ACCELERATED 3D MRI OF VOCAL TRACT SHAPING USING COMPRESSED SENSING AND PARALLEL IMAGING

Yoon-Chul Kim, Shrikanth S. Narayanan, and Krishna S. Nayak

Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA, USA

ABSTRACT

3D MRI of the upper airway has provided valuable insights into vocal tract shaping and data for the modeling of speech production. Small movements of articulators can lead to large changes in the produced sound, therefore improving the resolution of these datasets, within the constraints of a sustained sound (6-12 seconds), is an important area for investigation. This paper provides the first application of compressed sensing (CS) with parallel imaging to high-resolution 3D upper airway MRI. We use spatial finite difference as the sparsifying transform, and investigate the use of high-resolution phase information as a constraint during CS reconstruction. In a retrospective sub-sampling experiment with no sound production, 5x undersampling produced acceptable image quality when using phase-constrained CS reconstruction. The prospective use of this accelerated acquisition enabled 3D vocal-tract MRI during sustained production of English /s/,/ʃ/,/i/,/r/ with 1.33x1.33x1.33-mm^3 spatial resolution and 10-seconds of scan time.

Index Terms— speech production, compressed sensing MRI, vocal tract shaping, sensitivity encoding, phase constraint.

1. INTRODUCTION

Three-dimensional (3D) imaging of the upper airway during sustained sound production has recently emerged as a promising tool in speech production research as a means to capture the full geometry of the vocal tract. The diversity of tongue shapes and dynamics are made possible, at least in part, through different lingua-palatal bracing mechanisms [1-3] leading to complex airway geometries, the understanding of which is critical for investigations into the production of both normal and disordered speech. In addition to helping shed light on the intricate airway shaping mechanisms underlying the production of various linguistically-meaningful speech sounds, 3D imaging also lends itself to providing quantitative volumetric information of the airway regions. The shaping of the tongue and other articulators, and the temporal characteristics of their shaping, give rise to characteristic patterns of acoustic resonance behavior of the vocal tract that define the properties of human speech that can be modeled with such quantitative information.

Recent work has shown that three-dimensional tongue shape and the dynamics underlying shape formation are critical to understanding natural linguistic classes and issues of phonological representation as evidenced in speech motor control. Speech studies using MRI have focused on vowel, fricative, lateral sounds. 3D data were helpful in deriving meaningful acoustic models for these sounds.

These previous MRI studies were based on 2D multi-slice acquisitions, requiring multiple repetitions of the same sound and scan-times on the order of several minutes. These procedures are prone to data inconsistency, resulting from slightly different positions of the jaw, head, and tongue during each repetition. Compared to 2D multi-slice, 3D encoding provides superior contiguous coverage with the potential for thinner slices and improved signal-to-noise ratio (SNR) efficiency. 3D encoding with high spatial resolution currently requires prohibitively long scan times that exceed the normal duration of sustained sound production with minimal subject motion.

3D MRI scans may be accelerated by using time-efficient sampling [4], undersampling of k-space and advanced reconstruction methods that remove aliasing artifact. Rapid acquisition schemes based on spiral or echo-planar trajectories are prone to severe blurring artifacts or geometric distortions near the air-tissue interface in reconstructed images. Undersampled 3DFT acquisitions are relatively insensitive to off-resonance, but suffer severe aliasing artifacts in reconstructed images. Parallel imaging can effectively remove aliasing artifacts by the use of appropriate receiver coil arrays [5,6]. Compressed sensing (CS) reconstruction only exploits sparsity of the final reconstructed image in a transform domain [7-9].

We recently applied CS-MRI to 3D imaging of the upper-airway using a single-channel receiver coil and achieved 3x acceleration without substantial artifact [10]. In a follow-up study, we applied additional phase constraints, originally proposed by Lustig et al. [9], and achieved 5x acceleration without substantial artifact [11], by incorporating "high-resolution" phase information that was derived from a non-phase-constrained CS reconstruction. Our working hypothesis is that high-resolution phase information is important because air-tissue boundaries are the primary features of interest and experience the most substantial phase variation due to air-tissue susceptibility.

In this paper, we extend phase-constrained CS (PC-CS) reconstruction to imaging with multi-channel receiver coil arrays (parallel imaging). This combined use of compressed sensing and parallel imaging has been recently proposed by several groups [15,16], but the notion of incorporating high-resolution phase information is unique to this work. We present a two-stage reconstruction approach that first estimates phase maps for each coil element via conventional CS reconstructions, and then reconstructs the final image iteratively after incorporating the high-resolution phase maps and low-resolution magnitude coil sensitivity maps into a multi-coil CS reconstruction. This approach

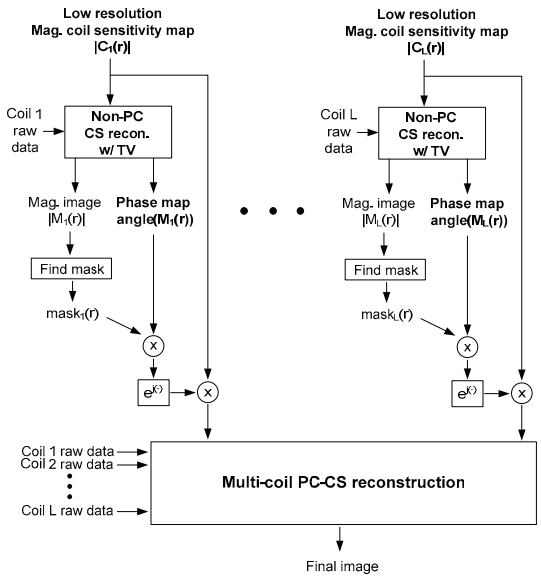

Fig. 1. Flowchart of the proposed reconstruction scheme.

is prospectively used to image the vocal tract during the production of sustained American English speech sounds with an acceleration factor of 5x.

2. METHODS

2.1 Data acquisition

Experiments were performed on a 3.0 T Signa Excite HD MRI scanner (GE Healthcare, Waukesha, WI) with gradients capable of 40mT/m amplitudes and 150mT/m/ms slew rates. The receiver bandwidth was set to ±125 kHz (4 μs sampling rate). The body coil was used for RF transmission, and 8-channel neurovascular array coil was used for signal reception (only 4 superior elements were used for reconstruction). The vocal tract region of interest (ROI) was imaged using a single thick midsagittal slab with 8 cm thickness in the right-left (R-L) direction. The readout direction was superior-inferior (S-I) and the phase encode directions were anterior-posterior (A-P) and right-left (R-L). A gradient echo (GRE) sequence was used with TE=2.3msec, TR=4.7msec, flip angle=10°, NEX=1, spatial resolution=1.33x1.33x1.33mm^3, and FOV=20x24x8cm^3.

Pseudo-random undersampling was implemented as follows: First, two independent and uniformly distributed random numbers corresponding to k-space radius and azimuthal angle were each generated to create pseudo random (k_y, k_z) location in polar form. From the randomly chosen samples, nearest (k_y, k_z) grids were selected as sampling points. Second, a low spatial frequency region was fully sampled. The outermost k-space radius of the fully sampled region was chosen to be 20% of the full k-space radius.

2.2 In-vivo experiments

Subjects were oriented in the supine position and their heads were immobilized by inserting foam pads between their ears and the receiver coil. A fully sampled dataset, without sound production, was acquired when one trained subject held the mouth open for 51 seconds, without swallowing. A total of 10800 (k_y, k_z) encodes, where the number of k_y and k_z encodes was 180 and 60, respectively, fully covered 3D k-space at the Nyquist rate.

The prospective accelerated acquisition was performed by imaging the vocal tract shaping during sustained sound production of American English /s/,/ʃ/,/i/,/r/. Scan time for each 5x accelerated acquisition took 10 seconds.

2.3 Image Reconstruction

Since all datasets were fully sampled along the readout (k_x) direction, data were first inverse-Fourier transformed along k_x. For each x position, fully sampled datasets were reconstructed using 2D inverse Fourier transform (IFT). For the retrospective and prospective undersampled acquisitions, un-acquired k-space locations were filled with zeros prior to IFT. IFT images from all four coil elements were root-sum-of-squared (RSS) to produce final image. For comparison, conventional conjugate-gradient-based un-regularized SENSE reconstruction was implemented based on the work of Pruessmann et al. [12].

The multi-coil PC-CS reconstruction is illustrated in Fig. 1. First, the phase map for the l^{th} coil element was calculated by taking the phase of the complex-valued image estimate $\hat{\mathbf{m}}$ obtained from a conventional CS reconstruction that is based on the following unconstrained convex optimization:

$$\min_{\mathbf{m}} \left\| \mathbf{s}_l - \Phi C_l \mathbf{m} \right\|_2^2 + \lambda \left\| \Psi \mathbf{m} \right\|_1, \qquad (1)$$

where $\left\| \cdot \right\|_p$ denotes the l_p-norm, $\mathbf{s}_l$ is the measured undersampled k-space data of the l^{th} coil element, Φ is the Fourier encoding matrix, and C_l is the diagonal matrix consisting of the magnitude coil sensitivity of the l^{th} coil element. Ψ is a sparsifying transform, e.g., wavelets, curvelets, or finite difference. In this work, we adopted the finite difference sparsifier that contains the horizontal and vertical gradients of the image. The l_1-norm of the finite difference of the solution is also known as Total Variation (TV) [13]. λ is a regularization parameter that controls the relative weight of sparsity and data fitting. Its value was chosen after visual inspection of reconstructed images representing a broad range of λ values. The phase estimate for the l^{th} coil element was masked by mask$_l$(r) (see Fig. 1) to contain only spatial locations where the magnitude image was greater than 20% of its maximum value.

Given the estimates of the phase and magnitude coil sensitivity maps, multi-coil PC-CS reconstruction was performed by solving the following optimization:

$$\min_{\mathbf{m}} \sum_{l=1}^{L} \left\| \mathbf{s}_l - \Phi P_l C_l \mathbf{m} \right\|_2^2 + \lambda \left\| \Psi \mathbf{m} \right\|_1, \qquad (2)$$

where P_l is the diagonal matrix consisting of the phase estimate from the non-PC CS reconstruction for the l^{th} coil element. The convex optimization solver adopted in this study was based on an iterative non-linear conjugate gradient algorithm [9].

2.4 Data Processing and Analysis

Vocal tract area functions were measured by 1) manually drawing the vocal tract midline on a midsagittal slice, 2) prescribing several cross-sectional slices orthogonal to the midline, from the lips to the glottis, and 3) calculating the vocal tract areas from each cross-sectional slice. All analysis was done using OsiriX software [14]. 3D visualizations of the tongue shape were constructed by manually segmenting the tongue ROI in coronal slices from each dataset, stacking the segmented slices, and generating a 3D volume rendering using the vol3d.m Matlab function (publicly available at http://www.mathworks.com).

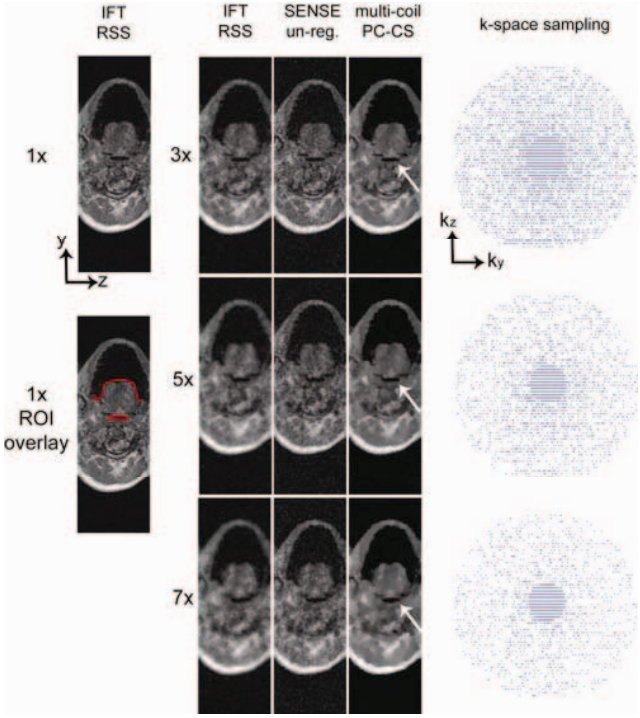

Fig. 2. Axial slice reconstructions in the retrospective experiment. The pseudo-random sampling patterns with fully sampled low-frequency region are shown on the rightmost column and were used throughout the study.

3. RESULTS

Figures 2, 3, and 4 contain experimental results obtained from one female American English speaker. Fig. 2 shows images from one axial slice extracted from 3D volume in the retrospective sub-sampling experiment. Images are reconstructed from IFT, un-regularized SENSE, and multi-coil PC-CS reconstructions of the datasets sub-sampled with different reduction factors. The IFT reconstructed images from the undersampled data exhibited incoherent aliasing artifacts and the image quality systematically degrades at higher reduction factors. Un-regularized SENSE reconstructed images show relatively noisy image quality for both 5x and 7x acceleration. The multi-coil PC-CS reconstruction exhibited the lowest overall noise level and clearest visualization of the air-tissue boundaries. Image quality degraded with higher reduction factors (see white arrows in Fig. 2).

Fig. 3 contains a midsagittal slice and cross sectional slices based on the slice prescriptions described in Fig. 3(b). Substantial signal loss in the hard palate was observed in the 7x case (see the white arrow in Fig. 3(e)). Both 5x and 7x reconstructions were not effective at resolving the small features, (e.g. see the black arrow in Fig. 3(c)). Overall, the depiction of the vocal tract areas in the 5x multi-coil PC-CS reconstruction was comparable to that of the fully-sampled reconstruction (compare Fig.3(c) and 3(d)).

Fig. 4 contains a midsagittal slice, vocal tract area function, and 3D visualization of tongue surface for /s/, /ʃ/, /i/, and /r/. Midlines (thin green lines in Fig. 4(a)) that were manually drawn have different shapes depending on the sound being produced. The midline tongue contour for /r/ was highly tortuous because of the large space from the sublingual cavity and the upward position of the tongue tip. Fig. 4(b) shows that the measured area functions are different for the different articulations. The fricative sounds /s/ and /ʃ/ have increased areas near the glottis region unlike the vowel sound /i/ because of the backward movement of the tongue root (compare the shaping of the epiglottis in /s/,/ʃ/,/i/ from the midsagittal in Fig. 4(a)). Fig. 4(c) suggests that 3D vocal tract geometry provides additional information such as the degree of the tongue grooving (compare the arrows in /s/,/ʃ/,/i/), cupping of the tongue (see the thick arrow in /r/) , and the volume of the sublingual cavity (see the thin arrow in /r/).

4. DISCUSSION

The major sources of the image phase in MRI are: 1) receiver coil phase, 2) resonance frequency offset due to field inhomogeneity and air-tissue magnetic susceptibility difference, and 3) gradient/DAQ timing delay. These may be estimated from separate calibration scans, or via self-calibration which was chosen in this study. Self-calibration is preferred because the vocal tract geometry could vary between calibration scans and accelerated imaging scans. The features of interest are air-tissue boundaries such as tongue surface, lips, hard palate, velum, and epiglottis which are coordinated for the generation of unique gestures depending on different articulation tasks. Even two separate productions of the same sound articulation could result in different vocal tract shaping.

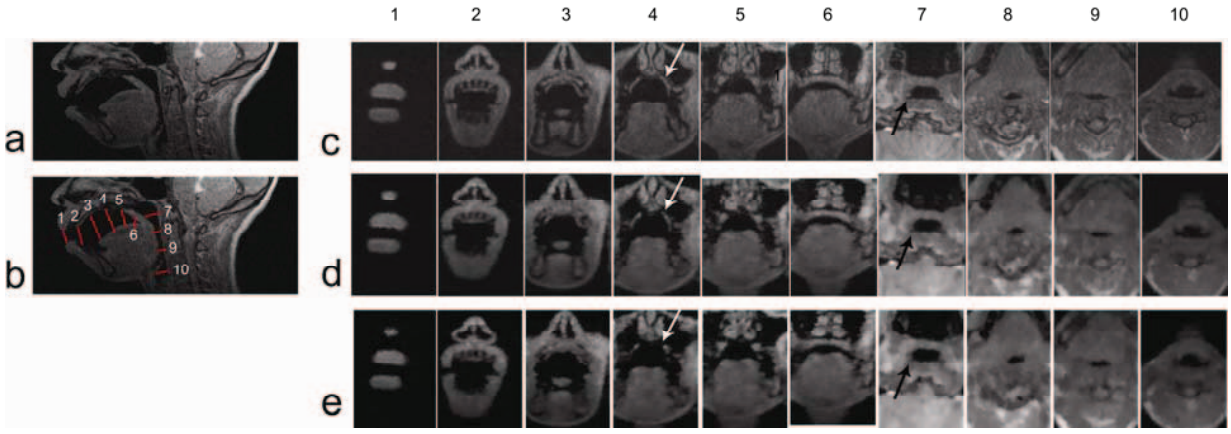

Fig. 3. Retrospective sub-sampling experiment from 3D data: **(a)** midsagittal slice, **(b)** slice prescription (red lines) overlaid onto the midsagittal image in (a). Cross sectional slices obtained via OsiriX from the prescription lines shown in (b) are ordered from 1 to 10 and are shown for **(c)** IFT-RSS reconstruction from fully sampled (1x) dataset as a ground truth, multi-coil PC-CS reconstruction from **(d)** 5x and **(e)** 7x undersampled dataset.

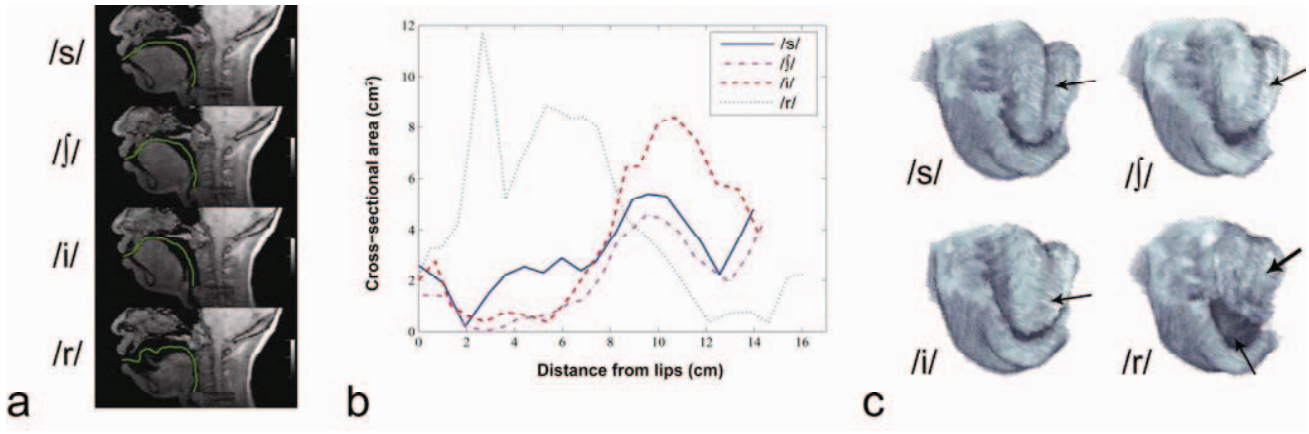

Fig. 4. The prospective use of accelerated 3D acquisition and multi-coil PC-CS reconstruction. **(a)** Reformatted midsagittal slices and their associated midlines drawn for cross-sectional slice prescription. **(b)** Area function plot. **(c)** 3D visualization of the tongue and lower jaw.

TV regularization was effective at improving the image quality of high contrast features such as the air-tissue boundaries and suppressing noise-like aliasing artifacts, and was more effective when combined with PC-CS reconstruction technique. These de-noising and edge-preserving characteristics will improve the performance of the subsequent image processing tasks. A drawback of the PC-CS method is that it requires many CS iterations, (i.e., four conventional CS reconstructions for phase estimation and one multi-coil PC-CS reconstruction).

In this study, the maximum rate of acceleration achieved without artifact (5x) was the same as that achieved in the single-coil case [11]. This could be due to lower SNR at the high spatial resolutions used in this study (1.33x1.33x1.33-mm^3 versus 1.5x1.5x2.0-mm^3 in Ref. [11]), and may be improved through an improved procedure for selecting λ or the use of a more appropriate sparsifying transform. This remains to be explored as future work.

Higher degrees of acceleration and improvements in spatial resolution are highly desirable. Fine linguistically relevant features such as the tongue tip constriction and epiglottis will be targeted in future studies. It may also be possible to measure the vocal tract area function with greater precision, therefore improving the accuracy of the quantitative analysis of vocal tract shaping in both normal and disordered speech production.

5. CONCLUSIONS

We have demonstrated a first application of compressed sensing MRI with parallel imaging to 3D imaging of vocal tract shaping during sustained sound production. A five-fold acceleration was achievable with this technique, with negligible loss of relevant tissue boundary information. Using this approach, 3D upper airway datasets could be obtained with 1.33x1.33x1.33-mm^3 spatial resolution in just 10-seconds, a duration practical for sustained sound production.

6. ACKNOWLEDGEMENTS

This work was supported by NIH Grant R01 DC007124-01. We acknowledge the support and collaboration of the Speech Production and Articulation kNowledge (SPAN) group at the University of Southern California.

7. REFERENCES

[1] T Baer, JC Gore, LC Gracco, and PW Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *J Acoust Soc Am*, 1991;90(2):799-828.

[2] SS Narayanan, AA Alwan, K Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *J Acoust Soc Am*, 1995;98(3):1325-1347.

[3] BH Story, IR Titze, "Vocal tract area functions from magnetic resonance imaging," *J Acoust Soc Am*, 1996;100(1):537-554.

[4] P Irarrazabal, DG Nishimura, "Fast three dimensional magnetic resonance imaging," *Magn Reson Med*, 1995;33:656-662.

[5] KP Pruessmann, M Weiger, MB Scheidegger, P Boesiger, "SENSE: sensitivity encoding for fast MRI," *Magn Reson Med*, 1999;42:952-962.

[6] MA Griswold, PM Jakob, RM Heidemann, M Nittka, V Jellus, J Wang, B Kiefer, A Haase, "Generalized autocalibration partially parallel acquisitions (GRAPPA)," *Magn Reson Med*, 2002;47:1202-1210.

[7] EJ Candes, J Romberg, T Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans Info Theory*, 2006;52(4):1289-1306.

[8] DL Donoho, "Compressed sensing," *IEEE Trans Info Theory*, 2006;52(4):1289-1306.

[9] M Lustig, D Donoho, JM Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn Reson Med*, 2007; 58:1182-1195.

[10] YC Kim, JF Nielsen, SS Narayanan, D Byrd, KS Nayak, "Application of compressed sensing to 3D imaging of the vocal tract for speech MRI," *Proc. ISMRM 16th Scientific Sessions,* Toronto, May 2008, p.2003.

[11] YC Kim, SS Narayanan, KS Nayak, "Accelerated three-dimensional upper airway MRI using compressed sensing," *Magn Reson Med*, 2009 (in press).

[12] KP Pruessmann, M Weiger, P Bornert, P Boesiger, "Advances in sensitivity encoding with arbitrary k-space trajectories," *Magn Reson Med*, 2001;46:638-651.

[13] LI Rudin, S Osher, E Fatemi, "Nonlinear total variation noise removal algorithm," *Physica D*, 1992;60(1-4):259-268.

[14] A Rosset, L Spadola, O Ratib, "OsiriX: an open-source software for navigating in multidimensional DICOM images," *J Digital Imaging*, 2004;(17):205-216.

[15] KF King, "Combined compressed sensing and parallel imaging," *Proc. ISMRM 16th Scientific Sessions*, Toronto, May 2008, p.1488.

[16] L Marinelli, CJ Hardy, DJ Blezek, "MRI with accelerated multi-coil compressed sensing," *Proc. ISMRM 16th Scientific Sessions*, Toronto, May 2008, p.1484.