Towards Dynamic 3D MRI of Speech

Yinghua Zhu[1], Yoon-Chul Kim[1], Michael Proctor[1], Shrikanth Narayanan[1], and Krishna Nayak[1]
[1]*University of Southern California, Los Angeles, California, United States*

Introduction: Our understanding of human speech production has been fundamentally changed thanks to the availability of MRI techniques such as 2D cine imaging [1, 2], 2D real-time imaging [3] and 3D static imaging [4]. In this work, we present our initial efforts towards 3D real-time imaging of the vocal tract, based on a stack-of-spirals data acquisition strategy. Data obtained using these imaging techniques are providing important new insights into the dynamics of tongue shaping in three dimensions – information which is critical to our understanding the goals of speech production. Dynamic vocal tract imaging in three dimensions is especially applicable to the study of segments articulated with complex vocal tract geometries, such as liquid and fricative consonants.

Methods and Results:

Sequence design: There is no widely accepted standard for spatial and temporal resolution needed for speech imaging. In fact, it depends largely on the articulators of interest, the sounds being studied, and in part, the language. Based on our experiences of 2D real-time speech imaging, we determined 2.5 mm isotropic spatial resolution and 150 ms temporal resolution as our initial goal. For pulse sequence design, we adopted a stack-of-spirals sampling [5] in which target imaging FOV was 20 x20 cm^2 in the in-plane and 5 cm in the slice direction. The parameters of five sequences are listed and compared in Table 1, along with the acceleration factors that are needed to achieve 150 ms temporal resolution. Data were reconstructed using 2D gridding within each kx-ky plane, followed by 3D inverse Fourier transformation.

Seq. Name	Num. of Interleaves	Tread (ms)	TR (ms)	Temp. Res. (s)	SNR in tongue	Required Acc. Factor
i7	7	4.6	8.5	1.19	48.0	7.9
i8	8	4.0	8.0	1.27	46.0	8.5
i9	9	3.6	7.5	1.35	50.5	9.0
i11	11	3.0	6.9	1.51	51.0	10.1
i13	13	2.5	6.5	1.68	48.5	11.2

Table 1. 3D stack-of-spiral sequences used for in-vivo experiments.

Experiments: Experiments were performed on a GE 1.5T MRI scanner, using a custom 4-channel upper airway receiver coil and a custom real-time imaging platform [6, 7]. Data were acquired using a mid-sagittal slab. Three native English speakers were imaged while producing VCV (vowel-consonant-vowel) sequences /*asa*/ and /*afa*/ using all five candidate sequences (see Table 1). Each speaker was instructed to sustain each consonant for about 3 seconds. The sustained productions enabled capture of the 3D vocal tract shape without any motion artifacts, and provided data for assessment of spatial blurring/distortion in the air-tissue boundaries of the articulators.

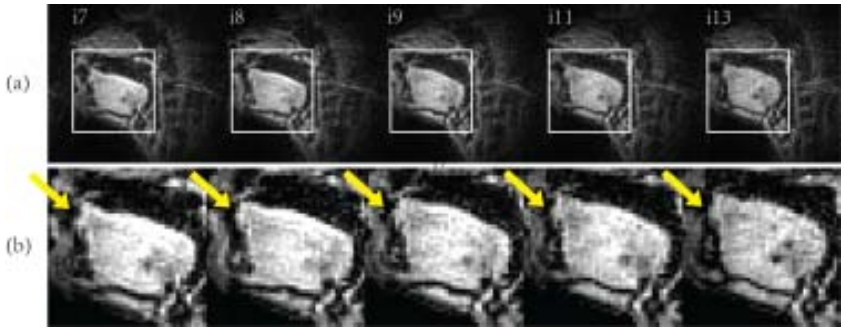

Fig 1. Mid-sagittal images captured from the production of /*s*/ in /*asa*/ using the sequences *i7*, *i8*, *i9*, *i11*, and *i13*. (a) Images showing full FOV of the vocal tract. (b) Zoomed-in images from the rectangular box in (a). The arrows indicate that the artifacts are the most prominent in the sequence *i7*, compared with the remaining sequences.

Results: Figure 1(a) shows mid-sagittal images using designed sequences during the production of /s/.We used identical shim value and center frequency in real-time acquisition, and observed that noise levels were almost the same among all sequences considered. The arrow points out the imaging artifact on the tongue tip observed using sequence *i7* in Figure1(b), compared with the remaining sequences. *i8* is the most time efficient sequence without image artifact and has comparable performance with *i9*, *i11* and *i13*, and is therefore used in the remainder of the study. Figure 2(a) shows the positions of coronal images (Figure 2(b)) and axial images (Figure 2(c)) that are reconstructed from the same 4D data set (3D + time) at the same time frame. These slices cover the region of interest and clearly demonstrate the characteristic tongue grooving (see arrows in Figure2(b)) and other important details of the vocal tract geometry (see arrows in Figure2(c)). Image SNR measurements in tongue were 48.0, 46.0, 50.5, 51.0, 48.5 for *i7*, *i8*, *i9*, *i11*, *i13*, respectively.

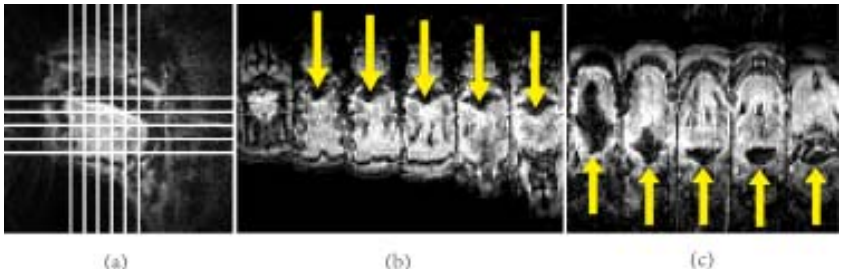

Fig 2. Three orthogonal slices from a time frame corresponding to the production of /*s*/ in /*asa*/ using sequence *i8*. (a) Midsagittal slice. Lines indicate slice locations corresponding to coronal (b) and axial (c) images. Arrows in (b) and (c) point out a critical feature of tongue shaping: the formation of a tongue groove.

Discussion: Our preliminary results suggest that 3D stack-of-spirals sampling with readout duration = 4.0 ms is a viable option for 3D speech imaging. This will require an additional 8.5x acceleration in order to provide dynamic information during natural speech, which may be accomplished using a combination of established methods, such as partial Fourier, parallel imaging, and compressed sensing. The assessment of the image quality using the accelerated techniques remains as future work. It is anticipated that the use of accelerating techniques will result in SNR loss by a factor of the square root of the acceleration factor.

Acknowledgments: This work is supported by NIH Grant R01 DC007124-01.

References: [1] Stone, et al., JSLHR:44:1026-1040, 2001; [2] Takemoto, et al., J.Acoust.Soc.Am.:119:1037-1049, 2006; [3] Narayanan, et al., J.Acoust.Soc.Am.:115:1771-1776, 2004; [4] Story, et al., J.Acoust.Soc.Am.:100:537-554, 1996; [5] Meyer, et al., ISMRM, p392, 1996; [6] Santos, et al., ISMRM, p468, 2002; [7] Bresch, et al., IEEE Sig.Proc.Mag.:25:123-132, 2008.