USC-TIMIT: A database of multimodal speech production data

Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee
Signal Analysis and Interpretation Laboratory, University of Southern California, 3740 McClintock Avenue, Los Angeles, California 90089-2564

Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu
Magnetic Resonance Engineering Laboratory, University of Southern California, 3740 McClintock Avenue, Los Angeles, California 90089-2564

Louis Goldstein, Dani Byrd
Department of Linguistics, University of Southern California, 3601 Watt Way, Los Angeles, California 90089-1693

Erik Bresch
Philips Research, High Tech Campus 5, 5656 AE, Eindhoven, Netherlands

Athanasios Katsamanis
School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytechniou Str., Athens 15773, Greece

Michael Proctor
MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith NSW 2751, Australia

Abstract

USC-TIMIT is a speech production database under ongoing development, which currently includes real-time magnetic resonance imaging data from five male and five female speakers of American English, and electromagnetic articulography data from five of these speakers. The two modalities were recorded in two independent sessions while the subjects produced the same 460 sentence corpus. In both cases acoustics were recorded in parallel with the articulatory data, and phonemically transcribed. The database, and companion techniques for reconstruction, processing and linguistic analysis, are freely available to the research community.
1 Introduction

Real-time magnetic resonance imaging (rtMRI) is an important emerging tool for speech research,\cite{1,2} providing dynamic information from the entire mid-sagittal plane of a speaker’s upper airway, or any other scan plane of interest, from a single utterance 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 – regions of the tract which cannot be monitored with other techniques. While sampling rates are currently lower than for Electromagnetic Articulometry (EMA)\cite{3} or X-Ray Microbeam (XRMB),\cite{4} rtMRI is a unique source of dynamic information about vocal tract shaping and global articulatory coordination.

We describe here an ongoing 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 (from \url{http://sail.usc.edu/span/usc-timit}). In parallel, we are collecting and making available 3D EMA data, from the same speakers using the same stimuli. EMA data complement rtMRI data by providing faster acquisition rates (but more partial spatial information), while these two modalities may be advantageously combined using co-registration techniques.\cite{5} We call this collection the USC-TIMIT database.

2 Real-time MRI acquisition and processing

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 through a mirror without moving their head. Sentences were presented one at a time, elicited at a natural speaking rate.

2.1 Tissue Sensing and Image Reconstruction

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.\cite{2,6} A body coil was used for radio frequency (RF) signal transmission. A custom 4-channel upper airway receiver coil array, with two anterior coil elements and two coil elements posterior to the head and neck, was used for RF signal reception. A 13-interleaf spiral gradient echo pulse sequence was used ($T_R = 6.164\ \text{msec}$, $\text{FOV} = 200 \times 200\ \text{mm}$, flip angle = $15^\circ$, receiver bandwidth = $\pm 125\ \text{kHz}$). The spiral fast gradient echo sequence consisted of slice-selective excitation, spiral readout, rewinder, and spoiler gradients. Slice thickness was 5 mm, located mid-sagittally; image resolution in the sagittal plane was $68 \times 68\ \text{pixels}$ ($2.9 \times 2.9\ \text{mm}$). Scan plane localization of the mid-sagittal slice was performed using RTHawk (HeartVista,
MR image reconstruction was performed using MATLAB (Mathworks, South Natick, MA). Image frame was produced by the use of gridding reconstruction from data sampled along spiral trajectories. Gridding reconstruction was performed on each individual anterior coil data. Root sum-of-squares of the reconstructed anterior coil images was taken to improve image signal-to-noise ratio as well as spatial coverage of the vocal tract. Reconstructed images from the posterior two coil elements showed spatial aliasing artifacts and thus were not considered for coil image combination. Sliding window technique was used to allow for view sharing and thus increase frame rate. Initial image frame was reconstructed from the first 13 consecutive spiral data. The next image frame was reconstructed from the 8 to 20 consecutive spiral data, and so on. The TR-increment for view sharing was 7, and end result was the generation of MRI movie with a frame rate of $1/(7*\text{TR}) = 1/(7*6.164 \text{ msec}) = 23.18 \text{ frames/sec}$. Alternatively, one can lower the TR-increment to 1. The 1-TR sliding window reconstruction maximizes frame rate to 162.23 frames/sec, but this increases file size of the movie as well as image reconstruction time.

2.2 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 based set up. 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.

Noise generated from the operation of the MRI scanners needed to be canceled satisfactorily in order to perform more detailed analyses of the audio for linguistic and statistical modeling purposes. Noise cancellation was performed using a custom adaptive signal processing algorithm that takes into account the periodic structure of the noise generated by the scanner consistent with the rtMRI acquisition set. Note that 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 with the experimenters via both the fiber-optic microphone setup as well as the in-scanner intercom system.

2.3 Phonetic alignment

Time-aligned phonetic transcriptions of all utterances in the database were generated from the audio recordings, using the freely available tool SailAlign. 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 USC-TIMIT data with greater accuracy.
<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Race</th>
<th>Age</th>
<th>Birthplace</th>
<th>EMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Male</td>
<td>White</td>
<td>29</td>
<td>Buffalo, NY</td>
<td>✓</td>
</tr>
<tr>
<td>M2</td>
<td>Male</td>
<td>White</td>
<td>33</td>
<td>Ann Arbor, MI</td>
<td>✓</td>
</tr>
<tr>
<td>M3</td>
<td>Male</td>
<td>White</td>
<td>26</td>
<td>Madison, WI</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>Male</td>
<td>White</td>
<td>26</td>
<td>St. Louis, MO</td>
<td>✓</td>
</tr>
<tr>
<td>M5</td>
<td>Male</td>
<td>White</td>
<td>27</td>
<td>Mammoth, CA</td>
<td></td>
</tr>
<tr>
<td>W1</td>
<td>Female</td>
<td>White</td>
<td>23</td>
<td>Commack, NY</td>
<td>✓</td>
</tr>
<tr>
<td>W2</td>
<td>Female</td>
<td>White</td>
<td>32</td>
<td>Westfield, IN</td>
<td></td>
</tr>
<tr>
<td>W3</td>
<td>Female</td>
<td>White</td>
<td>20</td>
<td>Palos Verdes, CA</td>
<td></td>
</tr>
<tr>
<td>W4</td>
<td>Female</td>
<td>White</td>
<td>46</td>
<td>Pittsburgh, PA</td>
<td></td>
</tr>
<tr>
<td>W5</td>
<td>Female</td>
<td>White</td>
<td>25</td>
<td>Brawley, CA</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: Study Participants – Demographic Details. Last column indicates if EMA data are available.

3 EMA acquisition and processing

EMA data were collected using the NDI Wave Speech Research System (NDI). These data comprise clean speech audio and 3D trajectories of the movement of sensors attached on articulators. Three sensors of the NDI Wave system were used as reference sensors, attached to the nasal bridge and behind left and right ears. Three sensors were glued on the midline of the tongue, the front-most sensor at 0.5 - 1 cm behind the anatomical tongue tip, and the rear-most sensor as far back as possible. The other three sensors were attached to the surfaces of the lower incisor, lower lip and upper lip. 3D spatial coordinates of sensors were recorded at a sampling rate of 100 Hz. After data collection, missing sensor tracking points were estimated by piecewise cubic Hermite interpolation, data were corrected for head movement and sensor trajectories were smoothed by low pass filtering at 20 Hz. The speech audio waveform was recorded simultaneously through a microphone at a sampling rate of 44.1 kHz, then it was downsampled to 16 kHz.

4 Database Description

To date, rtMRI data have been acquired from ten native speakers of General American English (Table 1), none of whom reported abnormal hearing or speaking development or pathologies. Figure 1 shows images of the vocal tracts of these speakers (single frames extracted from rtMRI data). EMA data have been acquired for five of these speakers (see Table 1).

In both rtMRI and EMA acquisitions, the corpus spoken by study participants was modeled after the 460-sentence MOCHA-TIMIT database. [11] 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. In addition to providing a phono-
logically comprehensive sample of English, this corpus was chosen to provide an additional resource for researchers who have previously made use of the MOCHA-TIMIT database. Figure 2 shows the sequence of rtMRI frames captured when subject M1 utters the phrase “Bright sunshine shimmers on the ocean”.

5 Database Analysis Tools

To facilitate use of the database, a number of tools are being developed for data inspection and analysis.

5.1 Data Inspection and Labeling

A graphical user interface has been developed to allow for audition, labeling, tissue segmentation, and acoustic analysis of the USC-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. Figure 3 shows a screenshot of this GUI.

5.2 Automatic Articulator Tracking

By identifying vocal tract tissues boundaries in rtMRI image frames, the location of articulators can be compared at different points in time, vocal tract apertures may 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 region segmentation of the upper airway, jaw and supraglottal articulators, which is suited for processing long sequences of MR images has been achieved by exploiting spatial representations of the MR data in the frequency domain. [12] 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. See Figure 4 for an example of air-tissue boundaries produced by this algorithm.

For other types of analysis, it is advantageous to define vocal tract outlines parametrically, with respect to a coordinate system anchored at anatomical landmarks. 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. [13] The method also allows for the use of reference boundaries and manual supervision to guide segmentation of anatomical features which are poorly imaged using magnetic resonance due to low signal-to-noise ratios and scarcity of soft tissue, such as dentition and the hard palate. An example of segmentation with this method can be seen in Figure 3.

5.3 Direct Image Analysis

rtMRI data is proving to be an ideal modality with which to study constriction kinematics, using computationally efficient methods of “direct image analysis,” which by-pass the need to first identify tissue boundaries in the upper airway. Pixel intensity in an MR image is directly proportional to the presence 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. [14] Further, 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. [15]

A number of tools have been developed to facilitate direct analysis of the USC-TIMIT image data. Coordinative relationships between articulators can be quantified by calculating pixel correlation [14], and kinematics of constriction formation and release can be estimated directly from regional pixel intensity variation in MR Image sequences.

6 Conclusion

The USC-TIMIT database currently consists of mid-sagittal rtMRI data from ten speakers producing the 460-sentence MOCHA-TIMIT corpus, with complementary EMA data for a subset of these speakers producing the same corpus, and a collection of supporting analysis tools, all made freely available to the research community. 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 mid-sagittal, including mid-lingual coronal cross-sections.
References


Figure 1: Extracted frames from rtMRI data, showing the vocal tracts of the ten speakers in the database.
Figure 2: Sequence of rtMRI data, while Subject M1 produces the sentence “Bright sunshine shimmers in the ocean”.
Figure 3: Graphical user interface allowing for audition, labeling, tissue segmentation, and acoustic analysis of the rtMRI data, displaying an example of parametric segmentation.
Figure 4: Region segmentation of articulators in rtMRI data: segment [kŋ] in utterance /waːkŋ/ produced by Subject M1.