Peer-Reviewed Technical Communication

Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)

Shrikanth Narayanan, Asterios Toutios,[a] Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, and Sungbok Lee
Signal Analysis and Interpretation Laboratory, University of Southern California, 3740 McClintock Avenue, Los Angeles, California 90089-2564

Krishna Nayak, Yoon-Chul Kim, and Yinghua Zhu
Magnetic Resonance Engineering Laboratory, University of Southern California, 3740 McClintock Avenue, Los Angeles, California 90089-2564

Louis Goldstein and Dani Byrd
Department of Linguistics, University of Southern California, 3601 Watt Way, Los Angeles, California 90089-1693

Erik Bresch
Philips Research, High Tech Campus 5, 5656 AE, Eindhoven, Netherlands

Prasanta Ghosh
Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka, 560012, India

Athanasios Katsamanis
School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytexneiou Street, Athens 15773, Greece

Michael Proctor
ARC Centre of Excellence in Cognition and its Disorders and Department of Linguistics, Macquarie University, New South Wales 2109, Australia

(Received 25 October 2013; revised 27 June 2014; accepted 2 July 2014)

USC-TIMIT is an extensive database of multimodal speech production data, developed to complement existing resources available to the speech research community and with the intention of being continuously refined and augmented. The database currently includes real-time magnetic resonance imaging data from five male and five female speakers of American English. Electromagnetic articulography data have also been presently collected from four of these speakers. The two modalities were recorded in two independent sessions while the subjects produced the same 460 sentence corpus used previously in the MOCHA-TIMIT database. In both cases the audio signal was recorded and synchronized with the articulatory data. The database and companion software are freely available to the research community. © *2014 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4890284]

PACS number(s): 43.70.Aj, 43.70.Jt, 43.70.Bk [ZZ] Pages: 1307–1311

I. INTRODUCTION

Real-time magnetic resonance imaging (rtMRI) is an important emerging tool for speech research,[1,2] directly providing dynamic information from the entire mid-sagittal plane of a speaker's upper airway, or any other scan plane of interest, from a single utterance with no need of repetitions. Mid-sagittal rtMRI captures not only lingual, labial, and jaw motion, but also articulation of the velum, pharynx, and larynx—regions of the tract which cannot be easily monitored with other speech articulation measurement techniques. While sampling rates are currently lower than for electromagnetic articulometry (EMA)[3] or x-ray microbeam (XRMB),[4] rtMRI is a unique source of dynamic information about vocal tract shaping and global articulatory coordination.

We describe here an ongoing initiative in which we are assembling a large-scale, multi-speaker rtMRI speech database and supporting toolset, with the aim of supporting speech research based on this modality, and making some of these resources available to the broader speech research community. In parallel, we are collecting and making available 3D EMA data, from the same speakers using the same stimuli. EMA data complement rtMRI data by providing faster acquisition rates (but sparser spatial information from a few key flesh points), while these two modalities may be advantageously combined using co-registration techniques.[5] We call this collection the USC-TIMIT database. It is freely available from the website http://sail.usc.edu/span/usc-timit.

[a]Author to whom correspondence should be addressed. Electronic mail: toutios@sipi.usc.edu

II. REAL-TIME MRI ACQUISITION

Subjects' upper airways were imaged while they lay supine in the MRI scanner. The sentence stimuli were presented in large text on a back-projection screen which subjects could read from within the scanner bore through a mirror without moving their head. Sentences were presented one at a time, elicited at a natural speaking rate. The experimenter interacted with the subject in the scanner through an intercom system.

MRI data were acquired at Los Angeles County Hospital on a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha, WI) with gradients capable of 40 mT/m amplitude and 150 mT/m/ms slew rate.[2,6] A body coil was used for radio frequency (RF) signal transmission. A custom four-channel upper airway receiver coil array, with two anterior coil elements and two coil elements posterior to the head and neck, was used for RF signal reception. A 13-interleaf spiral gradient echo pulse sequence was used ($T_R = 6.164$ ms, FOV = 200 × 200 mm, flip angle = 15°, receiver bandwidth = ±125 kHz). The spiral fast gradient echo sequence consisted of slice-selective excitation, spiral readout, rewinder, and spoiler gradients. Slice thickness was 5 mm, located mid-sagittally; image resolution in the sagittal plane was 68 × 68 pixels (2.9 × 2.9 mm). Scan plane localization of the mid-sagittal slice was performed using RTHawk (HeartVista, Inc., Los Altos, CA), a custom real-time imaging platform.[7]

MR image reconstruction was performed using MATLAB (Mathworks, South Natick, MA). Image frames were formed using gridding reconstruction[2,8] from data sampled along spiral trajectories. Gridding reconstruction was performed on each individual anterior coil data. Root sum-of-squares of the reconstructed anterior coil images was taken to improve image signal-to-noise ratio as well as spatial coverage of the vocal tract. Reconstructed images from the posterior two coil elements showed spatial aliasing artifacts and thus were not considered for coil image combination. Although new image data were acquired at a rate of 12.5 frames/s, sliding window technique was used to allow for view sharing and thus increase frame rate.[1] The initial image frame was reconstructed from the first 13 consecutive acquisitions of spiral data. The TR-increment for view sharing was seven acquisitions, meaning that the next image frame was reconstructed using acquisitions eight to 20 of the consecutive spiral data, and so on. The end result was the generation of MRI movies with a frame rate of 1/(7*TR) = 1/(7*6.164 ms) = 23.18 frames/s. The TR-increment value is a user-controlled parameter—for instance, one can also lower the TR-increment to 1. This 1-TR sliding window reconstruction would maximize frame rate to 162.23 frames/s, but would also increase the file size of the movie as well as image reconstruction time.

Audio was simultaneously recorded at a sampling frequency of 20 kHz inside the MRI scanner while subjects were imaged, using a fiber-optic microphone (Optoacoustics Ltd., Moshav Mazor, Israel) and custom recording setup. Synchronization with the video signal was controlled through the use of an audio sample clock derived from the scanner's 10 MHz master clock, and triggered using the scanner RF master-exciter unblank signal.

Noise generated from the operation of the MRI scanner required attenuation to a sufficient level in order to perform more detailed analyses of the audio for linguistic and statistical modeling purposes. Noise cancellation was performed using a custom adaptive signal processing algorithm that takes into account the periodic structure of the noise generated by the scanner consistent with the rtMRI acquisition set.[9] Note that subjects wore earplugs for protection from the scanner noise, but were still able to hear loud conversation in the scanner room and to communicate orally with the experimenters via both the fiber-optic microphone setup as well as the in-scanner intercom system.

TABLE I. Study participants, with subset of demographic details collected. All subjects in this set identified their race as white. Last column indicates if EMA data are available with the current data release.

ID	Gender	Age	Birthplace	EMA
M1	Male	29	Buffalo, NY	✓
M2	Male	33	Ann Arbor, MI	
M3	Male	26	Madison, WI	✓
M4	Male	26	St. Louis, MO	
M5	Male	27	Mammoth, CA	
F1	Female	23	Commack, NY	✓
F2	Female	32	Westfield, IN	
F3	Female	20	Palos Verdes, CA	
F4	Female	46	Pittsburgh, PA	
F5	Female	25	Brawley, CA	✓

III. EMA ACQUISITION AND PROCESSING

EMA data were collected using the NDI Wave Speech Research System (Northern Digital Inc., Waterloo, Ontario, Canada). Subjects were asked to read the stimuli while sitting upright. Speech audio was acquired along with the 3D trajectories of sensors attached to articulators.

The NDI wave system supports tracking of five or six degree-of-freedom (DOF) sensors. Six 5-DOF sensors were attached on articulators to capture their movements. Three sensors were attached on the midline of the tongue, the frontmost sensor at 0.5–1 cm behind the anatomical tongue tip, and the rear-most sensor as far back as possible. The other three sensors were attached to the surfaces of the lower incisor, lower lip and upper lip. For head movement correction, one 6-DOF sensor (for subjects M1 and F5 in Table I) or two 5-DOF sensors (M2, F1) were used as reference sensors. (Each 6-DOF sensor requires two channels out of the eight available in the standard NDI Wave system, while each 5-DOF sensor requires one channel.) Three-dimensional spatial coordinates of sensors were recorded at a sampling rate of 100 Hz.

In post-processing, missing sensor tracking points were estimated by piecewise cubic Hermite interpolation, data were corrected for head movement and sensor trajectories were smoothed by low pass filtering at 20 Hz. The speech audio waveform was recorded simultaneously through a microphone at a sampling rate of 44.1 kHz, and subsequently was down sampled to 16 kHz.

IV. DATABASE DESCRIPTION

To date, rtMRI data have been acquired from ten native speakers of General American English (Table I), none of

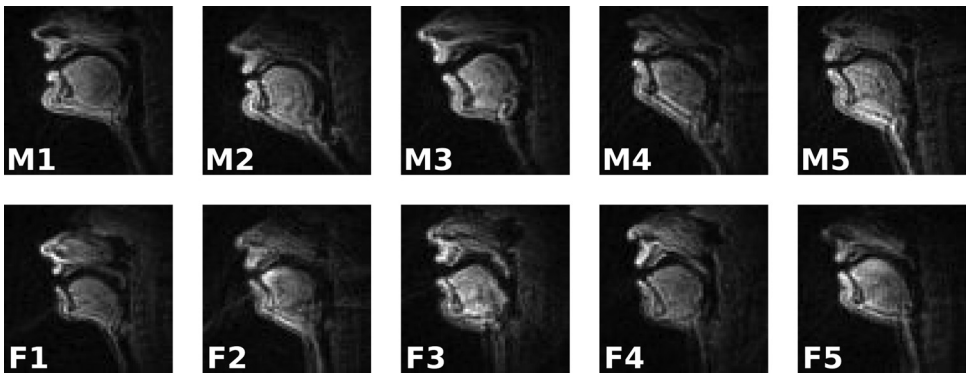

FIG. 1. Mid-sagittal images of vocal tracts for each of the ten speakers currently in the database. Indicative frames extracted from utterance "This was easy for us."

whom reported abnormal hearing or speaking development or pathologies. Figure 1 shows images of the vocal tracts of these speakers (single frames extracted from rtMRI data). EMA data have been acquired for four of these speakers (see Table I).

In both rtMRI and EMA acquisitions, the reading material spoken by study participants was the same 460 sentence set used in the MOCHA-TIMIT database.[10] This sentence set is designed to elicit all phonemes of American English in a wide range of prosodic and phonological contexts, with the connected speech processes characteristic of spoken English. In addition to providing a phonologically comprehensive sample of English, this corpus was chosen to provide an additional resource for researchers who have previously made use of the MOCHA-TIMIT database.

V. DATA ANALYSIS TOOLS

A graphical user interface has been developed to allow for audition, labeling, tissue segmentation, and acoustic analysis of the USC-TIMIT data (see Fig. 2). The primary purpose of this tool is to allow users to browse the database frame-by-frame, inspect synchronized audio and video segments in real-time or at slower frame rates, and label speech segments of interest for further analysis with the supporting tool set. The GUI facilitates automatic formant and pitch tracking, and integrates a method of rapid semi-automatic segmentation of rtMRI data for parametric analysis which seeks pixel intensity thresholds distributed along tract-normal grid-lines and defines airway contours constrained with respect to a tract centerline constructed between the glottis and lips.[11] The method also allows for the use of reference boundaries and manual supervision to guide segmentation of anatomical features which are poorly imaged using magnetic resonance due to low signal-to-noise ratios and scarcity of soft tissue, such as in regions of dentition and the hard palate.

For the analysis of EMA data, use of the software MVIEW is recommended, available from its developer (Mark Tiede, Haskins Laboratories). The EMA data are provided in a format that is readily compatible with MVIEW.

VI. FUTURE WORK

We envision this resource to be continually improved and updated. It is our immediate plan to provide with phonetic labels for the rtMRI and EMA data. We have undertaken a project to first force-align the data, and then correct manually the labels. In the longer term, we plan to provide the community with contours outlining the vocal-tract shape on the sequences of rtMRI images.[12]

We plan to enhance the present database by adding more types of data acquired from more speakers, and to expand the toolset to allow for more sophisticated inspection and analysis of these data. The database will initially be augmented with data from more speakers of General American English, but ultimately also with speakers of other varieties of English, and speakers of other languages. We intend to acquire video with higher frame-rates and improved SNR, and to incorporate data acquired from imaging planes other than mid-sagittal, including mid-lingual coronal cross sections.

VII. CONCLUDING REMARKS

By providing sequences of full mid-sagittal vocal-tract images during natural speech at a high enough spatio-temporal resolution, real-time MRI allows for the study of the spatio-temporal coordination of the speech organs toward linguistic[13] and also paralinguistic[14] goals in ways that were not previously possible, especially given that X-ray imaging has been abandoned as a tool for speech research due to safety and ethical concerns.

rtMRI, like most techniques for acquiring articulatory data, has its shortcomings. Its temporal resolution is not as good as that of EMA, and its spatial resolution not as good as that of static MRI. Subjects need to lay supine during data acquisition, which may perturb their speech production as compared to a natural upright position.[15] The MRI scanner is a very noisy environment, and subjects need to wear ear-plugs during acquisition, thus not having normal auditory feedback.

Nevertheless, we believe that these shortcomings do not overshadow rtMRI's utility for speech production research. One of the reasons that led us to the dissemination of the USC-TIMIT database is to enable the broad scientific community to explore its utility, and understand its limitations, from diverse scientific and technological perspectives. Importantly, one promising avenue for future effort is to find ways for combining rtMRI data with other speech production data that offer complementary advantages; indeed, the inclusion of EMA data was to facilitate such a goal. Significant technical challenges need to be overcome, and appropriate

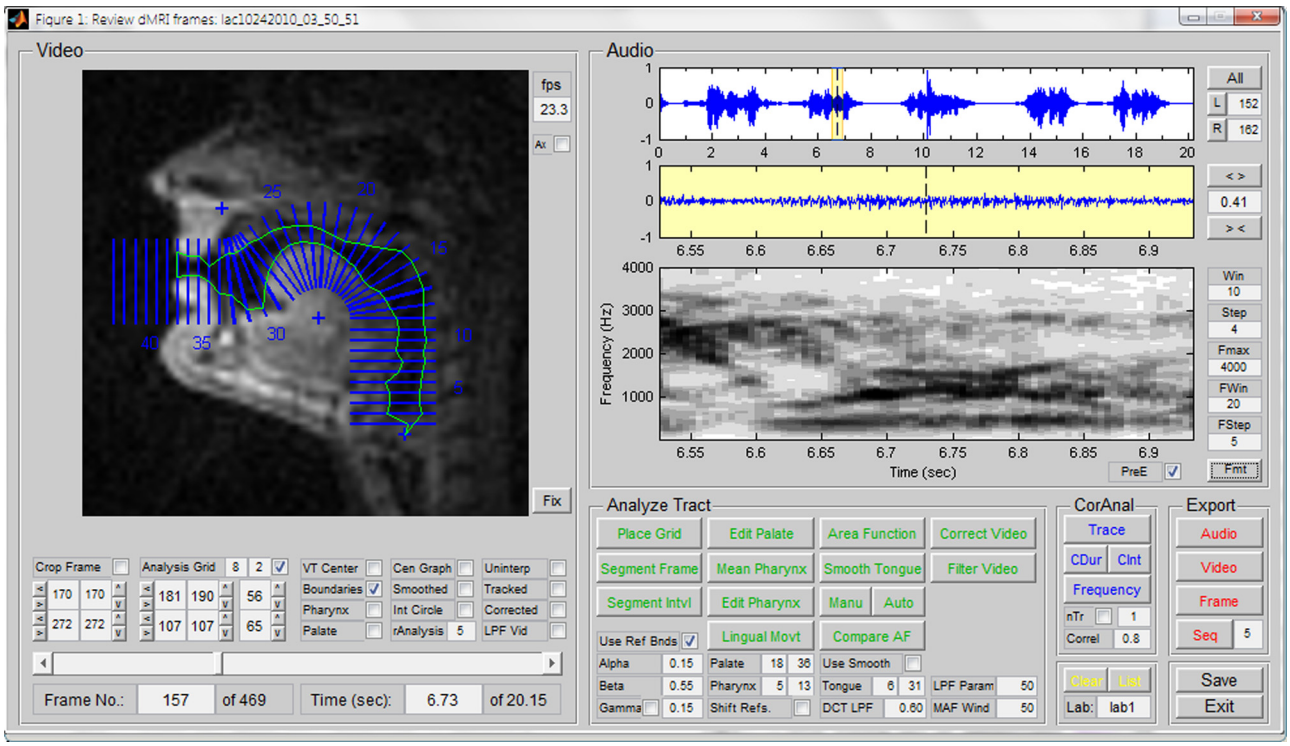

FIG. 2. (Color online) Graphical user interface allowing for audition, labeling, tissue segmentation, and acoustic analysis of the rtMRI data, displaying an example of parametric segmentation.

technical methods developed, to fully enjoy the potential benefits of such multimodal resources. For example, albeit using the exact same stimuli, rtMRI and EMA data have been collected in distinct sessions and under different conditions (the subjects lay supine during rtMRI acquisitions while were upright for EMA).

Real-time magnetic resonance imaging of the upper airway is an active research area.[16–19] The quest for improving the possible spatial and temporal resolution, and identifying effective trade offs for given application needs, is ongoing. However, with current imaging and audio acquisition capabilities, it is possible to collect large amounts of speech production data than ever before such as demonstrated by the current database. This opens up novel corpus-driven scientific research as well as technological efforts such as in automatic speech and speaker recognition.

ACKNOWLEDGMENT

This work was supported by National Institutes of Health grant R01 DC007124.

[1]S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776 (2004).
[2]E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]," IEEE Signal Process. Mag. **25**, 123–132 (2008).
[3]J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096 (1992).
[4]J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database user's handbook," Technical Report, Waisman Center on Mental Retardation and Human Development, University of Wisconsin (1994).
[5]J. Kim, A. C. Lammert, P. Kumar Ghosh, and S. S. Narayanan, "Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging," J. Acoust. Soc. Am. **135**, EL115–EL121 (2014).
[6]Y.-C. Kim, S. S. Narayanan, and K. S. Nayak, "Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order," Magn. Reson. Med. **65**, 1365–1371 (2011).
[7]J. Santos, G. Wright, and J. Pauly, "Flexible real-time magnetic resonance imaging framework," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA (2004), pp. 1048–1051.
[8]J. Jackson, C. Meyer, D. Nishimura, and A. Macovski, "Selection of a convolution function for Fourier inversion using gridding," IEEE Trans. Med. Imaging **10**, 473–478 (1991).
[9]E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791–1794 (2006).
[10]A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proceedings of the 5th Seminar on Speech Production*, Kloster Seeon, Bavaria (2000), pp. 305–308.
[11]M. I. Proctor, D. Bone, and S. S. Narayanan, "Rapid semi-automatic segmentation of real-time Magnetic Resonance Images for parametric vocal tract analysis," in *Proceedings Conference of the International Speech Communication Association*, Makuhari, Japan (2010), pp. 1576–1579.
[12]E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," IEEE Trans. Med. Imaging **28**, 323–338 (2009).
[13]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. **37**, 97–110 (2009).
[14]M. I. Proctor, E. Bresch, D. Byrd, K. S. Nayak, and S. S. Narayanan, "Paralinguistic mechanisms of production in human 'beatboxing': a real-time magnetic resonance imaging study," J. Acoust. Soc. Am. **133**, 1043–1054 (2013).

[15]M. Stone, G. Stock, K. Bunin, K. Kumar, M. Epstein, C. Kambhamettu, M. Li, V. Parthasarathy, and J. Prince, "Comparison of speech production in upright and supine position," J. Acoust. Soc. Am. **122**, 532–541 (2007).
[16]B. P. Sutton, C. Conway, Y. Bae, C. Brinegar, Z.-P. Liang, and D. P. Kuehn, "Dynamic imaging of speech and swallowing with MRI," in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN (2009), pp. 6651–6654.
[17]A. Scott, R. Boubertakh, M. Birch, and M. Miquel, "Towards clinical assessment of velopharyngeal closure using MRI: evaluation of real-time MRI sequences at 1.5 and 3 T," Br. J. Radiol. **85**, e1083–e1092 (2012).
[18]S. Zhang, A. Olthoff, and J. Frahm, "Real-time magnetic resonance imaging of normal swallowing," J. Magn. Reson. Imaging **35**, 1372–1379 (2012).
[19]A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction," Magn. Reson. Med. **69**, 477–485 (2013).