum use of the gradients (40 mT/m amplitude

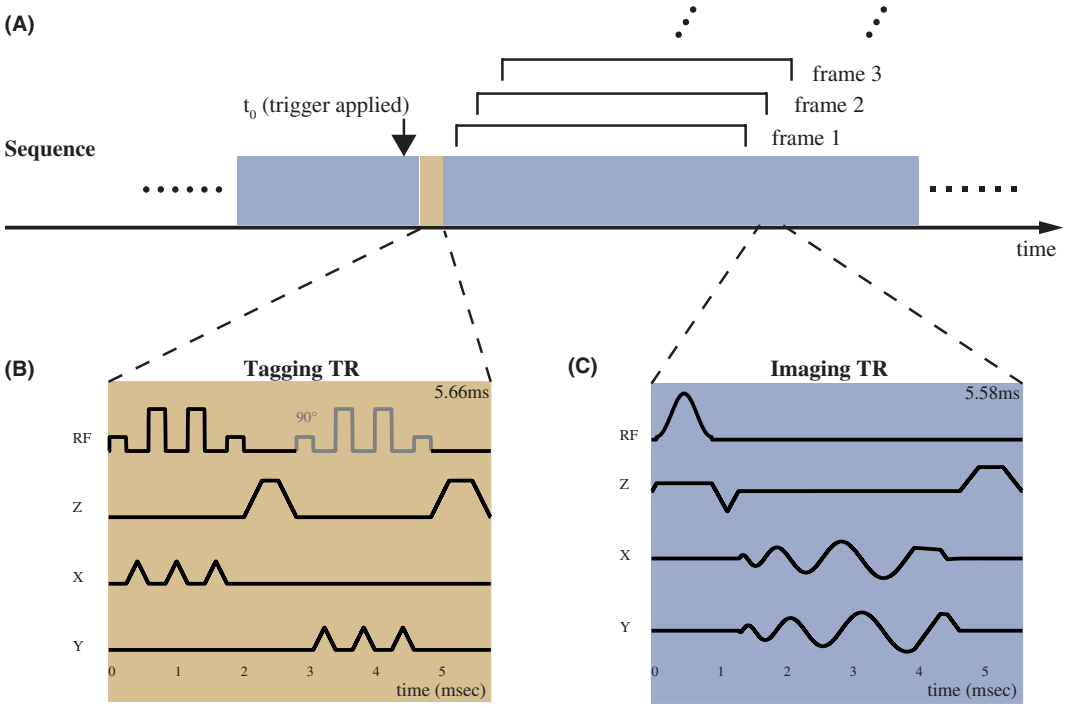

FIGURE 1 Speech RT-MRI with Intermittent Tagging. A, Overall acquisition timing. Continuous imaging is performed using interleaved spiral GRE imaging (C, blue block) with view-sharing reconstruction. 13-interleaves were used to fully sample k-space at each time frame using a bit-reversed interleaf order. Tag placement is performed using two 1-3-3-1 SPAMM pulses along x and y (B, yellow block). Note the second composite SPAMM pulse is shifted with a 90° relative phase and is with slightly larger crusher to avoid stimulated echo

and 150 mT/m/ms slew rate). The imaging parameters were: field of view 20 cm, slice thickness 7 mm, readout duration 2.49 ms, echo time/TR 0.71 ms/5.58 ms, 13-interleaves bit reversed view-ordering.

Coil-by-coil gridding reconstruction with view-sharing was performed on-the-fly during data acquisition. The Walsh method was used to estimate the sensitivity map for coil combining.[50] We used step size of 5 TRs for the sliding window, resulting in a nominal temporal resolution of 36 frames/s. The approximate end-to-end reconstruction latency was 27 ms. This setup enables the operator to observe the tagging lines deformation in real-time to monitor the subject completion of the designed articulation task, and to determine if the timing of triggering conformed to design. Concomitant fields correction[51] and image unwarping that accounts for gradient nonuniformity[52] were applied with gridding reconstruction.

2.2 | Selection of acquisition parameters

Tag persistence was quantitatively evaluated by analyzing the temporal evolution of contrast-to-noise ratio (CNR) CNR_{tag} as a function of time.[46] We assume that the steady state signal M_{ss} is reached before tagging. Immediately after tagging sequence, at time t_0, the longitudinal magnetization can be expressed as:

$$M_z\left(t_0\right) = M_{ss} Q\left(x, y\right), \tag{1}$$

where $Q(x, y)$ represents the modulation function due to the SPAMM sequence. The longitudinal magnetization immediately before the first RF at time t_1, considering T_1 recovery, is:

$$M_z\left(t_1\right) = M_{ss} Q\left(x,y\right) e^{-\frac{t_1}{T_1}} + M_0\left(1 - e^{-\frac{t_1}{T_1}}\right) = M_T + M_R. \tag{2}$$

The first term, denoted M_T, contains the fading tag information; the second term, denoted M_R, contains the recovery toward equilibrium magnetization M_0. We calculate the temporal evolution of tag contrast by considering n consecutive spiral GRE TRs, each with flip angle (FA) α. Each of such imaging RF will scale the magnetization with a factor of $\cos\alpha$. The M_T component immediately before the n^{th} RF excitation (at time t_n) can be expressed as:

$$M_T^{(n)} = M_{ss} Q\left(x,\, y\right) e^{-\frac{t_n}{T_1}} \prod_{j=1}^{n-1} \cos\alpha = M_{ss} Q\left(x,\, y\right) e^{-\frac{t_n}{T_1}} \left(\cos\alpha\right)^{n-1}, \tag{3}$$

and the M_R component can be recursively expressed as:

$$M_R^{(n)} = \left[M_R^{(n-1)} \cos\alpha - M_0\right] e^{-\frac{t_n - t_{n-1}}{T_1}} + M_0. \tag{4}$$

Applications of RFs during imaging contributes to reducing the tag information, as it consumes part of the longitudinal

magnetization. An optimal FA can be determined as described below.

The contrast in image is the part in $M_T^{(n)}$ (the peak-to-valley difference in magnetization) that tipped to the transverse plane by the imaging RF. The CNR_{tag} after the n^{th} RF excitation is defined as the ratio between the contrast in image and standard deviation of the image noise:

$$CNR_{tag} = \frac{M_T^{(n)} \sin \alpha}{\sigma}. \quad (5)$$

The tag persistence can be defined as the time span between the grids being placed and CNR_{tag} dropping below a certain threshold. Markl et al[53] suggested a CNR threshold of 6 for cardiac applications. Simulated σ is calculated as the simulated steady state signal divided by 15, as suggested by previous experiments.[12] A threshold time can be calculated as the time span between the tag being placed and when the CNR decrease below the threshold value.

Two healthy volunteers (27/M, 27/F) were scanned to verify tag persistence in the tongue and to identify the optimal imaging FA. Fifteen integer FAs ranging from 1° to 15° was used in the experiment. A wide tag spacing of 5 cm was used to mitigate partial volume effects in the post processing steps. The noise covariance matrix of the coils was measured with a separate scan with excitation RFs turned off. The measured noise covariance matrix was used to prewhiten the multicoil data and to calculate the standard deviation of the noise to normalize the result. For each FA, a separate scan was used to measure the steady state signal to properly scale between simulation and measurements. The subjects were instructed to keep their mouth in a closed neutral position and remain still during the scan to minimize off-resonance and motion artifacts. The peak and valley values were calculated by taking average over the manually selected regions of interest (ROIs). The peak ROI was drawn in two 4 × 6-pixel squares in the bright regions in the tongue; the valley ROI was selected over one 3 × 16-pixel stripe at the center of the dark tag lines.

2.3 | Triggering mechanism

In this study, we tested 3 different tag-triggering schemes to assess the best usage of the imaging window after the intermittent tagging sequence. Each involved a specific approach to coordinating the tag triggering by the operator with the speech production by a subject (who read linguistic stimuli projected on a screen).

2.3.1 | Manual triggering

In the manual triggering approach, the subjects were instructed to speak the linguistic stimuli (described below) 10 times with a full pause between each production (to ensure the intermediate return to a neutral vocal tract posture). The operator used the first 2-3 utterances to ascertain the token-to-token rhythm or pacing of the subject for application of the tagging module for the rest of the trials. The operator controlled both the button for the tagging module and the projector showing the stimuli 1 utterance at a time.

2.3.2 | Cued triggering

In the cued triggering approach, the MRI operator and the subject were instructed, respectively, to push the triggering button and to read the stimulus immediately upon its visual appearance on the projector screen.

2.3.3 | Periodic triggering

In the periodic triggering approach, an automatic triggering was implemented in the sequence system. The tagging module was applied every 182 TRs with a period of approximately 1015 ms, which is equivalent to 14 fully sampled frames when no view sharing is applied. The subjects were instructed to say the stimuli for 15 s with a pause between each individual speech item.

2.4 | Speech experiments

Four healthy volunteers (2M2F; 27-31 years), all native American English speakers, were scanned. The experiment protocol was approved by our Institutional Review Board, and informed consent was obtained from all volunteers. Audio recording and stimuli presentation were adapted from similar protocols successfully used in previous studies (e.g., Gerard et al[17]).

Table 1 shows the American English diphthong vowel stimuli used in this experiment.[54-56] Diphthongs are vowels in which the lingual postures, and their concomitant formant frequencies, require relatively large movements from 1 vowel target to another in the same syllable.[54] The diphthongs /aɪ/, /ɔɪ/ and /aʊ/ were chosen for this study because they involve substantial movement of tongue when gliding from initial to final vowel quality, and the duration of this movements (~180 ms to 300 ms)[56] can be thoroughly covered in the current imaging window.

The stimuli were placed both in carrier phrases and presented in isolation, so as to provide variation for investigating the proposed tagging sequence. Diphthong stimuli in isolation were the words/pseudo-words: “I”, “oy”, and “ow.” The stimuli in carrier phrases placed the diphthongs after labial consonants in the words: “buy,” “boy,” and “bow.” (for “ow”, subjects were instructed so as to ensure that their pronunciation rhymed with “now.”) A [b], a consonant made with lip rather than lingual closure, was used preceding and following the diphthong to minimize any coarticulation

TABLE 1 American English diphthong stimuli

Stimuli	Carrier phrase	Target diphthong	Starting posture description	Ending posture description	Approximate diphthong articulation duration
"I"	None	/aɪ/	Low back	High front	180 ms
"oy"		/ɔɪ/	Mid back	High front	300 ms
"ow"		/aʊ/	Low back	Mid/high back	180 ms
"A buy puppy"	a [·] puppy	/aɪ/	Mid central	High front	180 ms
"A boy puppy"		/ɔɪ/	Mid central	High front	300 ms
"A bow puppy"		/aʊ/	Mid central	Mid/high back	180 ms

Note. Starting/ending posture description refers to the approximate tongue position when tagging began (if in a carrier, during the "a") and ended.

TABLE 2 American English consonant stimuli

Stimuli	Target consonant	Articulation place and manner	Constriction area
"ara"	/ɹ/	Retroflex approximant	Lips, post-alveolar ridge, pharynx
"asha"	/ʃ/	Postalveolar fricative	Post-alveolar ridge
"acha"	/tʃ/	Postalveolar affricate	Post-alveolar ridge

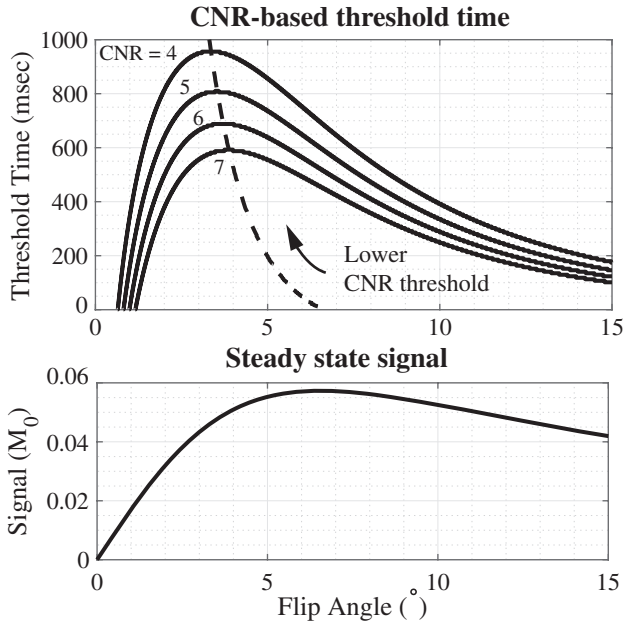

FIGURE 2 Simulation of tag persistence and steady-state signal as a function of imaging FA. Top: Threshold time is defined as the time span between the tag being placed and the tag CNR falling below the threshold value (shown for CNR cutoffs of 4, 5, 6, and 7). The dashed line marks the FAs that will deliver the longest threshold time for each CNR threshold. The longest persistence can be reached at a FA of 3-6.5°. Performance suffers quickly if the FA is too low, but less so if the FA is too high. Bottom: Steady state signal as a function of FA for the imaging TR = 5.58 ms and tongue T_1 = 850 ms at 1.5T. The Ernst angle in this case is 6.2°. The actual imaging FA was selected based on both tag persistence and steady-state tongue SNR

with other nearby lingual sounds. The tagging module was triggered in close temporal proximity with the onset of the diphthong. Different motion patterns and their relative timing during the transition between the component postures of the diphthongs were then imaged. The carrier phrase stimuli ("a buy/boy/bow puppy") are presented in this work.

Table 2 shows consonant stimuli used in the experiment. Stimuli occurred in the pseudo-words: "ara", "asha", and "acha," so as to place /ɹ/, /ʃ/ and /tʃ/ between 2 /ə/s having a relative neutral vocal tract posture. All of these target consonants are produced using a tongue constriction in the anterior oral hard palate area immediately posterior to the alveolar ridge. /ɹ/ (for this speaker) places the tongue tip in a retroflex posture (although other American English speakers are known to make /ɹ/ with a bunched, tip-down posture), and /ʃ/ and /tʃ/ raise the tongue tip and blade up toward the post-alveolar area; during /ʃ/ retains a small airway opening allowing turbulent airflow while /tʃ/ has a brief stop of airflow as the tongue fully contacts the palate followed by turbulent airflow as it draws away.

3 | RESULTS

3.1 | Acquisition parameters

Figure 2 shows CNR-based threshold time and signal intensity as functions of imaging FA. The longitudinal relaxation of the tongue muscle T_1 = 850 ms at 1.5T was measured by inversion recovery fast spin echo with multiple inversion times. This value agreed with previous literature.[6,57] Dashed lines in Figure 2A indicate CNR optimal FA that delivers the longest threshold time. The CNR optimal FA increases from 3° to 6.5° with higher threshold values providing shorter tag persistence. The Ernst angle for imaging tongue is $\alpha_E = 6.2°$

as showed in Figure 2B. The simulation shows a trade-off between CNR-based tag persistence and image signal to noise ratio (SNR) when choosing optimal excitation FA.

Figure 3 shows an in vivo experiment on tag persistence in human tongue. Measured signal of tag lines (center) and peak-to-valley contrast were plotted as functions of time with corresponding simulated curve. The curves were normalized by the standard deviation of noise measured in a separate scan. Only a subset of FAs (3°, 5°, 7° in 1-15°) are shown in the figure for illustrative purpose. The measured signal conformed to the simulation for all imaging FAs. The tag lines of FA = 3°, 5°, 7° recovered to the steady signal with SNRs of 13, 17, and 20 with decreasing times, respectively. Note that FA = 7° had the highest imaging SNR; however, the faster decay resulted in a CNR drop to 5 in only 600 ms. In contrast, the CNR by FA = 3° and 5° reached the threshold level in more than 650 ms, with the latter having 30% higher image SNR in the tongue compared with the former. In our experience, imaging using a very small FA ($\alpha < 5°$) was sensitive to B_1 inhomogeneity in the tongue, as the signal dropped dramatically when unintentionally decreasing the FA. As an overall result of the above considerations, we used FA of 5° with an imaging window of around 650-800 ms, with the ending CNR of 5-6. Figure 4 shows example images of tag fading.

3.2 | Triggering mechanism

Manual triggers were likely to miss the beginning of the diphthong even with the operator and the subject synchronized into the same rhythm with practice. The reflex delay of the human operator and the normal speech pacing and production variability of the subjects aggravated the miss rate. Furthermore, the operator's timing accuracy largely depended on the speech sound that came from the scanner, which was compromised by acoustic scanner noise.

Both cued and periodic triggering performed well. During cued triggering, the normal reflex delay of the subject between seeing the stimuli on the project and starting articulation was largely matched by the reflex delay of the MRI operator in executing the tagging button press, ensuring that the tag was reliably placed appropriately before the target tongue movement. Of interest, for the elicitation protocol of periodic tagging, the tagging module interrupted the acoustic sound of the readout gradient heard by the subjects and acted in effect as an auditory metronome for the subject, causing them to entrain to the tag triggering rhythm and thereby consequently aligning their productions with the tagging timing after the first 1-2 triggers. And, because there was no voluntary effort required by the operator on the triggering side, operator alignment errors were not an issue.

3.3 | Visualization of tongue deformation

Figure 5 uses American English Vowel Charts to provide a rough schema for understanding the tongue positioning. The blue curve in the chart marks the starting and ending points for the 3 diphthong vowels being studied. These vowels in

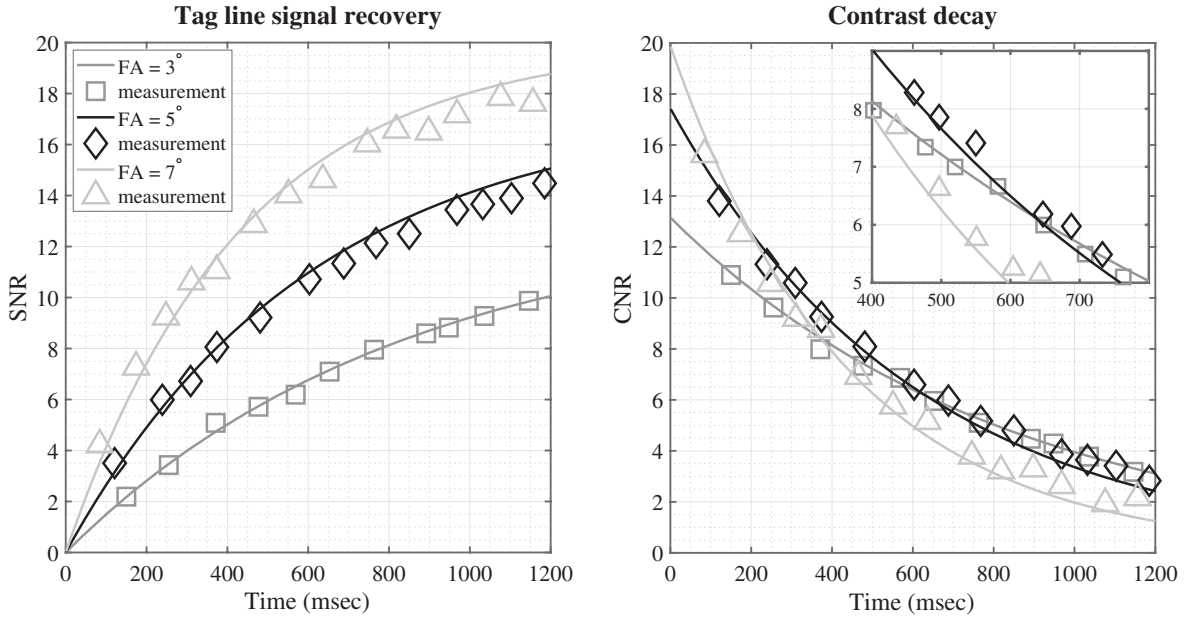

FIGURE 3 Tag persistence in human tongue at 1.5T. Left: simulation (line) and measurement (symbol) of the tag line signal for the first 1.2 s after the tag module was applied. Right: contrast decay after tag module being applied. Tongue T_1 = 850 ms was measured using an inversion recovery fast spin echo sequence with multiple inversion times. The signal and contrast were normalized by the standard deviation of noise, measured by a separate scan with RF excitation turned off

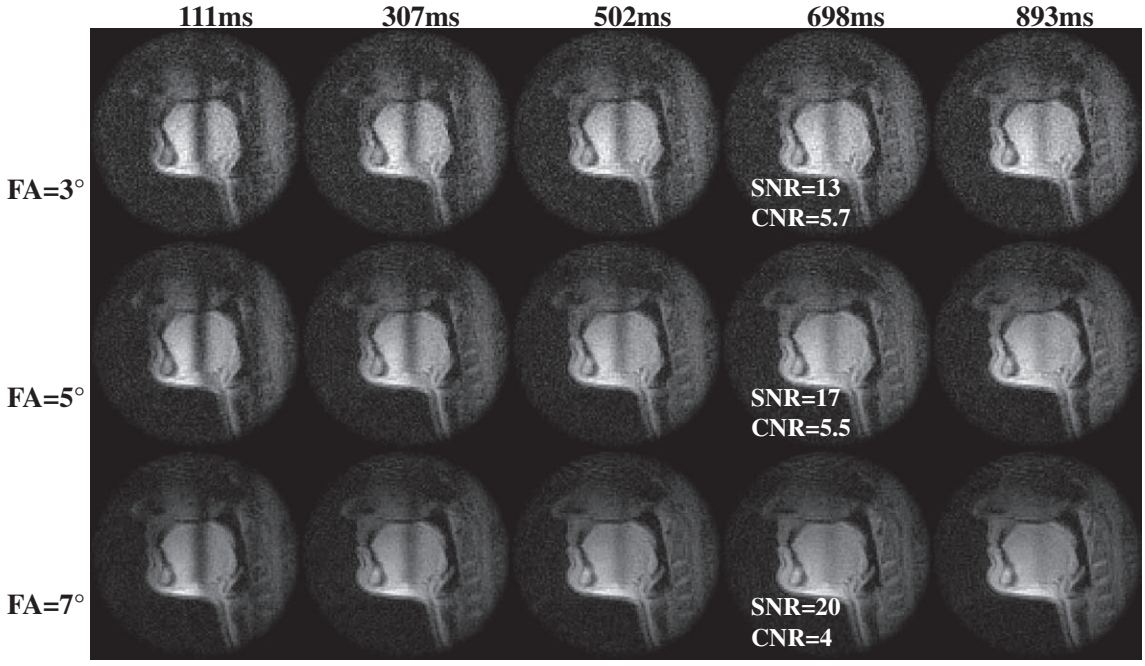

FIGURE 4 Example images of tag fading with imaging FA of 3°, 5°, and 7°. Wide tag spacing of 5 cm was used to mitigate partial volume effects. At around 700 ms (4th column), FA = 3°, 5° have similar CNR, while the latter has 30% higher SNR. As an overall consideration, we used FA of 5° with an imaging window of around 650-800 ms, with the ending CNR of 5-6

English are known to produce sweeping lingual motions that move the tongue upward from a depressed and/or retracted posture to a raised and fronted or raised and retracted posture as follows: in /ai/ from a low-back posture to a high-front posture, in /ɔɪ/ from a mid-back posture (with lip rounding) to a mid-high front posture, in /aʊ/ from a low posture to a high-back (lip rounded) posture. In these vowels, as in vowels generally, the tongue is generally more or less arched; it is not grooved or concave. Figure 6 reveals internal tongue movement during 3 American English diphthong articulation examples. The videos can be found in Supporting Information Video S1, which is avaible online. For orientation, note that /aɪ/ and /aʊ/ start with similar low and retracted tongue postures (note the pharyngeal narrowing); /aɪ/ and /ɔɪ/ end with similar postures of the tongue bunched up high in the palatal vault; and the starting posture of /ɔɪ/ is similar to the ending posture of /aʊ/ with the tongue high and retracted toward the velum (soft palate).

Figure 6 contains representative frames, illustrating multiple deformation patterns and capturing their relative timing. A shear between different parts of the tongue can be identified as square grids changing into parallelograms. Compression can be identified as square grids changing into bi-concave rectangles. Stretching and curving of the tongue can be identified by bended grid lines. Each of these types of deformations occurred during the course of diphthong articulation. Color arrows mark the start of 1 specific type of deformation in the representative frames.

In the case of /aɪ/ (top row), parallelograms emerge at 315 ms (cyan), indicating shear between the tongue body and tongue root. Also at this time, bi-concave rectangles can be observed at the top of the tongue body (magenta). These compressions move the tongue forward and somewhat higher. Compression of the tongue root happens later (frame 595 ms), further increasing the height of the tongue into the palatal vault (yellow).

In /ɔɪ/ (middle row), the tongue tip stretching forward was identified by the vertical tag lines in that area starting to curve (green). Then as the tongue moves forward and higher, upper-lower shear (cyan), compression in tongue body (magenta), and some tongue root fronting (yellow) is observed in the later frames.

In /aʊ/ (bottom row), we again see early compression and curving of the tongue tip (green). Shear (cyan) appears as the tongue retracting and bunching toward the pharyngeal wall. Compression in both tongue body (magenta) and tongue root (yellow) further move the tongue upward toward the velum.

The representative frames were chosen specifically to show the timing relations of these various tongue internal deformations, documented as the 4 colors distributing differently in time from left to right. For instance, in the top row deformation of tongue body (magenta) and tongue base (yellow) (which can be thought of as the tongue's 'undercarriage') is seen during /aɪ/, with the former happening earlier (~300 ms) than the latter (~590 ms). Another example

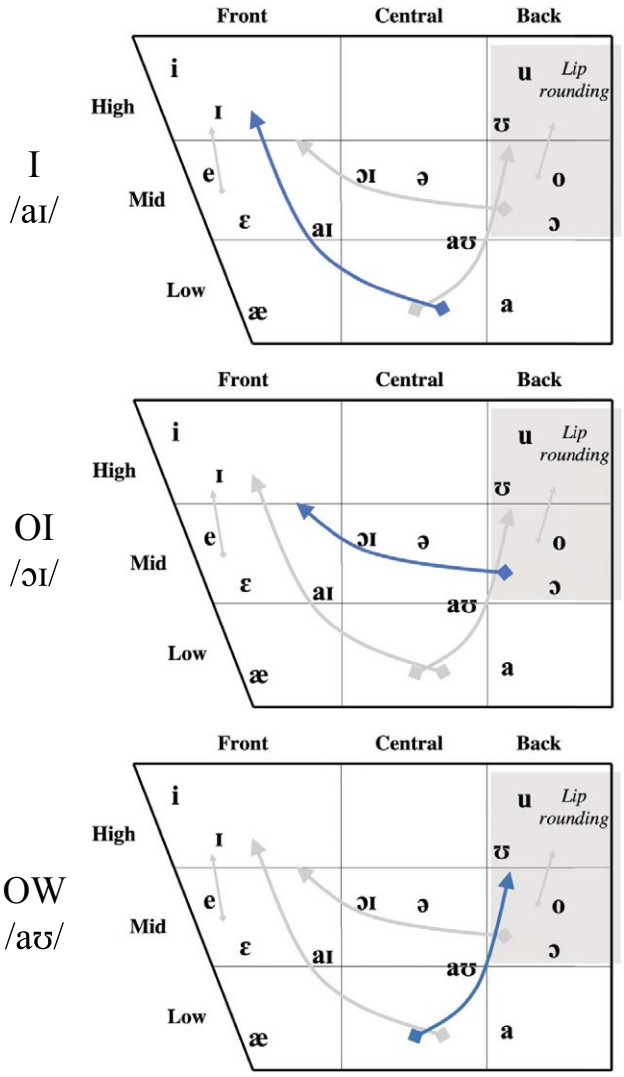

FIGURE 5 American English Vowel Charts illustrate a rough schema for understanding tongue position observed in the representative frames in Figure 6

is tongue tip deformation, which happened early in all diphthongs tested, indicated by green arrows on the left.

Figure 7 shows diphthongs in carrier phrases: (Figure 7A) "a buy puppy," (Figure 7B) "a boy puppy," and (Figure 7C) "a bow puppy." Supporting Information Video S2-S4 shows the 3 diphthong stimuli in carrier phrases with synchronized audio recording. Intensity-time (x-t) plots are shown in the top rows of (Figure 7A-C) and the moment at which the tagging module was applied is indicated at the very top of the figure and serves as the temporal alignment point for the figures. Six representative frames are zoomed out in the bottom rows with green and magenta dashed squares marking the start and end of the diphthong articulations. (Note that the representative frames in (Figure 7C) have a shorter time span compared with Figure 7A and B). The tag persisted from the beginning of the mid-central /ə/ that preceded the target word in the carrier phrase and successfully visualized deformation of the tongue for the entire course of the target diphthong.

The first frames in Figure 7A-C show the tag applied when the tongue started at a mid-central vowel /ə/ (the initial "a" of the carrier phrase), so that all of the deformations in the later frames are relative to this relatively neutral vocalic schwa posture. Note that while (Figure 7A, 4)/aɪ/ and (Figure 7C, 3)/aʊ/ start with similar low and retracted tongue postures marked by pharyngeal narrowing, differences in the internal tongue can be immediately visualized in the distinct grid deformations. This confirms subtle distinction between the starting position of /aɪ/ and /aʊ/, echoed by the American English Vowel Chart in Figure 5. Similarly, the deformational difference between the ending posture of /aʊ/ (Figure 7B, 4) and the starting posture of /ɔɪ/ (Figure 7C, 5) was clearly evident; more bi-concave rectangles exist in (Figure 7C, 5) in addition to parallelograms in both (Figure 7B, 4) and (Figure 7C, 5), indicating horizontal squeeze, which further packs the tongue up toward palatal vault. This is consistent with the placement in the second and third vowel charts in Figure 5.

With a relatively neutral schwa posture (frame 1's) as a reference, the deformations also indicate regional motion within the tongue: in (Figure 7B, 3-4) parallelograms in the middle of the tongue indicate shear serving to retract the tongue body back toward the pharyngeal wall; (Figure 7A, 6, Figure 7B, 6) indicate horizontal compression squeezing the tongue up toward the palate. Little or no deformation is observed during the maintenance of the most extreme postures such as (Figure 7A, 6; Figure 7B, 6; Figure 7C, 3).

Figure 8 shows different deformation patterns in 3 example consonant stimuli. In /ɹ/, curved tag lines in tongue tip (green) are evident, indicating the upward 'bending' deformation of the tongue front high into the palatal vault. Note that /ɹ/ has 3 constrictions: at the lips, in the post-alveolar region, and in the pharynx; while /ʃ/ and /tʃ/ only have 1 constriction, in the post-alveolar region. Thus in /ɹ/ vertical compression in the tongue body (yellow arrows in /ɹ/) is seen due to the tongue body and root is squeezed toward the pharyngeal wall. This vertical compression is not present in the other 2 consonant stimuli. In both /ʃ/ and /tʃ/, the x-t waveforms show there is a highly similar airway shape (i.e., tongue surface contour), as we would expect for the fricative portion (green dash). However, internal differences are visible, presumably arising from the pull-away characteristics of the blade that remains pressed or stabilized upward more so for /tʃ/ (magenta) than for /ʃ/ (green). Significantly, the tagged images show the tongue internal deformation differences even when tongue surface contours and vocal tract constriction locations are comparable.

4 | DISCUSSION

We have demonstrated intermittent tagging during RT-MRI as a potential means to visualize internal tongue motion

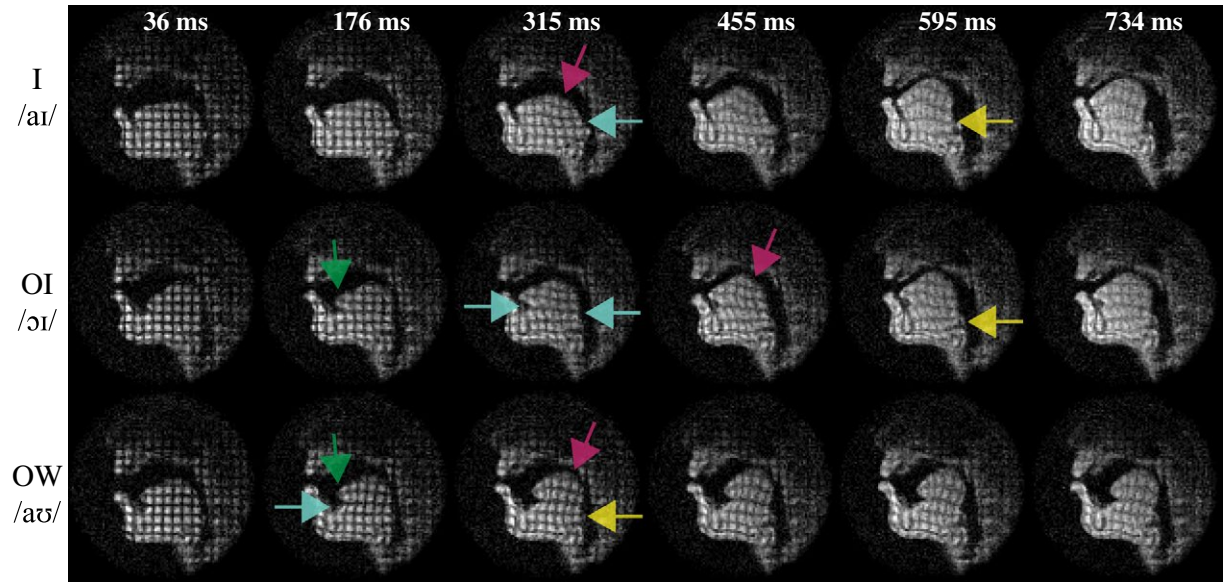

FIGURE 6 Tagged RT-MRI reveals internal tongue deformations and their relative timing during American English diphthong articulation. Each color indicates the start of a different motion pattern: (left to right) tongue tip deformation (green), shear (cyan), tongue body compression (magenta), and tongue root compression (yellow). Importantly, the relative timing of motion patterns is seen; for example, deformation of the tongue tip (green) was followed by shear (cyan) and finally compression of the tongue root (yellow)

during speech production. This approach eliminates the need for re-binning data using multiple repetitions and is suitable for investigations of natural speech production. We demonstrated a framework to select imaging parameters in consideration of image quality and tag persistence and achieved an imaging window of approximately 650-800 ms at 1.5T, with imaging SNR $\geq$ 17 and tag CNR $\geq$ 5 in human tongue. This work leverages mature speech RT-MRI techniques[12,45] to provide adequate spatiotemporal resolution for tagged imaging. The resulting method is able to capture tongue motion patterns and their relative timing as exemplified by internal tongue deformation during American English diphthong vowels and consonants. This method can also provide images for further quantification of internal tongue motion.[34,58-61]

The proposed method may provide insight into several open questions in speech science and linguistics. For instance, acoustic studies have shown that the vocalic formants of the initial and terminal portions of a diphthong are not necessarily the same as those found for the simple vowels in monophthongs used to describe them.[54] Hsieh et al[55] hypothesized that strong biomechanical coupling between starting and ending gestures truncates diphthong articulation, leading to less extreme [a] vowels (as compared with the corresponding monophthong). This study used constriction degree to examine diphthong articulation, by assuming that constrictions can be identified with higher signal intensity in an ROI. The proposed tagging method here can enable testing of this and similar hypotheses by directly examining the biomechanical subsystems in the tongue.

The proposed method may also serve to provide insights into disease states that affect speech production. CINE-tagging has been used by Lee et al[61] to assess tongue impairment in amyotrophic lateral sclerosis patients and by Stone et al[35] to investigate articulation variance between post-glossectomy patients and controls. For these applications, the requisite repeating motion required in CINE-tagging could be burdensome for some patients, aside from the fact that highly consistent repeatability, which is challenging in impaired speech, is required for re-binning data. Such challenge is demonstrated in Supporting Information Video S5. The proposed RT-MRI tagging method can substantially simplify the data acquisition and preclude errors from a re-binning process, by compromising resolution and/or SNR. Lastly, tongue muscle movement patterns in obstructive sleep apnea patients have been characterized in clinical studies for treatment evaluation.[62] The proposed method with automatic periodic tagging could potentially allow studies during natural sleep.

We investigated the performance of the proposed intermittent tagging with 3 varying triggering mechanisms. Cued and periodic tagging perform well for all 4 subject scans. Although there is variability in speech rate across subjects, the flexible nature of these intermittent-tagging protocols allows us to flexibly adjust the triggering timing.

As a feasibility effort, this work used a fairly simple tagging module. We used a 1-3-3-1 SPAMM tagging sequence, as established in the literature, and produced high quality visualization of tag grids in tongue. There exist many alternatives to SPAMM. Several cardiovascular MR tagging approaches can potentially be adapted

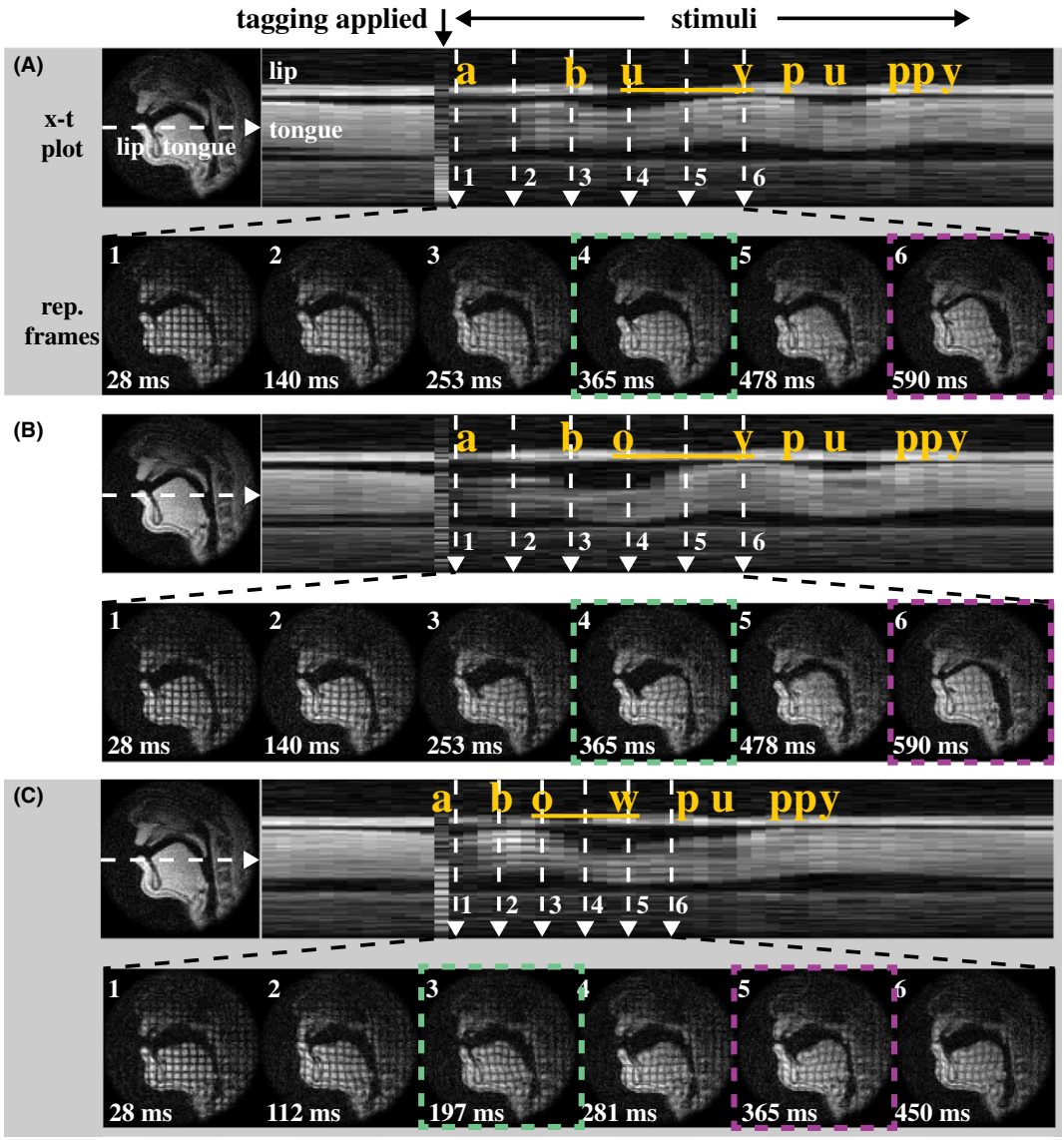

FIGURE 7 Tagged RT-MRI reveals deformation relative to the relatively neutral posture of the schwa /ə/ ("a") of the carrier sentence. Stimuli occurred in carrier phrases: (A) "a buy puppy," (B) "a boy puppy" and (C) "a bow puppy." The intensity-time (x-t) plots in top rows of (A-C) indicate tagging timing, and 6 representative frames are shown across time in each bottom row. Green and magenta dashed square mark the start and end gestures of the diphthong articulation. Note the deformation differences in internal tongue among the 3 diphthongs' starting postures and across their ending postures. (Such as start of /aɪ/ versus /aʊ/ as in a4 versus c3, start of /ɔɪ/ versus end of /aʊ/ as in B, 4 versus C, 5.)

for speech applications.[39] Particularly appealing options include HARmonic Phase (HARP)[63] and Displacement ENcoding with Stimulated Echoes (DENSE),[64] allowing faster and simpler post processing and analysis. HARP has been adapted for speech production in the CINE framework.[33,36,37,58,61] More rapid data acquisition implementation by echo planar imaging was proposed for cardiac HARP,[42] in which only the spectral peak of interest was acquired. DENSE provides higher sensitivity and spatial resolution. However, the technique is derived from stimulated echo acquisition mode (STEAM) sequence and suffers from low SNR. Phase contrast imaging has been shown for the application of tissue velocity mapping in myocardial motion[65] as well as in skeletal muscle contraction.[66] This technique encodes information about velocity into the phase of the detected signal. Note that all 3 of these alternatives are phase-sensitive methods; phase errors introduced by uncounted off-resonance need to be carefully considered when adapting to speech applications.[65,67-69]

The SPAMM parameters may also be optimized. We used grid spacing of 1 cm, but this spacing may need adjustment based on the size of the subject. For example, we expect a finer grid spacing will be required in small people, such as young children. The grid spacing may also need modification depending on the specific muscle groups or vocal tract subsystems of interest such that they are fine enough to

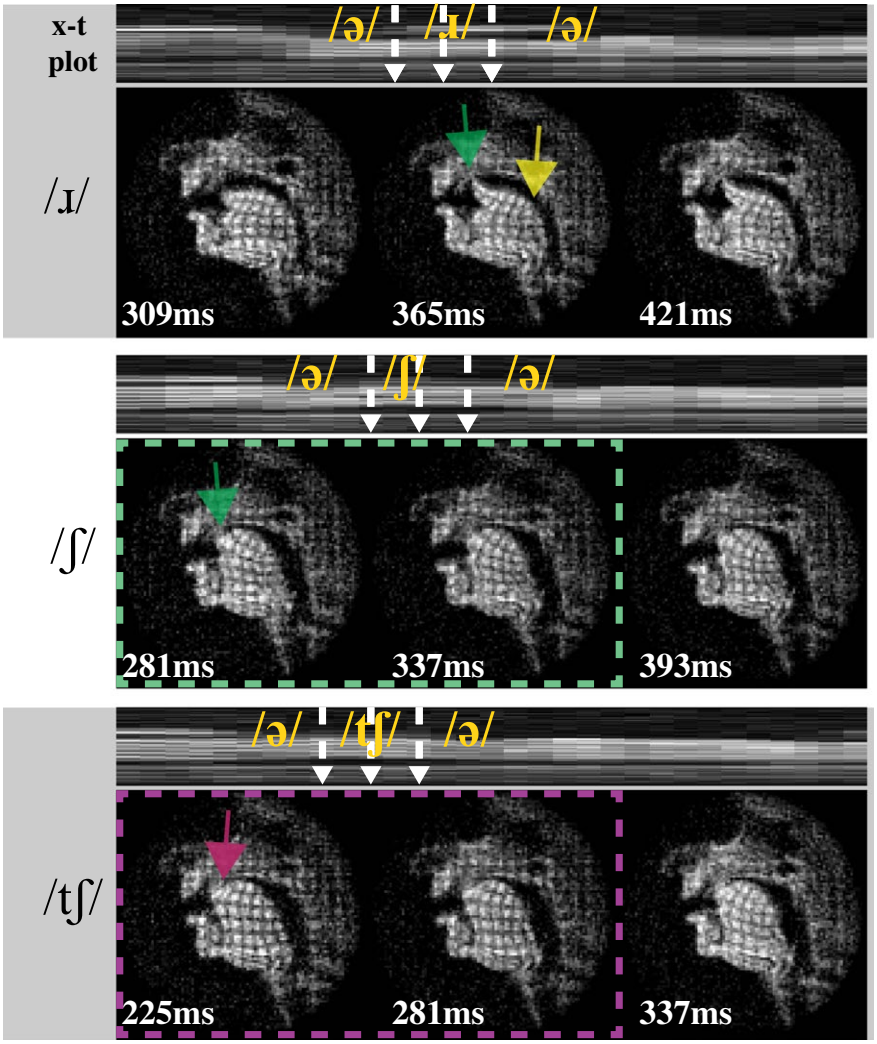

FIGURE 8 Tagged RT-MRI shows different deformation patterns (relative to preceding schwa postures) during the articulation of consonants /ɹ/, /ʃ/, and /tʃ/. The intensity-time (x-t) plots in top rows indicate tagging timing, and 3 representative frames are shown across time in each bottom row. All stimuli involve constriction with the tongue tip and/or blade (i.e., the tongue front) in the post-alveolar region of the vocal tract. Of interest, the tagged images show tongue internal deformation differences (magenta versus green) even when tongue surface contours and vocal tract constriction locations are comparable

distinguish the contractions and internal movements of the specific lingual muscle system(s) of interest such as for the tongue tip.

Improvement in tag persistence is also of interest. Variable FA has been used in spiral myocardial tagging to improve contrast throughout the entire cardiac cycle. Ryf et al[70] applied larger FA in the later stage of the imaging cycle to compensate the faded longitudinal magnetization. This topic remains as future work.

Motion artifacts exist in some of the current results. This is not surprising as Lingala et al[12] pointed out that fully sampled single slice RT-MRI cannot resolve all tongue movements, especially during faster pace speaking or those involving intrinsically faster subsystem movement such as by the tongue tip. These artifacts can be mitigated by undersampling and constrained reconstruction methods, which have yet to be explored in combination with tagging.

Imaging at 3T is of interest because it could provide longer tag persistence and higher SNR. We conducted all of our experiments at 1.5T field strength. Previous studies have compared imaging at 1.5T and 3T for cardiac applications for the SSFP sequence.[53] With the same imaging parameters, the tag persists approximately 25-30% longer due to slower T_1 relaxation and higher intrinsic imaging SNR in human tongue. This can be further improved by a smaller FA, considering the lower Ernst angle needed for longer T_1. However, stronger off-resonance emerges at higher field strength, especially at air-tissue boundaries with an amount of approximately 9.4 ppm.[71] This could cause blurring of the grid near the tongue surface, or even total disappearance in subtle structure such as the tongue tip. To mitigate the off-resonance artifacts, dynamic off-resonance can also be incorporated into the reconstruction pipeline to reduce artifacts.[21,29] Subjects with large proton density fat fraction at the base of the tongue (inferior-posterior) will also suffer from signal dephasing due to off-resonance of 3.5 ppm between fat and water.[72] This signal loss can be reduced by shortening the readout duration of spiral acquisition while trading-off temporal resolution, or by using another sampling pattern with short readout, such as radial sampling.[26]

5 | CONCLUSIONS

We have developed and demonstrated a method for intermittent tagging during real-time MRI of speech production to reveal internal deformations of the tongue. We incorporated 1-3-3-1 SPAMM tagging with rapid spiral GRE to reveal the internal tongue motion during articulation. We showed that this method can capture various motion patterns in the tongue and their relative timing using case examples of American English diphthongs and consonants. The proposed method can potentially provide tools to investigate muscle function or other applications of internal tissue movement in future scientific and clinical research.

ACKNOWLEDGMENTS

We thank Eric Peterson, William Overall, and Juan Santos at HeartVista, Inc. for supporting on RTHawk Research system. We acknowledge the support and collaboration of the Speech Production and Articulation kNowledge (SPAN) group at the University of Southern California, Los Angeles, California.

ORCID

Weiyi Chen https://orcid.org/0000-0001-5116-8645
Dani Byrd http://orcid.org/0000-0003-3319-5871
Shrikanth Narayanan http://orcid.org/0000-0002-1052-6204
Krishna S. Nayak http://orcid.org/0000-0001-5735-3550

REFERENCES

1. Delattre P, Freeman DC. A dialect study of american R'S by x-ray motion picture. *Linguistics*. 1968;6:29–68.
2. Perrier P, Boë LJ, Sock R. Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract castmodeling the transition with two sets of coefficients. *J Speech Lang Hear Res*. 1992;35:53–67.
3. Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, Jackson MT. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J Acoust Soc Am*. 1992;92:3078–3096.
4. Stone M, Shawker TH, Talbot TL, Rich AH. Cross-sectional tongue shape during the production of vowels. *J Acoust Soc Am*. 1988;83:1586–1596.
5. Bresch E, Kim Y-C, Nayak K, Byrd D, Narayanan S. Seeing speech: capturing vocal tract shaping using real-time magnetic resonance imaging. *IEEE Signal Process Mag*. 2008;25:123–132.
6. Scott AD, Wylezinska M, Birch MJ, Miquel ME. Speech MRI: morphology and function. *Phys Med*. 2014;30:604–618.
7. Demolin D, Hassid S, Metens T, Soquet A. Real-time MRI and articulatory coordination in speech. *C R Biol*. 2002;325:547–556.
8. Honda K, Takemoto H, Kitamura T, Fujita S, Takano S. Exploring human speech production mechanisms by MRI. *IEICE Trans Inf Syst*. 2004;E87:1050–1058.
9. NessAiver MS, Stone M, Parthasarathy V, Kahana Y, Paritsky A. Recording high quality speech during tagged cine-MRI studies using a fiber optic microphone. *J Magn Reson Imaging*. 2006;23:92–97.
10. Ventura SR, Freitas DR, Tavares JM. Application of MRI and biomedical engineering in speech production study. *Comput Methods Biomech Biomed Engin*. 2009;12:671–681.
11. Narayanan S, Nayak K, Lee S, Sethy A, Byrd D. An approach to real-time magnetic resonance imaging for speech production. *J Acoust Soc Am*. 2004;115:1771–1776.
12. Lingala SG, Zhu Y, Kim YC, Toutios A, Narayanan S, Nayak KS. A fast and flexible MRI system for the study of dynamic vocal tract shaping. *Magn Reson Med*. 2017;77:112–125.
13. Lingala SG, Sutton BP, Miquel ME, Nayak KS. Recommendations for real-time speech MRI. *J Magn Reson Imaging*. 2016;43:28–44.
14. Kier WM, Smith KK. Tongues, tentacles and trunks: the biomechanics of movement in muscular-hydrostats. *Zool J Linn Soc*. 1985;83:307–324.
15. Hiiemae KM, Palmer JB. Tongue movements in feeding and speech. *Crit Rev Oral Biol Med*. 2003;14:413–429.
16. Green JR. Mouth matters: scientific and clinical applications of speech movement analysis. *Perspect Speech Sci Orofac Disord*. 2015;25:6.
17. Gerard JM, Wilhelms-Tricarico R, Perrier P, Payan Y. A 3D dynamical biomechanical tongue model to study speech motor control. *Recent Res Dev Biomech*. 2003;arXiv:p;physics/0606148.
18. Buchaillard S, Perrier P, Payan Y. A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning. *J Acoust Soc Am*. 2009;126:2033–2051.
19. Toutios A, Narayanan SS. Articulatory synthesis of french connected speech from EMA data. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Lyon, France, 2013. pp. 2738–2742.
20. Fu M, Zhao Bo, Carignan C, et al. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. *Magn Reson Med*. 2015;73:1820–1832.
21. Sutton BP, Conway CA, Bae Y, Seethamraju R, Kuehn DP. Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T. *J Magn Reson Imaging*. 2010;32:1228–1237.
22. Burdumy M, Traser L, Burk F, et al. One-second MRI of a three-dimensional vocal tract to measure dynamic articulator modifications. *J Magn Reson Imaging*. 2017;46:94–101.
23. Zhu Y, Proctor MI, Kim Y-C, Narayanan SS, Nayak KS. Dynamic 3D Visualization of Vocal Tract Shaping during Speech. *IEEE Trans Med Imaging*. 2013;32:838–848.
24. Fu M, Barlaz MS, Holtrop JL, et al. High-frame-rate full-vocal-tract 3D dynamic speech imaging. *Magn Reson Med*. 2017;77:1619–1629.
25. Lim Y, Zhu Y, Lingala SG, Byrd D, Narayanan S, Nayak KS. 3D dynamic MRI of the vocal tract during natural speech. *Magn Reson Med*. 2019;81:1511–1520.
26. Niebergall A, Zhang S, Kunay E, et al. Real-time MRI of speaking at a resolution of 33 ms: undersampled radial FLASH with nonlinear inverse reconstruction. *Magn Reson Med*. 2013;69:477–485.

27. Lingala SG, Zhu Y, Lim Y, et al. Feasibility of through-time spiral generalized autocalibrating partial parallel acquisition for low latency accelerated real-time MRI of speech. *Magn Reson Med*. 2017;78:2275–2282.
28. Uecker M, Zhang S, Voit D, Karaus A, Merboldt KD, Frahm J. Real-time MRI at a resolution of 20 ms. *NMR Biomed*. 2010;23:986–994.
29. Lim Y, Lingala SG, Narayanan SS, Nayak KS. Dynamic off-resonance correction for spiral real-time MRI of speech. *Magn Reson Med*. 2019;81:234–246.
30. Kumada K, Niitsu M, Niimi S, Hirose H. A study on the inner structure of the tongue in the production of the 5 Japanese vowels by tagging snapshot MRI. *Ann Bull RILP*. 1992;26:1–5.
31. Niitsu M, Kumada M, Campeau NG, Niimi S, Riederer SJ, Itai Y. Tongue displacement: visualization with rapid tagged magnetization- prepared MR imaging. *Radiology*. 1994;191:578–580.
32. Napadow VJ, Chen Q, Wedeen VJ, Gilbert RJ. Intramural mechanics of the human tongue in association with physiological deformations. *J Biomech*. 1999;32:1–12.
33. 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.
34. Parthasarathy V, Prince JL, Stone M, Murano EZ, Nessaiver M. Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing. *J Acoust Soc Am*. 2007;121:491–504.
35. Stone M, Woo J, Zhuo J, Chen H, Prince JL. Patterns of variance in /s/ during normal and glossectomy speech. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2014;2:197–207.
36. Woo J, Xing F, Stone M, et al. Speech map: a statistical multimodal atlas of 4D tongue motion during speech from tagged and cine MR images. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2017;1–13.
37. Woo J, Lee J, Murano EZ, et al. A high-resolution atlas and statistical model of the vocal tract from structural MRI. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2015;3:47–60.
38. Shehata ML, Cheng S, Osman NF, Bluemke DA, Lima JA. Myocardial tissue tagging with cardiovascular magnetic resonance. *J Cardiovasc Magn Reson*. 2009;11:55.
39. Ibrahim el-SH. Myocardial tagging by cardiovascular magnetic resonance: evolution of techniques–pulse sequences, analysis algorithms, and applications. *J Cardiovasc Magn Reson*. 2011;13:36.
40. Douglas AS, Rodriguez EK, O'Dell W, Hunter WC. Unique strain history during ejection in canine left ventricle. *Am J Physiol*. 1991;260:1596–1611.
41. McVeigh ER, Epstein F. Myocardial tagging during real-time MRI. In: 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 2284–2285.
42. Sampath S, Derbyshire JA, Atalar E, Osman NF, Prince JL. Real-time imaging of two-dimensional cardiac strain using a harmonic phase magnetic resonance imaging (HARP-MRI) pulse sequence. *Magn Reson Med*. 2003;50:154–163.
43. Pan L, Stuber M, Kraitchman DL, Fritzges DL, Gilson WD, Osman NF. Real-time imaging of regional myocardial function using fast-SENC. *Magn Reson Med*. 2006;55:386–395.
44. Ibrahim E-S, Stuber M, Fahmy AS, et al. Real-time MR imaging of myocardial regional function using strain-encoding (SENC) with tissue through-plane motion tracking. *J Magn Reson Imaging*. 2007;26:1461–1470.
45. Santos JM, Wright GA, Pauly JM. Flexible real-time magnetic resonance imaging framework. In The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 1048–1051.
46. Fischer SE, McKinnon GC, Maier SE, Boesiger P. Improved myocardial tagging contrast. *Magn Reson Med*. 1993;30:191–200.
47. Axel L, Dougherty L. MR imaging of motion with spatial modulation of magnetization. *Radiology*. 1989;171:841–845.
48. Axel L, Dougherty L. Heart wall motion: improved method of spatial modulation of magnetization for MR imaging. *Radiology*. 1989;172:349–350.
49. Young AA, Axel L, Dougherty L, Bogen DK, Parenteau CS. Validation of tagging with MR imaging to estimate material deformation. *Radiology*. 1993;188:101–108.
50. Walsh DO, Gmitro AF, Marcellin MW. Adaptive reconstruction of phased array MR imagery. *Magn Reson Med*. 2000;43:682–690.
51. King KF, Ganin A, Zhou XJ, Bernstein MA. Concomitant gradient field effects in spiral scans. *Magn Reson Med*. 1999;41:103–112.
52. Markl M, Bammer R, Alley MT, et al. Generalized reconstruction of phase contrast MRI: analysis and correction of the effect of gradient field distortions. *Magn Reson Med*. 2003;50:791–801.
53. Markl M, Scherer S, Frydrychowicz A, Burger D, Geibel A, Hennig J. Balanced left ventricular myocardial SSFP-tagging at 1.5T and 3T. *Magn Reson Med*. 2008;60:631–639.
54. Lehiste I, Peterson GE. Transitions, glides, and diphthongs. *J Acoust Soc Am*. 1961;33:268.
55. Hsieh FY, Goldstein L, Byrd D, Narayanan S. Truncation of pharyngeal gesture in English diphthong [aI]. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Lyon, France, 2013. pp. 968–972.
56. Lee S, Potamianos A, Narayanan S. Developmental acoustic study of American English diphthongs. *J Acoust Soc Am*. 2014;136:1880–1894.
57. Valeti VU, Chun W, Potter DD, et al. Myocardial tagging and strain analysis at 3 Tesla: comparison with 1.5 Tesla imaging. *J Magn Reson Imaging*. 2006;23:477–480.
58. Woo J, Stone M, Suo Y, Murano EZ, Prince JL. Tissue-point motion tracking in the tongue from cine MRI and tagged MRI. *J Speech Lang Hear Res*. 2014;57:S626–S636.
59. Woo J, Lee J, Murano EZ, et al. A high-resolution atlas and statistical model of the vocal tract from structural MRI. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2015;3:47–60.
60. Xing F, Woo J, Gomez AD, et al. Phase vector incompressible registration algorithm for motion estimation from tagged magnetic resonance images. *IEEE Trans Med Imaging*. 2017;36:2116–2128.
61. Lee E, Xing F, Ahn S, et al. Magnetic resonance imaging based anatomical assessment of tongue impairment due to amyotrophic lateral sclerosis: a preliminary study. *J Acoust Soc Am*. 2018;143:EL248-EL254.
62. Brown EC, Cheng S, McKenzie DK, Butler JE, Gandevia SC, Bilston LE. Tongue and lateral upper airway movement with mandibular advancement. *Sleep*. 2013;36:397–404.
63. Osman NF, McVeigh ER, Prince JL. Imaging heart motion using harmonic phase MRI. *IEEE Trans Med Imaging*. 2000;19:186–202.
64. Aletras AH, Ding S, Balaban RS, Wen H. DENSE: Displacement encoding with stimulated echoes in cardiac functional MRI. *J Magn Reson*. 1999;137:247–252.

65. Nayak KS, Nielsen J-F, Bernstein MA, et al. Cardiovascular magnetic resonance phase contrast imaging. *J Cardiovasc Magn Reson*. 2015;17:71.
66. Mazzoli V, Gottwald LM, Peper ES, et al. Accelerated 4D phase contrast MRI in skeletal muscle contraction. *Magn Reson Med*. 2018;80:1799–1811.
67. Kuijer JP, Hofman MB, Zwanenburg JJ, Marcus JT, van Rossum AC, Heethaar RM. DENSE and HARP: two views on the same technique of phase-based strain imaging. *J Magn Reson Imaging*. 2006;24:1432–1438.
68. Haraldsson H, Sigfridsson A, Sakuma H, Engvall J, Ebbers T. Influence of the FID and off-resonance effects in dense MRI. *Magn Reson Med*. 2011;65:1103–1111.
69. Ryf S, Tsao J, Schwitter J, Stuessi A, Boesiger P. Peak-combination HARP: a method to correct for phase errors in HARP. *J Magn Reson Imaging*. 2004;20:874–880.
70. Ryf S, Kissinger KV, Spiegel MA, et al. Spiral MR myocardial tagging. *Magn Reson Med*. 2004;51:237–242.
71. Schenck JF. The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds. *Med Phys*. 1996;23:815–850.
72. Humbert IA, Reeder SB, Porcaro EJ, Kays SA, Brittain JH, Robbins J. Simultaneous estimation of tongue volume and fat fraction using IDEAL-FSE. *J Magn Reson Imaging*. 2008;28:504–508.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

VIDEO S1 Top: Movie of isolated American English Diphthong shown in Figure 4. This video shows top three rows listed in Table 1. Bottom: each color indicates the start of a different motion pattern: (left to right) tongue tip deformation (green), shear (cyan), tongue body compression (magenta), and tongue root compression (yellow). Importantly, the relative timing of motion patterns is seen in (b); for example, deformation of the tongue tip (green) was followed by shear (cyan) and finally compression of the tongue root (yellow)

VIDEO S2 Movie of tagged real-time MRI with a synchronized audio shown in Figure 5(A). This video shows "a buy puppy" listed in Table 1

VIDEO S3 Movie of tagged real-time MRI with a synchronized audio shown in Figure 5(B). This video shows "a boy puppy" listed in Table 1

VIDEO S4 Movie of tagged real-time MRI with a synchronized audio shown in Figure 5(C). This video shows "a bow puppy" listed in Table 1

VIDEO S5 Significance of tagged real-time MRI for speech. Top row: five trials of tagged real-time MRI during stimuli "a boy puppy". In each trial the tagging was triggered at the desired timing and successfully capture the deformation throughout the whole articulation. Bottom row: averaging over five trials shows high SNR but unacceptable quality, due to high intra-subject variability of speech production

How to cite this article: Chen W, Byrd D, Narayanan S, Nayak KS. Intermittently tagged real-time MRI reveals internal tongue motion during speech production. *Magn Reson Med*. 2019;82:600–613. https://doi.org/10.1002/mrm.27745