Rapid dynamic speech imaging at 3 Tesla using combination of a custom vocal tract coil, variable density spirals and manifold regularization

Rushdi Zahid Rusho¹, Abdul Haseeb Ahmed², Stanley Kruger³, Wahidul Alam¹, David Meyer⁴, David Howard⁵, Ingo Titze⁶, Mathews Jacob², Sajan Goud Lingala¹,³

¹ Roy J Carver Department of Biomedical Engineering, University of Iowa, Iowa city, IA, 52242
² Department of Electrical and Computer Engineering, University of Iowa, Iowa city, IA, 52242
³ Department of Radiology, University of Iowa, Iowa city, IA, 52242
⁴ Janette Ogg Voice Research Center, Shenandoah University, Winchester, VA, 22601
⁵ Department of Electronic Engineering, Royal Holloway, University of London, UK
⁶ National Center for Voice and Speech, Clearfield, Utah, 84112

Corresponding author
Sajan Goud Lingala, PhD
Assistant Professor
Roy J Carver Department of Biomedical Engineering
Department of Radiology
University of Iowa
sajangoud-lingala@uiowa.edu

Approximate word count: 5100
Figures: 10
Supplementary material: 4 figures, 1 table, and 8 videos.
Purpose: To improve dynamic speech imaging at 3 Tesla.

Methods: A novel scheme combining a 16-channel vocal tract coil, variable density spirals (VDS), and manifold regularization was developed. Short readout duration spirals (1.3 ms long) were used to minimize sensitivity to off-resonance. The manifold model leveraged similarities between frames sharing similar vocal tract postures without explicit motion binning. Reconstruction was posed as a SENSE-based non-local soft weighted temporal regularization scheme. The self-navigating capability of VDS was leveraged to learn the structure of the manifold. Our approach was compared against low-rank and finite difference reconstruction constraints on two volunteers performing repetitive and arbitrary speaking tasks. Blinded image quality evaluation in the categories of alias artifacts, spatial blurring, and temporal blurring were performed by three experts in voice research.

Results: We achieved a spatial resolution of 2.4mm²/pixel and a temporal resolution of 17.4 ms/frame for single slice imaging, and 52.2 ms/frame for concurrent 3-slice imaging. Implicit motion binning of the manifold scheme for both repetitive and fluent speaking tasks was demonstrated. The manifold scheme provided superior fidelity in modeling articulatory motion compared to low rank and temporal finite difference schemes. This was reflected by higher image quality scores in spatial and temporal blurring categories. Our technique exhibited faint alias artifacts, but offered a reduced interquartile range of scores compared to other methods in alias artifact category.

Conclusion: Synergistic combination of a custom vocal-tract coil, variable density spirals and manifold regularization enables robust dynamic speech imaging at 3 Tesla.
INTRODUCTION

Speech production involves an intricate coordination of several soft tissue articulators such as lips, tongue, soft-palate, epiglottis, vocal folds, and the pharyngeal wall. Safely visualizing the dynamics of both the deforming vocal tract air space and the neighboring articulators is a powerful means to explore and understand better the nuances of speech production. Dynamic magnetic resonance imaging has emerged as the modality of choice to visualize speech because of several advantages over competing modalities such as lack of ionizing radiation, ability to capture longitudinal measures, flexibility to image in arbitrary image planes, and the ability to visualize deep structures such as the vocal folds [1], [2]. It has been used in several speech science and clinical studies to better understand apraxia of speech [3], quantifying levator veli palatini muscle movements [4], answering open questions in phonetics (eg. understanding timing effects of nasal syllables [5]), and to better understand speech production in different languages such as French [6], [7], Arabic [8], Brazilian Portugese [9]. It has also been applied in studies aimed to understand mechanics of vocal tract shaping during singing [10], [11], beatboxing [12], [13], playing the horn musical instrument [14].

In the past decade, several schemes based on sparse k-space v.s time (k-t) sampling and model-based reconstruction have been applied to improve spatio-temporal resolutions, and/or vocal tract coverage at both 1.5 T and 3 T [9], [14]–[20]. Several constrained models have been applied, including those that exploit transform sparsity [14]–[17], joint low rank and transform sparsity [20], [21], and low rank plus transform sparsity [19]. However, these constraints tend to introduce motion blurring and loss of spatial and temporal features while modeling fast motion such as during speech. Spiral trajectories with short readouts (<2.5 ms) have been the natural choice at 1.5 T due to their robustness to encode fast motion [1]. However, at 3 T, adapting a 2.5 ms duration readout can induce more than 2 cycles of phase accrual, and cause significant blurring at air-tissue boundaries. It is noteworthy that, off-resonance can be significant at 3T, up to ~1200 Hz. Lesser time efficient radial trajectories are more commonly used due to their short readouts (~1.2 to 1.4 ms) [1]. Under-sampled Cartesian trajectories, combined with 3D spiral navigators have been applied at 3T for native 3D dynamic rapid imaging [20], [21].

In applications involving free breathing (eg. free breathing abdomen MRI), explicit motion binning strategies have shown significant promise to reduce motion blurring associated with classic low rank, and/or sparsity models [22]–[24]. These models rely on navigator data such as those from respiratory bellows, or data derived from k-space data to extract the breathing motion pattern. These data are then used to bin the dynamic data into image frames with similar motion state. The models benefit by
imposing sparsity/low rank assumption within each motion state, i.e., along this “extra dimension.” However, performing such explicit binning is infeasible in dynamic speech MRI because the motion can be arbitrary. Emerging manifold regularization approaches have shown significant promise to improve motion fidelity in free breathing and ungated cardiac MRI [25]–[27]. These schemes model image frames as points living on a low dimensional manifold embedded in a high dimension space. Similar image frames are mapped as neighbors on the manifold, even if they are distant in time. A key difference with extra dimension schemes is that it implicitly exploits similarities between image frames without the need of explicit motion binning. This makes it an attractive method to explore in the context of arbitrary speech production where explicit binning is not feasible.

In this work, we aim to improve the rapid imaging of dynamic speech at 3 T by synergistically exploiting the a) parallel imaging capabilities of a custom 16 channel vocal tract receive coil, b) self-navigation capabilities of short readout variable density spirals, and c) implicit motion resolved capabilities of manifold regularization. The custom vocal tract coil is designed to provide high sensitivity over several important anatomical regions relevant in speech production (e.g., lips, tongue, soft palate, vocal tract airspace including the infra glottic airspace). Variable density spirals that oversample the center of the k-space are used to estimate a graph Laplacian matrix, which determines the neighborhood relations between image frames. To minimize off resonance artifacts, our spirals are implemented with extremely short readouts (~1.3 ms); i.e., similar to readout duration of widely used radials at 3T. The reconstruction is posed as a SENSE-based $l_2$ norm non-local soft weighted temporal regularization scheme. We demonstrate the applicability of our approach to enable single slice (mid sagittal) imaging at 17.4 ms/frame; and concurrent three slice (1 mid, and 2 parasagittal) imaging at 52.2 ms/frame at a spatial resolution of 2.4 mm$^2$/pixel. A variety of speech tasks were imaged, including repeatedly producing vowel and consonant sounds, and producing fluent speech by two healthy volunteer speakers. We analyze the mechanics of the Laplacian matrix in these speech tasks, and qualitatively demonstrate implicit binning enabled by our approach. We compare our proposed reconstruction with two widely used constraints in speech MRI: $l_1$ norm-based finite difference transform sparsity constraint, and a low rank constraint. Through blinded ratings from three experts in vocal science and modeling research (co-authors D. Meyer, D. Howard, and I. Titze), we assess image quality of different reconstructions in the following categories: alias artifacts, spatial blurring, and temporal blurring.
METHODS

Custom vocal tract coil:

A custom 16 channel vocal tract receiver coil was employed on a 3T GE Premier scanner. This coil provided high spatial sensitivity over various regions of interest including tongue, soft palate, pharyngeal wall, larynx, and infra glottic airspace. The coil had two pieces with five elements on either side of the jaw close to the right and left cheeks; and a third piece with six elements placed near the chin and neck. These pieces were mounted on flexible paddles to facilitate flexible positioning for varying subject’s face and neck geometry (also see Fig. 1). The coil elements were constructed so that they could provide spatial diversity of the receiver coil sensitivity profiles along the superior to inferior and anterior to posterior directions.

Variable density spiral design:

A variable density spiral (VDS) sampling scheme was designed [28]. The design parameters were: slew rate = 140 mT/m/ms; maximum gradient amplitude = 80 mT/m; sampling time = 4 μs; spatial resolution = 2.4 mm x 2.4 mm; 27 spiral arms. Figure 2 shows the field of view (FOV) v./s the normalized k-space radius (k_r). When k_r was between (0 to 0.25); (0.25 to 0.5); (0.5 to 1), the FOV respectively changed linearly from (60 cm² to 30 cm²); (30 cm² to 20 cm²); and (20 cm² to 6.66 cm²). A 13 arm uniform density spiral (UDS) was also designed, where the FOV was maintained at 20 cm² for all k_r. The 13 arm UDS scheme is the widely used spiral design for dynamic speech MRI at 1.5 T [15]. VDS produced 1.3 ms long readouts with 335 readout points, while UDS produced approximately 2.69 ms long readouts with 674 readout points (also see Fig.2). For the single slice setting, successive spiral arms were interleaved by the golden angle (~222.379°). For concurrent multi-slice setting, golden angle increments occurred only after spiral arms at a specified angle were acquired from all the slices, similar to [15].

Manifold regularization:

We adapted a previously proposed manifold regularization approach [29], [30]. Figure 3 shows a manifold model schematic, where dynamic images are modeled as points on a low dimensional manifold embedded in a high dimensional ambient space. The dimension of the ambient space equals the number of pixels in the image. Image frames with a similar anatomical vocal tract posture are mapped as neighbors on the manifold, even if they occur arbitrarily in time. The shape of the manifold is
represented using a graph Laplacian matrix, which is learned from the measured data as described below. We represent dynamic image time series as a Casorati matrix, whose columns are the image frames: \( \mathbf{X}_{M \times N} = [x_1; x_2; \ldots; x_N] \); where \( N \) represents the number of frames, and \( M \) represents total number of pixels in each frame.

Manifold regularization exploits similarity of image frames in terms of proximity of points on the smooth manifold. We pose the recovery as an optimization problem, where the criterion is specified by

\[
\min_{\mathbf{X}} \| A(\mathbf{X}) - \mathbf{b} \|^2_2 + \lambda \text{ trace}(\mathbf{XLX}^\mathsf{H}); \quad (1)
\]

where \( A \) is an operator modeling coil sensitivity encoding and Fourier Transform operation on the spiral trajectory. \( \mathbf{b} \) is a vector containing the acquired under-sampled multi-coil kspace data for all time frames. Here, the graph Laplacian matrix \( \mathbf{L}_{N \times N} \) captures the structure of the manifold, and is defined as \( \mathbf{L}_{N \times N} = \mathbf{D} - \mathbf{W} \); where \( \mathbf{W} \) is the \( N \times N \) weight matrix containing the weights \( w_{ij} \); \( \mathbf{D} \) is the diagonal matrix, with diagonal entries, \( D_{ii} = \sum_j w_{ij} \). \( \lambda \) is a tunable parameter that balances data consistency and regularization. The above optimization criterion can also be expressed as

\[
\min_{\mathbf{X}} \| A(\mathbf{X}) - \mathbf{b} \|^2_2 + \lambda \sum_{i=1}^N \sum_{j=1}^N w_{ij} \| x_i - x_j \|^2_2; \quad (2)
\]

which is easier to understand intuitively. Note that the second term comprises of a weighted penalty where the weights \( w_{ij} \) determine the degree of similarity between the \( i^{th} \) and the \( j^{th} \) frame; and are inversely proportional to the distance between the corresponding points on the manifold. That is to say, similar frames are assigned higher weights, and vice-versa.

The second term above now may be viewed as an \( l_2 \) norm based non-local soft-weighted temporal regularization operation. Each row of the \( \mathbf{L} \) matrix would reveal a \( 1 \times N \) kernel containing coefficients used for that corresponding frame. Note, that the \( \mathbf{L} \) matrix is derived from the data itself (discussed below).

This is in stark contrast to the widely used local temporal finite difference kernel, where the \( \mathbf{L} \) matrix would be a predetermined block diagonal matrix with entries \([-1,2,1]\).
Estimating the Laplacian matrix:

We exploit self-navigating capability of the VDS trajectory to estimate the L matrix. We denote the central q% of the VDS k-space samples as \( b_{\text{low-res}} \). Starting with an initial guess of \( x_{\text{low-res}} \) obtained from the inverse nuFFT of \( b_{\text{low-res}} \), we determine L by iterating through the following steps:

\[
w_{ij} = \exp \left( -\frac{\|x_{\text{low-res},i} - x_{\text{low-res},j}\|^2}{\delta^2} \right); \quad (3)
\]

\[
L_{N \times N} = D - W; \quad (4)
\]

\[
\min_{X_{\text{low-res}}} \|A(X_{\text{low-res}}) - b_{\text{low-res}}\|_2^2 + \lambda \operatorname{trace}(X_{\text{low-res}}LX_{\text{low-res}}^H); \quad (5)
\]

The weights in (3) are determined through the Gaussian Kernel, where \( \delta^2 \) is a tunable parameter that controls the smoothness of the manifold. \( x_{\text{low-res}} \) is updated by solving (5) using a conjugate gradient algorithm. We iterate between (3) to (5) to ensure the L matrix is learned from artifact free data, as \( b_{\text{low-res}} \) typically contain few missing samples in the highly under-sampling scenario. The number of iterations were set to 5.

Reconstruction:

Once the L matrix is obtained, one can perform manifold regularized reconstruction of dynamic images by solving (1). However, for faster processing, we performed an eigen decomposition of L as \( L = \mathbf{V}\Sigma\mathbf{V}^H \), and used the \( r \) smallest eigen vectors \( \mathbf{V}_{N \times r} \) to approximate \( X = \mathbf{U}\mathbf{V}^H \). The matrix of \( r \) spatial basis images \( \mathbf{U}_{M \times r} \) were estimated by solving the following computationally simpler optimization problem using a nonlinear conjugate gradient algorithm [30]:

\[
\mathbf{U}^* = \min_{\mathbf{U}} \|A(\mathbf{U}\mathbf{V}^H) - \mathbf{b}\|_F^2 + \lambda \sum_{i=1}^{r} \sigma_i \|u_i\|_2^2 ; \quad (6)
\]

The final Casorati matrix was then obtained as \( X = \mathbf{U}\mathbf{V}^H \).

Reconstruction was implemented in MATLAB (The MathWorks, Inc., Natick MA) on a high-performance computing cluster at The University of Iowa, equipped with an Intel Xenon central processing unit with 28 cores at 2.40 GHz and 128 GB of memory, and a NVIDIA Tesla P100-PCIE graphical processing unit with 16GB memory. Coil elements 11,12,13 captured artifact energy from inferior heart anatomy outside the vocal tract FOV of interest, and were omitted in the reconstruction. Raw k-space data from the remaining 13 coils were coil compressed to 8 virtual coils via PCA-based coil compression. The
gpuNUFFT function was used to implement nuFFT operations in A [31]. Virtual coil sensitivity maps were estimated from time averaged data using the Walsh eigen decomposition approach.

In-vivo experiments on two speakers:

Two healthy adult volunteers (2 male; median age: 29) were scanned, and the study was approved by University of Iowa’s institutional review board where written consent was obtained from the subjects before scanning. Both UDS and VDS sampling schemes were implemented in a spiral gradient echo sequence with rewinding. Imaging parameters were TR=5.8 ms, flip angle=5°; total number of spiral interleaves = 2700; slice thickness = 6mm; spatial resolution = 2.4 mm\(^2\). Single slice acquisitions were performed in the mid-sagittal plane. Concurrent three slice acquisitions were performed in 1 mid-sagittal, and 2 para sagittal planes. Reconstruction was performed by combining every 3 interleaves/frame, which corresponded to a time resolution of 17.4 ms/frame for the single slice setting; and 52.2 ms/frame for the concurrent three slice setting. The volunteers produced the following speaking tasks: a) repetition of the phrase /za-na-za/; b) repetition of the phrase /loo: lee: la: za: na: za:/; c) producing arbitrary/fluent speech by counting numbers (starting from “one”) indefinitely through the scan duration. Note tasks in a) and b) involved interleaving of consonant and vowel sounds. All the above tasks were acquired using the VDS sequence. For one of the volunteers, the UDS and VDS sequences were compared when the volunteer sustained the production of the /a/ vowel. In this comparison, manifold reconstruction was performed using 13 arms/frame for the UDS and 27 arms/frame for the VDS scheme.

The proposed manifold regularized reconstruction was compared against a naïve inverse nuFFT reconstruction, an \(l_1\) norm sparsity-based temporal finite difference regularized scheme [15], and a nuclear norm-based low rank regularization scheme [32]. The sparsity-based finite difference regularized scheme was implemented in the Berkeley advanced reconstruction tool box (BART) computing environment; while the low rank scheme was implemented in MATLAB on the above Intel Xenon CPU. The dynamic Casorati matrix (X) to be reconstructed was of the size 160\(^2\)x900. Multiple slices in the concurrent three slice acquisition were reconstructed independently. Reconstruction times for the different schemes were: a) of the order of 94 minutes for manifold regularization. The L matrix estimation step was implemented on the CPU, which took ~83 minutes; and the U matrix estimation step was implemented on the GPU, which took ~11 minutes; b) of the order of 14.5 minutes for temporal finite difference regularization; and c) of the order of 25 minutes for low rank regularization.
**Tuning free parameters in the reconstruction:**

The proposed manifold regularized reconstruction depends on four parameters: a) the central q% k space samples in \( b_{\text{low-res}} \); b) manifold smoothness parameter \( \delta^2 \) in (3); c) regularization parameter \( \lambda \) in (6); d) number of eigen basis functions \( r \) during the eigen decomposition of the \( L \) matrix. We empirically determined these parameters by visually assessing tradeoffs amongst aliasing artifacts, spatial and temporal blurring. We performed this experimentation on two datasets (fluent speech and repetitive speech tasks), and determined the parameters as: \( q=18\% \); \( \delta^2 = 4.5 \); \( \lambda = 0.2 \); \( r = 30 \). We show image quality for these representative datasets for different parameters in the supplementary material. We observed this choice to be robust across all the datasets from the two speakers. Similarly, we empirically chose the free regularization parameter in the finite difference regularization scheme as 0.07, and the low rank regularization scheme as 0.015 for all the datasets. We note this parameter choice was in accordance with the L-curve heuristic that determines the best regularization parameter balancing energy between data consistency and regularization.

**Visualizing the mechanics of the Laplacian matrix:**

To demonstrate the implicit binning enabled by manifold regularization, we visualized the Laplacian matrix for two different speech tasks from one of the volunteers: a) fluent task of counting numbers indefinitely; and b) repetitive speaking task of repeating the phrase /loo: lee: la: za: na: za:/.

Here, we considered representative reconstructions from 3 slice acquisition for the fluent task, and single slice acquisition for the repetitive speaking task. We analyzed representative rows of the Laplacian matrix, which revealed underlying non-local kernel coefficients used during the proposed non-local weighted temporal regularization. Similar vocal tract postures from different frames corresponding to the peaks of the weights were analyzed. For a representative single slice reconstruction at 17.4 ms/frame, the temporal and spatial bases matrices, \( V \) and \( U \) were visualized and contrasted for the fluent counting and repetitive speaking tasks. These bases were plotted for the first five coefficients.

**Image quality assessment by expert raters:**

Image quality from four reconstruction algorithms (direct inverse nuFFT reconstruction, temporal finite difference regularization, low rank regularization, and manifold regularization) were evaluated in the categories of aliasing artifacts, spatial blurring, and temporal blurring. Three experts in
voice research (co-authors: D. Meyer, D. Howard, and I. Titze) performed blinded ratings of the images. The scoring criterion in different categories followed a four-point scale:

**Alias artifacts category:**
1 - unacceptable; strong alias artifacts present and hampers visualization of all articulators.
2 - adequate; moderate level alias artifacts, and moderate level of interference with articulator visualization.
3 - good; faint alias artifacts present but does not hamper interpretation of articulatory motion
4 - excellent; no alias artifacts.

**Spatial blurring category:**
1 - unacceptable; strong blurring at air-tissue interfaces.
2 - adequate; moderate level blurring; the air-tissue boundaries are blurred and hampers interpretation of articulator boundaries.
3 - good; faint blurring present but does not hamper interpretation of articulator boundaries.
4 - excellent; no perceivable blurring.

**Temporal blurring category:**
1 - unacceptable; strong motion artifacts/blurring; articulatory motion completely blurred.
2 - adequate; moderate level motion artifacts/blurring especially when articulator position change.
3 - good; faint motion artifacts present but does not hamper interpretation of articulatory motion
4 - excellent; no perceivable motion blurring/artifacts.

A total of 8 datasets from each reconstruction algorithm was scored. These were labelled as:

- **Dataset 1:** three-slice acquisition where subject 1 performed a fluent counting task without repetition.
- **Dataset 2:** three-slice acquisition where subject 1 repeated the phrase “loo-lee-la-za-na-za”.
- **Dataset 3:** three-slice acquisition where subject 2 repeated the phrase “loo-lee-la-za-na-za”.
- **Dataset 4:** three-slice acquisition where subject 2 performed a fluent counting task without repetition.
- **Dataset 5:** single slice acquisition where subject 2 repeated the phrase “za-na-za”.
• **Dataset 6**: single slice acquisition where subject 1 performed a fluent counting task without repetition.

• **Dataset 7**: single slice acquisition where subject 1 repeated the phrase “za-na-za”.

• **Dataset 8**: single slice acquisition where subject 2 performed a fluent counting task without repetition.

**Region of interest (ROI) time profile analysis:**

For the task of production of consonant and vowel sounds while producing the phrase /loo: lee: la: za: na: za:/, we performed a region of interest (ROI) analysis to analyze the constrictions of the vocal tract air space [33]. We considered three different ROIs along the vocal tract in the midsagittal reconstruction of the 3-slice acquisition from one representative subject. These were at the landmarks of a) apex (tip) of the tongue; b) lower lip; and c) velum. A mask was drawn to enclose each of these ROIs and kept fixed for all the reconstructed image frames in the dataset. We then calculated the mean of the pixel intensities within the ROI for each image time frame, and visualized it as a one-dimensional time varying signal. These ROI time profiles were compared amongst the low-rank, temporal finite difference, and manifold regularized reconstructions.

**RESULTS**

Figure 2 shows a representative frame of the manifold reconstruction obtained from a volunteer sustaining the /a/ sound obtained from the 13 arm UDS and 27 arm VDS schemes. Note that the 13 arm UDS scheme is the current widely used scheme at 1.5 Tesla. However, when implemented at 3 Tesla, we observed significant off-resonance induced artifacts with the ~2.69 ms long readouts. These were seen as signal loss, and blurring at air tissue boundaries especially at the velum, lips, and apex (tip) of the tongue (see arrows in Fig. 2). In contrast, the short readout duration of VDS scheme (~1.3 ms long readouts) offered robustness to these artifacts, and provided much clearer depiction of key articulators and air-tissue boundaries.

Figures 4 and 5 shows the Laplacian matrix respectively for the speaking tasks of: repetition of the phrase of /loo: lee: la: za: na: za:/, and fluent task of counting numbers. Representative rows of the L matrix are shown which corresponded to two postures of the vocal tract anatomy, where the tongue was raised, and the tongue was lowered. For better visualization, we highlighted entries which were above 10% of the maximum entry in each row by the red color, and superimposed onto the remaining entries in blue color. For the repetitive task in figure 4, the rows of the Laplacian matrix revealed large entries that occurred in a quasi-periodic manner. For the fluent task in figure 5, the rows of the L matrix
showed large entries that occurred less frequently, and arbitrarily in time. Note, the peaks in each row of the L matrix pointed to frames sharing similar vocal tract anatomy postures, even though they occurred distantly along the time dimension. This may be viewed as a soft weighted non-local temporal regularization operation.

Figure 6 shows the temporal bases (V) or the eigen vectors of the L matrix, and the estimated spatial coefficients (U) for the two speech tasks. We noted the temporal bases to be adaptive in nature, and captured the underlying motion dynamics of the speaking task. For the repetitive task of producing the phrase “za-na-za”, we observed clear quasi-periodic dynamics in the bases. While for the fluent counting task, we observed more arbitrary motion patterns corresponding to free speech.

Figure 7 shows a representative spatial frame and image temporal profile from all the reconstruction schemes for the fluent speech task of counting numbers indefinitely acquired using the concurrent 3-slice acquisition scheme. Reconstructions were performed using 3 spiral arms/frame, which corresponded to ~52.2 ms/frame. We qualitatively observed the inverse NUFFT-based reconstructions to depict strong aliasing artifacts. The low rank approach removes aliasing artifacts to an extent, however it demonstrated significant spatial and temporal blurring that hampered the visualization of articulatory boundary movements. Reconstruction with temporal finite difference regularization completely removed aliasing but demonstrated temporal stair casing artifacts. This manifested as motion blurring, especially between frames where articulators changed their position (eg. raising of the tongue tip, opening of lips). The manifold reconstruction showed faint aliasing artifacts. However, it demonstrated minimal spatial or motion blurring, and provided acceptable fidelity, capturing the arbitrary motion dynamics of the articulators.

Figure 8 shows qualitative comparisons of all the reconstruction algorithms from the 3-slice concurrent acquisition for the speech task of repeating the phrase “loo-lee-la-za-na-za”. Similar to figure 7, we observed, in comparison to the competing schemes, the proposed manifold regularization scheme captured articulatory motion with minimal spatial and temporal blurring. This is evident by the sharper image temporal profiles of the manifold regularization approach compared to the competing schemes. For example, the event of lowering the tongue tip from alveolar ridge while transitioning from the sound /n/ to /a/ was significantly blurred or artifacted with the NUFFT, low rank and temporal finite difference approaches. While it was robustly captured with the manifold regularization scheme.

Figure 9 shows three ROI-averaged time profiles from the mid sagittal reconstructed image of the 3-slice acquisition during the production of the sound /a/-/n/-/a/-/z/-/a/-/l/ for all the reconstruction algorithms. The reconstructed image frames for these methods are also shown for every
five frames. We observed poor depiction of articulatory motion in the low rank scheme. The temporal finite difference scheme showed reduced artifact level, but showed significant blurring at tongue tip, lips, and the velum boundaries. The manifold reconstruction produced clean depiction of the underlying motion patterns. For example, the tongue apex (tip) in ROI 1 and the lower lips in ROI 2 should be in the raised position near the hard palate at the beginning of the sounds /n/, /z/ and /l/ which is marked by a sharp increase of ROI time profile. This behavior is represented well in the manifold reconstruction but not in the low rank and temporal finite difference schemes. Similarly, the area between velum and airway in the ROI 3 changes very little for the sound /a/ which is depicted as flat lines in the ROI-averaged time profiles. However, unlike the pronunciation of sound /z/, pronunciation of sound /n/ involves the velum moving upward towards the nasal cavity, hence there is slight decrease in the area in ROI 3. During sound /z/, the tongue near the velum moves upwards, causing increased area in ROI 3. The manifold scheme successfully captured these articulatory patterns, but the low rank and temporal finite difference methods were not similarly successful.

Finally, reconstructions from all the algorithms for all the datasets in a video format are provided in the supplementary material.

Figure 10 shows the combined image quality expert ratings. Scores are displayed as a violin plot for each of the reconstruction schemes and for each categories. The median is highlighted by a circle and the interquartile range is shown as black vertical rectangle. The density of the violin plot at a particular score is proportional to the number of times that score was assigned. We also specify this number in the plots. We note consistent scores distributed in the 3’s (i.e, good) and 4’s (excellent) in the spatial blurring and temporal blurring categories for the manifold scheme compared to the other schemes. This consistent score distribution is also reflected when we compared the (mean ± standard deviation) across all the scored datasets. In the spatial blurring category: the manifold scheme produced a score of (3.58 ± 0.57) vs. low rank (1.79 ± 0.50) and temporal finite difference (2.92 ± 0.76) schemes. In the temporal blurring category, the manifold scheme produced a score of (3.67 ± 0.47) vs low rank (1.5 ± 0.64) and temporal finite difference (3.17 ± 0.80) schemes. In the aliasing artifacts category, the manifold scheme produced a score of (3.21 ± 0.56) vs low rank (1.83 ± 0.47) and temporal finite difference (3.13 ± 1.01) schemes. We noted that in the aliasing artifacts category, the manifold scheme had a smaller median of 3 compared to median of 3.5 for the temporal finite difference scheme, but had a smaller inter-quartile range. While the raters agreed in most categories, we note that rater 2 disagreed with the other raters on the aliasing artifact scores for the temporal finite difference reconstructions, where rater 2 considered spatial pixelations across air-tissue boundaries in the alias artifact category.
Finally, we report all the individual scores for each dataset from each of the raters in the supplementary material.

DISCUSSION

We demonstrated feasibility of rapid dynamic speech imaging at 3 Tesla by synergistically combining parallel imaging capabilities of a custom vocal tract coil, self-navigation capabilities of short readout variable density spirals, and implicit motion binning of manifold regularization. Our scheme achieved a temporal resolution of \(~17.4\) ms/frame for single slice and \(~52.2\) ms/frame for concurrent three slice imaging at a spatial resolution of 2.4 mm\(^2\)/pixel. Our 27 arm VDS scheme employed a \(~1.3\) ms readout compared to the widely used 13 arm UDS scheme at 1.5 T, which had a 2.69 ms readout length. We observed significantly fewer off-resonance induced artifacts at air-tissue boundaries along the vocal tract with the 27 arm VDS scheme compared to the 13 arm UDS scheme. We demonstrated improved reconstruction quality with our approach compared to competing algorithms such as low-rank, and temporal finite difference schemes for the tasks of fluent speaking, and repetitive speaking. We attribute these improvements to the non-local soft weighted temporal regularization enabled by the proposed approach. We analyzed the mechanics of the Laplacian matrix used in the manifold regularization which revealed implicit motion binning enabled by our approach. Through blinded image quality evaluation from three experts in voice production and modeling research, we observed the proposed manifold regularization approach to outperform the inverse nuFFT, low rank, and temporal finite difference regularization schemes in the categories of spatial blurring and temporal blurring. The proposed scheme showed faint alias artifacts, but had a smaller interquartile range of scores compared to other methods in alias artifact category.

This feasibility study has several noteworthy limitations. Firstly, we did not simultaneously acquire audio. Concurrent audio acquisition is typically needed to analyze timing events during speech production, and we plan to acquire concurrent audio in future studies. Second, our sequence had a low time duty cycle despite using short readout spirals: 4.4 ms of the 5.8 ms TR was spent in slice-selective excitation and gradient spoiling. In the future, we will explore time-optimized slice select excitation pulse design to improve the duty cycle. Third, we acquired slices only in the sagittal orientation. In the future, we will explore concurrent analysis of slices in arbitrary orientations to analyze vocal tract shaping in flexible planes. Finally, we performed our analysis on a small set of subjects (n=2) performing fluent counting and repetitive production of consonant and vowel sounds. Further analysis with regards
to multiple subjects performing a comprehensive set of speaking tasks are needed to fully generalize our approach to arbitrary speech.

Coil element numbers 11,12,13 captured significant artifact energy, which originated from portions of anatomy far outside the typical upper and infra glottic airway FOV of interest. Recently, this was shown to be due to a combination of gradient non-linearity, spiral sampling, and subject’s anatomy [34]. An approach to mitigate this artifact was also proposed. In the future we will explore this approach to effectively include measurements from all of our coil elements.

The proposed manifold regularization scheme can be improved in a number of ways. First, our implementation was not optimized for reconstruction speed. Future work will include leveraging emerging Python libraries and efficient code optimization on the GPU. Second, we consistently observed faint alias artifacts with our scheme. This is attributed to using a $l_2$ norm regularization on the L matrix, which has a well-known limitation of its inability to fully penalize artifact energy. To address this, we will explore the use of a $l_1$ penalty on the L matrix in the future. Third, we will explore the combined benefit of employing additional spatial regularization. Fourth, our manifold model assumed mapping of image frames as points on a low dimension manifold living in a high dimension ambient space. In the future, we will explore mapping smaller spatial patches from the image frames onto the higher dimension ambient space. This may be suited to model certain vocal tract shaping tasks where some articulators move at a rapid pace compared to the others (eg. glottic movement during beat boxing).

CONCLUSION
We proposed a novel scheme for improved rapid dynamic speech MRI at 3 Tesla. Our scheme leveraged parallel imaging capabilities of a custom 16 channel vocal tract coil, short readout duration and self-navigating capabilities of variable density spirals, and implicit motion binning capabilities of manifold regularization. We achieved single slice imaging at 17.2 ms/frame and concurrent 3-slice imaging at 52.2ms/frame at a spatial resolution of 2.4 mm². We demonstrated applicability of our scheme to image repetitive and fluent speaking tasks. Our approach showed better fidelity in capturing speech motion patterns compared to existing low rank or sparsity based reconstruction constraints.

ACKNOWLEDGMENT
This work was conducted on an MRI instrument funded by the NIH under 1S10OD025025-01.

DATA AND CODE AVAILABILITY STATEMENT
Example data and code from this work are available at: https://github.com/rushdi-rusho/manifold_speech.

REFERENCES

[1] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, “Recommendations for real-time speech MRI,” *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1. pp. 28–44, 2016, doi: 10.1002/jmri.24997.

[2] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, “Speech MRI: Morphology and function,” *Physica Medica*, vol. 30, no. 6. pp. 604–618, 2014, doi: 10.1016/j.ejmp.2014.05.001.

[3] C. Hagedorn et al., “Characterizing Articulation in Apraxic Speech Using Real-Time Magnetic Resonance Imaging,” *J. Speech, Lang. Hear. Res.*, vol. 60, no. 4, pp. 877–891, Apr. 2017, doi: 10.1044/2016_JSLHR-S-15-0112.

[4] C. M. Pelland, X. Feng, K. C. Borowitz, C. H. Meyer, and S. S. Blemker, “A Dynamic Magnetic Resonance Imaging–Based Method to Examine In Vivo Levator Veli Palatini Muscle Function During Speech,” *J. Speech, Lang. Hear. Res.*, vol. 62, no. 8, pp. 2713–2722, Aug. 2019, doi: 10.1044/2019_JSLHR-S-18-0459.

[5] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan, “Timing effects of syllable structure and stress on nasals: a real-time MRI examination,” *J. Phon.*, vol. 37, no. 1. pp. 97–110, 2009.

[6] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger, and P. A. Vuissoz, “Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers,” *Sci. Data 2021 81*, vol. 8, no. 1, pp. 1–9, Oct. 2021, doi: 10.1038/s41597-021-01041-3.

[7] C. Carignan, R. K. Shosted, M. Fu, Z. P. Liang, and B. P. Sutton, “A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French,” *J. Phon.*, vol. 50, pp. 34–51, May 2015, doi: 10.1016/J.WOCN.2015.01.001.

[8] R. K. Shosted, B. P. Sutton, and A. Benmamoun, “Using magnetic resonance to image the pharynx during Arabic speech: Static and dynamic aspects,” Accessed: Jul. 05, 2022. [Online]. Available: http://www.isca-speech.org/archive.

[9] M. Barlaz, M. Fu, J. Dubin, Z.-P. Liang, R. Shosted, and B. Sutton, “LINGUAL DIFFERENCES IN BRAZILIAN PORTUGUESE ORAL AND NASAL VOWELS: AN MRI STUDY.”
[10] F. Barbiera, A. Lo Casto, B. Murmura, G. Bortoluzzi, I. Orefice, and A. G. Gucciardo, “Dynamic Fast Imaging Employing Steady State Acquisition Magnetic Resonance Imaging of the Vocal Tract in One Overtone Male Singer: Our Preliminary Experience,” *J. Voice*, vol. 36, no. 2, pp. 170–175, Mar. 2022, doi: 10.1016/J.JVOICE.2020.05.016.

[11] M. Echternach *et al.*, “Vocal tract and register changes analysed by real-time MRI in male professional singers—a pilot study,” *http://dx.doi.org/10.1080/14015430701875653*, vol. 33, no. 2, pp. 67–73, 2009, doi: 10.1080/14015430701875653.

[12] M. Proctor, E. Bresch, D. Byrd, K. Nayak, and S. Narayanan, “Paralinguistic mechanisms of production in human ‘beatboxing’: a real-time magnetic resonance imaging study,” *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 1043–54, 2013, doi: 10.1121/1.4773865.

[13] T. Greer, R. Blaylock, N. Patil, and S. S. Narayanan, “How beatboxers produce percussion sounds: A real-time magnetic resonance imaging investigation,” *J. Acoust. Soc. Am.*, vol. 144, no. 3, p. 1827, Oct. 2018, doi: 10.1121/1.5068052.

[14] P. W. Iltis, J. Frahm, D. Voit, A. A. Joseph, E. Schoonderwaldt, and E. Altenmüller, “High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players,” *Quant. Imaging Med. Surg.*, vol. 5, no. 3, pp. 374–81, 2015, doi: 10.3978/j.issn.2223-4292.2015.03.02.

[15] S. G. Lingala, Y. Zhu, Y. C. Kim, A. Toutios, S. Narayanan, and K. S. Nayak, “A fast and flexible MRI system for the study of dynamic vocal tract shaping,” *Magn. Reson. Med.*, vol. 77, no. 1, pp. 112–125, 2017, doi: 10.1002/mrm.26090.

[16] A. Niebergall *et al.*, “Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction,” *Magn. Reson. Med.*, vol. 69, no. 2, pp. 477–485, 2013.

[17] M. Burdumy *et al.*, “Acceleration of MRI of the vocal tract provides additional insight into articulator modifications,” *J. Magn. Reson. Imaging*, vol. 42, no. 4, pp. 925–935, 2015, doi: 10.1002/jmri.24857.

[18] X. Feng, S. S. Blemker, J. Inouye, C. M. Pelland, L. Zhao, and C. H. Meyer, “Assessment of velopharyngeal function with dual-planar high-resolution real-time spiral dynamic MRI,” *Magn. Reson. Med.*, vol. 80, no. 4, pp. 1467–1474, Oct. 2018, doi: 10.1002/MRM.27139.
X. Feng, Z. Wang, and C. H. Meyer, “Real-time dynamic vocal tract imaging using an accelerated spiral GRE sequence and low rank plus sparse reconstruction,” *Magn. Reson. Imaging*, vol. 80, p. 106, Jul. 2021, doi: 10.1016/J.MRI.2021.04.016.

M. Fu et al., “High-resolution dynamic speech imaging with joint low-rank and sparsity constraints,” *Magn. Reson. Med.*, 2014.

M. Fu et al., “High-frame-rate full-vocal-tract 3D dynamic speech imaging,” *Magnetic Resonance in Medicine*, 2016.

L. Feng, L. Axel, H. Chandarana, K. T. Block, D. K. Sodickson, and R. Otazo, “XD-GRASP: Golden-angle radial MRI with reconstruction of extra motion-state dimensions using compressed sensing,” *Magn. Reson. Med.*, 2016, doi: 10.1002/mrm.25665.

A. G. Christodoulou et al., “Magnetic resonance multitasking for motion-resolved quantitative cardiovascular imaging,” *Nat. Biomed. Eng.*, 2018, doi: 10.1038/s41551-018-0217-y.

J. Y. Cheng et al., “Comprehensive Multi-Dimensional MRI for the Simultaneous Assessment of Cardiopulmonary Anatomy and Physiology,” *Sci. Rep.*, 2017, doi: 10.1038/s41598-017-04676-8.

S. Poddar, S. Member, M. Jacob, and S. Member, “Dynamic MRI using SmooThness Regularization on Manifolds (SToRM),” vol. 0062, no. 2, pp. 1–11, 2015, doi: 10.1109/TMI.2015.2509245.

A. H. Ahmed et al., “Free-breathing and ungated cardiac cine using navigator-less spiral SToRM,” Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.05542.

S. Poddar, Y. Q. Mohsin, D. Ansah, B. Thattaliyath, R. Ashwath, and M. Jacob, “Manifold Recovery Using Kernel Low-Rank Regularization: Application to Dynamic Imaging,” *IEEE Trans. Comput. Imaging*, vol. 5, no. 3, pp. 478–491, 2019, doi: 10.1109/tci.2019.2893598.

M. Lustig, S. J. Kim, and J. M. Pauly, “A fast method for designing time-optimal gradient waveforms for arbitrary k-space trajectories,” *IEEE Trans. Med. Imaging*, vol. 27, no. 6, pp. 866–873, Jun. 2008, doi: 10.1109/TMI.2008.922699.

S. Poddar and M. Jacob, “Dynamic MRI Using SmooThness Regularization on Manifolds (SToRM),” *IEEE Trans. Med. Imaging*, 2016, doi: 10.1109/TMI.2015.2509245.

A. H. Ahmed, R. Zhou, Y. Yang, P. Nagpal, M. Salerno, and M. Jacob, “Free-Breathing and Ungated Dynamic MRI Using Navigator-Less Spiral SToRM,” *IEEE Trans. Med. Imaging*, 2020, doi:
10.1109/tmi.2020.3008329.

[31] S. D. Knoll F, Schwarzl A, Diwoky C, “gpuNUFFT: an Open-Source GPU Library for 3D Gridding with Direct MATLAB Interface • Center for Advanced Imaging Innovation and Research,” Proc Intl Soc Mag Reson Med, 2014. https://cai2r.net/resources/gpunufft-an-open-source-gpu-library-for-3d-gridding-with-direct-matlab-interface/ (accessed Aug. 19, 2022).

[32] S. G. Lingala, E. DiBella, G. Adluru, C. McGann, and M. Jacob, “Accelerating free breathing myocardial perfusion MRI using multi coil radial k-t SLR.,” Phys. Med. Biol., vol. 58, no. 20, pp. 7309–27, 2013, doi: 10.1088/0031-9155/58/20/7309.

[33] A. Lammert, V. Ramanarayanan, M. Proctor, and S. Narayanan, “Vocal Tract Cross-Distance Estimation from Real-Time MRI using Region-of-Interest Analysis,” INTERSPEECH, pp. 959–962, 2013, Accessed: Aug. 23, 2022. [Online]. Available: https://www.iscaspeech.org/archive_v0/archive_papers/interspeech_2013/i13_0959.pdf.

[34] Y. Tian, Y. Lim, Z. Zhao, D. Byrd, S. Narayanan, and K. S. Nayak, “Aliasing artifact reduction in spiral real-time MRI,” Magn. Reson. Med., vol. 86, no. 2, pp. 916–925, Aug. 2021, doi: 10.1002/MRM.28746.
**FIGURES:**

*Figure 1:* The 16-channel custom vocal tract receiver coil. The coil consisted of three pieces mounted on flexible arms. The left and right cheek piece contained 5 elements each. The third neck and chin piece consisted of 6 elements. The flexible mounts allowed for close conformity of the coil elements to the subject’s face and neck. Also shown are the 16 individual coil reconstructed images generated via inverse NUFFT reconstruction using 27 spiral arms from the VDS acquisition. Note, the elements provide sensitivity to several upper airway and infra glottic airway structures (eg. lips, velum, tongue, pharynx, glottis), relevant in speech production.
Figure 2: Specifications of the 13 arm UDS (left column) and the 27 arm VDS (right column) design. Top row shows FOV v/s normalized k-space radius. Middle row shows the generated sampling trajectories, where the UDS trajectories had a readout duration of 2.69 ms, and VDS trajectories had a readout duration of 1.3 ms. Manifold regularized reconstructions using 13 arms/frame and 27 arms/frame respectively for UDS and VDS schemes are shown in the bottom row. The 13 arm UDS scheme depicted substantial off-resonance induced blurring and signal loss at several key air-tissue interfaces. In contrast, the 27 arm VDS scheme was less sensitive to these artifacts and had a clearer depiction of articulators (e.g. see the yellow arrows pointing to the velum, and the green arrows pointing to the lips).
Figure 3: Dynamic images can be modeled as points on a smooth nonlinear manifold embedded in a high dimensional ambient space. This is demonstrated in this schematic using image frames from a fluent speaking task. Image frames sharing similar vocal tract postures are mapped as neighbors on the 2D manifold even if they occur at different times (see red and green squares). Dissimilar images are distant on the 2D manifold even if they occur consecutively in time. Weights between frames determine the degree of similarity and are inversely proportional to the distance of the points on the manifold. The manifold regularization exploits the neighborhood relations of points on this manifold using a penalized optimization framework.

Figure 4: Visualizing the graph Laplacian matrix for the speaking task of repeating the phrase “loo-lee-la-za-na-za”. Representative midsagittal reconstruction from volunteer 1 imaged with the 3-slice
acquisition scheme was considered. (a) shows the estimated Laplacian matrix. (b) shows a representative row (row # 74) of this matrix, where entries that are greater than 10% of the maximum (or) minimum value of that row are highlighted in red color, and superimposed on the remaining entries in blue color. The 74th frame of this sequence had a raised tongue posture as seen in the highlighted red box in (b). While reconstructing this frame, similarity between frames corresponding to peaks in the 74th row of the L matrix are implicitly leveraged. These frames are shown in (b), and clearly depict the tongue raised posture. Similarly, (c) shows a representative row (row # 268) of this matrix, which corresponds to the 268th frame depicting a lowered tongue posture (see green box in (c)). Also, note how the peaks in (c) corresponded to image frames sharing the lowered tongue posture. We also note quasi-periodic peaks in the rows representative of the repetitive nature of the speaking task.

Figure 5: Visualizing the graph Laplacian matrix for the speaking task of fluently counting numbers indefinitely. Representative midsagittal reconstruction from volunteer 1 imaged with the single-slice acquisition scheme was considered. (a) shows the estimated Laplacian matrix. (b) shows a representative row (row # 60) of this matrix, where entries that are greater than 10% of the maximum (or) minimum value of that row are highlighted in red color, and superimposed on the remaining entries in blue color. The 64th frame of this sequence had a raised tongue posture as seen in the highlighted red box in (b). Similar to figure 4, we note similar frames that occur non-locally are identified by peaks of this row. Similarly, (c) shows a representative row (row # 118) of this matrix, which corresponds to the
118th frame depicting a lowered tongue posture (see green box in (c)), and note the similar frames sharing this posture being highlighted by the peaks. In contrast to Figure 4, we note the peaks in this task were arbitrary, which is representative of the fluent counting speaking task.

Figure 6: Illustration of temporal bases (V) or the eigen vectors of the Laplacian matrix; and the spatial coefficients (U) for the speaking tasks of fluently counting numbers and repeating the phrase za-na-za. For brevity, the first 5 bases and coefficients are shown. Note quasi periodic patterns are captured in the temporal bases of the repetitive za-na-za task, and more arbitrary patterns are captured in the temporal bases of the fluent speaking task.
Figure 7: Qualitative comparison of reconstructions from the inuFFT, low rank regularization, temporal finite difference regularization, and the proposed manifold regularization schemes. Reconstructions were performed from the concurrent 3-slice acquisition scheme with 3arms/frame, and a native time resolution of ~52.2 ms/frame. The speaker was producing the fluent task of counting numbers indefinitely. Shown are one frame of the reconstruction along with the temporal profile cuts (indicated by the white horizontal line). As expected, significant alias artifacts were observed in inuFFT. Motion blurring and residual alias artifacts were present in the low-rank scheme. Temporal finite difference scheme showed no aliasing artifacts, but had considerable inter-frame motion artifacts, and temporal stair casing artifacts in the temporal profiles (see arrows). The manifold scheme provided superior
spatial and temporal fidelity compared to competing algorithms as evident by crisper temporal profiles (see arrows).

Figure 8: Qualitative comparison of reconstructions from the inuFFT, low rank regularization, temporal finite difference regularization, and the proposed manifold regularization schemes for the task of repetitively producing the phrase “loo-lee-la-za-na-za”. The vertical yellow dotted lines indicate the timing of the various consonant and vowel sounds. Reconstructions were performed from the concurrent 3-slice acquisition scheme with 3arms/frame, and a native time resolution of ~52.2
ms/frame. Shown are one frame of the reconstruction along with the temporal profile cuts (indicated by the white horizontal line). Similar to figure 7, we observed alias artifacts in inuFFT; motion blurring and remaining alias artifacts in low-rank scheme; and temporal stair-casing artifacts in the temporal finite difference scheme. The proposed manifold regularized scheme showed improved reconstruction quality with better motion fidelity, as seen by sharper image time profiles (also see blue arrows).

**Figure 9:** Region of interest (ROI) time profile analysis on the mid-sagittal reconstructions from the concurrent 3-slice dataset for the speaking task of producing the phrase “lloo-lee-za-na-za”. We particularly zoom into image frames producing the sounds /a\-\n/\-\z/\-\a\-\l/. Three different ROIs were considered: airway near the tongue tip (ROI1), airway near the lower lip (ROI2), and airway behind the tongue and in front of the velum (ROI3). Mean pixel intensities in each of these ROIs are plotted as a function of time. Also shown are the reconstructed time frames at an interval of five frames. The tongue tip in ROI 1 and the lower lips in ROI 2 should be in the raised position at the beginning of the sounds /n/, /z/ and /l/ which should be reflected as a sharp increase of ROI time profile. This behavior is represented well in the manifold reconstruction but not in the low rank and temporal finite difference schemes (see black arrows). Similarly, the area between velum and airway in the ROI 3 should not change much for the sound /a/ which should be depicted as flat lines in the ROI-averaged time profiles,
and is represented well in the manifold scheme compared to temporal finite difference scheme (see magenta arrow).

*Figure 10:* Combined image quality scores from three expert raters. The score distribution across the 8 datasets from the 3 raters are shown as violin plots for the categories of alias artifacts (Fig. 10 A), spatial blurring (Fig. 10 B), and temporal blurring (Fig. 10 C). The median of the scores is indicated by the white circle, and the interquartile range is indicated by the black vertical box. The density of the violin plot at a particular score is proportional to the number of times that score was assigned (also listed as a number in the plot). We observed that the proposed manifold scheme consistently provided scores in the 3’s (good quality) and 4’s (excellent quality) for the spatial blurring and temporal blurring categories in comparison to the other schemes. In the alias artifacts category, the proposed scheme had a lower
median score compared to temporal finite difference scheme (median score of 3 v.s median score of 3.5). However, had a smaller inter-quartile range.

VIDEO CAPTIONS (SUPPLEMENTARY MATERIAL):

- **Video 1**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 1, which was a three-slice acquisition where subject 1 performed a fluent counting task without repetition.
- **Video 2**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 2, which was a three-slice acquisition where subject 1 repeated the phrase “loo-lee-la-za-na-za”.
- **Video 3**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 3, which was a three-slice acquisition where subject 2 repeated the phrase “loo-lee-la-za-na-za”.
- **Video 4**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 4, which was a three-slice acquisition where subject 2 performed a fluent counting task without repetition.
- **Video 5**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 5, which was a single slice acquisition where subject 2 repeated the phrase “za-na-za”.
- **Video 6**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 6, which was a single slice acquisition where subject 1 performed a fluent counting task without repetition.
- **Video 7**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 7, which was a single slice acquisition where subject 1 repeated the phrase “za-na-za”.
- **Video 8**: Reconstructions from the (1st row) inuFFT, (2nd row) low rank regularized, (3rd row) temporal finite difference regularized, and (4th row) manifold regularized schemes for Dataset 8,
which was a single slice acquisition where subject 2 performed a fluent counting task without repetition.