INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Illustrating the Production of the International Phonetic Alphabet Sounds using Fast Real-Time Magnetic Resonance Imaging

Asterios Toutios[1], *Sajan Goud Lingala*[1], *Colin Vaz*[1], *Jangwon Kim*[1], *John Esling*[2], *Patricia Keating*[3], *Matthew Gordon*[4], *Dani Byrd*[1], *Louis Goldstein*[1], *Krishna Nayak*[1], *Shrikanth Narayanan*[1]

[1]University of Southern California
[2]University of Victoria
[3]University of California, Los Angeles
[4]University of California, Santa Barbara

{toutios,shri}@sipi.usc.edu

Abstract

Recent advances in real-time magnetic resonance imaging (rtMRI) of the upper airway for acquiring speech production data provide unparalleled views of the dynamics of a speaker's vocal tract at very high frame rates (83 frames per second and even higher). This paper introduces an effort to collect and make available on-line rtMRI data corresponding to a large subset of the sounds of the world's languages as encoded in the International Phonetic Alphabet, with supplementary English words and phonetically-balanced texts, produced by four prominent phoneticians, using the latest rtMRI technology. The technique images oral as well as laryngeal articulator movements in the production of each sound category. This resource is envisioned as a teaching tool in pronunciation training, second language acquisition, and speech therapy.

Index Terms: phonetics, speech production, vocal-tract imaging, educational resource

1. Introduction

Real-time magnetic resonance imaging (rtMRI) is a tool for speech production research [1, 2] that provides dynamic information from the entire mid-sagittal plane of a speaker's upper airway, or any other scan plane of interest, from arbitrary, continuous utterances (with no need for repetitions of tokens to capture the production details of a given speech stimulus). Mid-sagittal 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, i.e., regions of the tract that cannot be easily or well observed using other techniques. RtMRI provides a rich source of information about articulation in connected speech, which can be valuable in the refinement of existing speech production models, or the development of new ones, with potential impact on speech technologies such as automatic recognition, speaker identification, or synthesis [3].

Recent advances in rtMRI technology at the University of Southern California have increased the spatiotemporal resolution and quality of rtMRI speech production data. The combination of a new custom eight-channel upper airway coil array, which has offered improved sensitivity to upper airway regions of interest, and a novel method for off-line temporal finite difference constrained reconstruction, has enabled the generation of vocal-tract movies at 83.33 frames per second, with an image resolution of 2.4 millimeters per pixel [4]. These numbers can be compared with the temporal resolution of 23.18 frames per second and spatial resolution of 3 millimeters per pixel, of earlier rtMRI data, such as those in the publicly released USC-TIMIT [5] and USC-EMO-MRI [6] databases.

This paper presents a new rtMRI resource that showcases these technological advances by illustrating the production of a comprehensive set of speech sounds present across the world's languages, i.e. not restricted to English, encoded as consonant and vowel *symbols* in the International Phonetic Alphabet (IPA), which was devised by the International Phonetic Association as a standardized representation of the sounds of spoken language [7]. These symbols are meant to represent unique speech sounds, and do not correspond to the orthography of any particular language. The IPA includes also *diacritics* that may be added to vowel and consonant symbols to indicate a modification or specification of their normal pronunciation. The project described herein addresses the normal pronunciation (i.e., without modification by diacritics) of the IPA symbols. In its most recent official version, updated in 2015[1], there are 107 symbols in the IPA, organized in a chart, and expansions to the basic set of symbols have been proposed to fill in notational gaps [8]. Any given spoken language uses a subset of the speech sounds codified in the IPA symbols, or speech sounds. English uses about 40 of them, depending on the dialect.

Given that even the basic set of IPA symbols comprises many more speech sounds than those of any one language, including many sounds not present in any Western languages, phonetic training is required for a speaker to be able to produce the set in full, or at least a large subset thereof. In the context of the work described herein, a set of IPA symbols, with supplementary words, sentences, and passages, were elicited from four distinguished phoneticians, who also self-assessed their productions, refining their accuracy. The collected data were used to develop a web resource, available on-line at `http://sail.usc.edu/span/rtmri_ipa/`. The present paper is meant to be a companion piece to that resource that complements numerous acoustic resources illustrating the production of the IPA, and a few articulatory resources [9, 10].We discuss some technical aspects of the data collection process, the development of the web resource from the data, and its outlook.

2. Data Collection

Four sessions of rtMRI data collections took place between June and September 2015 at the Los Angeles County Hospital. Subjects were four traditionally trained linguistics phoneticians (two women, two men): Professors Dani Byrd (DB); Patricia

[1] `http://www.internationalphoneticassociation.org/content/ipa-chart`

http://dx.doi.org/10.21437/Interspeech.2016-605

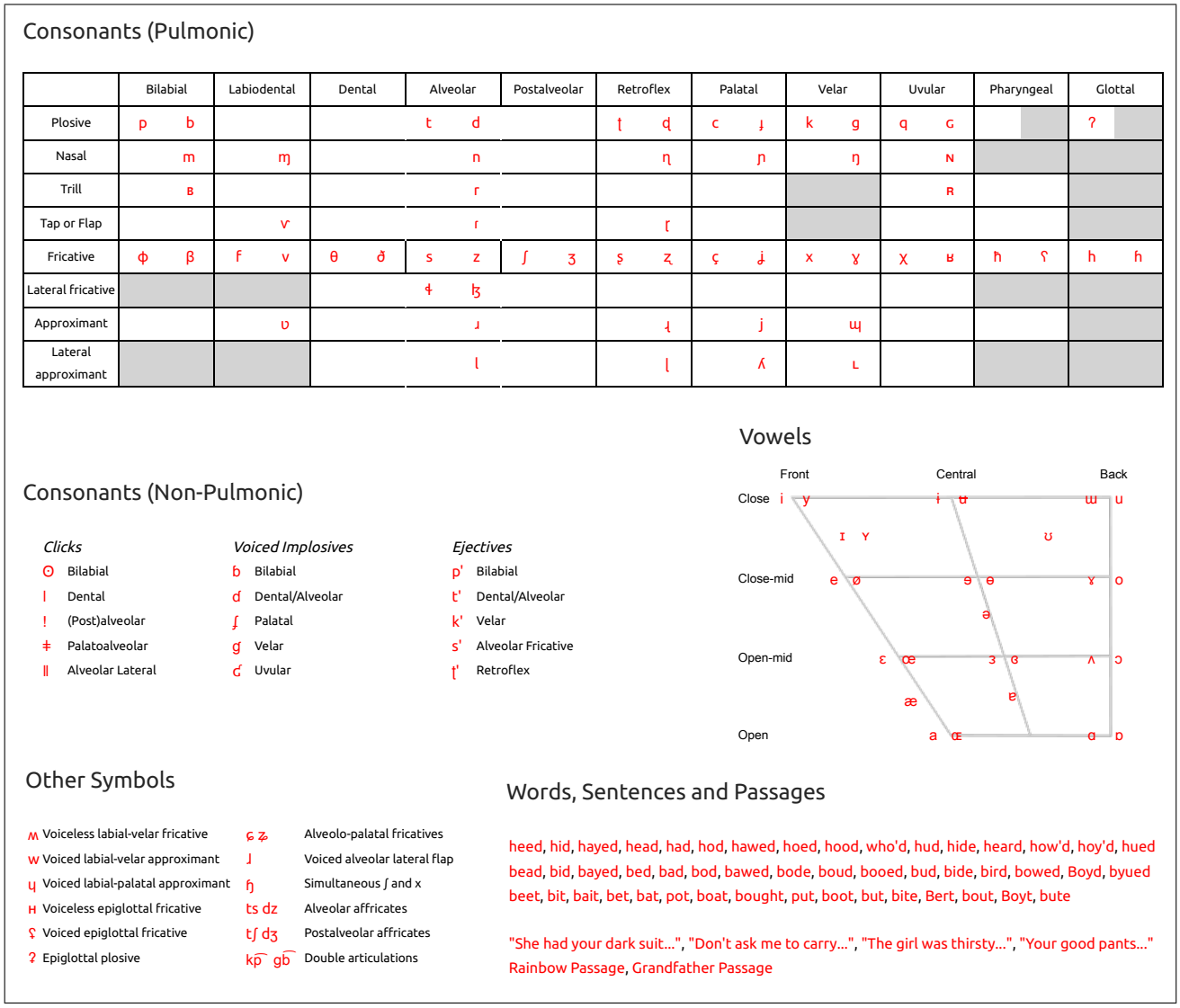

Figure 1: *Snapshot from the web resource, illustrating the stimulus set. On the web resource, symbols, words and phrases link to real-time MRI videos of their productions.*

Keating (PK); John Esling (JE); and Matthew Gordon (MG – note: these are also co-authors on this paper). The upper airways of the subjects were imaged while they lay supine in the MRI scanner. Subjects had their heads firmly but comfortably padded at the temples to minimize motion of the head. Stimuli were presented on a back-projection screen, from which subjects could read from inside the scanner via a specialized mirror setup without moving their head.

The core part of the stimuli comprised sounds from the *Pulmonic Consonants*, *Non-Pulmonic Consonants* and *Other Symbols* sections of the IPA chart, elicited in [aCa] context, and sounds from the *Vowels* section, elicited in isolation. If the subject had identified before the collection that they cannot produce confidently any of these sounds in a supine position, that sound was omitted from stimuli presentation. The stimulus set was supplemented by: three series of monosyllabic words including a full set of American English vowels and diphthongs in bVt, bVd, and hVd contexts; a set of four phonetically rich sentences (see Table 1); and the *Rainbow* and *Grandfather* passages commonly used in linguistic studies. Figure 1 shows a snapshot from the resulting web resource that illustrates this stimuli set. Note that it includes five symbols that are implicit, but not shown, in the official IPA chart.

Table 1: *Set of phonetically rich sentences that were elicited from subjects as part of the rtMRI data collections.*

- She had your dark suit in greasy wash water all year.
- Dont ask me to carry an oily rag like that.
- The girl was thirsty and drank some juice, followed by a coke.
- Your good pants look great! However, your ripped pants look like a cheap version of a K-mart special. Is that an oil stain on them?

MRI data were acquired on a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha WI) with gradients capable of

Figure 2: *From left to right, real-time MRI frames at the middle of the durations of sustained productions of vowels /i/, /e/, /a/, /o/, /u/, by subject MG. (There is a, commonly observed, light cardiac artifact in the shape of an arc that runs by the lower edge of the velum, the center of the tongue, and the lower lip.)*

Figure 3: *Production of palatoalveolar click by subject JE, in /a‡a/. The leftmost panel shows the acoustic waveform and marks the times at which the four real-time MRI frames were extracted.*

40 mT/m amplitude and 150 mT/m/ms slew rate. A body coil was used for radio frequency (RF) signal transmission. A novel, customized, eight-channel upper-airway receiver coil , with four elements on either side of the jaw, was used for RF signal reception. The coil's design enables high sensitivity over all salient speech articulators, thereby greatly increasing the signal-to-noise ratio in these regions (by a factor of 2-6 fold) in comparison with coils developed for other purposes, such as the neurovascular or head-and-neck coil [4].

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. In our earlier attempts, thirteen interleaved spirals together formed a single image. With each spiral acquired over 6 msec, every image comprised information spanning over 78 ms. However, a recently developed constrained reconstruction method, that exploits temporal finite difference sparsity of the dynamic image time series, has enabled the formation of images from only two interleaves, i.e. from information spanning over only 12 ms, which leads to videos with a frame rate of 83 frames/sec [4].

The imaging field of view is 200 $\times$ 200 mm, the flip angle is 15°, and the receiver bandwidth $\pm$125 kHz. Slice thickness is 6 mm, located mid-sagittally; image resolution in the sagittal plane is 84 $\times$ 84 pixels (2.4 $\times$ 2.4 mm). Scan plane localization of the mid-sagittal slice is performed using RTHawk (HeartVista, Inc., Los Altos, CA), a custom real-time imaging platform [11].

Audio is recorded concurrently with MRI acquisition inside the MRI scanner while subjects are imaged, using a fiber-optic microphone (Optoacoustics Ltd., Moshav Mazor, Israel) and a custom recording and synchronization setup [12]. The audio is recorded with a sampling rate of 100 kHz and is stored in a raw, uncompressed format to ensure no speech information loss. A post-processing step down-samples the audio to 20 kHz and enhances the recorded speech using customized de-noising methods [12, 13], in order to reduce the effect of loud scanner noise in the recording.

Table 2: *Number of IPA symbols in the web resource, produced by each subject.*

Subject	DB	PK	JE	MG
Consonants (Pulmonic)	50	57	59	59
Consonants (Non-Pulmonic)	10	14	14	15
Vowels	16	25	28	28
Other Symbols	6	12	16	16

3. Web Resource

After the data collections, an HTML page per subject was developed, following the general structure shown in Figure 1. Symbols, words, and phrases link to rtMRI videos (with sound) of the corresponding productions. The pages use Unicode entities for the IPA symbols, and are presented using the cascading style sheets of the Speech Production and Articulation kNowledge (SPAN) group website,[2] where they are hosted. Videos are displayed using HTML5 tags and an open-source lightbox environment.

The subjects assessed the speech production information in the HTML pages. If they were not satisfied with the production of a sound, that sound was removed from the page. This, combined with the fact that some symbols were added to the stimulus set between the first and last acquisitions (the order of the experiments was DB, PK, JE, MG), led to different counts of symbols in each subject's page. Table 2 summarizes this information.

Figures 2, 3, 4, and 5 present representative frames extracted from the dynamic rtMRI videos on the website. Figure 2 shows frames extracted from near the mid-point of sustained productions of the vowels /i/, /e/, /a/, /o/ and /u/ by subject MG; the depicted vocal-tract shapes agree well with standard phonetic knowledge. Figure 3 presents four frames during the production of the palatoalveolar click in an /a‡a/ sequence. As

[2] `http://sail.usc.edu/span/`

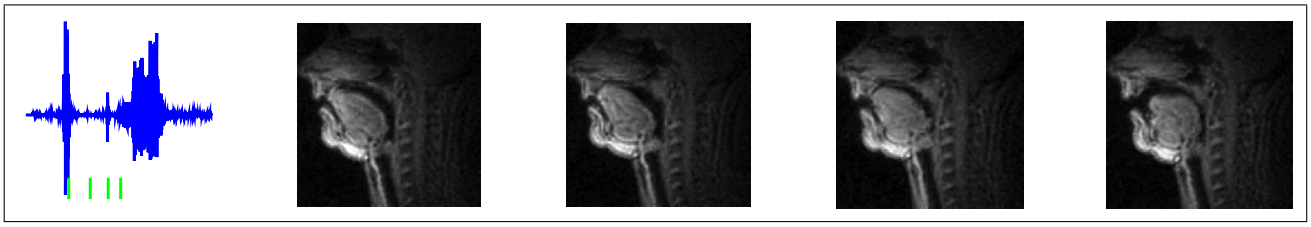

Figure 4: *Production of dental/alveolar ejective by subject DB, in /at'a/. The leftmost panel shows the acoustic waveform and marks the times at which the four real-time MRI frames were extracted.*

Figure 5: *Production of labial palatal approximant by subject PK, in /aɥa/. The leftmost panel shows the acoustic waveform and marks the times at which the four real-time MRI frames were extracted. (The artifact around the nasion is probably because of the subject's glasses, which she wore in the scanner.)*

expected, the tongue makes a broad contact across the roof of the mouth from the alveolar ridge to the palate at closure, and pulls toward the back of the mouth during release. Also, the first two frames show some distance between the arytenoid and the basis of epiglottis, indicating voicelessness due to abduction (during voiced segments the arytenoids come together and enter the thin mid-sagittal slice that is imaged with rtMRI [14]). Figure 4 shows frames from the production of /at'a/ by subject DB, where the dental/alveolar ejective is articulated with the blade of the tongue at the alveolar ridge, and the tip of the tongue behind the upper teeth. Note that the teeth themselves are never directly visible in MRI, because of their chemical composition. The second and third presented frames show the arytenoid cartilage in a higher position than in the first and fourth frames, which agrees with the expectation that air is forced out by pumping the glottis upward, thereby creating the ejective airstream mechanism. Finally, Figure 5 presents frames from the production of /aɥa/ by subject PK. The labial palatal approximant is produced by raising the body of the tongue toward the palate while compressing the lips. The arytenoid appears constantly in contact with the basis of the epiglottis, indicating voicing throughout.

4. Outlook

Teaching of pronunciation, in contexts like linguistics education, second language learning and speech therapy, usually relies on acoustics and visual information of the speaker's lips and front of the mouth. By allowing visual access to the *internal* articulators, rtMRI can offer very valuable supplementary information. It is our hope that the rtMRI web resource we present herein, covering an extended set of IPA sounds present in the world's languages, will be useful as a tool for such training.

The resource has been on-line since late autumn 2015. Even though there has been no formal outreach effort, we have noted increased web traffic to it, and links to it from various other sites, especially involved in the teaching of phonetics. This makes us very optimistic about the usefulness of the resource to the phonetics community.

Real-time MRI data have some shortcomings, which have been discussed in detail elsewhere [15, 3]. These include the fact that speech is produced in a supine position, the invisibility of the teeth, and the loud scanner noise (and use of earplugs). The latter is usually not of concern when rtMRI subjects produce normal speech in their native languages, but subjects in the IPA experiments reported some difficulties in sounds like clicks, ejectives, or non-native vowels, that they attributed to lack of normal or sufficient auditory feedback. Though these shortcomings should be kept in mind, we do not believe that they significantly diminish the usefulness and value of the resource.

The current web presentation of these data may not yet be optimized for teaching purposes. We plan to develop tools, building upon previous efforts [5], to allow for frame-by-frame examination of the rtMRI videos, and side-by-side comparisons between phones or subjects. Moreover, we have also started to apply automatic air-tissue boundary segmentation methods [16] to these data, and generate phonetic alignments.

Air-tissue boundary segmentations will enable, in turn, the application of recently developed methodologies for articulatory modeling [14, 17]. It is of particular interest to see how these models will behave when confronted with articulations of speech sounds used less frequently to date in creating such models.. Another exciting possibility is the expansion of articulatory synthesis models [18, 19] in order to account for non-pulmonic consonants, such as sounds involving other airstream mechanisms than pulmonic egressive. The data that have been collected in the context of the present work can be central to such efforts.

Of course, simple observation of the rtMRI videos can by itself be revelatory about aspects of articulation, especially for lesser-studied speech sounds.

5. Acknowledgment

Work supported by NIH grant R01DC007124 and NSF grant 1514544.

6. References

[1] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 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 Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

[3] A. Toutios and S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, p. e6, 2016.

[4] S. G. Lingala, Y. Zhu, Y.-C. Kim, A. Toutios, S. Narayanan, and K. S. Nayak, "A fast and flexible MRI system for the study of dynamic vocal tract shaping," *Magnetic Resonance in Medicine*, 2016.

[5] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, 2014.

[6] J. Kim, A. Toutios, Y.-C. Kim, Y. Zhu, S. Lee, and S. S. Narayanan, "USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging," in *International Seminar on Speech Production (ISSP)*, Cologne, Germany, 2014.

[7] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[8] J. Esling, "Phonetic notation," in *The Handbook of Phonetic Sciences, 2nd ed.*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Oxford: Wiley-Blackwell, 2010, pp. 678–702.

[9] E. Lawson, J. Stuart-Smith, J. M. Scobbie, and N. S., "Seeing Speech: an articulatory web resource for the study of Phonetics," http://seeingspeech.ac.uk, 2015, University of Glasgow. 1st April 2015.

[10] J. Abel, B. Allen, S. Burton, M. Kazama, B. Kim, M. Noguchi, A. Tsuda, N. Yamane, and B. Gick, "Ultrasound-enhanced multimodal approaches to pronunciation teaching and learning," *Canadian Acoustics*, vol. 43, no. 3, 2015.

[11] J. Santos, G. Wright, and J. Pauly, "Flexible real-time magnetic resonance imaging framework," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, 2004.

[12] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.

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

[14] A. Toutios and S. S. Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," in *International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, 2015.

[15] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, Jan. 2016.

[16] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323–338, 2009.

[17] T. Sorensen, A. Toutios, L. Goldstein, and S. S. Narayanan, "Characterizing vocal tract dynamics with real-time MRI," in *15th Conference on Laboratory Phonology*, Ithaca, NY, 2016.

[18] S. Maeda, "Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer," in *Sound Patterns of Connected Speech: Description, Models and Explanation*, A. Simpson and M. Pätzold, Eds., 1996, pp. 145–164.

[19] A. Toutios and S. Maeda, "Articulatory VCV Synthesis from EMA data," in *Interspeech*, Portland, Oregon, 2012.