Why the SQUARE vowel is the most variable in Sydney

Nhung Nguyen, Jason A. Shaw

The MARCS Institute, School of Humanities and Communication Arts, University of Western Sydney

nhung.nguyen@uws.edu.au, j.shaw@uws.edu.au

Abstract

Vowel variability is often explained in terms of linguistic and social factors. We have observed another factor that predicts vowel variability. Within four different corpora of Australian English vowels, we find a consistent relationship between the mean and standard deviation of formant values. For both F1 and F2, increases in mean formant values go hand in hand with increased variability. Given this observation, we propose that inferences about vowel variability take the mean formant values into account. Doing so changes conclusions about which vowels are most variable, undergoing change, or likely to reflect meaningful social variation.

Index Terms: vowel variability, acoustic phonetics, Australian English, forced alignment

1. Introduction

A typical method of visualizing vowel variability is to plot tokens in F1-F2 vowel space. The spread of tokens within this space, or sometimes ellipses summarizing the variance, indicate vowel variability [1, 2]. We exemplify this approach with a dataset from Cox's seminal paper on Australian vowels [1]. The vowel formant data from monophthongs (with the centring diphthongs /ɪə/ and /eː/ also classified as monophthongs [1]) produced by female speakers from that study (n= 60) are reproduced in Figure 1. Ellipses show two standard deviations from formant means. The shape of the ellipses indicates the spread of variation. Ellipses for FLEECE, NEAR, KIT, and GOOSE vowels [3], vowels with high F2 and low F1, are short and fat. Ellipses for TRAP, STRUT, START, and LOT vowels, vowels with high F1 and low F2, are tall and skinny. On the basis of these patterns, we might conclude that the TRAP, STRUT, START, and LOT vowels are most variable with regards to F1, and FLEECE, NEAR, KIT, and GOOSE vowels are most variable in relation to F2. This is a satisfactory conclusion. However, we observe that, on this conclusion, vowel variability is closely linked to the magnitude of mean formant values. The vowels with the highest mean values of F1 are the most variable in F1. The vowels that have the highest formant values of F2 are the most variable in F2. To highlight this point, Figure 2 plots the mean formant values from the same set of monophthongs shown in Figure 1 against the standard deviation. The result is clear. As mean formant values increase, so too does the standard deviation.

The strong correlation between the mean and standard deviation of formant values raises a number of questions about the proper characterization of vowel variation. This correlation may be crucial for interpreting vowel variability for sociolinguistic purposes such as differences across accents [2, 4], genders [1, 5], ages [6], and other social factors. In addition, knowledge of how much variability should be expected for a vowel given its mean formant values could inform the extent to which variability is tolerated in perception of accented speech [7] and in selective adaptation [8].

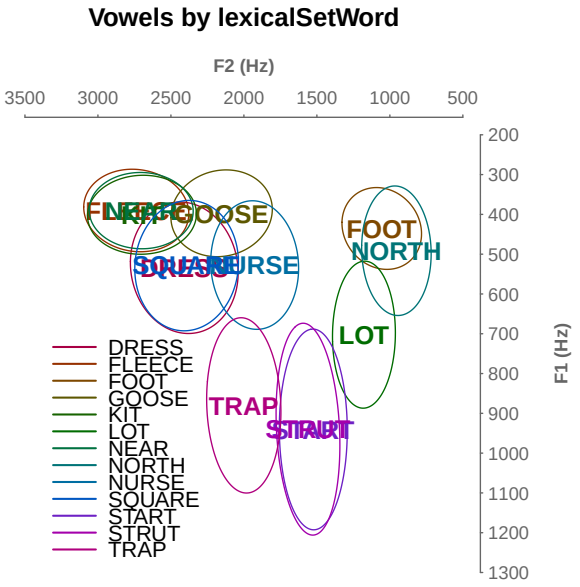

Figure 1: *Plots of Australian English monophthongs regenerated with random values sampled from means and standard deviations reported for female speakers in Cox [1]. Ellipses represent two standard deviations from the mean.*

In the remainder of this paper, we seek to assess the robustness of the correlation between the mean and standard deviation of Australian English vowel formants. We report the mean-standard deviation (mean-sd henceforth) relationship for vowel formants drawn from published sources [5], publically available corpora [9], and newly collected data. We find significant correlations in all of the datasets. Given this result, we suggest methods of assessing vowel variability that take into account the magnitude of mean vowel formants. In light of this suggestion, we reassess which Australian vowels are most variable.

2. Method

2.1. Descriptions of datasets

In addition to Cox [1], we looked for a mean-sd correlation in three additional corpora. We wanted first to replicate the result in a similar dataset. For this purpose, we searched through the AusTalk corpus [9] to find speakers from the Sydney region. We also wanted to know if the relationship emerges only across inter-speaker variation or if it can be found as well for

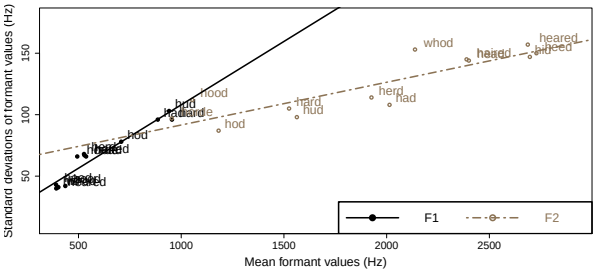

Figure 2: *Correlation between mean and standard deviation of 13 Australian English monophthongs in 13 /hVd/ words based on 60 female speakers from Sydney as reported in Cox [1].*

within-speaker variation. For this reason, we recorded several repetitions of each monophthong from another Sydney speaker. Lastly, we wanted to investigate whether the relationship is peculiar to the Sydney accent, so we explored a set of published results from Melbourne speakers [5]. We describe each of these corpora in greater detail below. For simplicity in presentation, we focus on just female speakers and on monophthongs produced in the /hVd/ environment. However, the main conclusions that we present generalize to male participants and diphthongs as well.

First, we sought to replicate the Cox result using speakers drawn from the publically available AusTalk corpus [9]. The data used for analyses were 13 monophthongs in the /hVd/ environment. We choose these words because they are in the same environment as the vowels reported in Cox [1]. We selected speakers from AusTalk based on the following criteria: female, aged 18 to 30, have Australian-born parents, live in the Sydney area, have low to no proficiency in other languages. We found five speakers that met this criteria (IDs: NSW10 Red-Squatter Pigeon, NSW28 Green-Cook's Petrel, NSW30 Gold-Terek Sandpiper, NSW19 Blue-Grey Grasswren, NSW25 Gold-Cape York Rock-wallaby). An exception to our criteria was made for ID NSW13 Blue-Laysan Albatross, who has a New Zealand-born father but typical Sydney vowels. A total of 78 observations from these six speakers were analyzed.

To evaluate whether the mean-sd correlation emerges from within-speaker variation, we recorded another female speaker from Sydney. The recording included the same 13 Australian English monophthongs embedded in the /hVd/ context. Data was collected in a sound-attenuated booth using a Shure SM10A-CN headset microphone. The speaker produced each /hVd/ words 10 times in random order, giving us 130 /hVd/ observations.

Lastly, the Melbourne dataset was taken from Billington's females [5]. Data collection approximated Cox's [1] with 13 female participants, each producing the 13 monophthongs twice in /hVd/ context. The means and standard deviations for F1 and F2 of each vowel are provided in [5]. The data are drawn from 338 total observations.

Neither of the three corpora analyzed here are as impressive as Cox's [1] in size, but they allow us to see if the trend observed in that data persists across a smaller sample of Sydney speakers, within a speaker, and in another region of Australia.

2.2. Vowel segmentation and formant extraction

Vowels in Cox's and Billington's datasets were hand-segmented and hand-measured. This work is labour-intensive, particularly given the size of those respective corpora. In addition to hand-segmenting the portion of the AusTalk data and single speaker data we report here, we also explored automatic segmentation using the Forced Alignment and Vowel Extraction (FAVE) program suite, web-based programs freely available through the University of Pennsylvania [10].

After segmentation, formants were extracted at 20%, 25%, 35%, 50%, 65%, and 80% of total vowel duration using the Mahalanobis method [11], which optimizes formant-tracking settings on a speaker-by-speaker basis. Initial settings were for five formants to be tracked with a maximum formant value of 5500Hz, a window size of 0.025s, and 12-point smoothing. Throughout this paper, we report the values of formants obtained at 25%.

To assess segmentation based on FAVE forced alignment, we correlated measurements of F1, F2, and duration based on hand- and machine-segmented values. For the AusTalk data, correlation coefficients were very high for F1 ($r = .97, p < .001$) and F2 ($r = 1.00, p < .001$). The correlation for duration was not as good ($r = .89, p < .001$). This indicates that, although there was some variation in where the boundary was placed by human versus machine segmentation, the variation in boundary placement had a negligable effect on formant extraction. Results for the single-speaker data were not quite as good: for F1 ($r = .86, p < .001$), for F2 ($r = .99, p < .001$), and for duration ($r = .76, p < .001$). Where the forced aligner deviated most from hand segmentation was for back rounded vowels (i.e., 'hod', 'hood', and 'horde'). For these vowels, the forced alignment occasionally placed the onset of the vowel boundary too early, including in the vowel some of the aperiodic energy in the /h/ of our /hVd/ context.

We think that the forced alignment results are reasonably promising, but we also note that it is subject to errors which seem to be systematic and could potentially influence the variability of vowel measurements. For this reason, we focus our discussion on the hand-segmented data and return to the human versus machine comparison in the discussion.

3. Results

For each of the four datasets under consideration, we plotted the mean formant values (F1 and F2) against the standard deviation and fit regression lines to F1 and F2. Figure 2 shows Cox's Sydney data (females only). Figure 3 shows the female Sydney speakers from AusTalk. Figure 4 shows our single Sydney speaker dataset. Figure 5 shows Billington's Melbourne data. For each of the datasets, the trend is the same. As the central tendency increases so too does the variance. Regression lines fit to the various corpora all indicate a significant positive relationship between the mean and standard deviation for both F1 and F2. The R^2 values are summarized in Table 1.

4. Discussion

4.1. The robustness of mean-sd correlation

We found a positive correlation between the mean and standard deviation of vowel formants for each of four different sets of vowel measurements. The linear relationship is stronger in Cox's data for both F1 and F2 than any of the other datasets. We take this to be an indication of the size and quality of Cox's

Table 1: *Mean-sd relationship across four datasets.*

dataset	formant	R^2	p
Cox's (Sydney)	F1	$R^2 = .93$	$p < .001$
	F2	$R^2 = .77$	$p < .001$
AusTalk (Sydney)	F1	$R^2 = .35$	$p < .05$
	F2	$R^2 = .70$	$p < .001$
Single-speaker (Sydney)	F1	$R^2 = .80$	$p < .001$
	F2	$R^2 = .40$	$p < .05$
Billington's (Melbourne)	F1	$R^2 = .78$	$p < .001$
	F2	$R^2 = .70$	$p < .001$

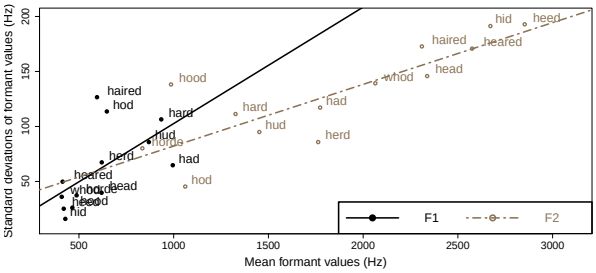

Figure 3: *Correlation between means and standard deviations of 13 Australian English monophthongs in 13 /hVd/ words, 6 female Sydney speakers, AusTalk dataset (hand-segmented).*

data. It is the largest dataset, composed of 60 female speakers. Moreover, the entire corpus was hand-segmented. The relationship is next strongest in Billington's data. Also hand-segmented and comparably large, the measurements in Billington are probably the next best estimation of population variance.

Formant values for the other datasets were extracted at a fixed percentage of vowel duration. This method may be less reliable for capturing the vowel target. Nevertheless, we find it intriguing that the mean-sd relationship persists despite this and other differences across corpora. Notably, we found the correlation regardless of whether variability was calculated within speakers or across speakers; whether in Sydney or in Melbourne; whether in adolescents (as in Cox's study and in Billington's study) or in young adults (as in the AusTalk speakers and our single speaker). Although we reported results here on just female speakers and on just monophthongs, the mean-sd

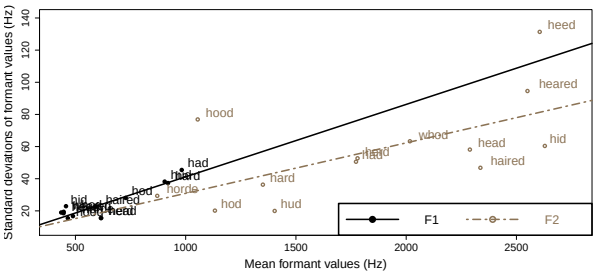

Figure 4: *Correlation between means and standard deviations of 13 Australian English monophthongs in 13 /hVd/ words, 1 female Sydney speaker, single-speaker dataset (hand-segmented).*

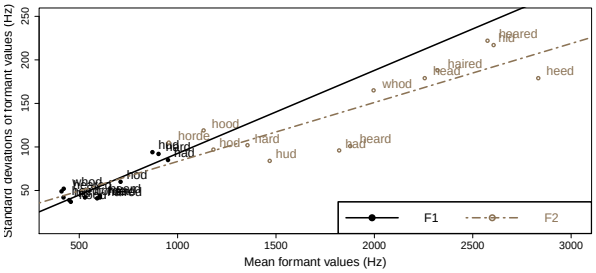

Figure 5: *Correlation between means and standard deviations of 13 Australian English monophthongs in 13 /hVd/ words, 13 female Melbourne speakers, Billington's dataset.*

relationship can also be found in male speakers in both Cox's and Billington's datasets, and it persists as well when diphthongs are included.

One thing that the four datasets reported on here have in common is that they all include vowels in the /hVd/ context. We would like to explore the mean-sd relationship in more diverse speech styles and contexts. Forced alignment and other modern corpus analysis tools could expedite such an analysis substantially. However, if our interpretation of the differences in goodness of fit between Cox's corpus and the AusTalk corpus is on the right track, then segmentation method might be very important to revealing mean-sd correlations. To underscore this point, we return to the discussion of forced alignment. We demonstrated reasonably strong correlations between measurements of vowels based on hand and machine segmentation. However, the correlation between the mean and standard deviation of vowel formants was weaker for datasets segmented using forced alignment. For the single-speaker dataset, mean-sd relationship weakened from $R^2_{F1} = .80$ (hand-segmented) to $R^2_{F1} = .59$ (machine-segmented) and $R^2_{F2} = .40$ (hand-segmented) to $R^2_{F2} = .34$ (machine-segmented). The mean-sd correlation for the AusTalk data also weakened from $R^2_{F1} = .35$ (hand-segmented) to $R^2_{F1} = .15$ (machine-segmented) and $R^2_{F2} = .70$ (hand-segmented) to $R^2_{F2} = .69$ (machine-segmented). It therefore appears that the strength of the mean-sd relationship reflects the quality of the measurements.

To sum up, the mean-sd correlation is robust across corpora. It surfaces in all of the datasets irrespective of vowel segmentation methods, formant measurement methods, regional varities, speaker age, and gender; however, it is strongest for large datasets measured carefully. This suggests the possibility of using the relationship to evaluate the quality of (automatic) vowel segmentation and formant measurement.

4.2. So which Australian vowel is most variable?

Early descriptions of Australian English vowels claimed that the FLEECE, GOOSE, FACE, PRICE, MOUTH, and GOAT vowels were the most variable, as these vowels marked the three Australian accent varieties (i.e., Broad, General, and Cultivated) in the 60s [12]. These vowels are amongst those that we noted in the introduction have a large standard deviation (i.e., large ellipses in Figure 1). FLEECE and GOOSE show more variation in F2 than do TRAP, STRUT, START, or LOT. However, if we take the central tendency of the formants into account, we would note that these vowels have just the amount of variation that is expected, based upon their mean values. Given the mean-

sd correlation, how then should we assess vowel variability? We conclude by suggesting a method that takes the mean formant value into account and apply the method to the four corpora of female Australian English vowels.

We suggest using the regression lines fit to the mean and standard deviation of formant values (e.g., Figures 2, 3, 4, 5) to establish a baseline for vowel variability. Datapoints above the regression lines (positive residuals) are variable relative to the magnitude of the mean. Values below the regression line (negative residuals) are stable relative to the mean. Table 2 shows lexical sets with the highest positive residuals.

Table 2: *F1 and F2 residuals for each dataset.*

dataset	F1 residual	F2 residual
Cox's (Sydney)	NORTH (10Hz) NURSE (9Hz) SQUARE (8Hz)	GOOSE (22Hz) FOOT (19Hz) NORTH (7Hz)
AusTalk (Sydney)	SQUARE (67Hz) LOT (48Hz) START (11Hz)	FOOT (57Hz) SQUARE (17Hz) KIT (15Hz)
Single-speaker (Sydney)	KIT (6Hz) TRAP (5Hz) NEAR (3Hz)	FLEECE (50Hz) FOOT (44Hz) NEAR (15Hz)
Billington's (Melbourne)	GOOSE (15Hz) STRUT (14Hz) NEAR (13Hz)	NEAR (32Hz) FOOT (27Hz) KIT (25Hz)

As can be seen from table 2, SQUARE has high F1 and F2 residuals in the AusTalk dataset. It also has a positive F1 residual in Cox's data. This is not a surprising result given the unclear extent of its monophthongization [2]. In addition, the Australian vowel system is undergoing change [13]. Due to the descent of the TRAP vowel, vowels surrounding TRAP, such as DRESS [13, 14], have also lowered. Since DRESS and SQUARE have been shown to be acoustically similar (except for duration) at least in the /hVd/ context [1], it is reasonable to speculate that SQUARE is also undergoing change. The residuals indicate that SQUARE is more variable than predicted by its mean in the inter-speaker Sydney datasets but not in the single-speaker dataset or in the Melbourne data. This indicates that there are comparably large differences in how the SQUARE vowel is produced across speakers in the Sydney region.

5. Conclusions

We observed a significant positive correlation between the variance of Australian vowel formant measurements and the central tendency of these measurements. The correlation emerged across corpora despite differences in measurement technique and speaker properties including age, gender, and region. The correlation also has consequences for interpreting vowel variability. When formant variability is considered relative to the mean, the SQUARE vowel emerges as the most variable in the Sydney region. This generalization is otherwise masked by the strong influence that mean formant values exert on formant variance.

6. Acknowledgements

The AusTalk corpus was collected as part of the Big ASC project (Burnham et al. 2009; Wagner et al. 2010; Burnham et al. 2011), funded by the Australian Research Council (LE100100211). See: https://austalk.edu.au/ for details. We would like to thank Jaydene Elvin for downloading the AusTalk data, Jaydene Elvin, Karen Mulak, Michael Tyler for help with recording material, Mona Faris for a PRAAT observation, and Jia Ong and Sarah Fenwick for stats discussion.

7. References

[1] Cox, F., "The Acoustic Characteristics of /hVd/ Vowels in the Speech of some Australian Teenagers", Australian Journal of Linguistics, 26(2):147–179, 2006.

[2] Harrington, J., Cox, F., and Evans, Z., "An acoustic phonetic study of broad, general, and cultivated Australian English vowels", Australian Journal of Linguistics, 17(2):155–184, 1997.

[3] Wells, J.C., Accents of English, Cambridge, 1982.

[4] Clermont, F., "Multi-speaker formant data on the Australian English vowels: A tribute to J.R.L. Bernard's (1967) pioneering research", Proc. 6th Australian Int. Conf. Speech Science and Technology, Adelaide, Australia, 145–150, 1996.

[5] Billington, R., "Location, Location, Location! Regional Characteristics and National Patterns of Change in the Vowels of Melbourne Adolescents", Australian Journal of Linguistics, 31(3):275–303, 2011.

[6] Millar, J., Vonwiller, J., Harrington, J., and Dermody, P., "The Australian National Database of Spoken Language", Proceedings of the International Conference on Acoustics Speech and Signal Processing, Adelaide, 2, 1994.

[7] Clarke, C.M., and Garrett, M.F., "Rapid adaptation to foreign-accented English", Journal of Acoustical Society of America, 116(6):3647–3658, 2004.

[8] Kleinschmidt, D., and Jaeger, T.F., "A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation", Proceedings of the 34th Annual Conference of the Cognitive Science Society, Sapporo, Japan, 2012.

[9] Estival, D., Cassidy, S., Cox, F. and Burnham, D., "AusTalk: an audio-visual corpus of Australian English". Online: https://www.academia.edu/6546894/AusTalk_an_audio-visual_corpus_of_Australian_English, accessed on 25 May 2014.

[10] Rosenfelder, I., Fruehwald, J., Evanini, K. and Jiahong, Y., "FAVE (Forced Alignment and Vowel Extraction) Program Suite". http://fave.ling.upenn.edu/index.html, 2011.

[11] Evanini, K., "Automatic vowel analysis", in K., Evanini, The permeability of dialect boundaries: a case study of the region surrounding Erie Pennsylvania, 50-94, PhD dissertation, 2009.

[12] Mitchell, A.G., and Delbridge, A., "The speech of Australian adolescents: A survey", Sydney: Angus and Robertson, 1965.

[13] Cox, F., "Australian English pronunciation into the 21st century", Prospect, 21(1):3–21, 2006.

[14] Cox, F. and Palethorpe, S., "A question of broadness", Paper presented at Australian Linguistics Society Conference, Newcastle, Australia, 2004.