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

Asterios Toutios[1]*, Sajan Goud Lingala*[1]*, Colin Vaz*[1]*, Jangwon Kim*[2]*, John Esling*[3]*, Patricia Keating*[4]*, Matthew Gordon*[5]*, Dani Byrd*[1]*, Louis Goldstein*[1]*, Krishna Nayak*[1]*, Shrikanth Narayanan*[1]

[1]University of Southern California, [2]Cobalt Speech and Language, Inc., [3]University of Victoria, [4]University of California, Los Angeles,[5]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 imaging, at very high frame rates (83 frames/seconds 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. This resource is envisioned to serve as a tool for teaching as well as pronunciation training, second language acquisition, and speech therapy.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Real-time magnetic resonance imaging (rtMRI) is a tool for speech production research [1, 2] which 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 of repetitions. 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.

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, have enabled the generation of vocal-tract movies at 83.33 frames per second, with an image resolution of 2.4 millimeters per pixel [3]. These numbers are to 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 [4] and USC-EMO-MRI [5] databases.

The present paper presents an new rtMRI resource, which 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 *letters* in the International Phonetic Alphabet (IPA), which was devised by the International Phonetic Association as a standardized representation of the sounds of oral language. [6] IPA letters are symbols meant to represent unique speech sounds, and are different from the orthography of any language. *Letters* is one of the two basic types of elements in the IPA, the second one being *diacritics*. Diacritics are added to IPA letters to indicate a modification or specification of that letter's normal pronunciation. The project described herein addresses the normal pronunciation (i.e. without modification by diacritics) of the IPA letters. In its most recent version, updated in 2005[1], there are 107 letters in the IPA, organized in a chart. Any given spoken language uses a subset of these letters, or speech sounds. English uses about 40 of them, depending on the dialect.

Given that the letters of the IPA comprise many more speech sounds than any given language, even sounds not present *any* Western language, special phonetic training is required for a speaker to be able to produce it in full, or at least a large subset thereof. In the context of the work described here, the letters of the IPA were elicited from four distinguished phoneticians, who also self-assessed their productions, assuring their correctness. The collected data were used to develop a web resource, available on-line at `http://sail.usc.edu/rtmri_ipa/`. The present paper is meant to be a companion piece to that resource, which complements numerous acoustic resources illustrating the production of the IPA, and at least one other articulatory resource, based on Ultrasound and an slower rtMRI system. In the rest of this paper, 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 in June and July 2015 at the Los Angeles County Hospital. Subjects were Professors Dani Byrd (DB), Patricia Keating (PK), John Esling (JE), and Matthew Gordon (MG). 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 without moving their head via a specialized mirror setup.

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

The core part of the stimuli comprised the letters in the *Pulmonic Consonants*, *Non-Pulmonic Consonants* and *Other Symbols* sections of the IPA chart, elicited in aCa context, and the letters of the *Vowels* section, elicited in isolation. If the subject had identified before the collection that they cannot produce confidently any of these letters, that letter was omitted from stimuli presentation. The stimuli was supplemented by: three series of monosyllabic words including a full set of American English vowels and diphtongs in bVt, bVd, and hVd contexts; a set of four phonetically rich sentences; and the commonly used in linguistic studies *Rainbow* and *Grandfather* passages.

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 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 in comparison with coils developed for other purposes, such as the neurovascular or head-and-neck coil.

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 form a single image. With each spiral acquired over 6 msec, every image comprised information spanning over 78 ms However, a recently developed reconstruction method, that exploits prior knowledge 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.

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 5 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 [7].

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 custom recording and synchronization setup. Speech in the recorded audio was then enhanced, using customized denoising methods, in order to reduce the effect of loud scanner noise.

3. Web Resource Development

4. Outlook

5. 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] 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. [Online]. Available: http://dx.doi.org/10.1002/mrm.26090

[4] 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.

[5] 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. [Online]. Available: ./pdfs/kimj2014emo.pdf

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

[7] 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, September 2004.