Faster 3D Vocal Tract Real-time MRI Using Constrained Reconstruction

Yinghua Zhu [1], *Asterios Toutios* [1], *Shrikanth Narayanan* [1,2], *Krishna Nayak* [1]

[1] Department of Electrical Engineering, University of Southern California, Los Angeles, USA
[2] Department of Linguistics, University of Southern California, Los Angeles, USA

yinghuaz@usc.edu, toutios@sipi.usc.edu, shri@sipi.usc.edu, knayak@usc.edu

Abstract

Real-time magnetic resonance imaging (rtMRI) is a valuable emerging tool for studying the dynamics of vocal production. Conventional 2D rtMRI typically images the midsagittal plane of the vocal tract, acquiring data from all the important articulators. Dynamic 3D MRI would be a major advance, as it would provide 3D visualization of the vocal tract shaping dynamics, especially for the modeling of complex vocal tract geometries, such as liquid and fricative consonants, which is not easily available from 2D rtMRI. In this paper, we present an approach to directly acquire data in 3D using a highly temporally-undersampled stack-of-spirals imaging sequence, and perform reconstruction using a partially separable model. We demonstrate visualization of vocal tract dynamics from midsagittal and coronal scan planes in the articulation of fricative consonants.

Index Terms: speech production, vocal tract shaping, dynamic 3D MRI, spiral imaging, English fricative

1. Introduction

Real-time magnetic resonance imaging (rtMRI) has proven to be a useful tool for speech research, and has fundamentally changed many aspects of our understanding of human speech production [1]–[4]. Conventional rtMRI of the upper airway typically involves imaging a single midsagittal slice, covering the entire vocal tract from the lips to the glottis. However, it does not satisfy the assessment of complex vocal tract geometries, such as liquid (/r/, /l/) and fricative (/s/, /ʃ/) consonants. Recent studies of vocal tract dynamics using rtMRI have primarily involved non-concurrent data acquisition along multiple planes (e.g., midsagittal and coronal scan planes) [5]. Using those, for example, these studies have investigated groove geometry and tongue shaping of sibilant fricatives, information otherwise not available from just one scan plane.

Dynamic imaging of the whole vocal tract in 3D is desirable to complement the 2D midsagittal plane views, providing a better understanding of the speech articulation. However, current MRI systems do not meet the requirements for capturing 3D vocal tract dynamics in real-time. Multi-planar imaging (including non-MRI approaches) has served as compromise solutions [6][7], but their results rely on the quality of the synchronization/alignment of the data acquired from different scans. A more direct attempt was made in capturing the 3D vocal tract using a stack-of-spirals [8], which required at least 6X acceleration to reach the minimal useful temporal resolution (150 msec).

To achieve the minimal temporal resolution with the current state of the art experiment setups, data need to be highly undersampled. A model-based method enables reconstruction from such incomplete data based on some necessary assumptions [9]. Recently, Fu applied a partially separable (PS) model to dynamic 3D vocal tract imaging [10] using a combination of spiral trajectory navigator and Cartesian trajectory for image encoding. The technique made it possible to monitor the dynamics of the vocal tract during speech, but the 8 mm left-right resolution was not sufficient for studying the speech articulations. Furthermore, the imaging encoding could be much more efficient by using spiral readouts.

In this paper, we apply the PS model to highly temporally-undersampled stack-of-spirals imaging data, in which both navigator and image encoding are spiral sequences. We demonstrate an effective reconstruction frame rate of 8.6 frames/sec (roughly 8X acceleration), and an effective spatial resolution of 3×3×3.1 mm^3. We illustrate the 3D visualization technique in a study of sibilant fricatives.

2. Methods

2.1. MR Imaging Setup

Experiments were performed at the Los Angeles County Hospital on a 1.5 T Signa Excite HD scanner (GE Healthcare, Waukesha, WI) with gradients supporting maximum amplitude of 40 mT/m and maximum slew rate of 150 mT/m/ms. The sampling period was set to 4 us. We used a body coil for radiofrequency (RF) transmission and a custom 4-channel upper airway receiver coil array for signal reception. Data from the two anterior elements of the coil were used for image reconstruction. MR imaging was performed with a custom real-time imaging framework, providing interactive control of scan parameters [11].

Each screened subject provided informed consent in accordance with institutional policy. Each subject was scanned in supine position with the head immobilized by foam pads between the head and the receiver coil. A spiral fast gradient echo pulse sequence was used for real-time imaging [2]. Imaging parameters were: 20×20×5 cm^3 field of view (FOV), 3×3×3.1 mm^3 resolution, 6.5 ms repetition time (TR), 15° flip angle. The sequence includes 16 kz phase encoding levels with 9 spiral interleaves per level. The order of the interleaves was designed for highly undersampling the data.

2.2. Partially Separable Model

The space from which MR data is acquired is usually referred to as the k-space, in which kx and ky are defined for 2D, and kz is added for 3D. In dynamic imaging, time augments a new dimension, forming the k-t space. MRI is traditionally slow due to the hardware limitations, so it is very challenging to acquire the data in k-space—especially in 3D—while satisfying the Nyquist sampling criterion. Reconstruction that is based on some assumptions (constraints) of the undersampled k-t space data has proven to be a powerful and efficient approach to overcome the problem.

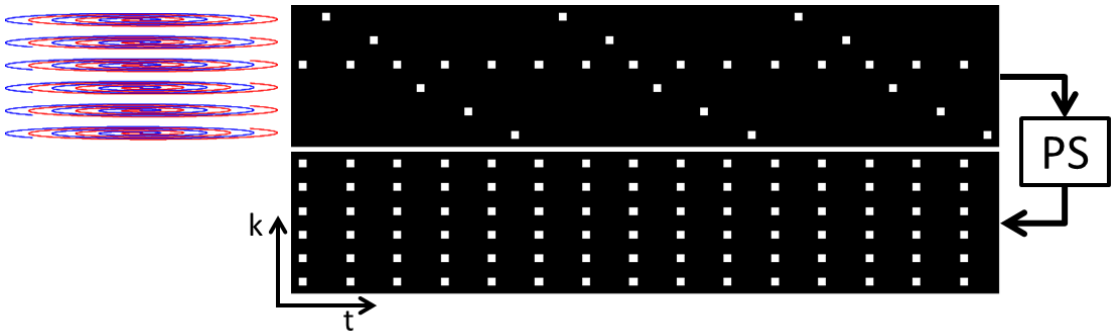


Figure 1: *Illustration of sampling pattern using a stack-of-spirals sequence with 6 kz-encoded planes, and the expected reconstructed data from PS model constrained method.*

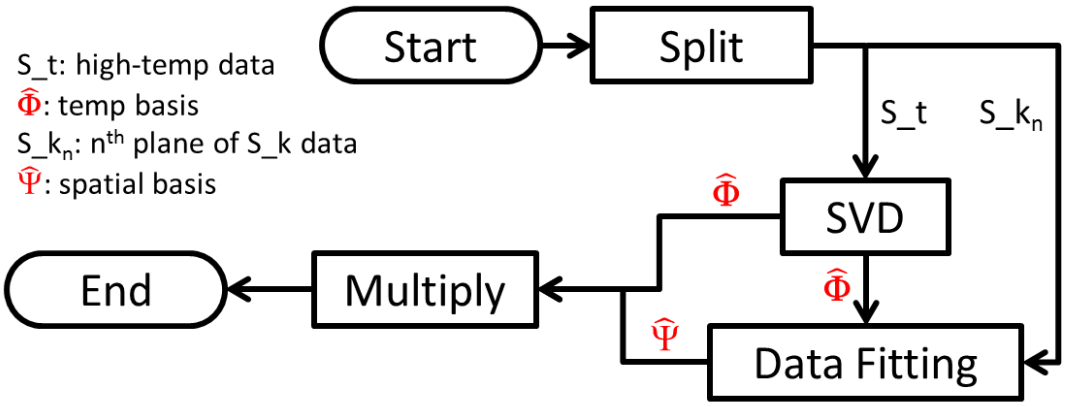


Figure 2: *Diagram of PS model reconstruction from the data acquired using sampling scheme in Figure 1. The PS model is applied pairwise to the frequently sampled plane and one of the remaining planes.*

The PS model [9] assumes that the matrix representation of the complete k-t space data $\mathbf{C}$ is low-rank, so $\mathbf{C}$ can be recovered using a few major components. The PS model expresses the k-t space as a 2D matrix, $\mathbf{C}(k,t) = \sum_{n=1}^{N} \psi_n(k)\varphi_n(t)$, where N is the order of the model, $\psi_n(k)$ and $\varphi_n(t)$ are the n-th spatial and corresponding temporal basis functions, respectively. Therefore, $\mathbf{C}$ can be factored as $\mathbf{C} = \mathbf{\Psi\Phi}$, in which the spatial basis $\mathbf{\Psi}_{i,j} = \psi_i(k_j)$, and the temporal basis $\mathbf{\Phi}_{i,j} = \varphi_j(t_i)$. Based on the low-rank assumption, a good estimate of $\mathbf{\Psi}$ and $\mathbf{\Phi}$ can recover $\mathbf{C}$ properly.

The estimated $\hat{\mathbf{\Phi}}$ and $\hat{\mathbf{\Psi}}$ can be determined using two subsets of $\mathbf{C}$ acquired using proposed sampling scheme (Figure 1), one with high temporal resolution but partial k-space coverage (S_t), the other with complete k-space coverage but low temporal resolution (S_k). Typically, $\hat{\mathbf{\Phi}}$ can be estimated from the vector along the temporal direction after singular value decomposition (SVD) of S_t. Assuming that the temporal basis of S_k is a subset of $\hat{\mathbf{\Phi}}$, S_k can be factored similar to $\mathbf{C}$. $\hat{\mathbf{\Psi}}$ can be estimated by solving $\text{S\_k} = \hat{\mathbf{\Psi}}\hat{\mathbf{\Phi}}(\text{t})$, where $\hat{\mathbf{\Phi}}(\text{t})$ is the temporal basis of S_t at the time points corresponding to S_k. The simplest minimum-norm least-square solution of $\hat{\mathbf{\Psi}}$ can be obtained using the pseudo inverse of $\hat{\mathbf{\Phi}}(\text{t})$. Finally, the estimated k-t space matrix $\hat{\mathbf{C}} = \hat{\mathbf{\Psi}}\hat{\mathbf{\Phi}}$ has S_t's temporal resolution and S_k's k-space coverage.

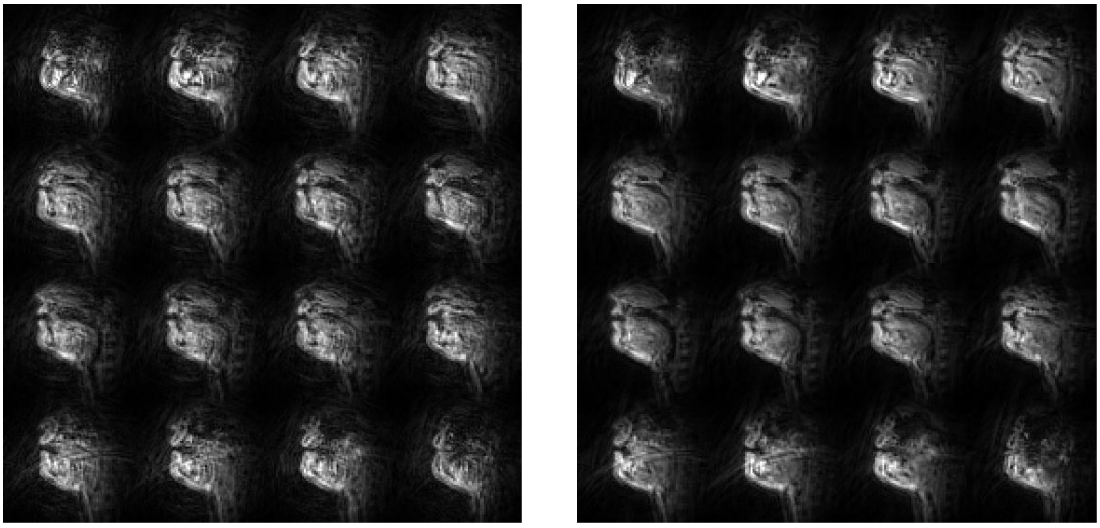
2.3. Data Acquisition

We acquired data from a midsagittal slab of 5 cm thickness on a kx-ky plane basis, using a stack-of-spirals sequence. The proposed sampling scheme included 16 kz-encoded planes, and repeatedly acquired data from one low-frequency plane, which is interleaved by one of the remaining planes in a sequential order. More specifically, the 8th kz-encoded plane was frequently sampled with the order [8, 1, 8, 2, …, 8, 16 (repeated)]. As described in Section 2.2, the PS reconstructed data has the temporal resolution of the 8th plane, and complete spatial coverage by all 16 planes. Figure 1 uses 6 kz-encoded planes to illustrate the sampling scheme, in which the 3rd plane is repeatedly sampled. Data from the other 5 planes are reconstructed from the PS model.

In our experiment design, there were 9 in-plane interleaves, thus the temporal resolution of the 8th plane is 6.5×9×2 = 117 ms, leading to an effective frame rate of 8.6 frames/sec after reconstruction. During each scan, the subject uttered a few repetitions of one typical vowel-consonant-vowel utterance such as /asa/ and /aʃa/ at normal speech rates.

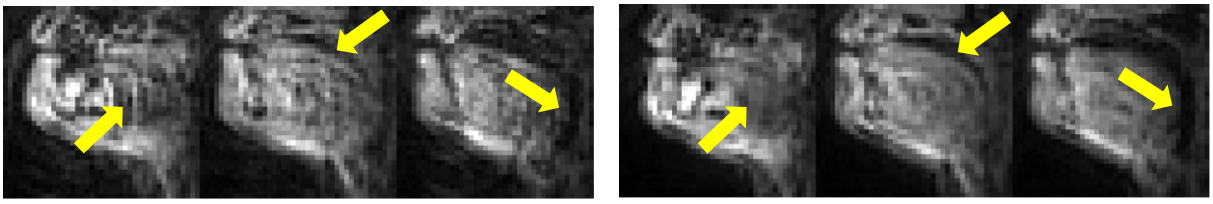
2.4. Video Reconstruction

Figure 2 demonstrates the flow of data processing. First, the acquired data were divided into a high temporal resolution but partial k-space coverage dataset (S_t, i.e., data from 8th plane only), and a high k-space coverage but low temporal resolution dataset (S_k). Secondly $\hat{\mathbf{\Psi}}$ and $\hat{\mathbf{\Phi}}$ were estimated separately by applying the PS model (of order 2) pairwise to S_t and one of the planes from S_k. Afterward, $\hat{\mathbf{C}}$ was obtained by multiplying $\hat{\mathbf{\Psi}}$ and $\hat{\mathbf{\Phi}}$, and images were reconstructed using 2D gridding in each kx-ky plane, followed



(a) **(b)**

Figure 3: *Comparison between reconstruction from static-of-spirals sequence with (a) conventional sliding window reconstruction, and (b) PS model constrained reconstruction. The proposed method generated more uniform pixel intensity and sharper boundaries.*



(a) **(b)**

Figure 4: *Cropped sagittal images at 2nd, 5th, and 8th frames of Figure 3 showing the tongue groove, using (a) conventional sliding winding, and (b) PS model constrained reconstruction. The arrows point out the motion artifact reduction.*

by a 3D inverse Fourier transformation. Video was generated at a frame rate of 8.6 frames/sec.

3. Results

Figure 3 displays the images reconstructed from the stack-of-spirals sequence and, compares the results. We chose the alveolar constriction of /s/ for the comparison. Figure 3(a) and 3(b) illustrate the images reconstructed using conventional sliding window reconstruction and proposed PS model methods, respectively. Images in Figure 3(a) clearly show spiral motion artifacts over all the scan planes, on both the tissues and the background. During the transition between /a/ and /s/, the motion artifacts were even more severe (not shown). Figure 3(b) displays results from the proposed sampling scheme and reconstruction method. Compared with the images in Figure 3(a), most of the motion artifacts were successfully removed, making the tissue intensity more uniform, and the boundaries sharper.

Figure 4 zooms in and crops the region of interest from the 2nd, 5th, and 8th frames shown in Figure 3(a) and 3(b). Arrows in Figure 4(a) point out the severe motion artifacts including those due to vibrating tissue intensities and blurry/jagged boundaries. Figure 4(b) highlights the reduction of motion artifacts from the proposed method: significantly more uniform tissue structure, cleaner and sharper boundaries.

Figure 5 compares the temporal characteristics of repeated productions of /asa/. The bright column in Figure 5(a) indicates the location where the data were extracted over time. Figure 5(b) and Figure 5(c) show the results of the conventional and proposed methods, respectively. Figure 5(b) shows that the conventional method fails to capture the temporal changes of the tongue surfaces, and resulted in appreciable residue noise in the vocal tract area. On the other hand, the PS model reconstruction portrays the temporal profile of the tongue motion much more precisely (see yellow arrows), and mitigates most of the motion blur seen in Figure 5(a). However, white arrows point out that the reconstruction of the stable alveolar ridge is affected by the motion of the tongue.

Figure 6 displays the coronal views of the same data when the groove of /s/ was formed. The arrow in Figure 6(a) indicates that the grooved tongue surface from the conventional method is unnaturally asymmetric. In contrast, the surface is clearly smoother in the PS model reconstruction (Figure 6(b)). The surface in the coronal images is in general blurrier than in the sagittal images.

4. Discussion

The first critical consideration of applying the PS model is the k-t space partitions of S_t and S_k. Researchers from the University of Illinois at Urbana-Champaign have tried the S_t/S_k combinations of Cartesian/Cartesian [12] and radial/radial [13] in 2D imaging, and spiral/Cartesian in 3D. Obviously there are more possibilities in 3D since one more k-space dimension is available. We have experimented with

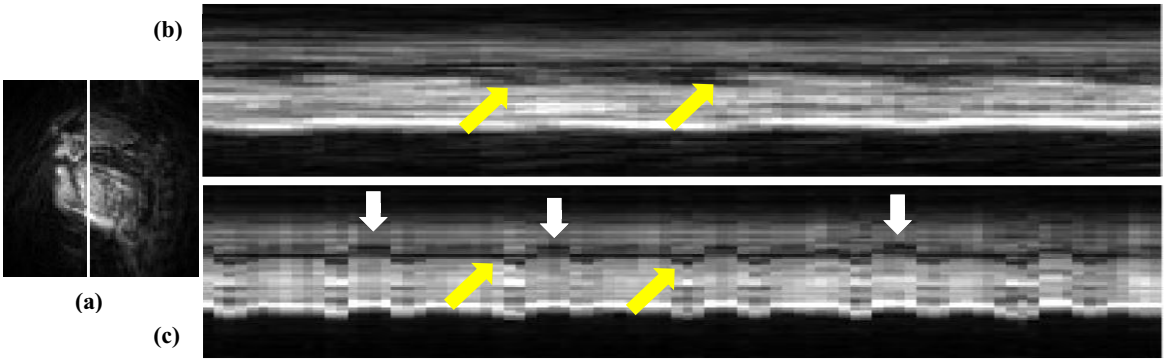


Figure 5: *(a) The bright column in an image from 8th plane marks the location of the temporal characteristics during 10 sec from (b) sliding window, and (c) PS model constrained reconstruction. The arrows point out the typical comparison of the temporal profiles of the tongue surfaces. Motion in (b) and (c) are not temporally aligned.*

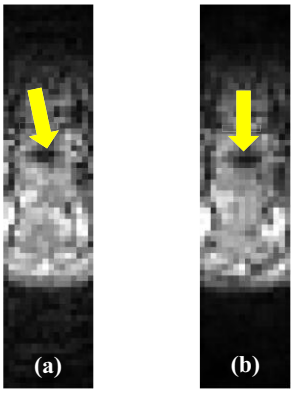


Figure 6: *Coronal views of the tongue groove of /s/, reconstructed from (a) sliding windowing reconstruction, and (b) PS model constrained reconstruction. The arrows point out the improved tongue surfaces of grooving.*

several choices of S_t/S_k combinations by adjusting the number of spiral interleaves in-plane and the number of planes used for S_t, for example, 3 interleaves from 4 low-frequency kz-encoded planes, and 1 interleaf from each of the kz-encoded planes. The proposed S_t that fully sampled 1 low-frequency showed the best results. That is because the reconstruction of data from a single low-frequency plane is the volume proton density projection onto a sagittal plane. In the particular application of human speech, there is much less motion across sagittal planes, making sagittal projection an optimal navigator to capture the temporal changes.

As a trade-off, the coronal image quality (Figure 6(b)) is lower than the sagittal image quality (Figure 4(b)) because the proposed S_t emphasizes sagittal motion, being less sensitive to motion in other directions.

The second critical issue of the PS model reconstruction is the model order. We have seen in the numerical phantom [14] experiments that as the order increases, the results improve until reaching a plateau threshold, which is dependent on the application. This is because the conventional PS model relies on a matrix inversion in data fitting, where errors and noises may be amplified. Two possible approaches that prevent the results from deteriorating include regularization reconstruction [15], and further partitioning the S_k before k-space reconstruction, in which a small order is sufficient. We have shown plane-wise reconstruction and retrospectively selected the model order and examined the results qualitatively to optimize the image quality. The unstable alveolar ridge in Figure 5(b) resulted from a low order (order of 2) model fitting, and can be mitigated using further k-space partitioning before PS model fitting. Exploring better k-space partitions combined with regularization reconstruction remains as future work.

The nominal frame rate was 8.6 frames/sec, but it is difficult to claim both true temporal and spatial resolution, because the k-space data were interpolated using a model assumption. A more objective assessment of the effective temporal and spatial resolution is desired in further development.

The current major limitation of the PS model applied to speech imaging is the requirement of repeated production during each data acquisition. Intuitively, the PS model interpolated k-space data in a smart and automatic way, by using a high-temporal data subset as navigator. It is therefore not realistic to expect the model to reconstruct motion that is only captured by a very small subset of interleaves. However, it is feasible to design a time-segmentation method to remove the dependence on the motion periodicity, if the stimuli are designed to image a geometrical feature (such as tongue groove) in multiple utterances. Our preliminary results (not shown) have exhibited the improvements using vocal tract motion cluster-based PS model reconstruction which we hope to further develop.

5. Conclusions

We have shown that the PS model has the potential to accelerate 3D stack-of-spirals speech imaging by up to a factor of 8. We have applied the PS model to dynamic 3D imaging of the vocal tract during speech, exploited the spatiotemporal correlations in k-t space, and obtained 3×3×3.1 mm^3 spatial resolution at 8.6 frames/sec. The proposed method captures 3D dynamics of the vocal tract shaping, and would be useful in the study of the speech production of complex vocal tract geometries such as liquid and fricative consonants at normal speech rates.

6. Acknowledgements

This work was supported by NIH Grant R01 DC007124-01. The authors thank Cornelius Brinegar for providing a demo of the PS model, and the SPAN (Speech Production and Articulation kNowledge) group at the University of Southern California for the support and collaboration.

7. References

[1] Demolin, D., Metens, T., Soquet, A., "Real-time MRI and articulatory coordinates in vowels speech production", in Proc. Speech Prod. Sem., 86-93, 2000.

[2] 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., 115(4):1771-1776, 2004.

[3] Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., Narayanan, S., "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging", IEEE Sig. Process. Mag: 25(3): 123-132, 2008

[4] Sutton, B., Conway, C., Bae, Y., Seethamraju, R., Kuehn, D, "Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T", J. Magn. Reson. Imaging, 32:1228-1237, 2010.

[5] Proctor, M., Lu, L.H., Zhu Y., Goldstein L., Narayanan, S., "Articulation of Mandarin Sibilants: a multi-plane realtime MRI study", in Proc. Speech Sci. Tech., 113-116, 2012.

[6] Yang, C., Stone, M., "Dynamic programming method for temporal registration of three-dimensional tongue surface motion from multiple utterances", Speech Commun., 38:201–209, 2002.

[7] Zhu, Y., Kim, Y.-C., Proctor, M., Narayanan, S., Nayak, K., "Dynamic 3D visualization of vocal tract shaping during speech", in Proc. ISMRM, 4355, 2011.

[8] Zhu, Y., Kim, Y.-C., Proctor, M., Narayanan, S., Nayak, K., "Toward dynamic 3D MRI of speech", in Proc. ISMRM, 294, 2012.

[9] Liang, Z.-P., "Spatiotemporal imaging with partially separable functions", in Proc. IEEE ISBI, 988-991, 2007.

[10] Fu, M., Christodoulou, A., Naber, A., Kuehn, D., Liang, Z.-P., Sutton, B., "High-frame-rate multislice speech imaging with sparse sampling of (k,t)-space", in Proc. ISMRM, 12, 2012.

[11] Santos, J., Nishimura, D., Yang, P., Pauly, J., "Adaptive architecture for real-time imaging systems," in Proc. ISMRM, 468, 2002.

[12] Brinegar, C., Zhang, H., Wu, Y.-J., Foley, L., Hitchens, K., Ye, Q., Pocci, D., Lam, F., Ho, C., Liang, Z.-P., "Real-time cardiac MRI using prior spatial-spectral information", in proc. IEEE EMBS, 4383-4386, 2009.

[13] Brinegar, C., Schmitter, S., Mistry, N., Johnson, A., Liang, Z.-P., "Improving temporal resolution of pulmonary perfusion imaging in rats using the partially separable fuctions model", Magn. Reson. in Med., 64:1162-1170, 2010.

[14] Zhu, Y., Narayanan, S., Nayak, K., Flexible dynamic phantoms for evaluating MRI data sampling and reconstruction methods, in Proc. ISMRM Workshop, Sedona, AZ, 2013.

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