Work note: Validating duration and formant measurements for English Diphthong Raising project

Elliott Moreton

2025 February 17 (M)

Goal: Comparison of raw automatic vs. human-supervised measurements of /aɪ/ and /eɪ/ from an experiment about English Diphthong Raising.

(1) Bird's-eye view of the project:

a. Many (maybe most) phonological patterns originate historically through *phonologization* of a phonetic precursor; e.g., vowel harmony from vowel-to-vowel coarticulation (Hyman, 1976).
b. *Research question*: So how do *non*-phonetic factors like UR, prosody, and morphology come to condition phonological patterns?
c. *Three competing hypotheses*: Late Abstractness, Early Abstractness, Abstract Phonetics. See Moreton 2021; Moreton et al. 2024 for details — not needed for today's talk.
d. Case study: English Diphthong Raising (Canadian Raising plus related patterns worldwide; e.g., Chambers 1973; Vance 1987, papers in Davis and Berkson 2021).

The project was presented in detail in P-Side on October 5, 2023.

(2) Experiment 1 (December 2022) collected pronunciation and judgement data from speakers across the U.S. The audio recordings then had to be processed to extract durations and formant measurements. This handout asks

a. whether human supervision and correction actually improved the quality of the acoustic measurements
b. whether the procedure followed by the humans was replicable, i.e., when different humans follow the same procedure, do they get the same results?
c. how much time does it take the humans, and is the time proportional to the improvement?

(3) This is a collaboration with Kelly Berkson, Stuart Davis, Jeff Lamontagne, and Monica Nesbitt at Indiana University, and Joe Pater at UMass-Amherst. Student RAs on this project include Abigail Amick (UNC-CH MA 2024), Erin Humphreys (UNC-CH MA 2024), and Brandon Osgan (currently UNC-CH first-year MA student). Supported in part by U.S. NSF BCS 1651105, "Inside phonological learning", to E. Moreton and K. Pertsova.

1 Experiment and post-processing

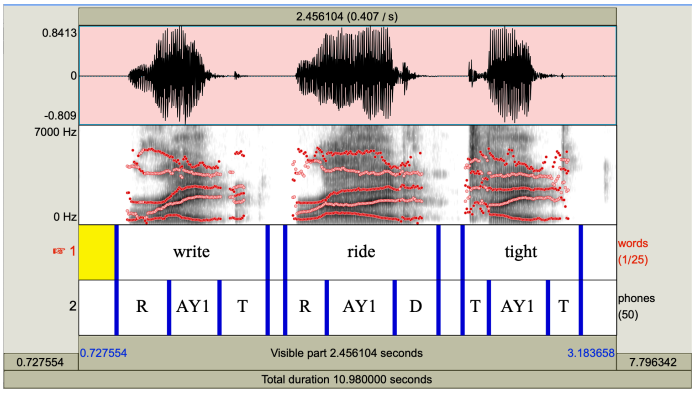
(4) Participants read and sorted five word lists:

List	Categories	*N*	Examples
1	/i/ vs. /æ/	18	beak, grass, flag, bead, …(practice)
2	aɪ /_ ±voice	12	tripe, tribe, fife, five, …
3	aɪ /_ ±voice	13	vibe, bite, strife, high, …
4	aɪ /_ ±voice	17	vibration, titanic, triumphant, dynamic, …
5	aɪ /_ ±voice	17	phytology, Fightology, rhizome, Pisces, …
6	eɪ /_ ±voice	18	ape, Abe, face, phase, …

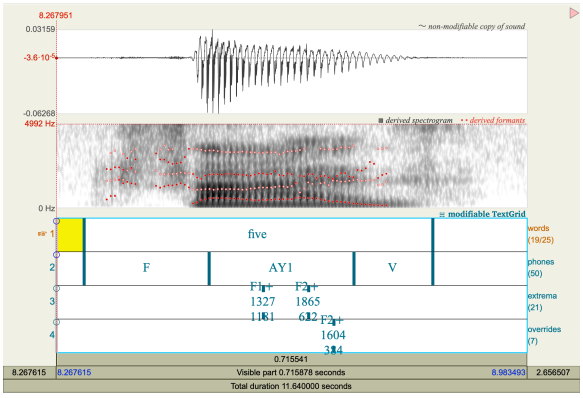
(5) *Data collection*: 217 participants run in December 2022. Stopping criterion was 10 "Perfect Sorters", i.e., 10 participants whose phonological raising index on the monosyllabic /aɪ/ words (Pages 2 and 3) was 1. (So about 5% of participant pool were Perfect Sorters.) The procedure yielded 1240 audio files and 8621 sort responses.

(6) *Data markup and scoring*: This is described in the "Procedure" handout. The gist is that

a. The lists participants were asked to read were checked by humans against the recordings of what they actually said, and the list were edited if necessary to match the words actually said.
b. The Montreal Forced Aligner generated Praat TextGrid files showing phoneme boundaries.



c. The TextGrids and sound files were viewed together in Praat, and the alignment and labelling were corrected by a human.
d. A human viewed each participant's Text Grids and sound files together, and chose formant-tracker settings for that particular speaker. Praat then extracted the formant tracks to Formant files.
e. Homebrew code marked the points of F_1 and F_2 maxima within the critical diphthong, and the values of F_1 and F_2 at those points in the TextGrid files.
f. A human again viewed the TextGrids and sound files together in Praat and corrected errors in the marked points.



There were two independent data streams. One was done entirely by EM all the way through. The other was done by UNC-CH RAs (Abbie Amick and Erin Humphreys for Steps a–d, and Brandon Osgan for Step f).

(7) Both EM and the RAs have done all of Lists 2 and 3, the monosyllabic /aɪ/ items, and List 6, the monosyllabic /eɪ/ items, so that is what this handout is about.

(8) The total amount of time spent by the RAs on the tasks was

RA task	Lists	Hours	Corrections made
Listening to sound files, correcting transcripts	1–6	71.5	349/1136 transcripts
Correcting phoneme boundaries	1–6	158.5	13689/32206 boundaries
Correcting extrema and formant measurements	2,3,6	76.4	(pending)
TOTAL		306	

(9) *Exclusions*: 214 participants got as far as saving their sorting responses at the end of the experiment. Of these, 201 completed the post-experiment questionnaire. For each of the three data sets (NONE, EM, RA) separately, participants were excluded based on the criteria of

a. *Audio*: Did the participant have formant measurements for at least one word? (Done by judgement of scorers.)
b. *Responses*: Did the participant have exactly 95 sort responses? (Done automatically.)
c. *Practice skipping*: Did the participant move at least one word when sorting the practice words (List 1)? (Done automatically.)
d. *Practice failure*: Did the participant correctly sort all of the practice words (List 1), except perhaps for *quiche* and *ant*? (Done automatically.)

```
[1] "Data scored by NONE : Excluded 62 participants, leaving 139 out of 201"
  X_audio        X_responses     X_skipped_practice X_failed_practice  include
 Mode :logical   Mode :logical   Mode :logical      Mode :logical     Mode :logical
 FALSE:194       FALSE:187       FALSE:189          FALSE:149         FALSE:62
 TRUE :7         TRUE :14        TRUE :12           TRUE :52          TRUE :139
[1] "Data scored by EM : Excluded 75 participants, leaving 126 out of 201"
  X_audio        X_responses     X_skipped_practice X_failed_practice  include
 FALSE:178       FALSE:187       FALSE:189          FALSE:149         FALSE:75
 TRUE :23        TRUE :14        TRUE :12           TRUE :52          TRUE :126
[1] "Data scored by RA : Excluded 68 participants, leaving 133 out of 201"
  X_audio        X_responses     X_skipped_practice X_failed_practice  include
 FALSE:185       FALSE:187       FALSE:189          FALSE:149         FALSE:68
 TRUE :16        TRUE :14        TRUE :12           TRUE :52          TRUE :133
```

(10) Alignment and formant-tracking errors often yield absurd values. Suppose we don't do any human supervision, but instead just throw out any token that contains an absurd value. How does that compare with a human? We added a data set called PLAUSIBLE, made from Data Set NONE by excluding any diphthong token that failed at least one of these criteria:

	Criteria			
	aɪ		eɪ	
Measurement	≥	≤	≥	≤
phone duration (s)	0.1	0.6	0.1	0.6
Nuclear F_1 (Hz)	400	1400	350	800
Nuclear F_2 (Hz)	750	2250	1500	3250
Offglide F_1 (Hz)	150	1250	200	600
Offglide F_2 (Hz)	1250	3250	1750	3250

(11) The four data sets ended up with the following numbers of main-stressed /aɪ/ and /eɪ/ tokens in Lists 2, 3, and 6:

	Diphthong	
Scorer	/aɪ/ (2, 3)	/eɪ/ (6)
NONE	3467	2478
PLAUSIBLE	2945	1463
EM	3093	2253
RA	2993	2056

Replacing humans (EM, RA) with plausibility checks (PLAUSIBLE) leads to about the same rate of rejection for /aɪ/ tokens (10–15%), but drastically over-winnows /eɪ/ tokens (40%, as opposed to 9% or 17%). The high rejection rate for /eɪ/ is due to wildly implausible F_1 values at both the nucleus and the offglide, which seems to be mainly caused by the MFA misaligning the initial and final word boundaries.

2 Comparison of effect sizes and standard errors: /aɪ/ monosyllables (Lists 2 and 3)

(12) To compare the quality of the raw automatic output against that of the two human-supervised streams, we consider a signal that we are virtually certain must be present in the data: the shorter duration and more-/i/-like F_1 and F_2 of the /aɪ/ vocoid in the pre-voiceless environment compared to the elsewhere environment in the monomorphemic monosyllables of Lists 2 and 3.

(13) *Dependent measures*: The following dependent measures were used for the model fitting:

`phone_dur` Time interval between the left and right boundaries of the /aɪ/.

`F1atF1ext` F_1 at the nucleus (time of F_1 maximum).

`F2atF1ext` F_2 at the nucleus (time of F_1 maximum).

`F1atF2ext` F_1 at the offglide (time of F_2 maximum).

`F2atF2ext` F_2 at the offglide (time of F_2 maximum).

Time was in seconds, and formants were in Hertz.

(14) *Model fitting*: For each of the three data sets (NONE, EM, RA), and for each of the dependent measures, a linear mixed-effects model was fit using the `lmer` method in the `lme4` package in R (Bates et al., 2015). Fixed effects were an intercept and a single coefficient for `category` (voiceless vs. elsewhere, with elsewhere being the reference category). Random effects by participant were included for the intercept and `category`. The model specification was thus `measure ~ 1 + category + (1 + category | participant)`.

(15) *Effects of voicelessness*: The table below shows the values of the `measure` coefficent for the three data sets, and the ratio of the effect sizes between the raw and human-corrected data sets:

	Effect of voiceless				Ratio of effect		
measure	NONE	PLAUSIBLE	EM	RA	PLAUSIBLE_NONE	EM_NONE	RA_NONE
1 phone_dur	-0.129	-0.118	-0.139	-0.138	0.91	1.09	1.07
2 F1atF1ext	-59.2	-72.1	-81.2	-67.0	1.22	1.37	1.13
3 F2atF1ext	62.5	79.2	96.9	97.7	1.27	1.55	1.56
4 F1atF2ext	-55.0	-81.1	-73.6	-60.8	1.47	1.33	1.11
5 F2atF2ext	158.	185.	198.	188.	1.18	1.25	1.19

For every dependent measure, the ratios are greater than 1, i.e., the effect size is larger in absolute terms after correction than before.

(16) *Standard errors of the estimates*: This next table shows the s.e.m.'s for the three data sets, and the ratio between their magnitudes in the raw vs. human-corrected data sets:

	s.e.m of voiceless effect				Ratio of s.e.m.s		
measure	NONE	PLAUSIBLE	EM	RA	PLAUSIBLE_NONE	EM_NONE	RA_NONE
1 phone_dur	0.00476	0.00460	0.00370	0.00340	0.97	0.77	0.72
2 F1atF1ext	14.4	7.05	6.20	9.56	0.49	0.43	0.66
3 F2atF1ext	16.7	9.20	7.32	10.8	0.55	0.44	0.64
4 F1atF2ext	17.3	7.47	5.29	10.5	0.43	0.31	0.61
5 F2atF2ext	14.3	12.7	11.4	12.3	0.88	0.79	0.86

For every dependent measure, the ratios are all less than 1, i.e., the standard errors are smaller after correction than before. They're also smaller after simply discarding implausible tokens.

(17) *t values*: If we express the effect in terms of the standard error of the estimate to get the t value (on which statistical significance depends), here's what it looks like:

measure	NONE	PLAUSIBLE	EM	RA
1 phone_dur	-27.2	-25.7	-37.6	-40.6
2 F1atF1ext	-4.10	-10.2	-13.1	-7.01
3 F2atF1ext	3.75	8.6	13.2	9.08
4 F1atF2ext	-3.18	-10.9	-13.9	-5.81
5 F2atF2ext	11.0	14.6	17.4	15.2

The shortening and offglide-fronting effects (`phone_dur` and `F2atF2ext`) are very robust. The other spectral effects are less so.

(18) Human correction thus increases the statistical power of the experiment to detect effects of voiceless environment on formants: It multiplies effect sizes by roughly 1.125 to 1.5, and standard errors by roughly 0.5 to 0.75, and so increases the standardized effect size (measured in standard errors) by a factor of roughly 1.5 to 3.

(19) This increase in power from human correction is approximately what we'd get by quadrupling the number of participants (which would leave the absolute effect size unchanged, while halving the standard error).

(20) Simply throwing out implausible tokens worsens performance on phone duration, but improves it on everything else. It sometimes even beats the RA.

(21) The test case here was the one where we expect the biggest effects (monomorphemic monosyllables with /aɪ/), and it used the largest number of tokens (two lists with 25 tokens in all). Once we either move to polysyllables with more subtle effects, or subdivide the data, we will start eating into that margin.

3 Comparison of effect sizes and standard errors: /eɪ/ monosyllables (List 6)

(22) The same comparison was repeated for the /eɪ/ monomorphemic monosyllables in List 6. Qualitatively, we expect the same shortening and peripheralization effects as for /aɪ/, though quantitatively less extreme (Moreton, 2004; Hualde et al., 2017).

(23) Here are the values of the `measure` coefficent for the three data sets, and the ratio of the effect sizes between the raw and human-corrected data sets:

	Effect of voiceless				Ratio of effect		
measure	NONE	PLAUSIBLE	EM	RA	PLAUSIBLE_NONE	EM_NONE	RA_NONE
1 phone_dur	-0.102	-0.0883	-0.123	-0.121	0.87	1.21	1.19
2 F1atF1ext	128.	-11.9	-23.2	-28.1	-0.093	-0.181	-0.220
3 F2atF1ext	79.8	28.9	32.9	48.2	0.362	0.412	0.604
4 F1atF2ext	109.	-11.7	-10.7	-12.1	-0.107	-0.098	-0.111
5 F2atF2ext	45.9	28.8	24.6	28.1	0.627	0.536	0.612

The spectral effect sizes are greatly overestimated by NONE compared to the other three, which agree with each other pretty well. NONE also gets the wrong sign for the F_1 effect. Since PLAUSIBLE agrees with the humans, it must be the case that the errors in NONE are caused mainly by gross formant-tracking errors.

(24) *Standard errors of the estimates*: This next table shows the s.e.'s for the three data sets, and the ratio between their magnitudes in the raw vs. human-corrected data sets:

	s.e. of voiceless effect				Ratio of s.e.s		
measure	NONE	PLAUSIBLE	EM	RA	PLAUSIBLE_NONE	EM_NONE	RA_NONE
1 phone_dur	0.00492	0.00595	0.00342	0.00358	1.21	0.70	0.79
2 F1atF1ext	23.2	3.40	2.28	12.7	0.15	0.10	0.55
3 F2atF1ext	15.8	9.23	7.65	11.8	0.58	0.48	0.75
4 F1atF2ext	23.3	2.76	2.52	13.6	0.12	0.11	0.58
5 F2atF2ext	10.4	7.58	5.24	7.3	0.73	0.50	0.70

NONE is again hopelessly worse than PLAUSIBLE, EM, and RA. The RAs had surprisingly large standard errors for the formants, but I haven't followed up to figure out how come.

(25) *t values*: Here are the coefficients divided by the corresponding standard errors:

measure	NONE	PLAUSIBLE	EM	RA
1 phone_dur	-20.7	-14.8	-36.	-34.0
2 F1atF1ext	5.52	-3.50	-10.2	-2.2
3 F2atF1ext	5.05	3.13	4.30	4.08
4 F1atF2ext	4.68	-4.24	-4.25	-0.89
5 F2atF2ext	4.41	3.80	4.69	3.85

We can ignore the NONE column because the coefficients themselves are garbage, so we don't care what the t values are. The main thing we can see here is that PLAUSIBLE is nearly as good as EM and RA for three of the spectral measures, but not for duration or nuclear F_1.

(26) The upshot here is that for /eɪ/ tokens, the raw machine-measured data (NONE) is hopeless. When tokens with implausible values are discarded (PLAUSIBLE) or corrected (EM, RA), the situation improves dramatically, although sensitivity to voicing effects on nuclear F_1 is markedly less for PLAUSIBLE and RA

than for EM, and sensitivity to voicing effects on duration is markedly less for PLAUSIBLE than for EM and RA.

(27) The t values for the /eɪ/ coefficients (in 25) are much smaller than those for the /aɪ/ coefficients (in 17), by a factor of perhaps 3, meaning that voicing effects are harder to detect in /eɪ/ than in /aɪ/. This is due not to differences in the standard errors (in 16 and 24), which are similar for both diphthongs, but to the smaller effect size in /eɪ/ than in /aɪ/ (15, 23).

If we're trying to see interactions between the voicing effect and other factors like morpheme boundaries, the number of cases per cell will go down, the standard errors will go up, and the coefficients will probably get smaller (since monomorphemic monosyllables are likely to exhibit the strongest possible effects). That could lead to finding "significant" effects in /aɪ/ but not in /eɪ/, unless the sample sizes are chosen on the basis of power calculations for /eɪ/ rather than for /aɪ/.

4 Conclusions

(28) Returning to the questions in (2) above:

a. whether human supervision and correction actually improved the quality of the acoustic measurements
   (i) For /aɪ/: yes; the standardized effect size increases by a factor of 1.5–3, as if the number of participants had been approximately quadrupled.
   (ii) For /eɪ/: yes; the uncorrected data is useless.

b. whether the procedure followed by the humans was replicable, i.e., when different humans follow the same procedure, do they get the same results?
   (i) For /aɪ/: the absolute effect sizes are very similar for EM and RA. The standard errors are somewhat larger for RA, but not hugely so.
   (ii) For /eɪ/: the absolute effect sizes are again similar for EM and RA. The standard errors are much larger for RA; don't know how come.

c. how much time does it take the humans, and is the time proportional to the improvement?
   (i) I didn't keep good track of my own hours. The RAs took 306 hours in all (15.3 20-hour workweeks; in effect, an entire semester).
   (ii) If we simply take the automatically-generated output and discard tokens with implausible duration or formant values, the result is nearly as good as humans for /aɪ/, and not drastically worse than humans for /eɪ/.

(29) All of the foregoing is for the monosyllables, which are the simplest case. The project calls for studying morphonologically- and prosodically-complex words (like in Lists 4 and 5), which are likely to be harder to measure because everything happens faster. RA work is in progress (thanks, Brandon!).

References

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Chambers, J. K. (1973). Canadian Raising. *Canadian Journal of Linguistics 18*, 113–135.

Davis, S. and K. Berkson (Eds.) (2021). *American Raising*. Number 106 in Publications of the American Dialect Society. Durham, North Carolina: Duke University Press.

Hualde, J. I., T. Luchkina, and C. D. Eager (2017). Canadian Raising in Chicagoland: the production and perception of a marginal contrast. *Journal of Phonetics 65*, 15–44.

Hyman, L. M. (1976). Phonologization. In A. Juilland (Ed.), *Linguistic studies offered to Joseph Greenberg: second volume: phonology*, pp. 407–418. Saratoga, California: Anma Libri.

Moreton, E. (2004). Realization of the English postvocalic [voice] contrast in F1 and F2. *Journal of Phonetics 32*(1), 1–33.

Moreton, E. (2021). Phonological abstractness in English Diphthong Raising. In S. Davis and K. Berkson (Eds.), *American Raising*, Number 106 in Publications of the American Dialect Society, Chapter 2, pp. 13–44. Durham, North Carolina: Duke University Press.

Moreton, E., J. Lamontagne, and M. Nesbitt (2024). Durational and spectral factors in judgements of American Raising. Abstract in: *Journal of the Acoustical Society of America* 155 (3, Part 2):A288–A289.

Vance, T. J. (1987). "Canadian Raising" in some dialects of the northern United States. *American Speech 62*, 195–210.