KU-ARTLEX: A SINGLE-SPEAKER EMA DATABASE FOR MODELING THE ARTICULATORY STRUCTURE OF THE LEXICON

Charles Redmon, Seulgi Shin, and Panying Rong

University of Kansas
redmon@ku.edu

ABSTRACT

Articulatory data (6-channel electro-magnetic articulography, EMA) was recorded on a 26,793-word database of English words – replicating the set used in the Massive Auditory Lexical Decision project [15] – along with two repetitions of 1,200 controlled CVC syllables based on the California Syllable Test [16], from a single male native speaker of Midwestern American English. The primary aim of this database is to serve as a window on the articulatory structure of the lexicon; i.e., what gestural profiles distinguish words in English, and what constitutes minimality (how are minimal pairs defined, if at all) from an articulatory standpoint? Further, using this open-access database, comparable perception experiments testing the predictability of word recognition patterns from articulatory and acoustic profiles can now be run by multiple research groups, providing greater clarity to broader theoretical debates on the nature of the encoding of the speech signal.

Keywords: electro-magnetic articulography, lexicon, open-access database, English

1. INTRODUCTION

Advances in the acquisition, storage, and sharing of articulatory data, particularly electro-magnetic articulography (EMA) [11], ultrasound [3], and real-time MRI [17, 11], have allowed for more precise and scalable specifications of what an articulatory encoding of the speech system might look like. Similarly, the recent release of a 26,793-word database of controlled productions of isolated words by a single speaker of Western Canadian English as part of the Massive Auditory Lexical Decision project [15] has allowed us to begin modeling the acoustic structure of the lexicon, as well as providing a common set of stimuli for different research groups to use in perception experiments designed to test current models spoken word recognition.

From the controlled lexical data in [15] we are capable of building predictive models of listener behavior on the basis of acoustic features, but given that there are also competing paradigms pursuing articulatory gesture-based models of speech perception [6, 4, 5], a more general database with parallel articulatory and acoustic data on a large and representative sample of the lexicon, produced under controlled context-equivalent conditions, is required. Further, given that much of the competing models of spoken word recognition (SWR) assume some kind of abstract phonemic or featural representation of words in the lexicon (e.g., the Neighborhood Activation Model [9], Shortlist [12, 13], TRACE [10]), in order for the articulatory structure-based model of SWR to emerge in a comparable way we need articulatory data that will permit gestural formulations of lexical encoding, contrast, phonological distance (neighborhood density), etc., that are not dependent on assumptions about the segmental phonological composition of words. In other words, we would like to be able to specify the encoding of the lexicon directly from the similarity and contrast structure of articulatory profiles of words in the lexicon, with no need for extrapolation from controlled syllable to word, extrapolation which is often mediated by canonical auditory impression-based abstract segmental transcriptions.[1]

In service of this end we constructed a parallel database to the 26,793-word stimulus set in [15], which was compiled from all unique word types in the Buckeye Corpus ($\sim$8000) [14], an additional 9,000+ words from the English Lexicon Project [2], 10,000 more words in the highest frequency set in COCA [7], and 1,252 compound words from CELEX [1]. Additionally, to provide a reference for more controlled articulations (less subject to lexical factors such as frequency and neighborhood density), a set of CVC syllables based on the California Syllable Test (CaST) [16], containing all combinations of 20 onset consonants, 3 corner vowels, and 20 coda consonants, was recorded. A single male native speaker of Midwestern American English produced all data in the corpus, and while the use of a single speaker necessarily limits the generalizability of the data, the lack of inter-speaker variation allows models of lexical structure to reduce (for the moment) variability to just that which serves to distinguish words in the lexicon. Further, since

any native speaker of a language, in order to be a functional member of a that linguistic community, must have an internally coherent system of phonetic contrasts (so as to not be misunderstood by other speakers), by modeling the articulatory data characterizing a single speaker's (production) lexicon we should be able to derive the critical articulatory information necessary for the encoding of the English speech system. By *internally coherent* we mean that the set of words in the lexicon must be perceptually differentiable, and thereby differentiable to some degree in acoustic and articulatory structure, within the speech of that any given speaker in order for them to be understood by other speakers of English (for this reason the application of such stimuli in perception experiments is vital). Nevertheless, we hope in the future to record additional speakers, as [15] are doing, to improve the generalizability of the database.

2. DATABASE

2.1. Speaker background

The first author, a 30-year-old male native speaker of the South Midland variety of American English [8], produced every token in the database. This speaker lived in Louisville, Kentucky up to age 18, after which he has lived in several locations, most notably for his idiolect including New Delhi, India for 1 year and Hyderabad, India for 2.5 years. He is a trained phonetician, though he attempted to produce all items as naturally as possible according to his dialect.[2]

2.2. Materials

Two sets of data were recorded: (1) a controlled set of 1,200 CVC syllables – 20 onset consonants /p, b, t, d, k, g, ʧ, ʤ, f, v, θ, ð, s, z, ʃ, h, m, n, l, ɹ/ × 3 corner vowels /i, ɑ, u/ × 20 coda consonants /p, b, t, d, k, g, ʧ, ʤ, f, v, θ, ð, s, z, ʃ, m, n, ŋ, l, ɹ/ – based on the California Syllable Test (CaST) materials [16], (2) a set of 26,793 isolated words based on the Massive Auditory Lexical Decision (MALD) project [15]. The CVC syllable set was repeated twice, while only a single repetition of each word in the model lexicon is provided at present, resulting in 29,193 total items.

2.3. Recording

Data were recorded in the Speech Science and Disorders Laboratory at the University of Kansas, in separate 1-3 hour sessions over the course of several months. The speaker produced items in isolation in random order (blocked by CVC syllable and real word sets) by controlling a slide presentation with a remote. EMA data were recorded on an NDI Wave system (Northern Digital Inc., Waterloo, Ontario, Canada) with 6 active sensors – (1) a tongue dorsum sensor placed as far back as possible on the center of the tongue (the exact position was then recorded and used on all subsequent trials), (2) a tongue tip sensor placed 1 cm behind the tip of the tongue, (3) a tongue center sensor equidistant between tip and dorsum sensors, (4-5) sensors on the upper and lower lips, and (6) a jaw sensor affixed to the lower incisors – all referenced to a 3D reference sensor attached to the forehead.

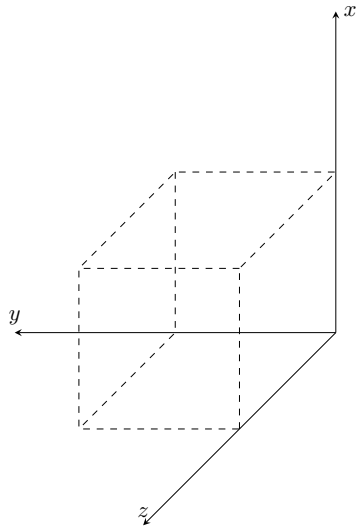

Figure 1: Coordinate system for EMA sensors, assuming a leftward-facing speaker.

EMA data were sampled at 100 Hz. Simultaneous acoustic data was recorded from a head-worn Differoid (cardioid) condenser microphone (Crown Audio Inc., Elkhart, Indiana, USA) positioned approximately 2 cm from the speakers lips, off-axis from the breath stream by approximately 45 degrees. Acoustic data was digitized through a Xenyx 802 mixer (Behringer, Bothell, Washington, USA) at 22050 Hz sampling and 16 bit quantization, where the gain was adjusted to be approximately 70% of the VU range.

The original coordinate system in which the data were recorded, relative to the reference sensor, has tongue height represented on an inverted x axis, tongue advancement on the y axis, and lateral tongue displacement along the z axis. In order to provide more interpretable data, the x axis was then inverted in post-processing so that upward movement is now in the positive direction. Figure 1 displays this coordinate system after post-processing.

2.4. Database format

Time-aligned acoustic and articulatory data are stored in separate files per word: WAV format for audio, and tab-delimited ASCII format (.tsv) for EMA. All data is presently hosted on Dataverse at *https://dataverse.harvard.edu/dataverse/artlex-american-english-midwest*. While data are typically smoothed and interpolated prior to analysis, only raw data is provided in the database, with the one exception being the inversion of the *x* axis as discussed above.

2.5. Phonetic feature summary

As a thorough sample of the English lexicon, instances of each vowel and consonant in the language are expected to occur in reasonably representative proportions for the language as a whole. Table 1 provides type-based counts and percentages of English phonemes in the model lexicon.

For the study of *contrast*, based on a phonemic transcription of the database, approximately 2,905 minimal pairs are present for stop consonants, 1,218 for fricatives, 24 for affricates, 217 for nasals, 394 for liquids, and 7,239 for (semi-)vowels.

2.6. Sample data and analysis

The primary goal of developing this database is to allow for the mapping of phonological structure in the lexicon from articulatory profiles of ensembles of contrastive words. Figures 2 and 3 display acoustic and articulatory data on the minimal pair: "sucker" /ˈsʌkəɹ/ vs. "supper" /ˈsʌpəɹ/. Of course, we have assumed here that we already know that a given pair of words is minimally distinct, the definition of which follows canonical phonological transcription of the two words. In the future we hope such a database will allow notions of *minimality* in phonetic contrast to be derived (possibly based on the gestural scores of words, as mapped from the data) purely from oppositions between words, at least in such cases where the EMA data is sufficient to capture the primary articulatory events in a word.

Figures 2 and 3, as expected, appear to differ primarily in the velar vs. labial/jaw gesture in the medial consonant interval, though these two words also serve to reinforce similarity between the non-contrastive phones in the two words, such as the similar tongue tip gestures for the onset sibilant fricative, and the tongue tip and torsum retraction for the final rhotic. These articulatory profiles also illustrate some of the challenges ahead, as gestures which are in some sense irrelevant during a given interval of

Phoneme	Count	Percent
ə	17268	9.61
t	12532	6.98
n	12013	6.69
ɹ	11777	6.56
s	11484	6.39
l	9802	5.46
r	9593	5.34
k	8416	4.69
d	7791	4.34
i	6171	3.44
p	5655	3.15
m	5559	3.10
z	5511	3.07
ɚ	5491	3.06
ɛ	5187	2.89
æ	4220	2.35
ɑ	4045	2.25
b	3673	2.05
eɪ	3599	2.00
f	3112	1.73
ŋ	2817	1.57
aɪ	2716	1.51
oʊ	2451	1.36
ʃ	2373	1.32
v	2363	1.32
g	2050	1.14
u	1842	1.03
w	1621	0.90
ʤ	1387	0.77
h	1326	0.74
ɔ	1172	0.65
j	1068	0.59
ʧ	1062	0.59
aʊ	784	0.44
θ	616	0.34
ʊ	481	0.27
ɔɪ	268	0.15
ʒ	165	0.09
ð	144	0.08

Table 1: Distribution of transcribed English phonemes (by count and percentage) in the model lexicon based on the stimuli in [15].

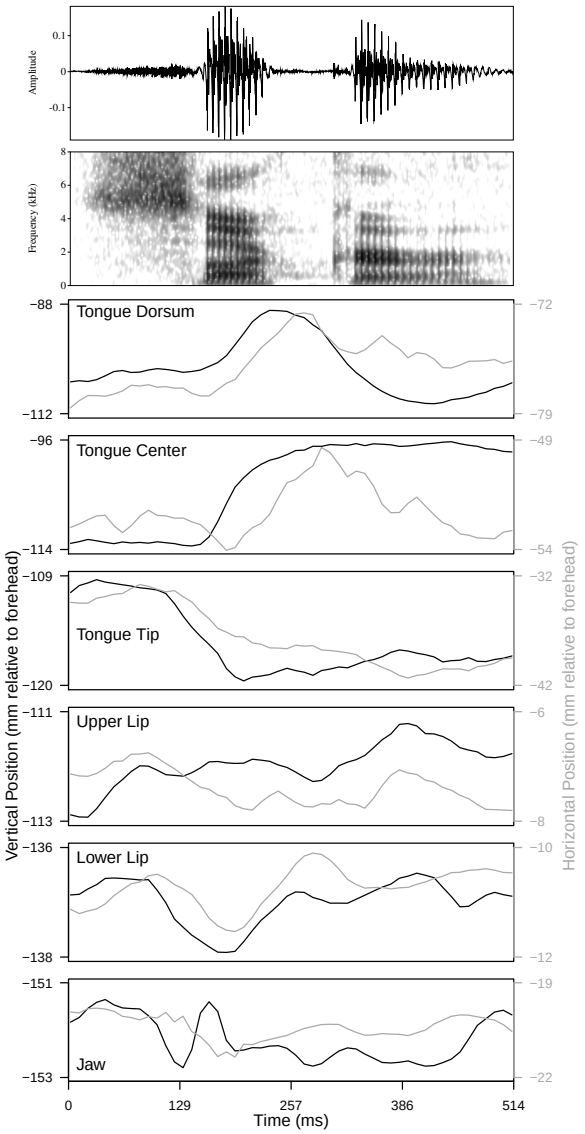

Figure 2: Acoustic and articulatory data (20 Hz low-pass filtered) on the word "sucker" /ˈsʌkəɹ/.

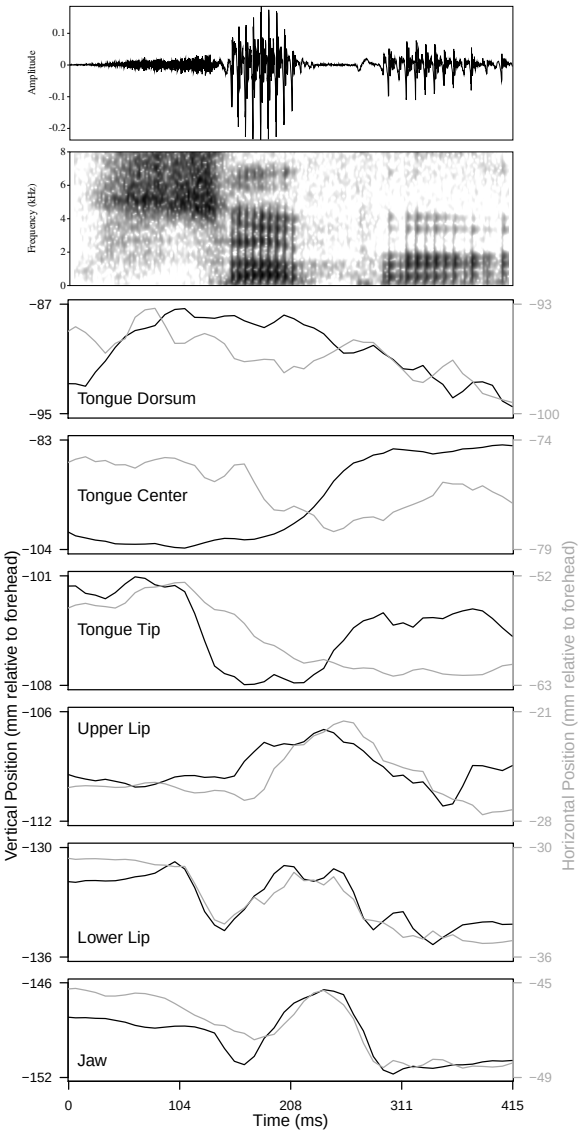

Figure 3: Acoustic and articulatory data (20 Hz low-pass filtered) on the word "supper" /ˈsʌpəɹ/.

the word (e.g., lower lip position for the velar plosive in "sucker"), must also be parsed as such. Interpreting what is signal and what is noise in determining the articulatory structure of the lexicon is fundamental to the enterprise, and ultimately is not unique to the articulatory domain. Acoustic models of the lexicon must also learn similar distinctions, and we expect with the availability of the database in [15] and our own parallel articulatory/acoustic database, progress in answering (and computationally demonstrating solutions to) these problems will be in the offing in the years to come.

3. DISCUSSION

We have presented an open-access database of acoustic and EMA data on a large sample of controlled productions of words and CVC syllables by a single native English speaker in the hope that the data will be used by independent researchers in quantifying various aspects of the articulatory structure of the lexicon. Ultimately, we seek a complete mapping of the articulatory topology of the lexicon, and hope that by making the data freely available this mapping will be more feasible and more robust to the oversights of any one research group.

In addition to the data on English, we are currently developing a parallel database of Korean, and hope to incorporate additional languages in the years to come. This work is part of a wider project to contextualize phonetic categories/gestures more broadly within the lexicon, and to provide data supportive of more thorough and scalable theory validation.

4. REFERENCES

[1] Baayen, R. H., Piepenbrock, R., Gulikers, L. 1995. The CELEX lexical database. *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA*.

[2] Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., Treiman, R. 2007. The English lexicon project. *Behavior Research Methods* 39(3), 445–459.

[3] Berkson, K., de Jong, K., Lulich, S. M., Cavar, M. E. 2018. Building a multilingual ultrasound corpus. *The Journal of the Acoustical Society of America* 144(3), 1717.

[4] Browman, C. P., Goldstein, L. 1989. Articulatory gestures as phonological units. *Phonology* 6(2), 201–251.

[5] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49(3-4), 155–180.

[6] Browman, C. P., Goldstein, L. M. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3, 219–252.

[7] Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2), 159–190.

[8] Labov, W., Ash, S., Boberg, C. 2006. *Atlas of North American English: Phonetics*. Mouton de Gruyter.

[9] Luce, P. A., Pisoni, D. B. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19(1), 1–36.

[10] McClelland, J. L., Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18(1), 1–86.

[11] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., others, 2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *The Journal of the Acoustical Society of America* 136(3), 1307–1311.

[12] Norris, D. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52(3), 189–234.

[13] Norris, D., McQueen, J. M. 2008. Shortlist b: a Bayesian model of continuous speech recognition. *Psychological Review* 115(2), 357–395.

[14] Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E. 2007. Buckeye Corpus of Conversational Speech (2nd release)[www. buckeyecorpus. osu. edu] columbus, oh: Department of psychology. *Ohio State University (Distributor)*.

[15] Tucker, B. V., Brenner, D., Danielson, K. D., Kelley, M. C., Nenadić, F., Sims, M. 2018. The Massive Auditory Lexical Decision database: Toward reliable, generalizable speech research. *Behavioral Research Methods* 1–18.

[16] Woods, D. L., Yund, E. W., Herron, T. 2010. Measuring consonant identification in nonsense syllables, words, and sentences. *Journal of Rehabilitation Research & Development* 47(3), 243–60.

[17] Wrench, A., Hardcastle, W. 2000. A multichannel articulatory database and its application for automatic speech recognition. *In Proceedings of the 5th Seminar of Speech Production* 305–308.

[1] We to not aim in this paper to outline what such a direct derivation of articulatory contrasts from lexical distinctions would entail, but rather aim to first present the kind of data necessary to make progress toward this end.

[2] While the speaker is from Louisville, Kentucky, which exhibits both Midland and Southern speech characteristics, the speaker does not exhibit any notable phonetic or phonological characteristics of Southern speech (one exception might be a greater proportion of /l/ velarization, though thorough analysis of the database is required to bear out this impression), and thus has been classified as more typical of the Midland variety.