The magnetic resonance imaging subset of the mngu0 articulatory corpus

Ingmar Steiner
INRIA / LORIA Speech Group
Bat. C, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, France
ingmar.steiner@inria.fr

Korin Richmond
Centre for Speech Technology Research, University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
korin@cstr.ed.ac.uk

Ian Marshall
Medical Physics & Medical Engineering, University of Edinburgh
Chancellor's Building, 49 Little France Crescent, Edinburgh, EH16 4SB, United Kingdom
ian.marshall@ed.ac.uk

Calum D. Gray
Clinical Research Imaging Centre, University of Edinburgh
Queen's Medical Research Institute, 47 Little France Crescent, Edinburgh, EH16 4TJ, United Kingdom
calum.gray@ed.ac.uk

Abstract: This paper announces the availability of the magnetic resonance imaging (MRI) subset of the mngu0 corpus, a collection of articulatory speech data from one speaker containing different modalities. This subset comprises volumetric MRI scans of the speaker's vocal tract during sustained production of vowels and consonants, as well as dynamic midsagittal scans of repetitive consonant-vowel (CV) syllable production. For reference, high-quality acoustic recordings of the speech material are also available. The raw data are made freely available for research purposes.

PACS numbers 43.70.Aj, 43.70.Jt [AL]
Date Received: November 28, 2011 **Date Accepted:** December 13, 2011

1 Introduction

Technology applications that use speech data, for example, text-to-speech (TTS) synthesis, automatic speech recognition (ASR), focus almost exclusively on acoustics. However, the acoustic signal produced by a human speaker crucially depends on the shape of the speaker's vocal tract and the movements of articulators such as the tongue or lips. Techniques such as motion capture or medical imaging can provide valuable articulatory data to supplement the acoustic speech signal. Applications such as TTS synthesis or ASR can exploit such data to improve their modeling of human speech production.

However, recording articulatory data is unfortunately not as straightforward as recording an acoustic signal. Specialist facilities and expertise are typically required, which make articulatory recordings more expensive. Indeed, the recording process itself can often be tricky, with practical complications that may also increase the burden on the subject. Consequently, the few articulatory corpora which have been made freely available (e.g. Munhall et al., 1995; Westbury, 1994; Wrench, 2000; Narayanan et al., 2011) have been well received and extensively used by the research community. Each of these contains articulatory and acoustic data for a range of speakers, but the data for any individual speaker may be insufficient for certain applications.

The mngu0 articulatory corpus addresses this shortfall and provides a large amount of articulatory data from a single speaker of British English. The corpus consists of multiple subsets

hit	ɪ	*fin*	f
pet	ɛ	*thin*	θ
hat	æ	*sin*	s
hot	ɒ	*shin*	ʃ
hut	ʌ	*mock*	m
put	ʊ	*knock*	n
heat	i	*thing*	ŋ
hoot	u	*ring*	ɹ
hurt	ɜ	*long*	l
hart	ɑ	*loch*	x
ought	ɔ	*sleep*	l
about	ə	⟨p⟩	p
there	ɛː	⟨t⟩	t
		⟨k⟩	k
		ball	ɫ

(a) Sustained production; vowels are shown on the left, consonants on the right. The prompts ⟨p, t, k⟩ represent the speaker holding the occlusion of the corresponding stop.

apa	*ipi*	*upu*	p
ata	*iti*	*utu*	t
aka	*iki*	*uku*	k
afa	*ifi*	*ufu*	f
atha	*ithi*	*uthu*	θ
asa	*isi*	*usu*	s
asha	*ishi*	*ushu*	ʃ
ara	*iri*	*uru*	ɾ
ala	*ili*	*ulu*	l
ama	*imi*	*umu*	m
ana	*ini*	*unu*	n
anga	*ingi*	*ungu*	ŋ
acha	*ichi*	*uchu*	x
ara	*iri*	*uru*	ɹ
awa	*iwi*	*uwu*	w
aya	*iyi*	*uyu*	j

(b) Dynamic production. Each prompt represents an individual scan, yielding the vocal tract configuration for the target phone in the vocalic context of [ɑ, i, u], respectively.

Table 1: Prompt lists for the MRI scanning session. The orthographic prompts are emphasized, with the underlined letter(s) corresponding to the target phone; the target phone itself is given in IPA notation.

of data acquired in different modalities, including a large number of sentences recorded using video and electromagnetic articulography (EMA), dental casts, and magnetic resonance imaging (MRI) scans of the speaker's vocal tract. Taken together, they provide both high-speed (200 Hz) articulatory movement data during running speech, and the speaker's vocal tract geometry in three dimensions, and offer a valuable resource to the speech community. While the EMA portion of the `mngu0` corpus has previously been described and made publicly available (Richmond et al., 2011), the purpose of *this* paper is to announce the availability of the corresponding MRI data for this speaker and describe the details of its acquisition.

2 MRI data

The purpose of MRI scanning the `mngu0` speaker was two-fold. First, we wanted to capture the 3D geometry of the speaker's vocal tract, as well as representative configurations of the speaker's articulators for producing a range of speech sounds. A series of volumetric MRI scans was performed for this purpose. Second, we wanted to investigate and capture the effects of coarticulation, for which a set of dynamic MRI scans was performed. Overall, therefore, by adapting the procedure described in Birkholz and Kröger (2006), a prompt list was designed that consisted of sustained vowels and consonants, as well as dynamic VCV transitions, elicited by repetitive production of consonant-vowel (CV) syllables.

The scanner used for this study was a GE Medical Systems Signa HDx 1.5T. The speaker was placed in the scanner in supine position and fitted with a head-and-neck radio frequency (RF) coil. All of the scans were completed within one 120 min session, with a short break between the sustained and dynamic scans.

The `mngu0` subject is a trained, professional speaker. As well as general advantages in terms of the level of performance obtained, this was beneficial for the MRI scans in particular, as he was able to sustain production of vowels for 20 s on average, and of repetitive utterances for around 10 s to 15 s. The speaker was therefore capable of producing each prompt over the entire duration of a scan.

As Figure 1 shows, the region of interest (ROI) is clearly visible in the resulting image data, extending from the lips to the rear wall of the pharynx in anterior-posterior direction, from the

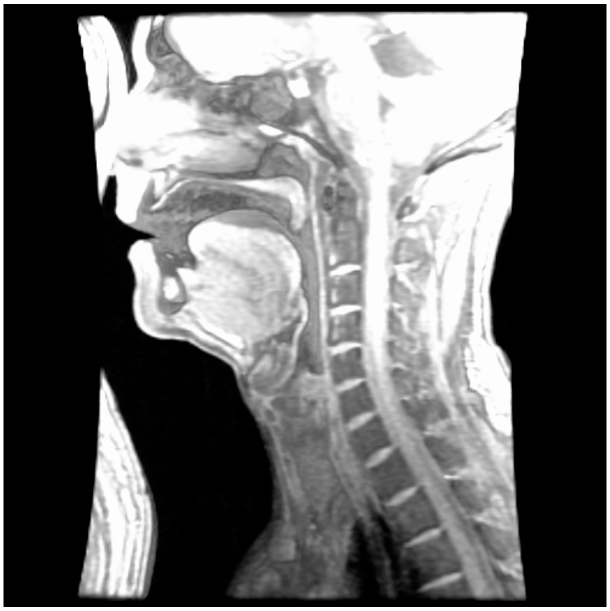

Figure 1: Cutaway volume rendering of raw volumetric data for [ɑ].

Mm. 1: Cutaway volume rendering of volumetric [ɑ] scan (360° horizontal rotation, static clip plane).

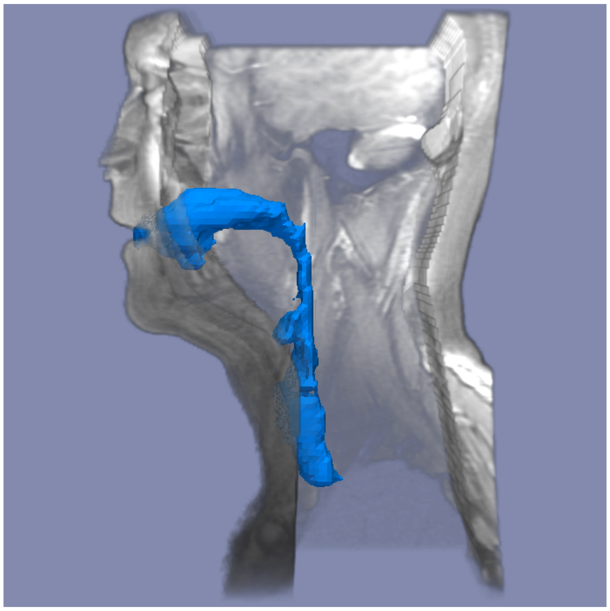

Figure 2: 3D vocal tract extracted from [ɑ] scan, shown with surface rendering.

Mm. 2: 3D vocal tract extracted from volumetric [ɑ] scan, with surface rendering (360° horizontal rotation).

larynx to the nasal cavity in inferior-superior direction, and (in the volumetric scans) laterally between the mandibular joints.

This figure also exhibits an aliasing artifact; a segment of the back of the speaker's head and neck appears at the anterior edge of the images, overlapping with the speaker's nose. This occurs when the imaged anatomy extends outside the field of view (FOV). But since we wanted to gain good coverage of the vocal tract, and the aliasing does not impact image quality in the ROI, we chose this configuration and the aliasing can safely be ignored.

Due to the acoustic conditions within the scanning chamber and the lack of a noise-canceling fiber-optic (FO) microphone, no simultaneous acoustic recordings were possible during the MRI session. For this reason, a separate acoustic recording session was conducted as well (see Section 3).

2.1 Volumetric scans

Static, 3D scans of 13 sustained vowels and 15 sustained consonants were acquired with a fast gradient echo sequence having 26 sagittal slices of 4 mm thickness, repetition time (TR) 51 ms, echo time (TE) 3 ms, flip angle 30°, and field of view (FOV) 280 mm reconstructed as 256 × 256 pixels.

The prompts are listed in Table 1a. For each scan, the speaker produced the corresponding prompt, maintaining audible production of the target phone during the entire acquisition.

Figure 1 and Mm. 1 illustrate one raw volumetric scan [ɑ], while Figure 2 and Mm. 2 show a 3D vocal tract mesh extracted from that scan, along with a surface rendering of the face and head (the aliasing artifact has been removed).

2.2 Dynamic mid-sagittal scans

Dynamic scans of 16 consonants in three vocalic contexts [ɑ, i, u] were acquired with a fast gradient echo sequence similar to the above, but with a single midline sagittal slice of thickness 10 mm, and TR 4 ms and TE 2 ms, enabling 40 consecutive time frames to be acquired in 10 s.

The prompts are listed in Table 1b. The speaker produced each prompt as repetitive CV syllables, synchronizing production of the target consonants to the scans by timing their stable phase with the noise emitted by the MRI scanner.

An example of the dynamic data for nasals [m, n, ŋ] is displayed in Figure 3. Each panel shows an overlay of 30 MRI frames, providing an "averaged" image that offers an enhanced view of the articulators.

2.3 Dental reconstruction

Since the teeth are invisible in MRI,[1] a final volumetric scan was acquired using blueberry juice (which has favorable NMR properties, due to its high manganese content) to distinguish oral cavity from teeth. This produced a negative 3D scan of the teeth for subsequent dental reconstruction.

For this dental scan, the speaker lay prone in the scanner, and filled his mouth with the juice by sucking it from a bottle through a flexible tube. While this did produce images clearly showing the teeth, there was no time left in the session for a long acquisition, and therefore the spatial resolution of the dental scan is no higher than that of the other volumetric scans.

3 Acoustic reference recordings

To compensate for not being able to simultaneously record the acoustic signal, and to give the speaker opportunity to familiarize himself with the prompt list and the general procedure in an informal, non-clinical environment, high-quality acoustic reference recordings were made on the day before the MRI session.

The acoustic recording session took place in a sound-proofed room at the Informatics Forum, University of Edinburgh. The prompts were recorded using a DPA Type 4035 microphone mounted on a headset. The microphone signal was captured directly to hard disk using an EDIROL UA-25 audio interface connected to a laptop computer. The recordings were made with a 96 kHz sampling rate at 24 bit quantization.

[1] their nuclear magnetic resonance (NMR) properties make them nearly indistinguishable from the surrounding air

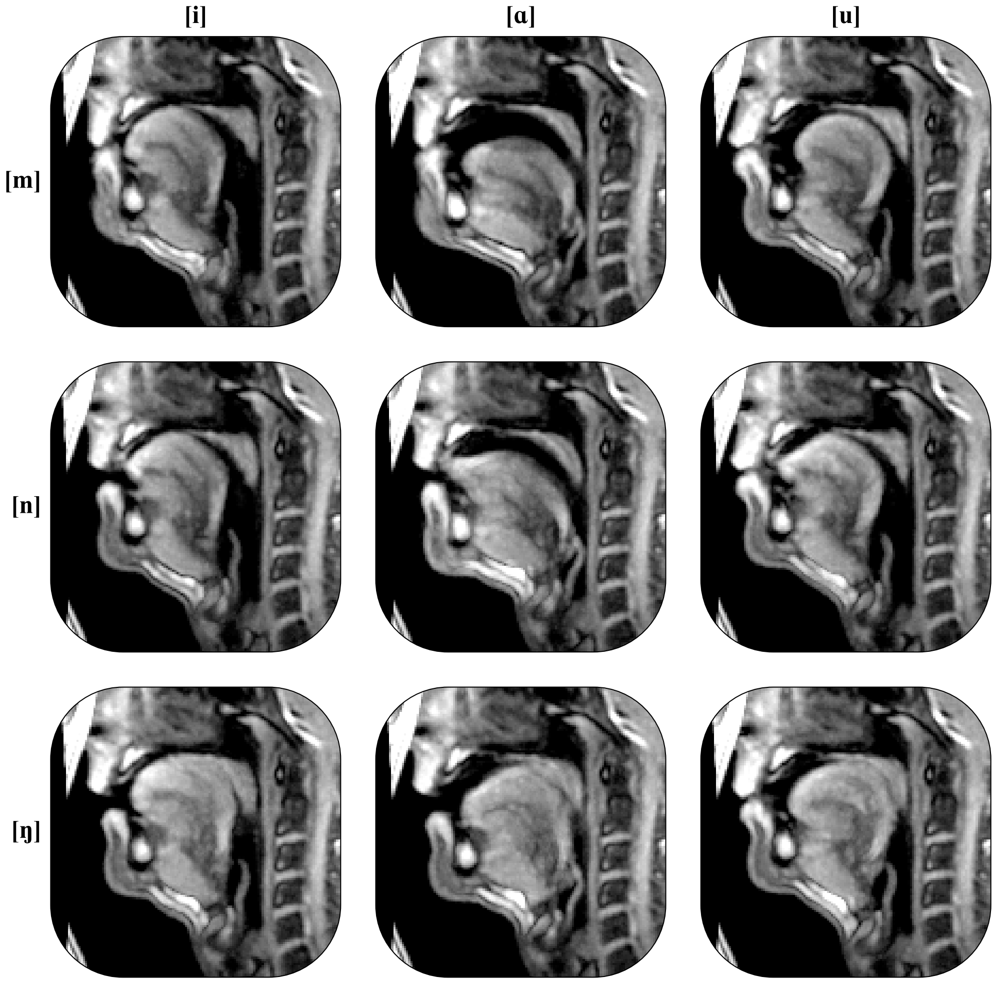

Figure 3: Overlay of 30 mid-sagittal frames of [m, n, ŋ] (rows) dynamically produced in vocalic context [i, ɑ, u] (columns). The critical articulators for each consonant (lips, tongue tip, and tongue dorsum, respectively) achieve occlusion, while the tongue body assumes a different target shape in each vocalic context. The velum is lowered in all conditions, allowing the speaker to sustain production of the nasal.

The speaker read out the prompt list twice, once standing upright, and again in supine position. This was done to allow comparison in the acoustic domain between these two postures. The supine recordings were made to ensure that the audio matched the articulatory configuration during the MRI scans, where gravity and posture can influence articulation (e.g. Kitamura et al., 2005).

The acoustic speech data has been manually segmented into prompts.

4 Distribution

The raw data from the MRI session has been anonymized to protect the privacy of the speaker, but is otherwise unmodified from the DICOM files produced at the MRI facility.

The volumetric scans and dynamic scans will be made freely available for research purposes, as separate downloads at a dedicated website, `http://mngu0.org/`. The acoustic data and the corresponding segmentation data will likewise be made available.

Acknowledgments

Imaging was carried out at the Brain Research Imaging Centre, Edinburgh (`http://www.bric.ed.ac.uk`), Division of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh, a core area of the Wellcome Trust Clinical Research Facility and part of the SINAPSE collaboration (`http://www.sinapse.ac.uk`).

This work was supported by Marie Curie Early Stage Training Site “EdSST” (MEST-CT-2005-020568) and EPSRC grant EP/E027741/1 (“ProbTTS”).

References and links

Birkholz, P., and Kröger, B.J. (**2006**). “Vocal tract model adaptation using magnetic resonance imaging,” in *Proceedings of the 7th International Seminar on Speech Production*.

Kitamura, T., Takemoto, H., Honda, K., Shimada, Y., Fujimoto, I., Syakudo, Y., Masaki, S., Kuroda, K., Oku-Uchi, N., and Senda, M. (**2005**). “Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner,” Acoust. Sci. Technol. **26**, 465–468, doi:10.1250/ast.26.465.

Munhall, K.G., Vatikiotis-Bateson, E., and Tohkura, Y. (**1995**). “X-ray film database for speech research,” J. Acoust. Soc. Am. **98**, 1222–1224, doi:10.1121/1.413621.

Narayanan, S., Bresch, E., Ghosh, P.K., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., and Zhu, Y. (**2011**). “A multimodal real-time MRI articulatory corpus for speech research,” in *Proceedings of Interspeech*, 837–840.

Richmond, K., Hoole, P., and King, S. (**2011**). “Announcing the electromagnetic articulography (day 1) subset of the `mngu0` articulatory corpus,” in *Proceedings of Interspeech*, 1505–1508.

Westbury, J.R. (**1994**). *X-Ray Microbeam Speech Production Database User’s Handbook Version 1.0*, University of Wisconsin, Madison, WI, USA.

Wrench, A.A. (**2000**). “A multi-channel/multi-speaker articulatory database for continuous speech recognition research,” PHONUS **5**, 1–14.