A Multimodal Real-Time MRI Articulatory Corpus for Speech Research

Shrikanth Narayanan[1,3], *Erik Bresch*[1], *Prasanta Ghosh*[1]
Louis Goldstein[3], *Athanasios Katsamanis*[1], *Yoon Kim*[2], *Adam Lammert*[1]
Michael Proctor[1,3], *Vikram Ramanarayanan*[1], *Yinghua Zhu*[2]

[1]Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles CA
[2]Magnetic Resonance Engineering Lab (MREL), University of Southern California, Los Angeles CA
[3]Department of Linguistics, University of Southern California, Los Angeles CA
http://sail.usc.edu/span shri@sipi.usc.edu

Abstract

We present MRI-TIMIT: a large-scale database of synchronized audio and real-time magnetic resonance imaging (rtMRI) data for speech research. The database currently consists of speech data acquired from two male and two female speakers of American English. Subjects' upper airways were imaged in the midsagittal plane while reading the same 460 sentence corpus used in the MOCHA-TIMIT corpus [1]. Accompanying acoustic recordings were phonemically transcribed using forced alignment. Vocal tract tissue boundaries were automatically identified in each video frame, allowing for dynamic quantification of each speaker's midsagittal articulation. The database and companion toolset provide a unique resource with which to examine articulatory-acoustic relationships in speech production.
Index Terms: speech production, speech corpora, real-time MRI, multi-modal database, large-scale phonetic tools

1. Introduction

There is a growing appreciation in the phonetics and the broader speech research communities of the importance of speech data acquired using multiple sensing modalities [2, 1]. A proper understanding of speech production and linguistic representation cannot be gained from acoustic signals or articulatory data alone [3, 4]. Phonetic transcriptions derived solely from acoustic data are inadequate for investigation of the underlying speech processes. Articulatory information also offers novel ways of uncovering prosodic differences due to speech rate, stress, duration, and affect, as shown by their effects on segmental timing, coarticulation, and tract posture [5, 6]. Incorporating articulatory knowledge into automatic speech recognition (ASR) technologies may also improve recognition of spontaneous speech, provide greater robustness to noise, and improve the interpretability of ASR models [7, 8].

Large-scale corpora of articulatory data have previously been compiled from a range of phonetic methodologies. The Wisconsin X-ray microbeam database (XRMB) [9] consists of parallel articulatory and acoustic data acquired from over 60 subjects, each of whom provide about 20 minutes of read speech and oral motor tasks – an invaluable resource for studying articulatory dynamics, but insufficiently rich in continuous speech data for most ASR applications [1].

The MOCHA-TIMIT database [1] contains synchronized electromagnetic midsagittal articulometry (EMA), laryngographic, electropalatographic (EPG), and acoustic data from 40 speakers of English, each reading a 460 sentence subset of the TIMIT Speech Corpus [10]. Because it was acquired using four different sensing modalities, MOCHA-TIMIT constitutes an unique source of real-time multi-modal speech data. However, as with all phonetic data acquired using these methods, our knowledge of articulation is restricted to the areas of the vocal tract where EPG, laryngographic, and EMA sensors can be placed: palate, glottis and anterior lingual fleshpoints.

Real-time magnetic resonance imaging (rtMRI) is an important emerging tool for speech research [11, 12], providing dynamic information from the entire midsagittal plane of a speaker's upper airway, or any other scan plane of interest. Midsagittal rtMRI captures not only lingual, labial and jaw motion, but also articulation of the velum, pharynx and larynx – regions of the tract which cannot be monitored with 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.

Here we describe an initiative in which we are assembling a large-scale, multi-speaker rtMRI speech database and supporting toolset, with the aim of advancing speech research based on this modality, and making some of these resources available to the broader speech research community. In Section 2 we detail the acquisition of the database, before outlining its structure and exemplifying the data in Section 3. In Section 4 we describe some tools developed for data access and analysis; and in Sections 5-6 we illustrate some phonetic use of these data, and suggest further applications for speech research.

2. Data Acquisition

2.1. Subjects

To date, data have been acquired from four native speakers of General American English (Table 1). None of the subjects spoke any other language fluently, or had lived outside the United States for a significant amount of time. Both parents of each of the subjects are native speakers of American English. None of the speakers reported abnormal hearing or speaking development or pathologies. Equal numbers of male and female subjects were, and will continue to be recruited, so as to provide a gender-balanced database of speakers.

2.2. Image Acquisition

MRI data were acquired at Los Angeles County Hospital on a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha WI) with gradients capable of 40 mT/m amplitude and 150 mT/m/ms slew rate. A custom 4-channel upper airway receiver coil array, with two anterior coil elements and two coil

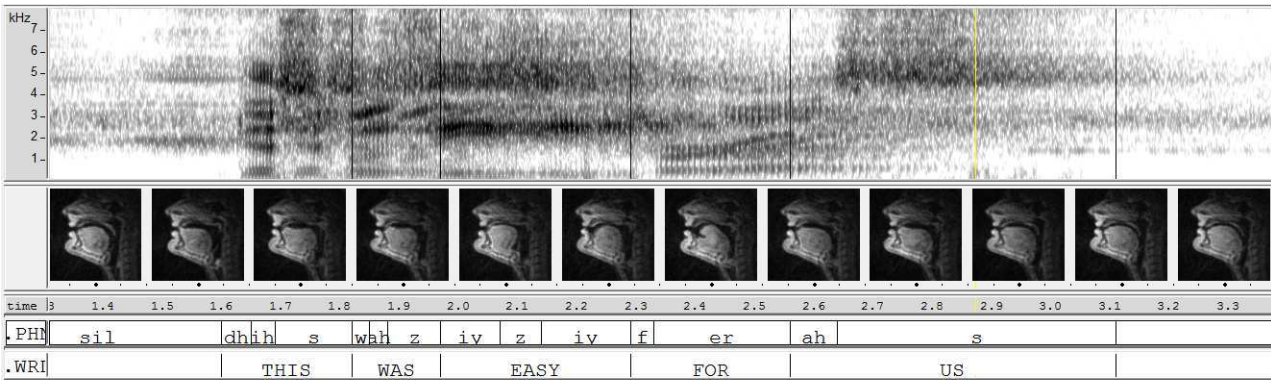

Figure 1: ***Example utterance from the MRI-TIMIT database****, showing audio spectrogram, synchronized rtMRI video (23.18 frames/sec), and forced-aligned phonemic transcription data. Illustrated sentence: "This was easy for us" uttered by Subject M1.*

ID	GENDER	ETHNICITY	AGE	BIRTHPLACE
M1	Male	White	29	Buffalo, NY
M2	Male	White	33	Ann Arbor, MI
W1	Female	White	23	Commack, NY
W2	Female	White	32	Westfield, IA

Table 1: *Study Participants – Demographic Details.*

elements posterior to the head and neck, was used for radio frequency (RF) signal reception. A 13-interleaf spiral gradient echo pulse sequence was used (T_R = 6.164 msec, FOV = 200 $\times$ 200 mm, flip angle = 15°). Scan slice thickness was 5 mm, located midsagittally; image resolution in the sagittal plane was 68 $\times$ 68 pixels (2.9 $\times$ 2.9 mm). New image data were acquired at a rate of 12.5 frames/second, and reconstructed as 23.18 frames/sec. video using a sliding window technique. More details about the rtMRI acquisition can be found in [12].

Subjects' upper airways were imaged while they lay supine in the MRI scanner. Stimuli were presented in large text on a back-projection screen which subjects could read from within the scanner bore without moving their head. Sentences were presented one at a time, elicited at a natural speaking rate. Participants were trained in the task before entering the scanner, and were paid for their time. The average recording time for each subject, including calibration and pauses between utterances, was 2 hours.

2.3. Audio Acquisition

Audio was simultaneously recorded at a sampling frequency of 20kHz inside the MRI scanner while subjects were imaged, using a custom fiber-optic microphone noise-cancelling system. Synchronization with the video signal was controlled through the use of an audio sample clock derived from the scanner's 10MHz master clock, and triggered using the scanner RF master-exciter unblank signal. More details about audio acquisition, the noise cancellation technique, and audio-video synchronization can be found in [13]. Subjects wore ear plugs for protection from the scanner noise, but were still able to hear loud conversation in the scanner room and to communicate orally and aurally with the experimenters via an in-scanner intercom system.

2.4. Phonetic alignment

Time-aligned phonetic transcriptions of all utterances in the database were generated from the audio recordings, using the freely available tool *SailAlign* [14]. Given the special audio recording conditions, automatic phonetic alignment proved to be especially challenging, and generic, one-pass, Viterbi-based implementations failed to provide sufficiently accurate results. By using an environment-adaptive, iterative alignment procedure, *SailAlign* proved to be more robust, and has allowed us to transcribe the MRI-TIMIT data with greater accuracy. An illustrative example of a phonetically-aligned utterance is given in Fig. 1.

3. Database Description

3.1. Corpus

The corpus spoken by study participants was modeled after the 460-sentence MOCHA-TIMIT database [1]. The sentence set is designed to elicit all phonemes of American English in a wide range of prosodic and phonological contexts, with the connected speech processes characteristic of spoken English, including assimilations, lenitions, deletions and mergers. As well as providing a phonologically comprehensive sample of English, this corpus was chosen to allow for systematic comparison of the MRI data with mid-sagittal EMA data [1], and to provide an additional resource for researchers who have previously made use of the MOCHA-TIMIT database.

A total of 98 minutes of speech data have been acquired from the four speakers currently in the database (Table 2). An example utterance is illustrated in Fig. 1. A full list of sentences in the corpus is provided at the MRI-TIMIT project page: `http://sail.usc.edu/span/mri-timit/index.php`.

4. Database Analysis Tools

To facilitate use of the database, a number of tools are being developed for inspection and analysis of these data. Brief descriptions of some of the major tools are given below; further details are found at the MRI-TIMIT project page: `http://sail.usc.edu/span/mri-timit/index.php`, where software, documentation, and example utterances may also be downloaded.

ID	#SENT	#PHON	MEAN T_{Sent}	h:mm:ss
M1	464	14,312	2.73 sec	0:21:05
M2	460	14,194	2.59 sec	0:19:50
W1	460	14,189	2.55 sec	0:19:31
W2	463	14,181	2.52 sec	0:19:25
TOTAL	1,847	56,876	2.59 sec	1:19:51

Table 2: *Total number of sentences, total number of phones, mean sentence durations, and total duration of all utterances, for each subject in the MRI-TIMIT database.*

4.1. Data Inspection and Labeling

A graphical user interface has been developed to allow for audition, labeling, tissue segmentation, and acoustic analysis of the MRI-TIMIT 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 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.

4.2. Automatic Articulator Tracking

A robust tool has been developed for unsupervised region segmentation of the upper airway, jaw and supraglottal articulators, which is more suitable for processing long sequences of MR Images in the database. The segmentation algorithm [15] 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 (Fig. 2).

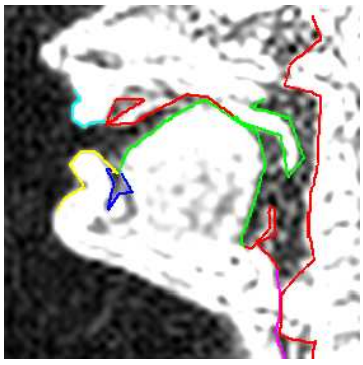

Figure 2: ***Region segmentation of articulators in rtMRI data****: segment [kɲ] in utterance /ˈwɜˠːkɪŋ/ produced by Subject M1.*

4.3. Direct Image Analysis

While tissue segmentation is invaluable in the study of tongue shaping and vocal tract dynamics, other phonetic phenomena are better examined in rtMRI data with direct image analysis, which makes direct use of pixel intensities and their variation over time, without appealing to conventional image processing features such as edges [16]. Direct analysis of MRI sequences can provide robust estimation of articulatory dynamics, without the need for pre-processing, anatomical delineation, or tissue tracking – techniques which are susceptible to segmentation error, and are typically labor- and computationally-intensive.

A number of tools have been developed to facilitate direct analysis of the MRI-TIMIT image data. Coordinative relationships between articulators can be quantified by calculating pixel correlation [16], and kinematics of constriction formation and release can be estimated directly from regional pixel intensity variation in MR Image sequences. These approaches have been used to automatically determine constriction location and kinematic differences between stops [17], and intergestural timing relationships in consonants composed of multiple gestures [18], and are directly applicable to the data in MRI-TIMIT.

5. Illustrative Applications

As a large-scale, multi-speaker corpus in which the entire upper airway is imaged dynamically, the MRI-TIMIT database creates opportunities for research in a number of areas which have not been easily explored using traditional speech corpora.

5.1. Phonetic and Phonological Analysis

5.1.1. Deriving Representations from Data

Because it provides global information about the articulatory space compared to data obtained using flesh-point sensing modalities, rtMRI can offer new insights into articulatory representations. Enhanced articulatory representations derived from rtMRI data have the potential to inform work in phonetic and phonological theory, and speech and speaker modelling.

5.1.2. Role of Variability

The inclusion of data from multiple speakers allows for the study of subject-specific production strategies, production invariants, and the influence of individual speaker morphology on articulatory and acoustic goals of production. rtMRI data are especially well suited to these topics because the morphological characteristics of speakers are simultaneously captured along with their production kinematics, and because speakers are not required to pose their articulators, as with static MRI. We have begun to use MRI-TIMIT to quantify individual differences in the size, shape and relative proportions of the various articulators [19], and this continues to be a topic of active investigation.

5.1.3. Characterizing Place of Articulation

Phonetic methodologies other than palatography and X-ray are inherently unsuitable for characterizing place of articulation in fine detail. The utility of EMA for analyzing coronal articulation, for example, depends critically on the way that sensor coils are located on the front of the tongue; the optimal location will vary between speakers and utterances, is not fixed, and cannot be known beforehand.

Not only does rtMRI interfere less with articulation than flesh-point tracking, it provides a rich source of information about the way in which the tongue contacts the passive articulators for different consonants, speakers, speech styles, and in different vowel contexts. An algorithm developed to automatically detect the constriction center for VCV sequences in rtMRI data [20] is providing new insights into consonant kinematics and phonological representations [17], and is directly applicable to the data in MRI-TIMIT.

5.2. Modeling and Applications

5.2.1. Articulatory-acoustic Mappings

A central problem in phonological theory is characterizing the many-to-one mapping from representations in the speech articulatory space to acoustic space [21, 22]. The problem is compounded by our incomplete knowledge of the articulatory goals of production, but rtMRI provides 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.

5.2.2. Dynamic Articulatory Modeling

rtMRI speech production data can facilitate research in the dynamic articulatory modeling of the full vocal tract shape. In our ongoing work, we investigate the application of statistical graphical models that can capture the spatio-temporal dependencies between various articulators in a data-driven manner [23, 24]. Findings in this area can potentially also inform theories of speech production. To this end, the MRI-TIMIT database represents a novel experimental platform with which to link speech theory with realistic observations.

5.2.3. Automatic Speech Recognition

Dynamic articulatory data have the potential to inform approaches to automatic speech recognition (ASR) [7, 8]. Because it provides such a rich source of global information about vocal tract dynamics during speech production, it is worth investigating the discriminatory power of rtMRI-derived production features in ASR approaches. Additionally, examining the extent to which production-oriented features can provide complementary information to that provided by acoustic features, will offer further insights into the role of articulatory knowledge in automatic speech recognition [24].

6. Future Directions

The MRI-TIMIT database currently consists of midsagittal data from four speakers, and a collection of supporting tools. The goal of this project is to build on this foundation by adding more types of data acquired from more speakers, and to expand the toolset to allow for more sophisticated inspection and analysis of these data. The database will initially be augmented with data from more speakers of General American English, but ultimately also with speakers of other varieties of English, and speakers of other languages. We intend to acquire video with higher frame-rates and improved SNR, and to incorporate data acquired from imaging planes other than midsagittal, including mid-lingual coronal cross-sections.

7. Acknowledgements

Research supported by NIH Grant R01 DC007124-01.

8. References

[1] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th SSP*, Kloster Seeon, 2000, pp. 305–308.

[2] R. Rose, J. Schroeter, and M. M. Sondhi, "The potential role of speech production models in automatic speech recognition," *JASA*, vol. 99, pp. 1699–1709, 1995.

[3] A. M. Liberman and D. H. Whalen, "On the relation of speech to language," *Trends in Cog. Sci.*, vol. 4, no. 5, pp. 187–196, 2000.

[4] C. Y. Espy-Wilson, "Articulatory strategies, speech acoustics and variability," in *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, J. Slifka, S. Manuel, and M. Mathies, Eds. MIT, 2004, pp. B62–B76.

[5] T.-P. Jung, A. Krishnamurthy, S. Ahalt, M. Beckman, and S. Lee, "Deriving gestural score from articulator-movement records using weighted temporal decomposition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 2–18, 1996.

[6] S. Lee, E. Bresch, and S. S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," in *Proc. ISSP*, Ubatuba, Brazil, 2006, pp. 11–17.

[7] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *JASA*, vol. 121, no. 2, pp. 723–742, 2007.

[8] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Trans. Audio, Speech and Lang. Proc.*, in press.

[9] J. R. Westbury, G. Turner, and J. Dembowski, "X-Ray microbeam speech production database user's handbook," Waisman Center, University of Wisconsin, Tech. Rep., 1994.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[11] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *JASA*, vol. 115, pp. 1771–1776, 2004.

[12] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time MRI," *IEEE Sig. Proc. Mag.*, vol. 25, no. 3, pp. 123–132, 2008.

[13] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime MRI scans," *JASA*, vol. 120, no. 4, pp. 1791–1794, 2006.

[14] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for VLSPR*, 2011.

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

[16] A. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-Driven Analysis of Realtime Vocal Tract MRI using Correlated Image Regions," in *Proc. Interspeech*, Makuhari, Japan, Sep 2010.

[17] C. Hagedorn, M. Proctor, and L. Goldstein, "Automatic Analysis of Geminate Consonant Articulation using Real-time Magnetic Resonance Imaging," in *Proc. 9th ISSP*, Montreal, Canada, 2011.

[18] M. Proctor, A. Lammert, L. Goldstein, and S. Narayanan, "Temporal analysis of articulatory speech errors using direct image analysis of rtMRI," *JASA*, vol. 128, no. 4, pp. 2289–2289, 2010.

[19] A. Lammert, M. Proctor, and S. Narayanan, "Morphological Variation in the Adult Vocal Tract: A Study Using rtMRI," in *Proc. 9th ISSP*, Montreal, Canada, 2011.

[20] M. I. Proctor, N. Katsamanis, L. Goldstein, C. Hagedorn, A. Lammert, and S. Narayanan, "Direct estimation of articulatory dynamics from real-time magnetic resonance image sequences," in *Proc. Interspeech*, Florence, Italy, Aug 2011.

[21] B. Atal, J. Chang, M. Mathews, and J. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *JASA*, vol. 63, pp. 1535–1555, 1978.

[22] P. K. Ghosh and S. S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *JASA*, vol. 128, no. 4, pp. 2162–2172, 2010.

[23] E. Bresch, A. Katsamanis, L. Goldstein, and S. Narayanan, "Statistical multi-stream modeling of real-time MRI articulatory speech data," in *Proc. Interspeech*, Makuhari, Japan, 2010.

[24] A. Katsamanis, E. Bresch, V. Ramanarayanan, and S. Narayanan, "Validating rt-MRI based articulatory representations via articulatory recognition," in *Proc. Interspeech*, Florence, Italy, Aug 2011.