ARTICULATION OF ENGLISH VOWELS IN RUNNING SPEECH: A REAL-TIME MRI STUDY

Michael Proctor[1], Chi Yhun Lo[1], Shrikanth Narayanan[2]

Macquarie University[1], University of Southern California[2]
michael.proctor@mq.edu.au

ABSTRACT

16,105 vowels produced by three speakers of American English were examined using real-time MRI, to provide further insights into tongue shaping and articulatory contrast in stressed and unstressed positions in fluent speech. High front vowels were found to maintain characteristic lingual postures in prosodically weak environments. Non-high vowels were articulated with a more raised dorsum and different pharyngeal constrictions in unstressed positions, compared to their stressed counterparts. The data reveal complex patterns of reduction, influenced by individual speaker vocal tract morphology, that resist simple characterization as 'centralization'.

Keywords: vowel production, English prosody, stress, tongue, real-time MRI

1. INTRODUCTION

The phonological characterization of unstressed vowels in English has been a long-standing topic of debate. In early generative frameworks, vowels occurring in unstressed positions are universally transcribed as schwa [4], reflecting a view that "all unstressed vowels will surface as . . . [ə]" [5]. Two reduced vowels are commonly distinguished in English: both central, but distinguished by height [26] – a view reflected by Ladefoged [15], who transcribes all reduced vowels as [ə] or [ɪ]. Hayes [11] observes that at least three vowel qualities other than schwa ([i]–[ɪ]–[oʊ]) may contrast in unstressed positions in different environments in English.

The lack of consensus about the characterization of unstressed vowels reflects differences of opinion about levels of representation [3], and different approaches to transcription of a gradient phenomenon [19, 20, 21], but also results in part from phonological description uninformed by phonetic data on vowel production in prosodically weak positions. Studies examining the acoustic distribution of English unstressed vowels [1, 8] have shown that they cluster into at least two groups differing in height, consistent with Kondo's [14] conclusion that schwa only has an F1 target value [6].

Another complicating factor is that there are lexical [12] and morphological influences on the ways that English vowels reduce: Browman & Goldstein distinguish between 'lexical' schwas, appearing to have an articulatory target [25, 7], and 'epenthetic' schwas, which may be targetless [2, 10, 16, 17].

The goal of this study is to shed more light on mechanisms and patterns of English vowel reduction, building on previous work by:

1. examining articulation of the *full set* of American English vowels, across a *wider range of prosodic environments*,
2. exploiting a large multi-speaker articulatory-acoustic corpus of *running speech*, containing multiple tokens of each vowel,
3. making use of real-time MRI to examine *shaping of the whole tongue*, and articulation of the whole vocal tract

2. METHOD

Data were obtained from MRI-TIMIT: a freely-distributed large-scale database of synchronized audio and real-time magnetic resonance imaging (rtMRI) data for speech research [23, http://sail.usc.edu/span/usc-timit/]. Because it provides dynamic information from the entire mid-sagittal plane of a speaker's upper airway, rtMRI is a unique source of information about vocal tract shaping during vowel production [22]; furthermore, because the accompanying acoustic recordings are phonemically transcribed using forced alignment, MRI-TIMIT allows targeted investigation of vowel articulation in specific phonological environments.

2.1. Corpus Database

Vowel production was examined in three native speakers of General American English (2 female). Each vowel uttered by each speaker in the production of the full corpus of 460 unique sentences was analyzed – a total of 16,105 vowel tokens in 59 minutes (3,526s) of speech (Table 1).

2.2. Encoding Prosodic Information

Time-aligned phonetic transcriptions of the MRI-TIMIT utterances (created using a custom forced-alignment algorithm [13]), were first augmented with prosodic information to allow stress-sensitive

Table 1: Corpus Details: Subject IDs, numbers of sentences, words, segments & vowels produced, and total duration of utterances (sec).

SUBJ	SENT	WORD	SEG	VOW	DUR
M2	460	3,449	14,194	5,363	1,190s
W1	460	3,456	14,189	5,367	1,171s
W2	463	3,449	14,181	5,375	1,165s
Total	1,383	10,354	42,564	16,105	3,526s

searches. Each word was encoded with the corresponding phonemic transcription from the CMU Pronouncing Dictionary [27] to include lexical stress marking, alongside the existing phonetic data.

In the resulting enhanced transcriptions, each vowel is marked as underlyingly primary-stressed (e.g. IH1: *'swing'*), secondary-stressed (IH2: *'millionaires'*), or unstressed (IH0: *'distress'*), while the output of the forced alignment indicates the specific vowel quality realized in the actual utterance (Fig. 1). The augmented transcriptions therefore encode three types of information about the vowels produced by each speaker in the corpus: (i) underlying vowel quality, (ii) lexical stress, and (iii) phonetic realization.

Figure 1: Enhanced time-aligned phonetic transcription. Prosodically-sensitive transcription of an example utterance by Speaker W1: start & end times (s) of each segment, phonetically-transcribed phone, phonemic lexical transcription containing stress marking, and context utterance.

```
4.064620,6.406960,sil,,
6.406960,6.477030,ah,AH0 L AW1, ALLOW LEEWAY HERE
6.477030,6.587140,l,AH0 L AW1, ALLOW LEEWAY HERE
6.587140,6.947500,aw,AH0 L AW1, ALLOW LEEWAY HERE
6.947500,6.977530,sil,, ALLOW LEEWAY HERE
6.977530,7.007560,l,L IY1 W EY2, ALLOW LEEWAY HERE
7.007560,7.047600,iy,L IY1 W EY2, ALLOW LEEWAY HERE
7.047600,7.077630,w,L IY1 W EY2, ALLOW LEEWAY HERE
7.077630,7.127680,ey,L IY1 W EY2, ALLOW LEEWAY HERE
7.127680,7.157710,hh,HH IH1 R, ALLOW LEEWAY HERE
7.157710,7.327880,ih,HH IH1 R, ALLOW LEEWAY HERE
7.327880,7.578130,r,HH IH1 R, ALLOW LEEWAY HERE
7.578130,8.769320,sil,,
```

2.3. Corpus Search Tools & Method

Search tools were developed to facilitate context-sensitive searches of the corpus data, using regular expressions, so that specific phones could be targeted in the phonological environments of interest. Unconstrained searches were first conducted to establish the total number of occurrences of each vowel in each prosodic environment (Table 2).

Having identified the prosodic distribution of vowels in the data, targeted searches were used to locate primary stressed vowels in phonological environments which best revealed their intrinsic phonetic properties. 5-gram searches (5-segment sequences) were found to be necessary to select prototypical exemplars of each vowel, because of the pervasive coarticulatory influence of surrounding segments. Stressed vowels produced in bilabial and glottal contexts, or adjacent to intervals of silence were selected; target vowels produced within 3-segment's proximity to liquid consonants or competing primary stressed vowels were excluded, due to the strong coarticulatory effects these were found to have on the vowel of interest.

To ensure that the resulting subset of 'prototypical' vowels selected was a representative sample, a criterion was established for targeted searches to return at least 5% of the total number of primary stressed tokens. If this criterion was not met, the search was widened to include vowels produced in the vicinity of labiodental, then dental, then alveolar segments, until sufficient tokens were found.

2.4. Image Analysis

For each vowel of interest located in the acoustic transcriptions, the corresponding image frame was identified in the MRI data, to show the configuration of the speaker's vocal tract at the same point in time. Image frames were selected at the midpoint of the two timestamps delineating the vocalic interval, to capture the tongue shape of the speaker as close as possible to the articulatory target of the vowel.

Because of the high speech rate (12.1 segs/sec), and the relatively slower default video rate (23.18 f.p.s), image frames were reconstructed from the raw MRI data using a sliding window technique, with a temporal resolution of 6.164 ms (162 f.p.s.), so that the lingual posture could be captured more accurately with respect to the acoustic landmark.

Composite images constructed from individual MRI frames were then generated to show mean lingual postures. For each vowel produced in each prosodic position, a mean image was constructed from the set of all individual MR images capturing one token of that vowel. Composite images were cubic-interpolated to 5 $\times$ original resolution (340 x 340 px), and contrast was adjusted to maximally enhance lingual resolution.

3. RESULTS

The frequency of each vowel produced by speaker W1, and their distribution across prosodic positions are shown in Table 2. Because vowels distributed with near identical frequencies for the other two speakers in the study, the patterns of production observed for W1 will be discussed, with reference to individual speaker variation where relevant.

The data reveal large differences in vowel frequency in American English. More than a third

Table 2: Distribution and frequency of vowels: speaker W1. No. occurrences of each vowel in primary, secondary, and lexically unstressed positions. Column 6 indicates occurrences of vowels realized with a phonetic quality differing from the underlying lexical representation.

ARPA	IPA	PRI	SEC	UNS	OTH	TOTAL
IH	ɪ	378	24	362	285	1049
AH	ʌ	207	14	655	104	980
IY	i	262	12	198	24	496
ER	ɝ	93	5	276	92	466
AE	æ	270	33	16	51	370
EH	ɛ	265	16	14	20	315
AO	ɔ	183	19	3	63	268
EY	eɪ	188	48	5	27	268
UW	u	202	14	23	8	247
AY	aɪ	191	29	19	3	242
AA	ɑ	198	23	12	7	240
OW	oʊ̯	136	29	29	7	197
UH	ʊ	50	6	3	31	90
AW	aʊ̯	82	5	0	0	87
OY	oɪ	39	6	2	1	48
Total		**2740**	**283**	**1617**	**723**	**5363**

of all vowels in the corpus (37.7%) use one of the two qualities [ɪ] or [ʌ], and the majority of vowels (55.6%) use one of the four most frequent vowel qualities ([ɪ]-[ʌ]-[i]-[ɝ]). Only 7.9% of vowels in the corpus use one of the four least frequent qualities ([oɪ]-[aʊ]-[ʊ]-[oʊ]).

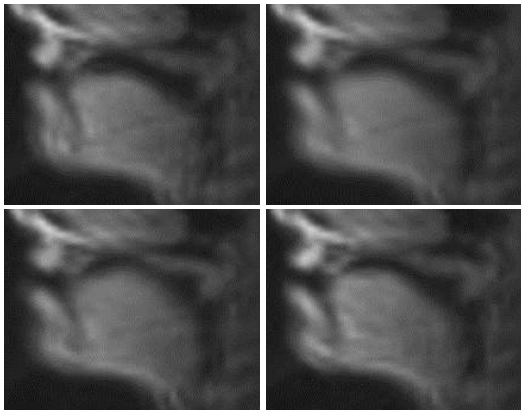
3.1. Vowel Articulation

All vowels in the corpus are realized with different mean articulatory postures in different prosodic environments, although some vowels show larger differences between their stressed and unstressed realizations than others. All unstressed vowels are realized with some aspect of their midsagittal lingual posture more 'centralized', compared to their underlying stressed counterparts; however, the exact way in vowels are articulated in unstressed positions depends on their underlying quality.

3.1.1. Front Vowels

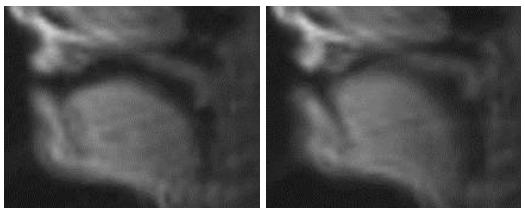
Articulation of the mid front vowel /ɛ/ by speaker W1 across prosodic environments is compared in Fig. 2. 'Prototypical' [ɛ] posture is illustrated by the image in the top left panel, constructed from the subset of primary stressed vowels occurring in positions least influenced by coarticulation with surrounding segments. The top-right image shows the mean posture of *all* 266 [ɛ] vowels in primary stressed positions. Mean tongue posture in unstressed positions (13 tokens) reveals a slightly raised (−2.4 mm) dorsal apex, compared to the prototypical articulation of [ɛ]. Aperture of the midsagittal palatal constriction above the tongue blade is 9.4 mm in the mean stressed vocal tract configuration, and 7.1 mm in the unstressed posture.

Figure 2: Mean articulatory posture: /ɛ/, produced in prototypical (top-left), primary-stressed (top-right), secondary-stressed (bottom-left), and unstressed (bottom-right) positions (Subject W1).



Similar patterns are observed in the different realizations of the low front vowel /æ/ (Fig. 3): vowels appearing in unstressed positions are articulated with a raised tongue dorsum (−2.4 mm) and a wider pharyngeal cavity (+4.1 mm above the epiglottis), compared to their stressed [æ] variants.

Figure 3: Mean articulatory posture: /æ/, produced in stressed (L), and unstressed (R) positions (Subject W1).



3.1.2. High Front Vowels

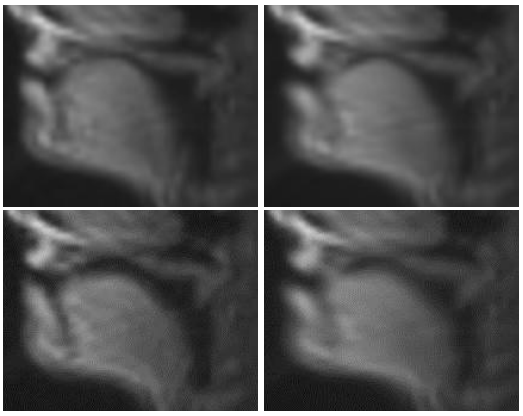
High front vowels showed the least difference in midsagittal articulatory posture between primary-stressed and unstressed positions (Fig. 4). The same basic lingual configuration was maintained across prosodic environments, but both /i/ and /ɪ/ were realized with a slightly less open pharynx in unstressed positions.

3.1.3. Back Vowels

The high back vowel /u/ shows large differences between prototypical and unstressed articulations, which are much more fronted (Fig. 5). The mean stressed lingual posture observed for [u] also more closely resembles the mean posture observed in unstressed environments than the prototypical posture, which suggests that this vowel may be more susceptable to coarticulatory influences than others.

Non-high back vowels exhibited the pattern of reduction exemplified in Fig. 5 (bottom row): un-

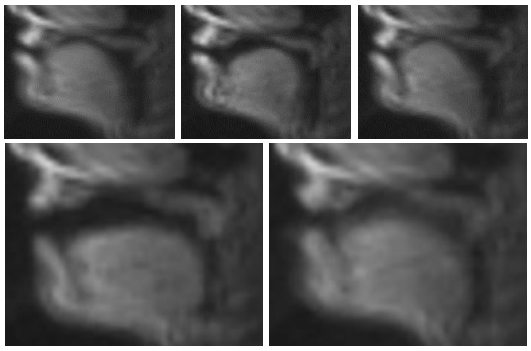
Figure 4: Mean tongue posture, high front vowels: /i/ (top row) and /ɪ/ (bottom row), produced in stressed (L), and unstressed (R) positions (Subject W1).



stressed vowels were realized with a less constricted pharynx, and and advanced, raised tongue dorsum, compared to their stressed counterparts.

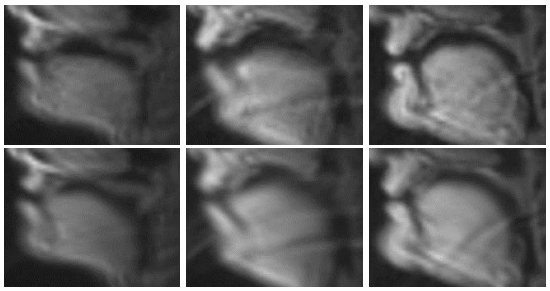
As expected from the vowel quality most commonly associated with schwa in American English, considerable variation in tongue posture was observed for the mid back vowel /ʌ/ across stress positions. Prototypical (primary stressed) and mean unstressed realizations are compared in Fig. 6. The reduced variant of /ʌ/ is realized by all three speakers with a raised dorsal apex. For all three speakers, stressed [ʌ] shows more pharyngeal constriction than the unstressed vowel.

Figure 5: Mean articulatory posture, back vowels: /u/ (top row) and /ɑ/ (bottom row) produced in stressed (L), protypically stressed (C), and unstressed (R) positions (Subject W1).



The data in Fig. 6 also reveal that articulatory patterns of vowel reduction are speaker-specific, and may be influenced by a speaker's vocal tract morphology. Stressed [ʌ] vowels realized by speakers W1 and W2, for example, involve more lingual elongation (tongue-tip to tongue root) than their unstressed counterparts, while both types of [ʌ] are articulated by speaker M2 with a more bunched tongue

Figure 6: Mean articulatory posture: mid-back monophthong /ʌ/ produced in prototypical (top row), and unstressed (bottom row) positions, by speakers W1, W2 & M2 (L–R).



posture, that raises and fronts, without reshaping, in unstressed positions.

4. DISCUSSION

This study demonstrates the utility of the MRI-TIMIT database as a tool for large-scale articulatory phonetic investigation, capable of providing new insights into vowel production, individual speaker variation, and the phonetic correlates of prosody. Mean tongue postures associated with vowel targets in runnning speech were constructed from hundreds of exemplars, using phonetically-aligned real-time MRI data. The data reveal robust patterns of vowel articulation in prosodically strong and weak positions.

High front vowels /i/ and /ɪ/ were found to be articulated with similar midsagittal lingual postures in all prosodic environments, consistent with characterizations of these vowels as phonetically unreduced in unstressed positions [11]. Non-high vowels were found to be articulated with a raised dorsum and variations in pharyngeal constriction in unstressed positions, consistent with their description as centralized [26]; however, the data reveal complex patterns of reduction that resist simple characterization, as they also appear to be influenced by individual speaker vocal tract morphology [18]. Differences between prototypical and stressed lingual postures observed for some vowels – especially /u/ – suggest that these vowel may be more prone to interaction with surrounding segments than other vowels [24], and that running speech may be characterized by pervasive target undershoot [9].

Further insights into the details of vowel reduction will follow from closer analysis of these data, including more detailed quantification of differences in tract shaping, jaw movement, and acoustic analysis.

5. ACKNOWLEDGEMENT

This work was supported by National Institutes of Health grant R01 DC007124.

6. REFERENCES

[1] Bates, S. A. R. 1995. *Towards a Definition of Schwa: an acoustic investigation of vowel reduction in English*. PhD thesis Univ. Edinburgh.

[2] Browman, C. P., Goldstein, L. M. 1992. 'targetless' schwa: An articulatory analysis. In: Docherty, G., Ladd, R., (eds), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge: CUP 26–56.

[3] Burzio, L. 2007. Phonology and phonetics of English stress and vowel reduction. *Language Sciences* 29(2-3), 154–176.

[4] Chomsky, N., Halle, M. 1968. *The Sound Pattern of English*. New York: Harper & Row.

[5] Crosswhite, K. 2004. Vowel reduction. In: Hayes, B., Kirchner, R., Steriade, D., (eds), *Phoneticallly-based Phonology*. Cambridge: CUP 191–231.

[6] Davidson, L. 2006. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *J.Phon.* 34, 104–137.

[7] Flemming, E. 2005. Deriving natural classes in phonology. *Lingua* 115(3), 287–309.

[8] Flemming, E., Johnson, S. 2007. Rosa's roses: reduced vowels in American English. *JIPA* 37, 83–96.

[9] Gay, T., Ushijima, T., Hirose, H., Cooper, F. S. 1974. Effect of speaking rate on labial consonant-vowel articulation. *JASA* 55(2), 385–385.

[10] Gick, B. 2002. An X-Ray Investigation of Pharyngeal Constriction in American English Schwa. *Phonetica* 59(1), 38–48.

[11] Hayes, B. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago: Univ. Chicago Press.

[12] Jurafsky, D., Bell, A., Gregory, M., Raymond, W. D. 2000. Probabilistic relations between words: Evidence from reduction in lexical production. In: Bybee, J., Hopper, P., (eds), *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins 229–254.

[13] Katsamanis, A., Black, M., Georgiou, P., Goldstein, L., Narayanan, S. 2011. SailAlign: Robust long speech-text alignment. *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.

[14] Kondo, Y. 1994. Targetless schwa: is that how we get the impression of stress-timing in English? *Proc. Edinburgh Linguistics Dept. Conf.* 63–76.

[15] Ladefoged, P. 2001. *A course in phonetics*. Fort Worth: Harcourt College Publ.

[16] Lammert, A., Bresch, E., Byrd, D., Goldstein, L., Narayanan, S. S. 2009. An articulatory study of lexicalized and epenthetic schwa using real time magnetic resonance imaging. *JASA* 125(4), 2569–2569.

[17] Lammert, A., Goldstein, L., Ramanarayanan, V., Narayanan, S. to appear. Gestural control in the English past-tense suffix: an articulatory study using real-time MRI. *Phonetica*.

[18] Lammert, A., Proctor, M., Narayanan, S. 2013. Interspeaker variability in hard palate morphology and vowel production. *JSLHR* 56(6), S1924–S1933.

[19] Lindblom, B. 1963. Spectrographic study of vowel reduction. *JASA* 35(11), 1773–1781.

[20] Moon, S., Lindblom, B. 1994. Interaction between duration, context, and speaking style in English stressed vowels. *JASA* 96(1), 40–55.

[21] Mooshammer, C., Hoole, P., Geumann, A. 2007. Jaw and order. *Language and Speech* 50(2), 145–176.

[22] Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D. 2004. An approach to real-time magnetic resonance imaging for speech production. *JASA* 115(4), 1771–1776.

[23] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., , Proctor, M. 2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *JASA* 136(3), 1307–1311.

[24] Recasens, D. 1984. Vowel-to-vowel coarticulation in Catalan VCV sequences. *JASA* 76(6), 1624–1635.

[25] Smorodinsky, I. 2001. Schwas with and without active gestural control. *JASA* 109(5), 2446–2446.

[26] Trager, G. L., Bloch, B. 1941. The syllabic phonemes of English. *Language* 17(3), 223–246.

[27] Weide, R. 1994. CMU pronouncing dictionary.