SIP (2016), vol. 5, e6, page 1 of 12 © The Authors, 2016.

doi:10.1017/ATSIP.2016.5

OVERVIEW PAPER

Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research

ASTERIOS TOUTIOS AND SHRIKANTH S. NARAYANAN

Real-time magnetic resonance imaging (rtMRI) of the moving vocal tract during running speech production is an important emerging tool for speech production research providing dynamic information of a speaker's upper airway from the entire midsagittal plane or any other scan plane of interest. There have been several advances in the development of speech rtMRI and corresponding analysis tools, and their application to domains such as phonetics and phonological theory, articulatory modeling, and speaker characterization. An important recent development has been the open release of a database that includes speech rtMRI data from five male and five female speakers of American English each producing 460 phonetically balanced sentences. The purpose of the present paper is to give an overview and outlook of the advances in rtMRI as a tool for speech research and technology development.

Keywords: Speech production, Real-time MRI, Vocal tract shaping, Articulation

Received 9 April 2015; accepted 24 February 2016

I. INTRODUCTION

A long-standing challenge in speech research is obtaining accurate information about the movement and shaping of the vocal tract. Dynamic vocal tract-imaging data are crucial for investigations into phonetics and phonological theory, where they afford insights into the nature and execution of speech production goals, the relationship between speech articulation and acoustics, and the mechanisms of speech motor control. Such data are also important for advancing the knowledge and treatment of speech pathologies, and to inform models used in speech technology applications, such as machine speech recognition and synthesis.

A number of techniques are available for the acquisition of data on the kinematics of speech production. Electromagnetic articulography (EMA) [1] uses electromagnetic fields to track the positions of markers attached on articulators in two or three dimensions with sampling rates up to 400 Hz. X-ray microbeam (XRMB) [2] generates a very narrow beam of high-energy X-ray, and rapidly directs this beam, under high-speed computer control, to track the motions of 2–3 mm diameter gold pellets glued to articulators with rates up to 160 Hz. Electropalatography (EPG) [3] uses an artificial palate with embedded electrodes to record linguopalatal patterns of contact, typically at 100–200 Hz. Ultrasound [4, 5] can be used to image the tongue, and X-ray [6–9] or videofluoroscopy [10] to image the sagittal projection of the entire vocal tract at frame rates typically between 10 and 50 Hz. Synchronized-sampling (repetitive) MRI can be used to reconstruct tongue motion in two-dimensional (2D) or 3D from multiple repetitions of an utterance [11, 12].

Nevertheless, it is still difficult to safely obtain information about the location and movement of speech articulators in all parts of the vocal tract (like the tongue, velum, and larynx, hidden from plain view) and at sufficiently high sampling rates with respect to their natural movement speed during speech. All aforementioned speech production data acquisition technologies are limited in one sense or the other. EMA and XRMB both provide rich data about the movement of sensors or markers attached on lingual and labial fleshpoints, but such sensors/markers cannot be easily placed at posterior locations on the tongue, on the velum, in the pharynx, or in the larynx; hence these technologies are limited in terms of the spatial coverage of the complex vocal tract geometry. EPG is restricted to contact measurements at the palate. Ultrasound cannot consistently or reliably image the tongue tip, the pharyngeal surface of the tongue (because of the obscuring effect of the hyoid bone), or the opposing surfaces such as the hard and soft palate (and hence the airway shaping). X-ray and videofluoroscopy expose subjects to unacceptable levels of radiation.

Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California (USC), 3740 McClintock Avenue, Los Angeles, CA 90089, USA

Corresponding author:
A. Toutios
Email: toutios@usc.edu

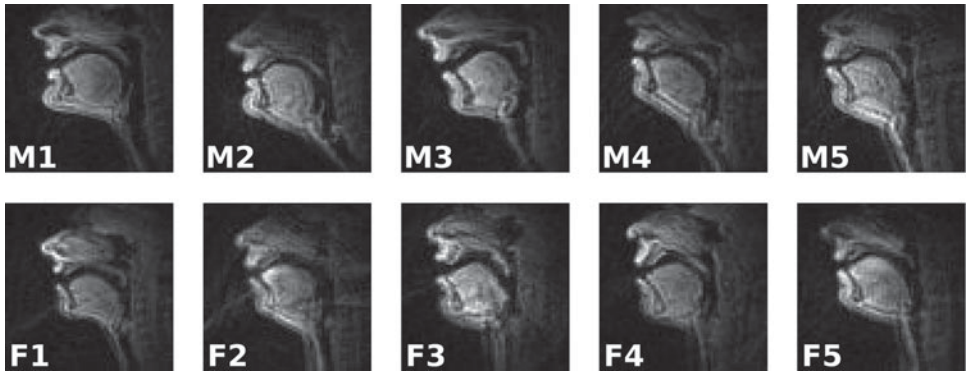

Fig. 1. Example rtMRI frames from the ten speakers in the USC-TIMIT database (top row, male; bottom row, female).

In early development, magnetic resonance imaging (MRI) has been used to capture images of static configurations of the vocal tract, but while subjects sustain continuant speech sounds over unnaturally long periods of time. In synchronized-sampling cine-MRI (or repetitive cine-MRI), articulatory dynamics of connected speech may be reconstructed from large numbers of repetitions (which should be identical) of short utterances.

Real-time magnetic resonance imaging (rtMRI) is an important emerging tool for speech production research [13, 14], providing dynamic information from the entire midsagittal plane of a speaker's upper airway, or any other scan plane of interest, from arbitrary, continuous utterances with no need of repetitions. Midsagittal rtMRI captures not only lingual, labial, and jaw motion, but also articulation of the velum, pharynx and larynx, and structures such as the palate and pharyngeal wall – regions of the tract that cannot be easily or well observed using other techniques. While sampling rates are currently lower than for EMA or XRMB, rtMRI is a unique source of dynamic information about vocal tract shaping and global articulatory coordination. Because rtMRI allows unparalleled views of the state of articulation in regions of the tract from which it has previously proven problematic to obtain accurate data, this technique is beginning to offer new insights into the the goals of production of coronal [15], pharyngeal [16] and nasal [17] segments, and the coordination of articulators during the production of multi-gestural segments in speech [18–20]. Most importantly, rtMRI data also provide a rich source of information about articulation in connected speech, which is proving to be valuable in the refinement of existing speech models and the development of new models of representation for automatic speech recognition (ASR) and other speech processing applications. RtMRI of the upper airway (a definition that also includes studies of other functions of the vocal tract besides speech, such as swallowing) is an actively growing research area [21–27].

The present paper provides an overview of rtMRI for speech research that is particularly being developed by an interdisciplinary team at the University of Southern California (USC). It summarizes their advances in creating and refining rtMRI acquisition methods, developing analysis tools, collecting multilingual speech and vocal production data, and using them to address scientific and technology problems of interest. This includes the public release of a unique corpus of articulatory data, called the USC-TIMIT database [28], available from http://sail.usc.edu/span/usc-timit/, which includes rtMRI data from ten speakers, each uttering the same 460 sentences used in the context of the popular MOCHA-TIMIT database [29] of EMA, EPG, and electroglottographic (EGG [30]) data. This set of sentences was designed to elicit all phonemes of English in a wide range of prosodic and phonological contexts, with the connected speech processes characteristic of spoken English, including assimilations, lenitions, deletions, and mergers. USC-TIMIT also includes EMA data collected separately from four of the subjects. See Figs 1 and 2 for example images from the database.

The rest of this paper elaborates on some technical aspects of rtMRI data acquisition at USC (Section II); describes associated tools for data analysis (Section III); reviews illustrative applications (Section IV), and discusses challenges and future directions (Section V).

II. DATA ACQUISITION

The first two subsections of this section briefly discuss some technical details of the acquisition and reconstruction protocols that have been used most extensively at USC, including for the USC-TIMIT corpus. The third subsection discusses some alternative protocols and recent developments. Note that several details are shared among protocols. This will be implied unless otherwise noted.

A) Imaging

The upper airways of the subjects are imaged while they lay supine in the MRI scanner. Subjects have their heads firmly but comfortably padded at the temples to minimize motion of the head. Stimuli are presented in large text on a back-projection screen, from which subjects can read from inside the scanner bore without moving their head. The nature of the experiment and the protocol is explained to subjects before they enter the scanner, and subjects are paid for their time upon completion of the session. The overall recording time for each subject includes calibration and breaks in-between stimuli. The USC Institutional Review Board has previously approved the data collection procedures.

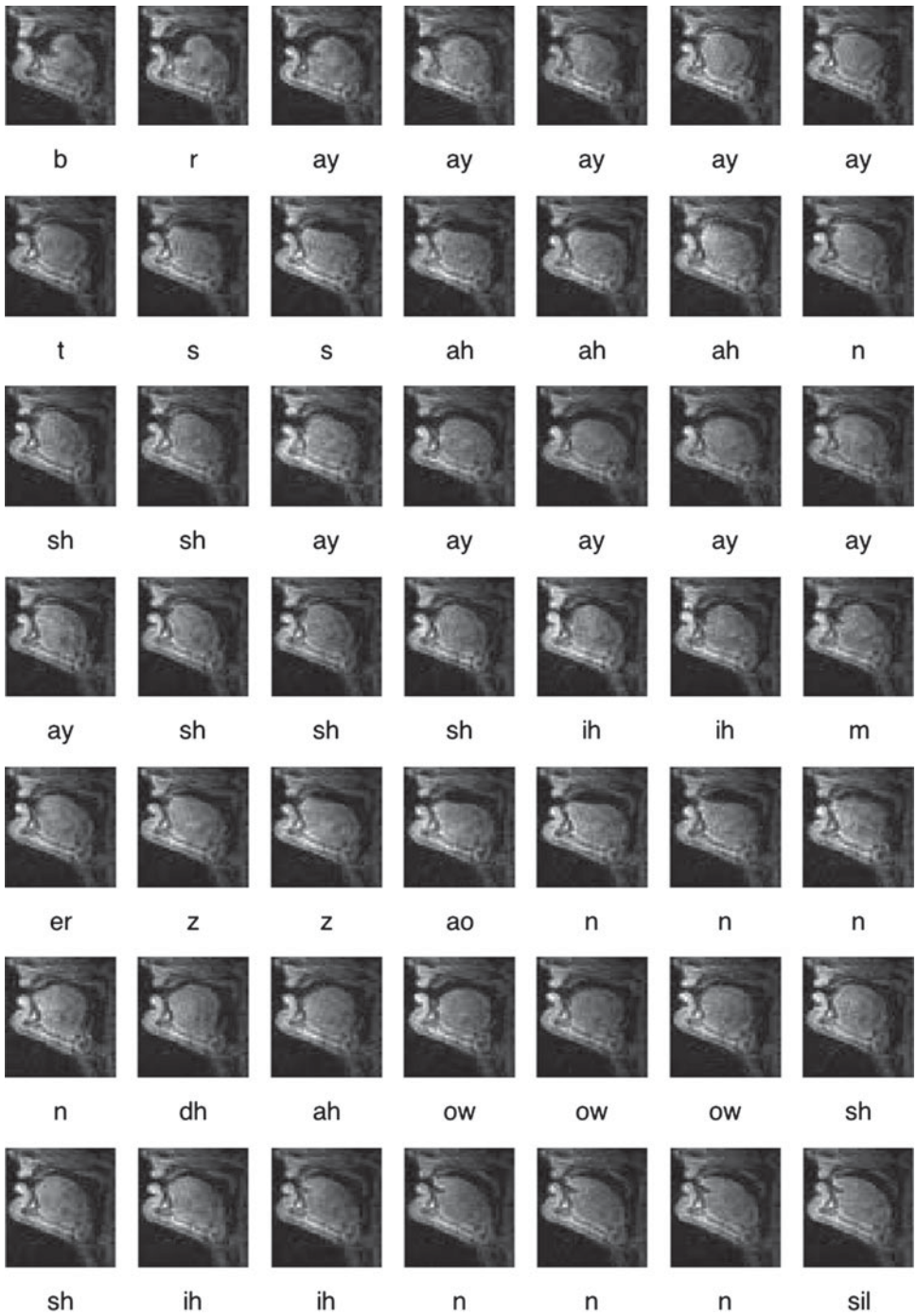

Fig. 2. Example rtMRI sequence from the USC-TIMIT database. A male subject utters the sentence "Bright sunshine shimmers on the ocean" (one of the 460 MOCHA-TIMIT sentences included for each subject). Note that there is a zoom into the frames, as compared to Fig. 1. The phonetic labels are a result of automatic alignment. The symbol "sp" stands for "space" and "sil" for "silence".

Data are 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. A body coil is used for radio frequency (RF) signal transmission. A custom upper airway receiver coil array is used for RF signal reception. This four-channel array includes two anterior coil elements and two coil elements posterior to the head and neck. However, only the two anterior coils are used for data acquisition. The posterior coils are not used because they have been previously shown to result in aliasing artifacts.

The rtMRI acquisition protocol is based on a spiral fast gradient echo sequence. This is a scheme for sampling the spatial frequency domain (k-space) in which data are acquired in spiraling patterns. Thirteen interleaved spirals together form a single image. Each spiral is acquired over 6.164 ms (repetition time (TR), which includes slice excitation, readout, and gradient spoiler) and thus every image comprises information spanning $13 \times 6.164 = 80.132$ ms. A sliding window technique is used to allow for view sharing and thus increase frame rate [13]. The TR-increment for view sharing is seven

Table 1. Technical details of four extensively used rtMRI sequences

	Sequence 1	Sequence 2	Sequence 3	Sequence 4
Magnetic field strength	1.5 Tesla	1.5 Tesla	1.5 Tesla	1.5 Tesla
Gradients	ZOOM	ZOOM	ZOOM	ZOOM
Spatial gradient Max. amplitude	22 mT/m	40 mT/m	40 mT/m	40 mT/m
waveform design Max. slew rate	77 T/m/s	150 T/m/s	150 T/m/s	150 T/m/s
Coil	4-channel	4-channel	4-channel	8-channel
Slice thickness	5 mm	5 mm	5 mm	6 mm
Readout duration (T_{read})	2.552 ms	2.520 ms	2.584 ms	2.520 ms
Repetition time (TR)	6.164 ms	6.004 ms	6.028 ms	6.004 ms
Field of view (FOV)	20 cm × 20 cm	20 cm × 20 cm	20 cm × 20 cm	20 cm × 20 cm
Spatial resolution	3.0 mm × 3.0 mm	2.4 mm × 2.4 mm	3.0 mm × 3.0 mm	2.4 mm × 2.4 mm
Pixel dimension	68 × 68	84 × 84	68 × 68	84 × 84
Number of interleaves	13	13	9	2 (Golden angle interleaving)
Relative SNR efficiency	1.00	0.63	0.83	(Not assessed)
Time to acquire full image	80.1 ms	78.1 ms	54.3 ms	12 ms
View-sharing TR-increment	7	7	5	(No view-sharing)
Reconstruction frame-rate	23.18 fps	23.79 fps	33.18 fps	83 fps

acquisitions, which results in the generation of an MRI movie with a frame rate of 1/(7 × TR) = 1/ (7 × 6.164 ms) = 23.18 frames/s [13, 14, 31].

The imaging field of view is 200 mm × 200 mm, the flip angle is 15°, and the receiver bandwidth ±125 kHz. Slice thickness is 5 mm, located midsagittally; image resolution in the sagittal plane is 68 × 68 pixels (2.9 mm × 2.9 mm). Scan plane localization of the midsagittal slice is performed using RTHawk (HeartVista, Inc., Los Altos, CA), a custom real-time imaging platform [32].

MR image reconstruction is performed using MATLAB (Mathworks, South Natick, MA). Images from each of the two anterior coils of the four-channel coil array are formed using gridding reconstruction [14, 33]; and the two images are combined by taking their root sum-of-squares in order to improve image signal-to-noise ratio (SNR) and spatial coverage of the vocal tract.

B) Audio acquisition

Acquiring and synchronizing the acoustic signal with the MRI data – which is crucial in order to facilitate the interpretation and analysis of the articulatory information in the speech production videos – presents numerous technical challenges. In the deployed system, audio is simultaneously recorded at a sampling frequency of 20 kHz inside the MRI scanner while subjects are imaged, using a fiber-optic microphone (Optoacoustics Ltd., Moshav Mazor, Israel) and custom recording and synchronization setup. The audio signal is controlled through the use of a sample clock derived from the scanner's 10 MHz master clock, and triggered using the scanner RF master-exciter un-blank signal, which is a TTL (Transistor–Transistor Logic) signal synchronous to the RF pulse.

Apart from synchronization, another challenge to acquiring good quality audio is the high noise level generated by the operation of the MRI scanner. It is important that this noise be canceled satisfactorily in order to perform further detailed analyses of the audio for linguistic and statistical modeling purposes. For the sequences in Table 1, the MRI noise has a specific periodic structure, which enables noise cancellation using a custom adaptive signal processing algorithm which exactly takes into account this periodic structure [34]. See Fig. 3 for an example of noise cancellation.

Note that subjects wear earplugs for protection from the scanner noise, but are still able to hear loud conversation in the scanner room and to communicate effectively with the experimenters via both the fiber-optic microphone setup as well as the in-scanner intercom system.

C) Alternative protocols

Three more rtMRI acquisition protocols based on spiral fast gradient echo sequences have been extensively used, according to the purpose of the specific experiment. The technical details of the sequences employed are summarized in Table 1. Sequence 1 in the table is the one described in the previous subsections. Sequences 2 and 3, like Sequence 1, make use of the four-channel coil array already discussed. The more recent Sequence 4 makes use of an eight-channel array that has four elements on either side of the jaw. Sequence 4 combines fast spirals with sparse sampling and constrained reconstruction, enabling frame rates of up to 83-frames/s and multi-slice imaging [35].

Sequence 1 is the most efficient in terms of SNR, i.e. it provides clearer images than, at least, Sequences 2 and 3. The SNR of Sequence 4 is very difficult to quantify, as this is coupled with constrained reconstruction through a nonlinear process. Visual inspection of data collected with Sequence 4 shows no degradation of the image quality compared with Sequence 1. Imaging of the area around the glottis is improved as a result of the eight-channel coil array configuration.

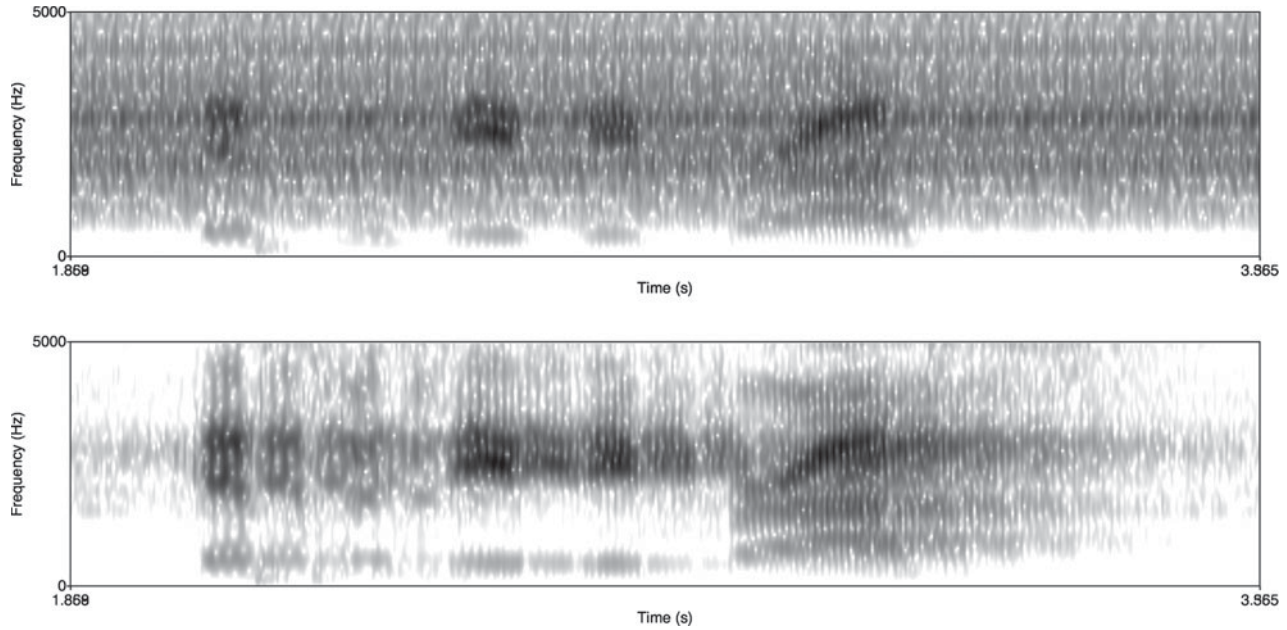

Fig. 3. Spectrograms of the audio, recorded concurrently with the rtMRI data, for the utterance "This was easy for us" spoken by a female subject before (top) and after (bottom) de-noising.

Audio de-noising for Sequences 2 and 3 is done using the same method as that for Sequence 1. However, Sequence 4 does not exhibit the same periodic structure as the other sequences. To achieve its de-noising, an audio enhancement method using dictionary learning and wavelet packet analysis that does not rely on periodicity has been recently developed [36].

We finally note that the USC team has also developed a protocol for accelerated static volumetric upper-airway MRI acquisition, which captures the 3D volume of the upper airway in as fast as 7 s [37, 38]. This has enabled capturing the 3D articulation of the full set of continuant English phonemes from several subjects, with no particular difficulty in sustaining the speech sounds for the required amount of time.

III. DATA ANALYSIS TOOLS

While some speech production phenomena may be studied by manually inspecting the raw rtMRI data and measuring the timing of articulatory events identified in these image sequences [e.g., 19], many other aspects of speech production require additional signal processing and analysis. A number of tools to aid inspection and analysis of rtMRI data have been developed at USC.

A) Data inspection and labeling

A graphical user interface (GUI) has been developed to allow for audition, labeling, tissue segmentation, and acoustic analysis of rtMRI data. 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 also facilitates automatic formant and pitch tracking, and rapid semi-automatic segmentation of the upper airway in sequences of video frames, for visualization of tongue movement, or as a precursor to dynamic parametric analysis of vocal tract shaping. Fig. 4 shows a screenshot of this GUI.

B) Automatic articulator tracking

By identifying air-tissue boundaries in rtMRI images, the position and configuration of articulators can be compared at different points in time. Vocal tract cross-distances may also be calculated, and changes in lingual posture can be examined during the production of different speech segments. For many types of speech, vocal tract outlines may be tracked using semi-automatic or fully automatic identification of tissue boundaries in rtMRI data.

Unsupervised segmentation of regions corresponding to the mandibular, maxillary, and posterior areas of the upper airway has been achieved by exploiting spatial representations of these regions in the frequency domain, the native domain of MRI data [39]. The segmentation algorithm uses an anatomically informed object model, and returns a set of tissue boundaries for each frame of interest, allowing for quantification of articulator movement and vocal tract aperture in the midsagittal plane. The method makes use of alternate gradient vector flows, non-linear least-squares optimization, and hierarchically optimized gradient descent procedures to refine estimates of tissue locations in the vocal tract. Thus, the method is automatic and well suited for processing long sequences of MR images. Fig. 5 shows an example of air-tissue boundaries produced by this algorithm. Obtaining such vocal tract contours enables the calculation of vocal-tract midsagittal cross-distances, which in turn can

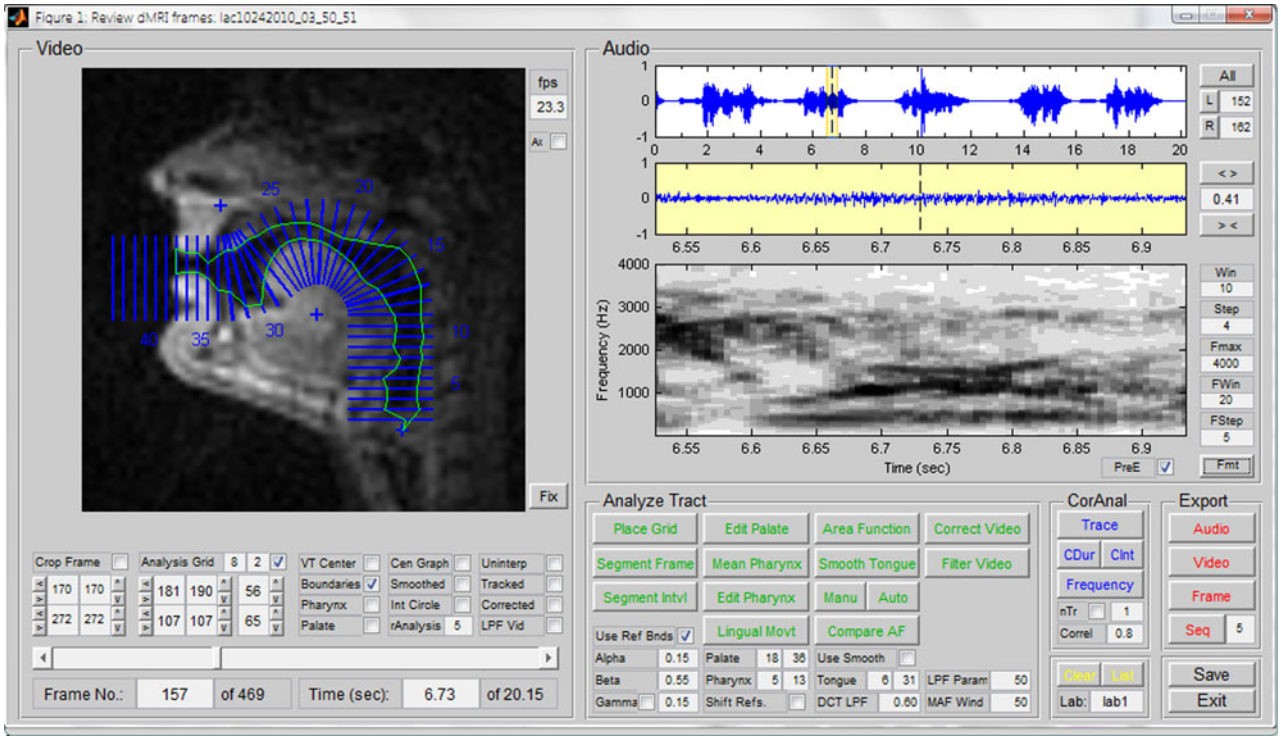

Fig. 4. GUI allowing for audition, labeling, tissue segmentation, and acoustic analysis of the rtMRI data, displaying an example of parametric segmentation.

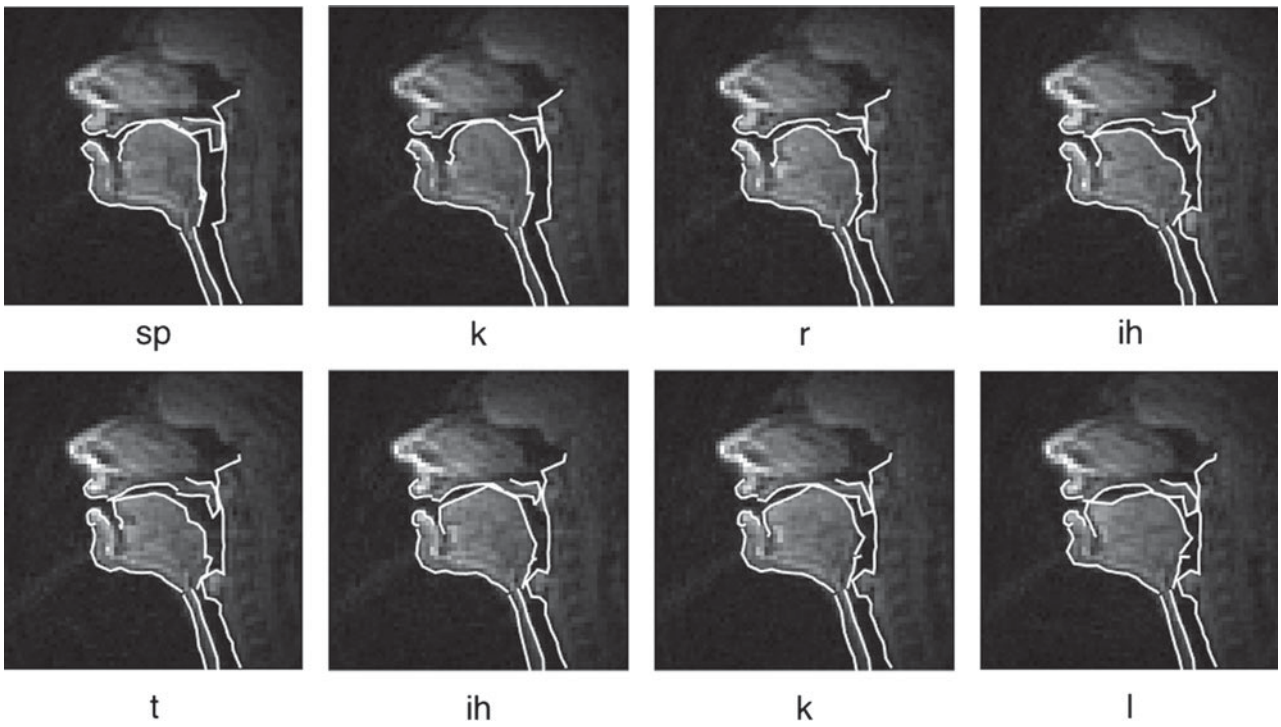

Fig. 5. Example of region segmentation (white outlines) of articulators in rtMRI data. The word uttered by the female subject is "critical". The symbol "s" stands for "space".

be used to estimate area functions, via standard reference sagittal-to-area transformations [40–42].

The above segmentation method requires significant computational resources. As a faster (yet less accurate) alternative, a method of rapid semi-automatic segmentation of rtMRI data for parametric analysis has been developed, 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 [43, 44]. A version of this method has been integrated in the aforementioned GUI (see Fig. 4).

An optional pre-processing step before the application of these segmentation algorithms is the correction of any

brightness gradient in the rtMRI sequences, which is a result of the coil configuration. To this end, a thin-plate spline-based intensity correction procedure [45] is applied, to obtain an estimate of the combined coil sensitivity map, which is constant for all images contained in the sequence. Thus, corrected maximally flat magnitude images can be obtained [39].

C) Direct image analysis

While boundary detection is important for capturing the posture of individual articulators at different points in time, it is often enough to observe the dynamics of the formation and release of constrictions in different regions of the vocal tract [46]. Pixel intensity in an MR image is indicative of the presence or absence of soft tissue; as a result, tissue movement into and out of a region of interest in the upper airway may be estimated by calculating the change in mean pixel intensity in the vicinity of that region. Using this concept, a direct image analysis method has been developed that by-passes the need to first identify tissue boundaries in the upper airway [47, 48]. Constriction location targets may be automatically estimated by identifying regions of maximally dynamic correlated pixel activity along the palate and at the lips, and closure and release gesture timings may be estimated from landmarks in the velocity profile derived from the smoothed intensity function [49].

IV. APPLICATIONS

The capability of vocal tract rtMRI data acquisition creates research opportunities for new and deeper insights in a number of areas. The promise held by these data and methods has already begun to be realized in a number of domains, from phonetics and phonological theory research to speech technology research. In this section, some findings of the USC team and applications that showcase the utility of rtMRI as an emerging tool for speech research are briefly summarized.

A) Compositionality of speech production

The USC team has been combining the rtMRI technology with linguistically informed analysis of vocal tract constriction actions in order to investigate the production and cognitive control of the compositional action units of spoken language. Of particular interest is the framework of Articulatory Phonology [50], which provides a theoretical foundation for the team's work. Note that this effort has required the collection of specifically tailored rtMRI data, besides general-purpose data, such as those of the USC-TIMIT database.

Speech is dynamic in nature: it is realized through time-varying changes in vocal tract shaping, which emerge lawfully from the combined effects of multiple constriction events distributed over space (i.e. subparts of the vocal tract) and over time. Understanding this dynamic aspect is fundamental to linguistic studies and is intended through the USC team's research to be added to the fields current – essentially static – approach to describing speech production.

RtMRI allows pursuing such a goal through examining the decomposition of speech into such cognitively controlled vocal tract constriction events, or gestures. Of specific interest are: (i) the compositionality in space, i.e. the deployment of concurrent gestures distributed spatially, over distinct constriction effectors within the vocal tract; (ii) the compositionality in time, i.e. the deployment of gestures temporally; and (iii) the characterization of articulatory setting, i.e. the set of postural configurations that the vocal tract articulators tend to be deployed from and return to in the process of producing fluent and natural speech.

An example study on the compositionality of speech production in space examined retroflex stops and rhotics in Tamil [51]. The study revealed that in some contexts these consonants may be achieved with little or no retroflexion of the tongue tip. Rather, maneuvering and shaping of the tongue in order to achieve post-alveolar contact varies across vowel contexts. Between back vowels /a/ and /u/, post-alveolar constriction involves curling back of the tongue tip, but in the context of the high front vowel /i/, the same constriction is achieved by tongue bunching. Results supported the notion that so-called retroflex consonants have a specified target constriction in the post-alveolar region, but that the specific articulations employed to achieve this constriction are not fixed.

An example line of research on the compositionality in time examined the coordination of velic and oral gestures for nasal consonants. For English /n/ [18], it was found that near-synchrony of velum lowering and tongue tip raising characterizes the timing for onsets, while temporal lag between the gestures is characteristic for codas, supporting and extending previous findings for /m/ [52]. In French, which, unlike English, uses nasal vowels, the coordination of velic and oral gestures was found to be more tightly controlled, to allow the distinction between nasal vowels and consonants [17]. But, while the nature of the coordinative relation was different between French and English, the timing of the corresponding gestures varied in the same way as a function of prosodic context.

Regarding the characterization of articulatory setting, research at USC supported the hypothesis that pauses at major syntactic boundaries (i.e. grammatical pauses), but not ungrammatical (e.g. word search) pauses, are planned by a high-level cognitive mechanism that also controls the rate of articulation around these junctures [53]. The hypothesis was that postures adopted during grammatical pauses in speech are more mechanically advantageous compared to postures assumed at absolute rest, i.e. that equal changes in articulatory posture result to greater changes in the space of speech tasks. This hypothesis was verified using locally weighted linear regression to estimate the forward map from low-level articulator variables to high-level task variables [54]. The analysis showed that postures assumed during grammatical pauses in speech, as well as speech-ready

postures, are significantly more mechanically advantageous than postures assumed during absolute rest.

B) Speaker specificity

Speakers have diverse vocal-tract morphologies, which affect their speech production (note, for example, the different vocal tracts of the ten USC-TIMIT speakers in Fig. 1). The USC team has started using rtMRI data, collected from diverse speakers, to study how individual vocal morphological differences are reflected in the acoustic speech signal and what articulatory strategies are adopted in the presence of such morphological differences to achieve speech invariance, either perceptual or acoustic. The capability of the USC team to collect large volumes of data from diverse speakers is crucial to this effort.

Initial work with rtMRI has focused on individual differences in the size, shape, and relative proportions of the hard palate and posterior pharyngeal wall. Specific aims were: to characterize such differences [55]; to examine how they relate to speaker-specific articulatory and acoustic patterns [56]; and to explore the possibility of predicting them automatically from the acoustic signal [57].

The long-term objective of this ongoing work is to improve scientific understanding of how vocal-tract morphology and speech articulation interplay and explain the variant and invariant aspects of speech signal properties within and across talkers.

This line of research may benefit automatic speaker recognition technology. State-of-the-art automatic speaker-recognition methods yield strong results over a range of read and spontaneous speech domains, utterance lengths, and noise conditions [58–60]. In several studies, the technology performs better than even trained human listeners [61]. Despite considerable success in automatic speaker recognition, technologies are not informative about articulatory differences between speakers. RtMRI data can be used to improve the interpretability of such systems by associating acoustic differences to articulatory ones [62].

C) Articulatory-acoustic maps

Benefits from rtMRI are also expected in the context of studying the forward map from articulation to acoustics (or, articulatory synthesis) and the inverse (acoustic-to-articulatory) mapping. Note that these problems have been classically addressed without taking into account speaker variability.

Characterizing the many-to-one mapping from representations in the articulatory space to those in the acoustic space is a central problem in phonological theory [63, 64]. The problem is compounded by our incomplete knowledge of the articulatory goals of production. Data from rtMRI provide a rich new source of information, which can inform research in this domain. This, in turn, can simplify the modeling of the articulatory-acoustic map and lead to more accurate estimates of articulatory features from the acoustic signal in acoustic-to-articulatory inversion. Since rtMRI provides rich information of the speech production process, an analysis of the non-uniqueness in articulatory-to-acoustic mappings using various rtMRI derived features can be performed to provide insight into the relationship between various articulatory features and the non-uniqueness in the mapping.

An important tool to be in place in order to achieve the above research goals is an articulatory synthesizer, i.e. a simulation of the articulatory-to-acoustic relationship in the vocal tract. Work has been done using Maeda's time-domain vocal tract simulation [65] to synthesize speech on the basis of EMA data [66], with the full midsagittal vocal tract profile being inferred from EMA using Maeda's articulatory model [40]. RtMRI, on the other hand, readily provides (i.e. after segmentation described in [39]) the full midsagittal profiles, and ongoing work aims at using rtMRI information for articulatory synthesis. Note that the synthesizer addresses the problem of synthesizing running (co-articulated) speech, and can be adapted to reflect different vocal-tract morphologies.

D) The potential for ASR

Dynamic articulatory data have the potential to inform approaches to ASR [67, 68]. Since it provides such a rich source of global information about vocal tract dynamics during speech production, the discriminatory power of rtMRI-derived production features may help realize this potential in ASR. Additionally, examining the extent to which production-oriented features can provide information complementary to that provided by acoustic features can offer further insights into the role of articulatory knowledge in ASR [69, 70].

From a more theoretical viewpoint, there have been several well-known hypotheses regarding the relation between production and perception systems in human speech communication [71, 72]. Quantitatively modeling these relationships in order to develop better models of automatic speech and speaker recognition is a very challenging task that can benefit vastly from the availability of rich speech production data. For example, using mutual information as a metric, it has been shown in a data-driven manner that the non-uniform auditory filterbank in the human ear (receiver) is optimal in providing least uncertainty in decoding articulatory movements in the human speech production system (transmitter) [73]. This finding indicates that the design of the filterbank for speech recognition systems needs to be optimally designed with respect to the characteristics of the speech production system.

More such computational models need to be developed in order to understand the effect of speaker dependence, language effect, pathologies and paralinguistic features in speech and speaker recognition tasks, particularly to discover robust recognition models. RtMRI data may be central to such an effort.

V. CONCLUDING REMARKS

The present paper has discussed several advances in rtMRI technology and data analysis methods, with ongoing and envisioned lines of research based on these advances. With current imaging and audio acquisition capabilities, it is possible to collect: (i) data tailored to the goals of specific linguistic studies; and (ii) large amounts of general-purpose speech production data that open up novel corpus-driven scientific research as well as technological efforts such as in automatic speech and speaker recognition. The USC-TIMIT database, which consists of midsagittal rtMRI data from ten speakers who produce the 460-sentence MOCHA-TIMIT corpus, with complementary EMA data from four of these speakers producing the same corpus, and a collection of supporting analysis tools, has been made freely available to the research community.

Recent developments continue to increase the spatiotemporal resolution of rtMRI. The novel Sequence 4 has a temporal resolution at 12 ms, which is sufficiently fine to capture accurately fast aerodynamic events, like those in the production of trills, and latencies involved in interarticulator coordination. The nominal frame rates of Sequences 1–3 (in Table 1) are adequate for visualization of articulatory postures and movements, especially in the context of studying compositionality in space. Note that these frame rates can be increased by changing the TR-increment for view sharing down to one TR (which nevertheless does not change the time needed to acquire a full image) for better exploring compositionality in time [17]. It is also imaginable to leverage the much higher temporal resolution of EMA data, either via co-registration, or by using EMA to animate models built from rtMRI data.

RtMRI is not restricted to imaging dynamically the midsagittal slice of the vocal tract but can also image other slices of interest to the study of speech production, such as parasagittal, coronal, axial or oblique. We have recently demonstrated the possibility of acquiring, in parallel, images from multiple slices of the vocal tract [20, 35]. Our goal is to build upon the foundation of the USC-TIMIT database, by adding data from slices of interest other than the midsagittal, with higher spatio-temporal resolutions, acquired from more speakers both of English and other languages, and to expand the toolset to allow for more sophisticated inspection and analysis of these data.

RtMRI for speech production research presents some shortcomings, most of which are open research topics for the USC team. First, rtMRI is currently done in a supine position, which is not a natural posture for speech, almost exclusively performed in the upright position. Much literature has been devoted to the assessment of differences in speech articulation between the two positions [74–77], and it has been suggested that the differences seem limited and that compensatory mechanisms, at least in healthy subjects, appear to be sufficiently effective to allow the acquisition of meaningful speech data in the supine position [26]. The potential use of upright, or open-type, scanners would fully remove this consideration, and there have been a few studies that demonstrate the utility of such scanners in upper-airway MRI [78, 79].

The MRI scanner is a very noisy environment, and subjects need to wear earplugs during acquisition, thus not having natural auditory feedback. Though it may be reasonable to expect that the subjects would speak much louder than normal, or that their articulation would be significantly affected as a result, it was observed on our site that, in practice, these statements held true only for rare cases of subjects. It is possible that somatosensory feedback compensates for the shortage of auditory feedback [80, 81]. Expert phoneticians that participated as subjects in rtMRI data collections at USC reported that the lack of auditory feedback presented a problem only when they tried to produce certain speech sounds *not* present in their native languages.

Because of the magnetic fields involved, people need to be excluded from being subjects in speech MRI research if they have prosthetics such as pacemakers or defibrillators, which are identified in a screening process [82]. People with a history of claustrophobia need also be excluded [83]. Otherwise, subject comfort is usually not an issue for adult healthy subjects, and for observed scan durations (overall time spent in the scanner) of less than 90 min.

Dental work is not a safety concern, but may pose problems in imaging. However, the disruptions associated with it do not consistently degrade image quality. In general, image quality is subject-dependent and in some cases it can be difficult to even maintain constant quality throughout the speech sample [84]. We have seen on our site that the impact of dental work appears to be more prominent when such work resides on the plane that is imaged, and often localized around the dental work: for example, orthodontic permanent retainers at the upper incisors result in loss of midsagittal visual information from a small circle (typically with diameter up to 3 cm) around the upper incisors.

The teeth themselves are invisible in MRI, because of their chemical composition. Various methods have been used to superimpose the teeth onto MRI images, including using data from supplementary CT imaging [85], dental casts [86, 87], or MRI data acquired using a contrast agent in the oral cavity such as blueberry juice [88] or ferric ammonium citrate [89], leaving the teeth as signal voids. Superimposing the teeth on rtMRI sequences would be useful for the exact modeling of anterior fricative consonants. At the time of writing, the data disseminated to the research community by the USC team do not include information on the teeth.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institutes of Health under award number R01DC007124. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] Schönle, P.W.; Gräbe K.; Wenig, P.; Höhne, J.; Schrader, J.; Conrad, B.: Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain Lang., **31** (1) (1987), 26–35.

[2] Westbury, J.; Turner, G.; Dembowski, J.: X-ray Microbeam Speech Production Database User's Handbook. Technical Report, Waisman Center on Mental Retardation and Human Development, University of Wisconsin, 1994.

[3] Hardcastle, W.; Jones, W.; Knight, C.; Trudgeon, A.; Calder, G.: New developments in electropalatography: a state-of-the-art report. Clin. Linguistics Phonetics, **3** (1) (1989), 1–38.

[4] Whalen, D. *et al.*: The Haskins optically corrected ultrasound system (HOCUS). J. Speech, Lang. Hearing Res., **48** (3) (2005), 543–553.

[5] Stone, M.: Imaging and measurement of the vocal tract, in Brown, K. (ed.), *Encyclopedia of Language and Linguistics*, Elsevier, Oxford, 2006, 526–539.

[6] Delattre, P.: Pharyngeal features in the consonants of Arabic, German, Spanish, French, and American English. Phonetica, **23** (1971), 129–155.

[7] Wood, S.: X-ray and model studies of vowel articulation, PhD thesis, Lund University, 1982.

[8] Munhall, K.; Vatikiotis Bateson, E.; Tohkura, Y.: X-ray film database for speech research. J. Acoust. Soc. Am., **98** (2) (1995), 1222–1224.

[9] Badin, P. *et al.*: Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model, in *Int. Congress on Acoustics*, Trondheim, Norway, June 1995.

[10] Giles, S.; Moll, K.: Cinefluorographic study of selected allophones of English /l/. Phonetica, **31** (3–4) (1975), 206–227.

[11] Stone, M. *et al.* Modeling the motion of the internal tongue from tagged cine-MRI images. J. Acoust. Soc. Am., **109** (6) (2001), 2974–2982.

[12] Takemoto, H.; Honda, K.; Masaki, S.; Shimada, Y.; Fujimoto, I.: Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. J. Acoust. Soc. Am., **119** (2) (2006), 1037–1049.

[13] 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) (2004), 1771–1776.

[14] Bresch, E.; Kim, Y.-C.; Nayak, K.; Byrd, D.; Narayanan, S.: Seeing speech: capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. IEEE Signal Process. Mag., **25** (3) (2008), 123–132.

[15] Hagedorn, C.; Proctor, M.; Goldstein, L.: Automatic analysis of singleton and geminate consonant articulation using real-time magnetic resonance imaging, in *Interspeech*, Florence, Italy, August 2011.

[16] Israel, A.; Proctor, M.; Goldstein, L.; Iskarous, K.; Narayanan, S.: Emphatic segments and emphasis spread in Lebanese Arabic: a real-time magnetic resonance imaging study, in *Interspeech*, Portland, OR, September 2012.

[17] Proctor, M.; Goldstein, L.; Lammert, A.; Byrd, D.; Toutios, A.; Narayanan, S.: Velic coordination in French nasals: a real time magnetic resonance imaging study, in *Interspeech*, Lyon, France, August 2013.

[18] Byrd, D.; Tobin, S.; Bresch, E.; Narayanan, S.: Timing effects of syllable structure and stress on nasals: a real-time MRI examination. J. Phonetics, **37** (1) (2009), 97–110.

[19] Proctor M.; Walker, R.: Articulatory bases of English liquids. In Parker, S. (ed.), The Sonority Controversy, volume 18 of Studies in Generative Grammar, *De Gruyter*, Berlin, 2012, 285–312.

[20] Kim, Y.-C.; Proctor, M.; Narayanan, S.; Nayak, K.: Improved imaging of lingual articulation using real-time multislice MRI. J. Magn. Reson. Imaging, **35** (4) (2012), 943–948.

[21] Sutton, B.; Conway, C.; Bae, Y.; Brinegar, C.; Liang, Z.-P.; Kuehn, D.: Dynamic imaging of speech and swallowing with MRI, in *Int. Conf. IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN, September 2009.

[22] Uecker, M.; Zhang, S.; Voit, D.; Karaus, A.; Merboldt, K.-D.; Frahm, J.: Real-time MRI at a resolution of 20 ms. NMR Biomed., **23** (8) (2010), 986–994.

[23] Scott, A.; Boubertakh, R.; Birch, M.; Miquel, M.: Towards clinical assessment of velopharyngeal closure using MRI: evaluation of real-time MRI sequences at 1.5 and 3 T. The Br. J. Radiol., **85** (1019) (2012), e1083–e1092.

[24] Zhang, S.; Olthoff, A.; Frahm, J.: Real-time magnetic resonance imaging of normal swallowing. J. Magn. Reson. Imaging, **35** (6) (2012), 1372–1379.

[25] Niebergall, A. *et al.*: Real-time MRI of speaking at a resolution of 33 ms: undersampled radial FLASH with nonlinear inverse reconstruction. Magn. Reson. Med., **69** (2) (2013), 477–485.

[26] Scott, A.D.; Wylezinska, M.; Birch, M.J.; Miquel, M.E.: Speech MRI: morphology and function. Phys. Med., **30** (6) (2014), 604–618.

[27] Iltis, P.W.; Frahm, J.; Voit, D.; Joseph, A.A.; Schoonderwaldt, E.; Altenmüller, E.: High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. Quant. Imaging Med. Surg., **5** (3) (2015), 374–381.

[28] Narayanan, S. *et al.*: Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). J. Acoust. Soc. Am., **136** (3) (2014), 1307–1311.

[29] Wrench, A.; Hardcastle, W.: A multichannel articulatory speech database and its application for automatic speech recognition, in *5th Seminar on Speech Production*, Kloster Seeon, Germany, 2000, 305–308.

[30] Childers, D.; Krishnamurthy, A.: A critical review of electroglottography. Crit. Rev. Biomed. Eng., **12** (2) (1984), 131–161.

[31] Kim, Y.-C.; Narayanan, S.; Nayak, K.: Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order. Magn. Reson. Med., **65** (5) (2011), 1365–1371.

[32] Santos, J.; Wright, G.; Pauly, J.: Flexible real-time magnetic resonance imaging framework, in *Int. Conf. IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, September 2004.

[33] Jackson, J.; Meyer, C.; Nishimura, D.; Macovski, A.: Selection of a convolution function for Fourier inversion using gridding. IEEE Trans. Med. Imaging, **10** (3) (1991), 473–478.

[34] Bresch, E.; Nielsen, J.; Nayak, K.; Narayanan, S.: Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. J. Acoust. Soc. Am., **120** (4) (2006), 1791–1794.

[35] Lingala, S.; Zhu, Y.; Kim, Y.-C.; Toutios, A.; Narayanan, S.; Nayak, K.: High spatio-temporal resolution multi-slice real time MRI of speech using golden angle spiral imaging with constrained reconstruction, parallel imaging, and a novel upper airway coil, in *Int. Society of Magnetic Resonance in Medicine Scientific Sessions*, Toronto, Canada, May 2015.

[36] Vaz, C.; Ramanarayanan, V.; Narayanan, S.: A two-step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis, in *Interspeech*, Lyon, France, August 2013.

[37] Kim, Y.-C.; Narayanan, S.; Nayak, K.: Accelerated 3D upper airway MRI using compressed sensing. Magn. Reson. Med., **61** (6) (2009), 1434–1440.

[38] Kim, Y.-C. *et al.*: Toward automatic vocal tract area function estimation from accelerated three-dimensional magnetic resonance imaging, in *ISCA Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, August 2013.

[39] Bresch, E.; Narayanan, S.: Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. IEEE Trans. Med. Imaging, **28** (3) (2009), 323–338.

[40] Maeda, S.: Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. in Hardcastle, W.; Marchal, A. (eds.), Speech Production and Speech Modelling, *Kluwer Academic Publisher*, Amsterdam, 1990, 131–149.

[41] Soquet, A.; Lecuit, V.; Metens, T.; Demolin, D.: Mid-sagittal cut to area function transformations: direct measurements of mid-sagittal distance and area with MRI. Speech Commun., **36** (2002), 169–180.

[42] McGowan, R.S.; Jackson, M.T.-T.; Berger, M.A.: Analyses of vocal tract cross-distance to area mapping: an investigation of a set of vowel images. J. Acoust. Soc. Am., **131** (2012), 424–434.

[43] Proctor, M.; Bone, D.; Narayanan, S.: Rapid semi-automatic segmentation of real-time Magnetic Resonance Images for parametric vocal tract analysis, in *Interspeech*, Makuhari, Japan, September 2010.

[44] Kim, J.; Kumar, N.; Lee, S.; Narayanan, S.: Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data, in *Int. Seminar on Speech Production*, Cologne, Germany, May 2014.

[45] Liu, C.; Bammer, R.; Moseley, M.E.: Parallel imaging reconstruction for arbitrary trajectories using k-space sparse matrices (kSPA). Magn. Reson. Med., **58** (6) (2007), 1171–1181.

[46] Browman, C.; Goldstein, L.: Towards an articulatory phonology. Phonol. Yearbook, **3** (1986), 219–252.

[47] Lammert, A.; Proctor, M.; Narayanan, S.: Data-driven analysis of real-time vocal tract MRI using correlated image regions, in *Interspeech*, Makuhari, Japan, September 2010.

[48] Lammert, A.; Ramanarayanan, V.; Proctor, M.; Narayanan, S.: Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis, in *Interspeech*, Lyon, France, August 2013.

[49] Proctor, M.; Katsamanis, A.; Goldstein, L.; Hagedorn, C.; Lammert, A.; Narayanan, S.: Direct estimation of articulatory dynamics from real-time magnetic resonance image sequences, in *Interspeech*, Florence, Italy, August 2011.

[50] Goldstein, L.; Fowler, C.A.: Articulatory phonology: a phonology for public language use, in Meyer, A.S.; Schiller, N.O. (eds.), *Phonetics and phonology in language comprehension and production: differences and similarities*, 2003, 159–207.

[51] Smith, C.; Proctor, M.; Iskarous, K.; Goldstein, L.; Narayanan, S.: Stable articulatory tasks and their variable formation: Tamil retroflex consonants, in *Interspeech*, Lyon, France, August 2013.

[52] Krakow, R.: Nonsegmental influences on velum movement patterns: syllables, sentences, stress, and speaking rate. Nasals, Nasalization, and the Velum, **5** (1993), 87–118.

[53] Ramanarayanan, V.; Lammert, A.; Goldstein, L.; Narayanan, S.: Are articulatory settings mechanically advantageous for speech motor control? PLoS ONE, **9** (8) (2014), e104168.

[54] Lammert, A.; Goldstein, L.; Narayanan, S.; Iskarous, K.: Statistical methods for estimation of direct and differential kinematics of the vocal tract. Speech Commun., **55** (2013), 147–161.

[55] Lammert, A.; Proctor, M.; Narayanan, S.: Morphological variation in the adult hard palate and posterior pharyngeal wall. J. Speech, Lang., Hearing Res., **56** (2) (2013), 521–530.

[56] Lammert, A.; Proctor, M.; Narayanan, S.: Interspeaker variability in hard palate morphology and vowel production. J. Speech, Lang., Hearing Res., **56** (6) (2013), S1924–S1933.

[57] Li, M.; Lammert, A.; Kim, J.; Ghosh, P.; Narayanan, S.: Automatic classification of palatal and pharyngeal wall shape categories from speech acoustics and inverted articulatory signals, in *ISCA Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, August 2013.

[58] Martin, A.F.; Greenberg, C.S.; Howard, J.M.; Doddington, G.R.; Godfrey, J.J.; Stanford, V.M.: Effects of the new testing paradigm of the 2012 NIST speaker recognition evaluation, in *Odyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014, 1–5.

[59] Mccree, A. *et al.*: The NIST 2014 speaker recognition i-vector machine learning challenge, in *Odyssey 2014*, 2014, 224–230.

[60] Kinnunen, T.; Li, H.: An overview of text-independent speaker recognition: from features to supervectors. Speech Commun., **52** (1) (2010), 12–40.

[61] Shen, W.; Campbell, J.; Schwartz, R.: Human error rates for speaker recognition. The J. Acoust. Soc. Am., **130** (4) (2011), 2547–2547.

[62] Rose, P.: Technical forensic speaker recognition: evaluation, types and testing of evidence. Comput. Speech Lang., **20** (2–3) (2006), 159–191.

[63] Atal, B.; Chang, J.; Mathews, M.; Tukey, J.: Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Acoust. Soc. Am., **63** (5) (1978), 1535–1555.

[64] Ghosh, P.; Narayanan, S.: A generalized smoothness criterion for acoustic-to-articulatory inversion. The Journal of the Acoustical Society of America, **128** (4) (2010), 2162–2172.

[65] Maeda, S.: A digital simulation method of the vocal-tract system. Speech Commun., **1** (3–4) (1982), 199–229.

[66] Toutios, A.; Narayanan, S.: Articulatory synthesis of French connected speech from EMA data, in *Interspeech*, Lyon, France, August 2013.

[67] King, S.; Frankel, J.; Livescu, K.; McDermott, E.; Richmond, K.; Wester, M.: Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Am., **121** (2) (2007), 723–742.

[68] Mitra, V.; Nam, H.; Espy-Wilson, C.; Saltzman, E.; Goldstein, L.: Articulatory information for noise robust speech recognition. IEEE Trans. Audio, Speech Lang. Process., **19** (7) (2011), 1913–1924.

[69] Katsamanis, A.; Bresch, E.; Ramanarayanan, V.; Narayanan, S.: Validating rt-MRI based articulatory representations via articulatory recognition, in *Interspeech*, Florence, Italy, August 2011.

[70] Ghosh, P.; Narayanan, S.: Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am., **130** (4) (2011), EL251–EL257.

[71] Lindblom, B.: Role of articulation in speech perception: clues from production. The J. Acoust. Soc. Am., **99** (3) (1996), 1683–1692.

[72] Wilson, S.; Saygin, A.; Sereno, M.; Iacobini, M.: Listening to speech activates motor areas involved in speech production. Nat. Neurosci., **7** (7) (2004), 701–702.

[73] Ghosh, P.; Goldstein, L.; Narayanan, S.: Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures. J. Acoust. Soc. Am., **129** (6) (2011), 4014–4022.

[74] Tiede, M.K.; Masaki, S.; Vatikiotis-Bateson, E.: Contrasts in speech articulation observed in sitting and supine conditions, in *Proc. 5th Seminar on Speech Production*, Kloster Seeon, Bavaria, 2000, 25–28.

[75] Kitamura, T. *et al.*: Difference in vocal tract shape between upright and supine postures: observations by an open-type MRI scanner. Acoust. Sci. Technol., **26** (5) (2005), 465–468.

[76] Stone, M. *et al.*: Comparison of speech production in upright and supine position. J. Acoust. Soc. Am., **122** (1) (2007), 532–541.

[77] Traser, L.; Burdumy, M.; Richter, B.; Vicari, M.; Echternach, M.: Weight-bearing MR imaging as an option in the study of gravitational effects on the vocal tract of untrained subjects in singing phonation. PLoS ONE, **9** (11) (2014), e112405.

[78] Honda, Y.; Hata, N.: Dynamic imaging of swallowing in a seated position using open-configuration MRI. J. Magn. Reson. Imaging, **26** (1) (2007), 172–176.

[79] Perry, J.L.: Variations in velopharyngeal structures between upright and supine positions using upright magnetic resonance imaging. Cleft Palate-Craniofacial J., **48** (2) (2010), 123–133.

[80] Katseff, S.; Houde, J.; Johnson, K.: Partial compensation for altered auditory feedback: a tradeoff with somatosensory feedback? Lang. Speech, **55** (2) (2012), 295–308.

[81] Lametti, D.R.; Nasir, S.M.; Ostry, D.J.: Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. J. Neurosci., **32** (27) (2012), 9351–9358.

[82] Kalin, R.; Stanton, M.S.: Current clinical issues for MRI scanning of pacemaker and defibrillator patients. Pacing Clin. Electrophysiol., **28** (4) (2005), 326–328.

[83] Murphy, K.J.; Brunberg, J.A.: Adult claustrophobia, anxiety and sedation in MRI. Magn. Reson. Imaging, **15** (1) (1997), 51–54.

[84] Lingala, S.G.; Sutton, B.P.; Miquel, M.E.; Nayak, K.S.: Recommendations for real-time speech MRI. J. Magn. Reson. Imaging, **43** (1) (2016), 28–44.

[85] Story, B.; Titze, I.; Hoffman, E.: Vocal tract area functions from magnetic resonance imaging. J. Acoust. Soc. Am., **100** (1) (1996), 537–554.

[86] Narayanan, S.; Alwan, A.; Haker, K.: Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals. J. Acoust. Soc. Am., **101** (2) (1997), 1064–1077.

[87] Alwan, A.; Narayanan, S.S.; Haker, K.: Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II: The rhotics. J. Acoust. Soc. Am., **101** (2) (1997), 1078–1089.

[88] Takemoto, H.; Kitamura, T.; Nishimoto, H.; Honda, K.: A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions. Acoust. Sci. Technol., **25** (6) (2004), 468–474.

[89] Ng, I.W.; Ono, T.; Inoue-Arai, M.S.; Honda, E.; Kurabayashi, T.; Moriyama, K.: Application of MRI movie for observation of articulatory movement during a fricative /s/ and a plosive /t/. Angle Orthodont., **81** (2) (2011), 237–244.

Asterios Toutios is a Research Associate with the Signal Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC), where he coordinates the interdisciplinary Speech Production and Articulation kNowledge (SPAN) group. Prior to USC he was at LORIA and TELECOM ParisTech, both in France, from 2007 to 2012. He obtained the Diploma in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece, the M.Sc. in Information Systems and the Ph.D. in Applied Informatics, both from the University of Macedonia, Thessaloniki, Greece. His research focuses on vocal tract imaging and articulatory speech synthesis.

Shrikanth (Shri) Narayanan is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995 to 2000. At USC he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biomedical problems, and applications with direct societal relevance. Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor for the Computer Speech and Language Journal and an Associate Editor for the IEEE Transactions on Affective Computing, APSIPA Transactions on Signal and Information Processing, and the Journal of the Acoustical Society of America. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000–2004), IEEE Signal Processing Magazine (2005–2008), and the IEEE Transactions on Multimedia (2008–2011). He is a recipient of a number of honors including Best Transactions Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee), and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010–2011. Papers co-authored with his students have won awards at Interspeech 2013 Social Signal Challenge, Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2013 and 2010, InterSpeech 2009-Emotion Challenge, IEEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005, and ICSLP 2002. He has published over 600 papers and has been granted 16 U.S. patents.