Supplemental Materials:

Automatic measurement of vowel duration via structured prediction

Yossi Adi[a)], Joseph Keshet[b)]

Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel, 52900

Emily Cibelli[c)], Erin Gustafson[d)]

Department of Linguistics, Northwestern University, Evanston, IL, 60208

Cynthia Clopper[e)]

Department of Linguistics, Ohio State University, Columbus, OH, 43210

Matthew Goldrick[f)]

Department of Linguistics, Northwestern University, Evanston, IL, 60208

[a)]e-mail: adiyoss@cs.biu.ac.il

[b)]e-mail: joseph.keshet@biu.ac.il

[c)]e-mail: emily.cibelli@northwestern.edu

[d)]e-mail:egustafson@u.northwestern.edu

[e)]e-mail:clopper.1@osu.edu

[f)]e-mail: matt-goldrick@northwestern.edu

Clopper and Tamati (2014) examined how degree of vowel contrast was influenced by two factors: lexical contrast (the presence vs. absence of a word in English with the other member of the vowel); and dialect (Northern speakers have a smaller contrast than Midland speakers for /ɛ~æ/, whereas the pattern is reversed for /ɑ~ɔ/).

Vowel contrast was quantified by the distance in Bark (Traunmüller, 1990) between the first and second formant (F1/F2) of vowels from contrasting categories. These were estimated at vowel midpoint using Praat (Boersma and Weenink, 2001). Vowel distance was defined as the median of the Euclidean distances in the F1-F2 Bark space from a given vowel to each corresponding vowel in the same competitor condition. For example, if the target word was *dead*, the relevant set of distances would be to /æ/ target words with lexical competitors (e.g., *dad*). The effect of lexical competitor and dialect on these distance values was modeled via contrast-coded fixed effects in a mixed-effects regression model Clopper and Tamati (2014, 2015). Lexical neighborhood density and log word frequency were centered and included as control factors. Random intercepts for word and talker were included, along with uncorrelated random slopes for density and frequency by talker.

Our analysis followed that of the original paper. Using each algorithm's segmentation of the vowel, Praat was used to calculate F1 and F2 values. Tokens with F1 or F2 values more than 3 standard deviations from the mean of each vowel were considered to be outliers. As the authors in the original paper replaced outlier tokens with manual measurements, we

employed a similar approach; for each token identified as an outlier (5.7% of tokens), the distance measurements were replaced with the manual measurements corresponding to those tokens.

1 Significance of fixed effects

Following the HG analyses, models were refit after excluding observations with standardized residuals exceeding 2.5 (Baayen, 2008) and the significance of fixed-effects predictors was assessed via the likelihood ratio test (Barr et al., 2013).

The parameter estimates for model fit to the subset of the vowel distance data contrasting /ae~ɛ/ are shown in Table 1. The manual model had a significant effect of dialect ($\chi^2 = 5.05$, $p = 0.025$) but no effect of lexical competitor ($\chi^2 = 0.11$, $p = 0.074$). The interaction of dialect and competitor was not significant ($\chi^2 = 0.09$, $p = 0.7631$).

The algorithms largely recovered the same effects as found in the manual data. The DCM model recovered the significant main effect of dialect ($\chi^2 = 5.04$, $p = 0.025$); this parameter was marginal in the DCM (no classifier) model ($\chi^2 = 3.09$, $p = 0.079$) and the HMM model ($\chi^2 = 3.66$, $p = 0.056$). The interaction did not reach significance for any of the algorithmic models, but was marginal in the DCM model ($\chi^2 = 2{,}83$, $p = 0.093$). As in the manual model, all algorithmic models found no significant or marginal effects for either control variable ($p > 0.10$).

	Frequency	*t*	Density	*t*	Dialect	*t*	Lexical Competitor	*t*	Dialect * Competitor	*t*
Manual	0.026 (0.085)	0.302	-0.001 (0.007)	-0.115	**-0.345 (0.144)**	**-2.390**	0.039 (0.114)	0.338	0.114 (0.078)	1.462
DCM	0.0357 (0.119)	0.301	0.001 (0.011)	0.054	**-0.355 (0.148)**	**-2.396**	0.091 (0.175)	0.519	*0.138 (0.0818)*	*1.687*
DCM (no classifier)	0.014 (0.123)	0.113	0.001 (0.011)	0.112	*-0.277 (0.151)*	*-1.828*	0.111 (0.182)	0.60	0.131 (0.083)	1.586
HMM	*0.1345 (0.102)*	*1.322*	-0.003 (0.009)	-0.298	*-0.257 (0.128)*	*-2.003*	-0.121 (0.144)	-0.839	-0.007 (0.089)	-0.074

Table 1: Estimates and t-values for fixed effects in regression model, Clopper & Tamati dataset (standard error of estimate in parentheses), /ae~ɛ/ data. Significant effects ($p <= 0.05$, as assessed by likelihood ratio tests) are bolded; marginal effects ($0.05 > p < 0.10$) are italicized.

The parameter estimates for model fit to the subset of the data contrasting /ɑ~ɔ/ are shown in Table 2. The manual model had a significant effect of lexical competitor ($\chi^2(1) = 12.40$, $p < 0.0001$), but the effect of dialect was not significant ($\chi^2(1) = 2.48$, $p = 0.115$). The interaction of dialect and competitor just missed reaching significance ($\chi^2(1) = 3.65$, $p = 0.056$).

Models from all algorithmic methods produced a similar, but not identical, pattern of fixed effects as those found in the manual model. The effect of lexical competitor was significant for all methods (all $\chi^2(1) > 7$, $p < 0.05$). The effect of dialect did not reach significance in the HMM model ($\chi^2(1) = 1.14$, $p = 0.285$) but was significant in the DCM model ($\chi^2(1) = 5.07$, $p = 0.024$) and marginal in the DCM (no classifier) model ($\chi^2(1) = 3.54$, $p = 0.060$). While the interaction of dialect and competitor just missed reaching significance in the manual model it was significant in the models of all three algorithmic methods (all $\chi^2(1) > 3.9$, $p < 0.04$).

2 Comparison of model predictions

We performed leave-one-out validation on the /ae~ɛ/ and /ɑ~ɔ/ models. Comparisons of the predictions of each algorithmic model fit to the manual data predictions are shown in Figure 1.

For the /ae~ɛ/ data predictions, the mean squared error relative to the manual data

	Frequency	t	Density	t	Dialect	t	Lexical Competitor	t	Dialect * Competitor	t
Manual	-0.056 (0.071)	-0.791	-0.009 (0.009)	-0.943	-0.218 (0.134)	-1.627	**0.555 (0.132)**	**4.194**	*0.160 (0.083)*	*1.925*
DCM	-0.059 (0.070)	-0.841	-0.007 (0.010)	-0.714	**-0.256 (0.107)**	**-2.404**	**0.385 (0.129)**	**2.977**	**0.190 (0.089)**	**2.147**
DCM (no classifier)	-0.065 (0.067)	-0.963	-0.009 (0.009)	-0.941	*-0.220 (0.112))*	*-1.969*	**0.481 (0.125)**	**3.842**	**0.180 (0.090)**	**2.003**
HMM	-0.025 (0.064)	-0.387	-0.007 (0.008)	-0.807	-0.156 (0.144)	-1.083	**0.377 (0.117)**	**3.223**	**0.215 (0.096)**	**2.249**

Table 2: Estimates and t-values for fixed effects in regression model, Clopper & Tamati dataset (standard error of estimate in parentheses), /ɑ~ɔ/ data. Significant effects ($p <= 0.05$, as assessed by likelihood ratio tests) are bolded; marginal effects ($0.05 > p < 0.10$) are italicized.

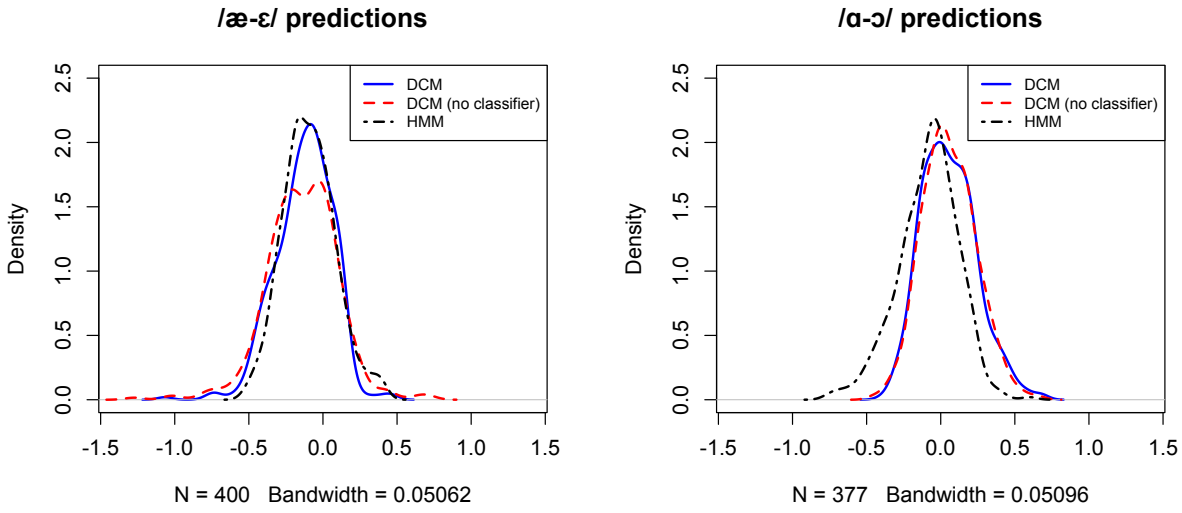

Figure 1: Comparisons of leave-one-out model predictions of vowel distances for the Clopper & Tamati (2014) dataset. In each panel, the deviance of each algorithmic method as compared to the manual predictions (manual-algorithmic) are shown. Predictions for the /ae~ɛ/ distances are presented on the left, and /ɑ~ɔ/ distances on the right.

predictions was lowest for the HMM model (0.045 Bark^2), followed by the DCM model (0.047 Bark^2) and the DCM model without classifier (0.069 Bark^2). Bootstrap confidence intervals of the differences between each system's predictions found that the HMM system had a smaller mean squared error than the DCM system without classifier (95% CI, [-0.033, -0.015] Bark^2); the DCM predictions also had a smaller mean squared error than the DCM without classifier predictions (95% CI, [-0.027, -0.017] Bark^2). There was no difference between the HMM and DCM predictions (95% CI, [-0.010, 0.005] Bark^2).

Turning to the /ɑ~ɔ/ data predictions, the mean squared error relative to the predictions of the manually annotated data was lowest for the bootstrap confidence intervals of the differences between each algorithm's predicted values found no differences between the predictions of any of the algorithmic methods; all confidence intervals contained 0 (DCM vs. HMM 95% CI: [-0.019, 0.004] Bark ^{2}; DCM no classifier vs. HMM 95% CI: [-0.018, 0.008] Bark ^{2}; DCM vs. DCM no classifier 95% CI: [-0.002, 0.006] Bark ^{2}).

Acknowledgements

Research supported by NIH grant 1R21HD077140 and NSF grant BCS1056409.

REFERENCES

R. H. Baayen. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press, 2008.

D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68 (3):255–278, 2013.

P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *[Computer program]*, 2001. URL `http://www.praat.org/`.

C. G. Clopper and T. N. Tamati. Effects of local lexical competition and regional dialect on vowel production. *Journal of the Acoustical Society of America*, 136(1):1–4, 2014.

C. G. Clopper and T. N. Tamati. Erratum: Effects of local lexical competition and regional dialect on vowel production. *The Journal of the Acoustical Society of America*, 138(2):570, 2015.

H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1):97–100, 1990.