AGE VECTORS VS. AXES OF INTRASPEAKER VARIATION IN VOWEL FORMANTS MEASURED AUTOMATICALLY FROM SEVERAL ENGLISH SPEECH CORPORA

Jeff Mielke*, Erik R. Thomas*, Josef Fruehwald†, Michael McAuliffe‡,
Morgan Sonderegger‡, Jane Stuart-Smith§, and Robin Dodsworth*

*North Carolina State University, †University of Edinburgh,
‡McGill University, §University of Glasgow
jimielke@ncsu.edu

ABSTRACT

To test the hypothesis that intraspeaker variation in vowel formants is related to the direction of diachronic change, we compare the direction of change in apparent time with the axis of intraspeaker variation in F1 and F2 for vowel phonemes in several corpora of North American and Scottish English. These vowels were measured automatically with a scheme (tested on hand-measured vowels) that considers the frequency, bandwidth, and amplitude of the first three formants in reference to a prototype. In the corpus data, we find that the axis of intraspeaker variation is typically aligned vertically, presumably corresponding to the degree of jaw opening for individual tokens, but for the North American GOOSE vowel, the axis of intraspeaker variation is aligned with the (horizontal) axis of diachronic change for this vowel across North America. This may help to explain why fronting and unrounding of high back vowels are common shifts across languages.

Keywords: vowels, formants, variation, change, automatic

1. INTRODUCTION

We examine vowel formant variation in several natural speech corpora of North American and United Kingdom English. We compare the direction of change in apparent time with the axis of intraspeaker variation, for several vowel phonemes, in order to examine the idea that a speaker's tokens of a particular vowel will be aligned along an axis coinciding with the direction that vowel is shifting diachronically in a given community. A frontward progression of the GOOSE vowel from [u] to [ʉ] and sometimes to [y] (and often with a small degree of diphthongization) is well documented in North American and Scottish English [7, 9, 18, 20]. and various sources [8, 9] have shown that GOOSE typically shows an elongated distribution of tokens that coincides with its direction of diachronic movement. This fact suggests that it would be possible to test whether vowels typically shift along their axes of distribution, which in turn could provide a potential motivation for some kinds of vowel shifting. We will see that intraspeaker variation does not align with diachronic change for most vowels. However, high back and central vowels vary mostly in the F2 dimension, and we suggest a connection between this and the crosslinguistic frequency of /u/ fronting. Comparing variation and change across dialects in this way requires the ability to measure vowels in the same way across corpora representing diverse language varieties. As such, this also an opportunity to test and/or demonstrate the use of ISCAN [12] for large-scale vowel analysis.

ISCAN is an open-source software system for Integrated Speech Corpus ANalysis, which enables automated acoustic phonetic analysis across spoken corpora of diverse formats and sizes. This system is meant to overcome the significant practical and methodological barriers to conducting essentially the same study across corpora, including necessary technical skills and non-comparability of results using non-standardized measures. Our first step is to address a source of error in automatic formant measurement, namely pervasive false F2 measurements which occur particularly in LPC-based formant tracking of front vowels and which would obscure interesting patterns of intraspeaker variation if they are not corrected.

2. IMPROVING AUTOMATED VOWEL FORMANT MEASUREMENTS

Our starting point for automated vowel formant measurements is based on FAVE-extract [4, 14]. In that formant measurement scheme, each vowel is measured several times with different numbers of LPC coefficients, and the candidate measurements are compared against a prototype consisting of mean formant frequencies and bandwidths and a covariance matrix. The candidate that is selected as the most probable is the one with the smallest Mahalanobis distance from the prototype.[1]

To evaluate the accuracy of the automatic measurements, a set of manually-computed measurements of vowels for a set of seven subjects from northeastern Ohio was utilized as a reference point. Those measurements were taken from subjects' readings of two stories in a study of regional variation in Ohio [19].[2] To identify measurement points for automated measurements, the test recordings were force-aligned using the Penn Phonetics Lab Forced aligner (P2FA) [21]. Our initial (**basic**) implementation of the automated formant measurement scheme uses six measures: the frequencies (in Hz) and bandwidths (in log_{10} Hz) of the first three formants, measured 0.33 of the way into the vowel interval. Initial prototypes were generated for each vowel phoneme in each corpus on the basis of previous measurements of the same datasets that were corrected and pruned to eliminated obvious formant tracking errors, most of which involved underestimating F2 in front vowels by tracking a false formant between F1 and F2. For the northeastern Ohio test dataset, the prototypes were based on the Raleigh corpus [2].

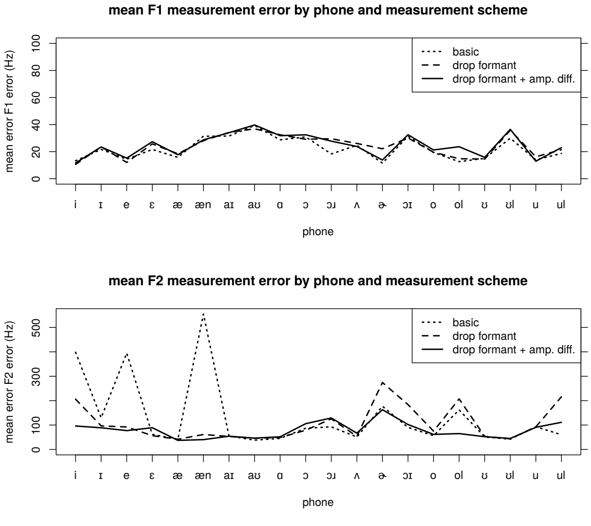
Every stressed vowel token was measured at 0.33 of the vowel's duration using a maximum formant frequency of 5500 Hz for females and 5000 Hz for males and 8-14 LPC coefficients, yielding seven sets of candidate measurements for each of 5486 stressed vowel tokens. The candidate for each token that was closest to the prototype was retained. For each speaker, new prototypes (means and covariance matrices) were then generated based on these measurements, and new candidates were selected on the basis of these prototypes. This process was repeated until the selected candidates did not change from one iteration to the next, with a limit of 20 iterations. For this test dataset, the mean number of iterations required was 6.0 and the maximum was 13.

As will be seen below, this procedure frequently underestimates F2 in high vowels. These underestimates can become self-reinforcing as the procedure iterates and the prototype's mean F2 gets lower and lower. In some cases, there is no number of LPC coefficients that tracks F1 and F2 consistently without also tracking a false formant in between. To address this problem, we implemented a modified procedure (**drop formant**) which considers candidates in which the tracked F1, F2, or F3 can be dropped and replaced by the next highest measured formant. For example, if there is a candidate in which the four lowest formant tracks are F1, a false formant, F2, and F3, an additional candidate could also be included that retains the first, third and fourth tracks as F1, F2, and F3 (dropping the false formant track). For the drop formant scheme, the range of LPC coefficients considered was extended to 8-16. As a filter to prevent real formants from being dropped, a linear regression was performed on the formant amplitudes (dB) and the log_2 formant frequencies, and candidates were created only if they dropped formants whose amplitudes were lower than the model estimate for the formant's frequency. This is based on the assumption that formants have peaks in amplitude that are roughly a linear function of frequency because the source spectrum is roughly linear in this way.

To further encourage the tracking of real formants, a third scheme (**drop formant** $+$ **amplitude difference**) was designed to take into account the fact that pairs of well-measured formants have characteristic amplitude differences. For example A1-A2 and A2-A3 are both typically positive values. If the measured F2 track is a false formant (which likely has lower amplitude than surrounding real formants), then A2-A3 is likely to be negative. A1-A2 and A2-A3 were added to the formant frequencies and bandwidths as the seventh and eighth phonetic measures included in the prototypes.

Formant measurements produced by the automatic procedures were then compared against the manually-computed measurements. Of the 5486 automatically measured vowels, 1019 had been measured earlier by hand.[3]

Figure 1: Mean formant measurement error by phone, for three automatic measurement schemes



mean F1 measurement error by phone and measurement scheme

mean F2 measurement error by phone and measurement scheme

Figure 1 shows the mean absolute F1 and F2 measurement error for each vowel, for each formant measurement scheme. Mean F1 error is below 40 Hz for every phone. The largest portion of wide discrepancies involved F2 of three front vowels (/i/, /e/, and prenasal (raised) /æ/) for the basic measurement scheme. Note that categorically different allophones in American English are split for the purposes of data analysis, but they were not distinguished by the measurement scheme. Small deviations in formant frequencies were regarded as unimportant, as they could have been due to small differences in the precise time point at which a measurement was taken.

Allowing the measurement algorithm to drop measured formant tracks (the **drop formant** scheme) largely eliminated these F2 errors but introduced F2 errors in /ɚ/, /ɔɪ/, and pre-lateral (unfronted) /u/. Including amplitude difference in prototypes (the **drop formant + amplitude difference** scheme) eliminated these F2 errors and also improved pre-lateral (unfronted) /u/. The vowels with the highest error after both of these refinements are diphthongs, whose measurements are sensitive to variation in measurement time point, and pre-lateral /u/, which shares a prototype with a much fronter allophone (non-pre-lateral /u/).

3. STUDY OF AGE VECTORS AND AXES OF VARIATION IN ENGLISH DIALECTS

The data are for the study of age vectors and axes of variation are stressed vowels drawn from six speech corpora: the Sounds of the City corpus [16], representing Glasgow, Scotland; the Scottish Corpus of Texts and Speech [1], representing Scotland more broadly; the Canadian subset of the International Corpus of English (ICE-CAN) [6], representing Canada; the Buckeye corpus [13], representing the U.S. North Midland; the Raleigh corpus [2], representing the U.S. urban South; and subsets of the Santa Barbara corpus [3] that represent Western U.S. and the Northern Cities Shifted (Inland North) region of the U.S.. The selected words belong to the following lexical classes (meaning they have the same vowel as these representative words): FLEECE ([i]), BLEW ([ʉ], [u], [jʉ], or [ju]), GOOSE ([ʉ] or [u]), KIT [ɪ], FOOT ([ʉ] or [ʊ]), WAIST ([e] or [eɪ]), STRUT ([ʌ]), GOAT ([o] or [oʊ]), DRESS ([ɛ]), TRAP ([a] or [æ]), LOT ([ɔ] or [ɑ]), and THOUGHT ([ɔ]). Vowels after /j/ or before a nasal or /ɹ/ were excluded, and GOOSE, FOOT, GOAT, LOT, and THOUGHT were split into pre-lateral and non-pre-lateral sets (e.g., GOOSE vs. POOL). The analysis is limited to words that are not known to have been involved in context-sensitive change in any of the dialects under study, determined using UNISYN [5], which was also used for assigning words to lexical classes.

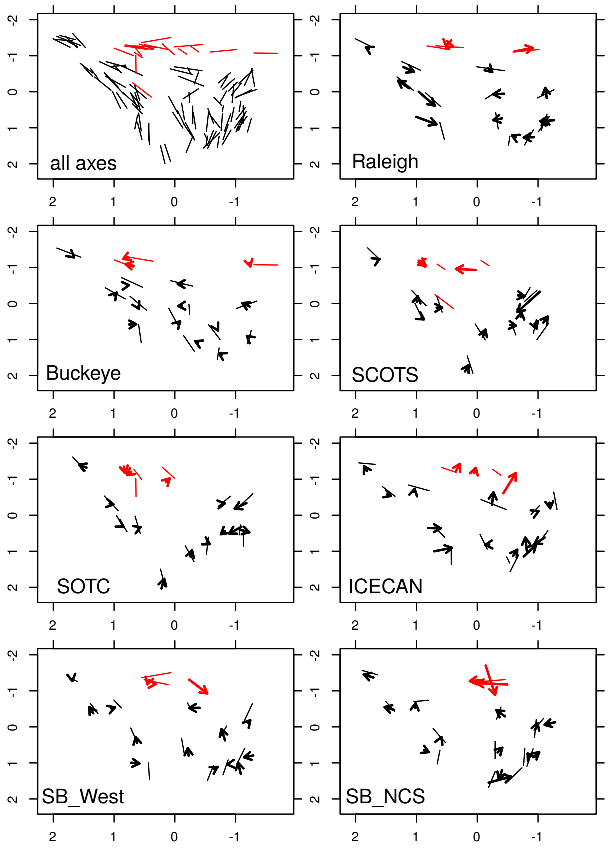
ISCAN was used to measure F1 and F2 at the nucleus (0.33 of the vowel duration) of 547,344 stressed vowels. Optimal formant measurements for each token were selected using the **drop formant + amplitude difference** scheme described in the previous section. These measurements were then normalized using the Lobanov method [10].

Age vectors were calculated using the mean normalized F1 and F2 measurements for the oldest and youngest generation within each corpus (young vs. old for Buckeye, birth year before 1950 vs. after 1967 for Raleigh and SCOTS, all older speakers and 1980s middle-aged speakers vs. 1980s-2000s young speakers for Sounds of the City, birth year before vs. after 1950 for ICE-CAN, and age at recording over vs. under 35 for Santa Barbara). In general, the dividing line between older and younger generations is between 1950 and 1960 for all of these corpora.[4] Age vectors are shown in Figure 2 as thick arrows. To measure the axes of intraspeaker variation, a principal component analysis was performed for F1 and F2 of each speaker-vowel combination with at least 20 tokens. The loadings were used to calculate the angle of the main axis of variation for each speaker-vowel combination. These were averaged across speakers within each regionally-defined group. Mean axes of variation are shown in Figure 2 as thin line segments. The length of each line segment reflects the magnitude of vowel variation along the axis.

In the majority of cases, there is no obvious connection between age vectors and axes of intraspeaker variation, as Figure 2 shows. For low vowels, the axis of intraspeaker variation is ordinarily aligned vertically, presumably corresponding to the degree of jaw opening for individual tokens. The mid and high front vowels exhibit lower angles of orientation, likely resulting from their positions along the front margin of the vowel envelope causing them to parallel the margin rather than vary vertically like the central and back vowels.

The GOOSE vowel in North American English differs in showing generally horizontal orientations. These GOOSE orientations show a striking disconformity with the largely vertical alignments of mid back vowels. They also differ from the nearly vertical slopes observed for GOOSE in the two Scottish corpora, consistent with GOOSE lowering found in Glasgow [15, 17].[5]

Figure 2: Age vectors (reflecting change in apparent time; arrows) and axes of intra-speaker variation (thin lines) for vowels across seven groups of speakers from six corpora. *x*-axis is normalized F2, *y*-axis is normalized F1. GOOSE (plain and pre-lateral) and BLEW are in red. Arrows with zero or near-zero length (meaning virtually no diachronic change) appear as arrowheads only.

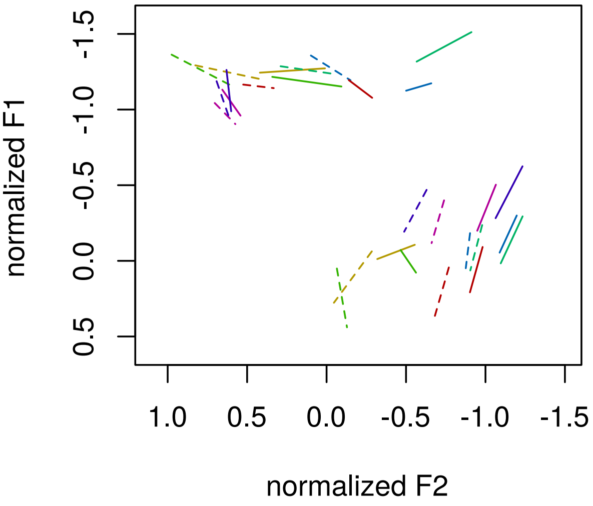


Since a preceding coronal is known to condition GOOSE fronting [8], it is conceivable that the horizontal intra-speaker variation in this vowel is primarily driven by the presence or absence of a preceding coronal consonant. Figure 3 shows the GOOSE and GOAT vowels split according to whether the vowel is preceded by a coronal consonant, revealing that GOOSE's horizontal alignment is present in both post-coronal and non-post-coronal subgroups. GOAT, which also undergoes fronting, is vertically aligned.

4. CONCLUSIONS

Comparison of automatic measurements of vowels with manually measured readings laid bare the deficiencies of raw automatic readings. The deviations, particularly for F2 of front vowels, proved pervasive and thus are likely to plague any LPC-based automated formant extraction routine. Two refinements of the process, a procedure to drop apparent false formants and the inclusion of formant relative amplitude in the prototypes, eliminated nearly all of the gross deviations. In the majority of cases, there is no obvious connection between age vectors and axes of intra-speaker variation. The horizontal orientations of GOOSE in North American varieties, particularly their disconformity with the alignments of other back vowels, are salient. This anomaly may be related to the fact that fronting and unrounding of high back vowels are common shifts across languages. The correlation of the GOOSE orientation with its diachronic development differs from that of lower vowels, which appear to show little or no correlation. The detection of this pattern also demonstrates the value of large corpora involving large numbers of tokens per speaker in that the alignments would have gone undiscovered without utilization of the full potential of these corpora. The observed differences in alignment between North American and Scottish GOOSE alignments probably reflect the fact that fronting of GOOSE is an old, completed process in Scottish English [20], placing GOOSE at the front of the articulatory space where it can behave like other front vowels, whereas the fronting is currently ongoing in North American English and thus GOOSE is not yet completely a front vowel.

Figure 3: Axes of intra-speaker variation for GOOSE (upper group) and GOAT (lower group) vowels. Dashed lines are post-coronal, solid lines are non-post-coronal. red=SB_West, yellow=Raleigh, green=Buckeye, cyan=SB_NCS, light blue-ICECAN, dark blue=SOTC, violet=SCOTS.



5. ACKNOWLEDGMENTS

This work is part of the SPeech Across Dialects of English (SPADE) project (https://spade.glasgow.ac.uk/), funded by the Digging into Data Challenge/Trans-Atlantic Platform via contributions from the NSF (SMA-1730479) in the USA, ESRC and AHRC in the UK, and SSHRC/CRSH and NSERC/CRSNG in Canada.

6. REFERENCES

[1] Anderson, J., Beavan, D., Kay, C. 2007. SCOTS: Scottish corpus of texts and speech. In: *Creating and digitizing language corpora*. Palgrave Macmillan UK 17–34.

[2] Dodsworth, R., Kohn, M. 2012. Urban rejection of the vernacular: The SVS undone. *Language Variation and Change* 24(2), 221–245.

[3] Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., Martey, N. 2000. Santa Barbara Corpus of Spoken American English. *Linguistic Data Consortium*. CD-ROM.

[4] Evanini, K. 2009. *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*. PhD thesis University of Pennsylvania.

[5] Fitt, S. 2000. Documentation and user guide to UNISYN lexicon and post-lexical rules. Technical Report, Centre for Speech Technology Research, University of Edinburgh.

[6] Greenbaum, S., Nelson, G. 1996. The international corpus of English (ICE) project. *World Englishes* 15(1), 3–15.

[7] Kurath, H., McDavid, R. I., Jr. 1961. *The Pronunciation of English in the Atlantic States*. Ann Arbor: University of Michigan Press.

[8] Labov, W. 1994. *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell.

[9] Labov, W., Ash, S., Boberg, C. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. De Gruyter Mouton.

[10] Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America* 49(2B), 606–608.

[11] Mahalanobis, P. C. 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* 2(1), 49–55.

[12] McAuliffe, M., Coles, A., Goodale, M., Mihuc, S., Wagner, M., Stuart-Smith, J., Sonderegger, M. 2019. ISCAN: a system for integrated phonetic analyses across speech corpora. *Proceedings of the 19th Congress of Phonetic Sciences* Melbourne.

[13] 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].

[14] Rosenfelder, I., Fruehwald, J., Evanini, K., Yuan, J. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite.

[15] Scobbie, J. M., Stuart-Smith, J., Lawson, E. 2013. Back to front: A socially-stratified ultrasound tongue imaging study of Scottish English /u/. *Italian Journal of Linguistics* 24(1), 103–148.

[16] Stuart-Smith, J. 2014. Fine phonetic variation and sound change: A real-time study of Glaswegian. Final Report: RPG-142 (Sounds of the City).

[17] Stuart-Smith, J., José, B., Rathcke, T., Macdonald, R., Lawson, E. 2017. Changing sounds in a changing city: An acoustic phonetic investigation of real-time change over a century of Glaswegian. In: Montgomery, C., Moore, E., (eds), *Language and a Sense of Place: Studies in Language and Region*. Cambridge: Cambridge University Press 38–65.

[18] Thomas, E. R. 2001. *An Acoustic Analysis of Vowel Variation in New World English*. Durham, N.C.: Duke University Press. Publication of the American Dialect Society 85.

[19] Thomas, E. R. 2010. A longitudinal analysis of the durability of the Northern/Midland dialect boundary in Ohio. *American Speech* 85, 375–430.

[20] Wells, J. C. 1982. *Accents of English*. Cambridge, UK: Cambridge University Press. 3 vols.

[21] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08* 5687–5690.

[1] Mahalanobis distance [11] measures the distance of a vector from the mean of a multidimensional distribution, taking into account correlation between dimensions.

[2] Vowels in a predetermined list of words from the stories had been measured in Praat by marking off the onset and offset of the vowel visually (in some cases with auditory input) and obtaining values of F1-F4 with LPC at three points within the vowel (35 ms after onset, midpoint, and 35 ms before offset). The manual procedure allowed the practitioner to vary the LPC settings as needed in order to yield the most accurate formant estimates. Judgments of whether appropriate formant readings were obtained were based on visual inspection of spectrograms with superimposed LPC formant tracks. The maximum formant value was always set at 5500 Hz, but the number of LPC coefficients was varied from 8 to 18 as needed.

[3] Since hand and automatic measurements could be made at different times, we excluded 29 tokens whose automatic measurements were over 50 ms before or after the hand measurements. For vowels with two or more hand measurements, we used linear interpolation to estimate the formant values at the automatic measurement time.

[4] Note that this method of assigning 'age' was a practical way of dealing with the apparent-time corpora which were recorded at different time points, and the real-time corpus (SOTC) which here has recordings made over three decades.

[5] GOOSE fronting is observed as an age vector in the SCOTS corpus but not SOTC because the SCOTS corpus contains more standard Scottish speakers, plus a large range of Scottish regional accents. SOTC is Glasgow vernacular, where GOOSE is already very front.