Order of Operations in Sociophonetic Analysis

Joseph A. Stanley*

1 Background

In modern sociophonetic data analysis, a researcher makes numerous decisions when processing their data. Many of the "big" decisions include how data is collected, how audio is transcribed, how these transcriptions are aligned to the audio, and how acoustic measurements are extracted. Once a spreadsheet of numbers has been generated, further decisions must be made regarding normalization, detecting and handling outliers, and excluding tokens irrelevant to the research question. This paper takes a closer look at this second group of decisions—those made after acoustic measurements have been extracted—and shows that the order in which they are applied can have significant effects on the final conclusions drawn from the data.

Recent sociophonetic research on North American vowels typically involves at least half a dozen or so procedures that are applied to a spreadsheet of formant measurements before a particular metric can be generated. Such steps include detecting (and removing) outliers, removing stopwords (i.e. function words and other highly frequent words), removing unstressed vowels, normalizing the data, isolating midpoints from vowel trajectories, removing diphthongs and other vowel classes irrelevant to the study, and excluding allophones such as vowels in preliquid environments. The researcher's task is to decide how and when each of these procedures should be carried out.

Within many of these steps, there are competing methods that aim to accomplish the same task. Some studies specifically pit these techniques against each other, often comparing them to some gold standard, and decide which technique is best based on some predetermined set of criteria. The following examples illustrate such conversation:[1]

- Vowel overlap and merger is often measured with Pillai scores (Nycz and Hall-Lew 2013), though Johnson (2015) argues that Bhattacharyya's Affinity is better on actual vowel data. This led Kelley and Tucker (2020) to compare them and other overlap measures on the same dataset to see which performed better.
- While Adank et al (2004) compare 12 different methods for vowel normalization and conclude that the Lobanov (1971) is most suitable for sociolinguistic data, Barreda (2021b) shows that Adank et al. miscalculated the Lobanov normalization and that there are theoretical and cognitive reasons for why sociolinguists should avoid it.
- VarBrul (Sankoff 1975) and its successor, Goldvarb (Sankoff, Tagliamonte and Smith 2005), were once the standard for statistical analysis in variationist sociolinguistics, but Johnson (2009; 2014) argues that mixed-effects models are more ideally suited. Nevertheless, Tagliamonte and D'Arcy (2017) argue that GoldVarb can still be useful.
- FAVE (Rosenfelder et al. 2014; see also Evanini 2009; Labov, Rosenfelder, and Fruehwald 2013) makes decisions regarding Praat settings and where along the duration of a vowel the measurements should be taken. However Kendall and Vaughn (2020) show that small changes in these parameters can have a large effect on the results. Furthermore, Fast Track (Barreda 2021a) uses data-internal heuristics to determine which of many more candidate tracks is best.

These meta-analyses of linguistic methodology compare multiple techniques that solve the same problem. However, such studies are conducted more or less independently of each other: any

* I am grateful for the audience at NWAV49 for their comments and feedback on this paper. I also graciously acknowledge the contributions by Thomas Kettig and Rich Ross for helping determine the order of operations recommended here.

[1] To be clear, such discussion over competing forms is good! Thanks to such papers, the field has mostly abandoned techniques that produced poor results. Methods sections in papers are clearer too since researchers have a better idea about which steps to run and which of the many options they chose to do.

data processing outside of the scope of the task being analyzed is considered irrelevant. Consequently, there is little regard to the *order* in which these steps should be applied. As will be shown in this paper, this is an unfortunate oversight because each step is connected to the other in an analysis pipeline. For example, a procedure that detects outliers will find one set of tokens if stopwords have already been excluded, but a different (albeit overlapping) set if stopwords are still being considered. So, when this order is modified, it can change the overall outcome of the analysis.

This paper highlights the importance of the order of operations in sociophonetic data analysis and illustrates that changes in the pipeline can have non-negligible effects on the overall results.

2 Methods

2.1 Data

Data for this study come from formant measurements extracted from 53 speakers from the American Mountain West: Idaho, Utah, Montana, Wyoming, and Colorado. These speakers read sentences in their homes using their own recording devices, their audio was transcribed manually, the transcriptions were force aligned to the audio with the Montreal Forced Aligner (McAuliffe et al. 2017), and formant measurements were extracted with Fast Track (Barreda 2021a), resulting in a spreadsheet of acoustic measurements from 278,387 vowel tokens, each sampled at 11 equidistant timepoints along the duration of the vowel. While much could be said about the analytical decisions made so far, for the purposes of this paper I will call the resulting spreadsheet good so that I can focus on the subsequent analysis steps. It is beyond the scope of this paper to discuss the pipeline up to this point in the analysis.

2.2 Linguistic Phenomena

As a case study, this study focuses on a recent a set of vowel changes now widespread across North American English. The retraction of /ɑ/, often resulting in a merger with /ɔ/ (*i.e.* the low back, or *cot-caught*, merger), is argued to have triggered the lowering and retraction of the front lax vowels /æ/, /ɛ/, and /ɪ/ (Becker 2019a). A related phenomenon is the "prenasal split" where prenasal allophones of /æ/ (as in *ban*, *sand*, or *ham*) raise while pre-obstruent allophones of /æ/ (as in *bat*, *sad*, or *happy*) lower, resulting in an increased perceptual distance between the two allophones. Of the many names in circulation,[2] I will refer to this set of changes as the Low-Back-Merger Shift (LBMS; Becker 2019b).

To adequately measure merger and vowel shift, several metrics have been used to compare relative vowel positions across speakers and studies. This paper focuses on three. Experiment 1 measures overlap between /ɑ/ and /ɔ/ and separation between /æN/ and /æ/ using Pillai scores (Nycz and Hall-Lew 2013). Experiment 2 measures the degree of shifting in /æ/ and /ɛ/ by comparing them to "benchmarks" derived from the *Atlas of North American English* (Labov, Ash and Boberg 2006). Experiment 3 assesses the degree of shifting of all three vowels using the Low-Back-Merger Shift Index (LBMS Index), as defined by Becker (2019b).

2.3 Steps in the Pipeline

Going from a spreadsheet of formant measurements to interpretable output involves various steps, which may vary by project, variety, linguistic phenomena under investigation, and researcher preferences. For the purposes of this paper, seven steps were chosen to be part of the experiments.

- Outlier removal: For each vowel/allophone for each speaker, Mahalanobis (1936) distances were calculated from their centroid in the F1-F2 space, and if the square root of that distance was greater than 2, then the observation was removed. Note that all measurements from timepoints for all tokens for that allophone and speaker were pooled together. To my knowledge, there are many ways that automatic outlier detection could work with trajectory data and it is

[2] Such names include the *California Vowel Shift* and the *Canadian Vowel Shift*. See Stanley (2020) for a recent review of the half a dozen other names found in recent sociophonetic research.

currently not clear which one is best or most common.

- Remove stopwords: Words that were part of a standard list of stopwords were removed.
- Remove unstressed vowels: Vowels that did not bear primary stress, based on their stress assignments in the CMU Dictionary (Lenzo 2013), were removed.
- Normalization: The data were normalized using either the Lobanov (1971) procedure or the *Atlas of North American English* procedure (Labov, Ash and Boberg 2006; cf. Nearey 1978), depending on the experiment. These two were chosen since they are among the more common methods in recent sociophonetic work.
- Defining allophones: This step of the pipeline defines relevant allophones (prenasal, preliquid, and elsewhere) for the relevant vowels[3] and makes it so that they are treated as distinct vowel categories for the remainder of the analysis. For example, prior to this step, all tokens containing /æ/ are treated as a single cluster; following this step, /æN/ is separated from pre-obstruent /æ/, which is particularly important for detecting and removing outliers.
- Remove pre-sonorants: Pre-sonorant tokens (prelaterals, prerhotics, and prenasals) are excluded since most the phenomena investigated here typically only apply to pre-obstruent allophones of these vowel. This is done so that phenomena that may be in present like the *Mary-merry-marry* merger, the *fail-fell* merger, or the *pin-pen* merger do not have any potential influence. If this step occurs before the "defining allophones" step, the latter does nothing.
- Isolating midpoints: All of the metrics used in this study are based on single-point measurements (*i.e.*, midpoints) from each vowel token. However, formant extraction tools like FAVE (Rosenfelder et al. 2014) and Fast Track (Barreda 2021a) return measurements from multiple time points along vowels' durations. Since only single-point measurements are desired, this step discards the trajectory information.

2.4 Methods

To test the effect of order of operations, the dataset was sent through 5,040 unique pipelines, representing all possible permutations of these seven steps.[4] The resulting 5,040 spreadsheets were then each analyzed for each of the three experiments. All experiments used identical functions when processing this dataset, unless specified otherwise. Within each experiment, the data is analyzed using identical code so that the only modification between trials is the order in which those procedures are applied. To be clear, the exact same input spreadsheet was used for all permutations and the exact same functions were applied to that spreadsheet. The only thing that changed was the order that those functions were applied to the spreadsheet.

3 Experiment 1: Pillai Scores

Quantifying vowel merger and/or separation has never been a straightforward task, partially because voice quality, length, trajectory, and speaker intuition are all involved in discriminating vowels (Labov 1994; Faber and Di Paolo 1995). However, one method uses Pillai scores to measure the degree of overlap between two vowel classes in the acoustic space (Nycz and Hall-Lew 2013; Kelley and Tucker 2020). These values, which are the result of a MANOVA model, range from 0 (implying complete overlap) to 1 (implying complete separation).

For this experiment, two vowel pairs are measured using Pillai scores. The low back merger is

[3] In theory, a large number of allophones could be defined for many vowels, such as prelateral, prerhotic, post-coronal, etc. Even if, say, /u/ is not relevant for studying front vowels, a careful division of the phoneme into its various allophones could affect an analysis of front vowels. For example, when compared to the distribution of the entire phoneme, a particularly backed token of /u/ (such as in *school*) may be considered an outlier and removed, which could affect vowel normalization. But if prelateral allophones of /u/ are treated independently of other allophones of /u/, that observation may not be an outlier and not removed, which again could affect vowel normalization. Admittedly, the effect is likely small. So, for simplicity, only the three allophones (prenasal, preliquid, and elsewhere) for the vowels relevant to that particular phenomenon are defined. It is beyond the scope of this paper to list all allophones that should be defined, especially since it will vary across dialects. Researchers should carefully consider which allophones may be relevant for their study.

[4] $7! = 7\times6\times5\times4\times3\times2\times1 = 5{,}040$

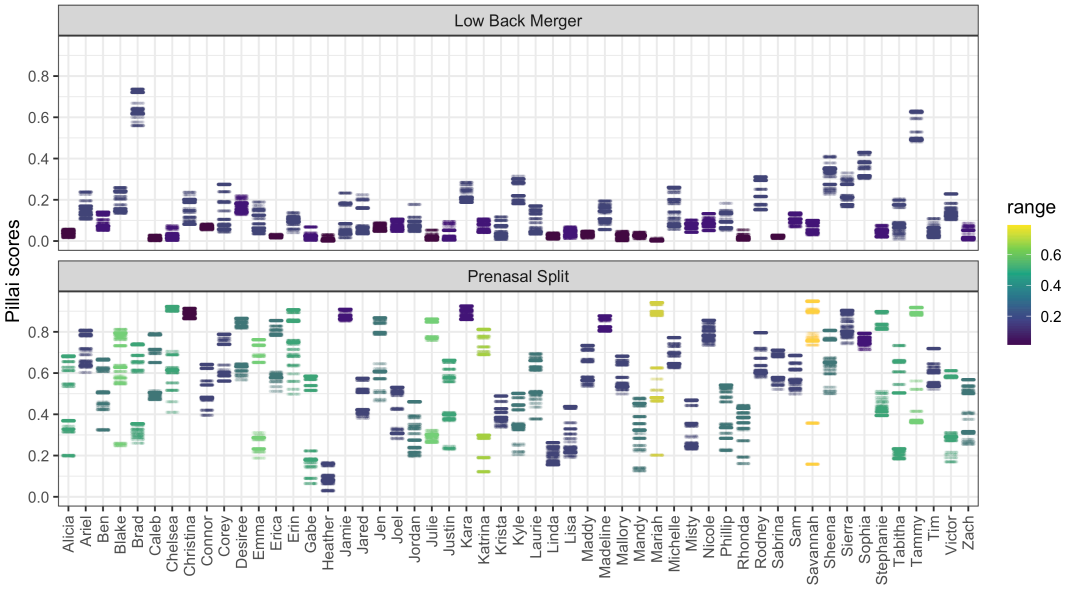

Figure 1: All Pillai scores (5,040 results per speaker, per vowel pair)

expected to have low Pillai scores. The prenasal split is also quantified by taking the Pillai score between /æ/ and /æN/. People with a greater split are expected to have high Pillai scores.

All seven steps were included in the pipeline, rearranged into 5,040 permutations. For this experiment, the ANAE normalization technique was used. Since a prenasal allophone is of interest when calculating overlap between /æ/ and /æN/, the only nasal that the "remove pre-sonorants" step excluded was /ŋ/. These pre-velar-nasal allophones (as in *bang*) were excluded since they often have a different raising pattern than the /æN/ allophone investigated here. In the pipelines for the low back merger, all prenasal tokens were excluded in the "remove pre-sonorants" step. Once the data was processed through the 5,040 pipelines, Pillai scores for both linguistic phenomena were then extracted from each speaker for each permutation, resulting in 267,120 scores. Figure 1 displays all results. For each of the two vowel pairs, there were between 94 and 223 unique Pillai scores per person. How spread out those were though varied by person and by vowel pair.

For the low back merger, the distribution of results by speaker was somewhat narrow. The average range of Pillai scores was 0.098 and the average standard deviation was 0.030. Some people had very tight clusters so that regardless of the permutations, the interpretation is always the same. Even the widest ranges were not too alarming: there are relatively few people who would be considered merged in some permutations but unmerged in others. So, overall, it seems like the results for this analysis might not be affected that much by the order of operations.

For the prenasal split though, the range of values was wider. This time, the average range of Pillai scores was 0.320 and the average standard deviation was 0.113. Some had very high Pillai scores in some permutations and low Pillai scores in other permutations. The most extreme case was Savannah, whose Pillai scores ranged from 0.157, indicating high overlap, to 0.945, indicating virtually no overlap. An analysis that happens to use one of the 16 permutations that produced the low score would classify Savannah as having little prenasal split and perhaps a lagger in this linguistic change. Meanwhile an analysis that happens to use any one of 174 permutations that produced the highest score and would classify her as being a leader in this change. The interpretation of Savannah's participation in this change is therefore dependent on the order of operations.

4 Experiment 2: *ANAE* Benchmarks

As with establishing whether two vowels are merged, concluding definitively whether a vowel is shifted or the degree to which it is shifted in a speaker is difficult. The vowel's relative position to other vowels may provide some clue but only if those reference vowels themselves are not shifting (which may be hard to determine). And a definitive cutoff value in raw Hz is nonsensical because

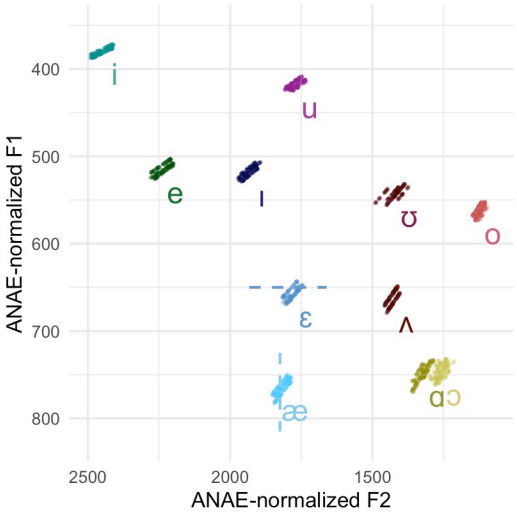

Figure 2: Corey's vowel space

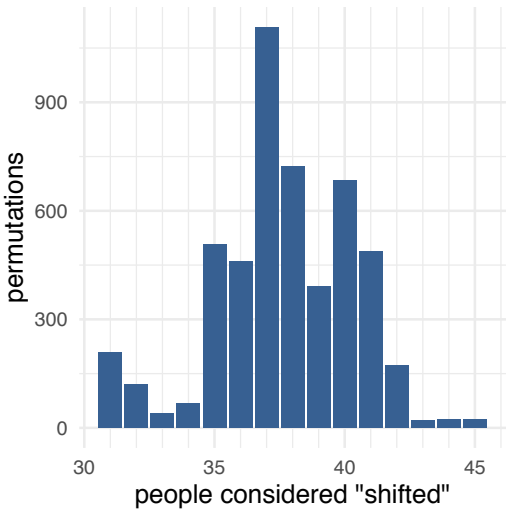

Figure 3: People shifted and permutations

of speaker-specific anatomical differences. Nevertheless, many have interpreted numbers in the legends of some of the maps in the *Atlas of North American English* (*ANAE*; Labov, Ash and Boberg 2006) as "benchmarks" that can be then applied to their own data.[5] For this paper, benchmarks from Map 15.4 (Labov, Ash and Boberg 2006: 217), which illustrates the Canadian Shift, will be used. Specifically, if the F1 of /ɛ/ is higher than 650 Hz and if the F2 of /æ/ is lower than 1825 Hz, those two vowels are considered shifted. Of course, these numbers only make sense if the data is normalized using the same procedure as what is described in the *ANAE* (Dinkin 2018).

Again, all seven steps were included in the pipelines. Most steps were identical to those used in Experiment 1, with two exceptions. First, the data was normalized using the method described in the *ANAE* (Labov, Ash and Boberg 2006; Nearey 1978) so that the results are more comparable. Second, the "defining allophones" step in this experiment meant defining preliquid, prenasal, and preobstruent allophones of /æ/, /ɛ/, and /ɪ/. Allophones of other vowels could very well have been included, but having a small set was sufficient for the purposes of this study. These seven steps were rearranged into 5,040 permutations and, after being processed by each pipeline, the mean F1 of /ɛ/ and the mean F2 of /æ/ were compared to the *ANAE* "benchmarks." Due to limitations in space, it is not feasible to display results for all speakers, but Figure 2 shows an example speaker whom I call "Corey," a man born in Montana in 1990. There are 5,040 dots per vowel, each representing the mean F1-F2 value for a single permutation. Each of Corey's vowels has 151–199 unique outcomes, which was typical for this sample.

As seen in Figure 2, the benchmarks, represented by dashed lines, go through Corey's /ɛ/ and /æ/ vowel clouds. In 77.8% of the permutations, /ɛ/ was past the benchmark. Perhaps that should lead a researcher to conclude that his vowel is shifted. But a simple majority may not be the deciding factor because one of the 1124 permutations that put his /ɛ/ above the threshold may be the best one methodologically and theoretically. Besides, should researchers run all permutations to decide the majority interpretation? His /æ/ was more ambiguous since it was considered shifted in 46.3% of the permutations, which makes interpreting that vowel even more difficult.

While not all speakers in this sample were as problematic as Corey, he was certainly not unique. Of the 53 speakers, 14 had the same results every time, meaning average measurements of /æ/ and /ɛ/ were always on the same side of the benchmark in all 5040 permutations. The rest of the 39 people would be considered shifted in some permutations and not shifted in others. The result is an inconsistent number of people are considered shifted. Some permutations would suggest that only 31 people in this sample have the shift while other permutations suggest that 45 people have it

[5] I do not believe these "benchmarks" were at all intended to be used in this way; in fact, I believe they should *not* be used in this way (see more discussion in Stanley 2020: 88–90). Nevertheless, they will be used here because of their preponderance in recent sociophonetic literature.

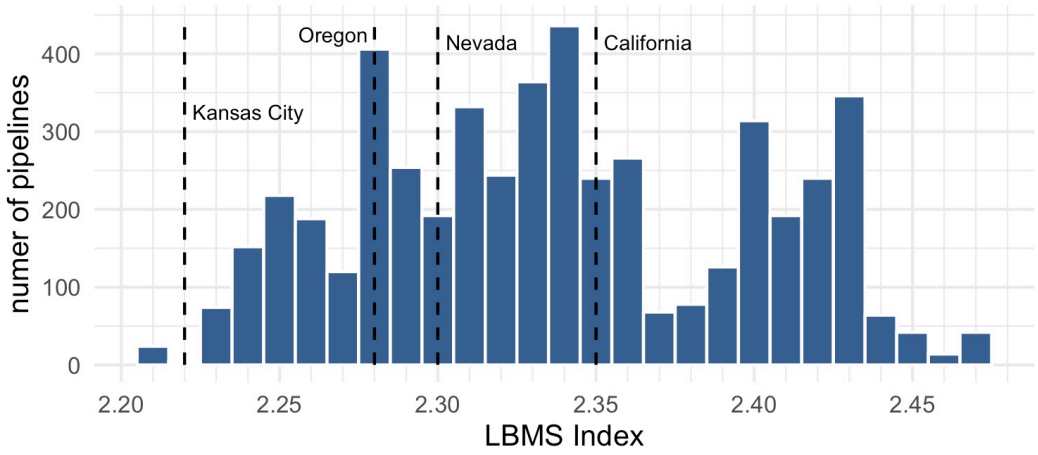

Figure 4: Jen's Distribution of LBMS Indices

(Figure 3). So just a change in the order can shift the overall interpretation of this dataset.

5 Experiment 3: The LBMS Index

As an alternative to the reliance upon *ANAE* "benchmarks," the LBMS Index has been recently proposed to quantify the LBMS as a continuous measure (Becker 2019b). It is defined as the average of three Euclidean distances: between /i/ and /ɪ/, /i/ and /ɛ/, and /i/ and /æ/. Larger numbers suggest greater shifting (specifically lowering and/or retraction) in one or more vowels. While it is not without its issues, such as reducing movement in F1 and F2 across three vowels into a number that itself is difficult to interpret, the chapters within Becker (2019a) illustrate its use in several speech communities across North America and provide reference values for future researchers to compare their data against. Becker offers suggestions on what data processing should happen before calculating the LBMS Index, but an order is not specified.

For this experiment, the seven steps and 5,040 pipelines from Experiment 2 were also used here. There were two changes to reflect Becker's guidelines. First, the "removing pre-sonorants" step also removed tokens that were followed by /g/. Second, the Lobanov normalization procedure was used. This resulted in 5,040 LBMS Indices for each of the 53 speakers.

Again, while it is not feasible to show the results for all speakers, Figure 4 shows the distribution of LBMS Indices for "Jen," a woman from Colorado born in 1981. Of the 5,040 permutations, Jen had 243 unique LBMS Indices, which is the median in this sample. The plot shows that there was wide variation in Jen's LBMS Indices. Overlayed on the plot in Figure 4 are some of the reference values from areas in or near the Rockies, as reported in Becker (2019a). Is Jen more shifted than what Fridland and Kendall (2019) found in California? Or is she perhaps somewhat conservative, with lower LBMS Indices than what Strelluf (2019) found in Kansas City? Is Jen on the forefront of change, a lagger, or somewhere in the middle? What a naïve researcher may conclude will depend on the order of operations that happens to be employed.

Let's for a moment treat the Rockies region generally as a single speech community and take the median LBMS Index across these 53 speakers for each permutation. Figure 5 shows the distribution of these medians. The same issue of trying to decide how this sample compares to other samples persists. The lowest median was 2.23, which is the lowest reported LBMS Index in the West, suggesting that the Rockies are the most linguistically conservative area of the West. Meanwhile, the highest value was 2.53, suggesting shifting in the Rockies is on par with the Vancouverites in Swan's (2019) sample. The amount of variation that results in changing the order of operations is on par with the magnitude of sociolinguistic variation that some would consider significant. The problem is that this is not meaningful sociolinguistic variation: they are differences in data processing.

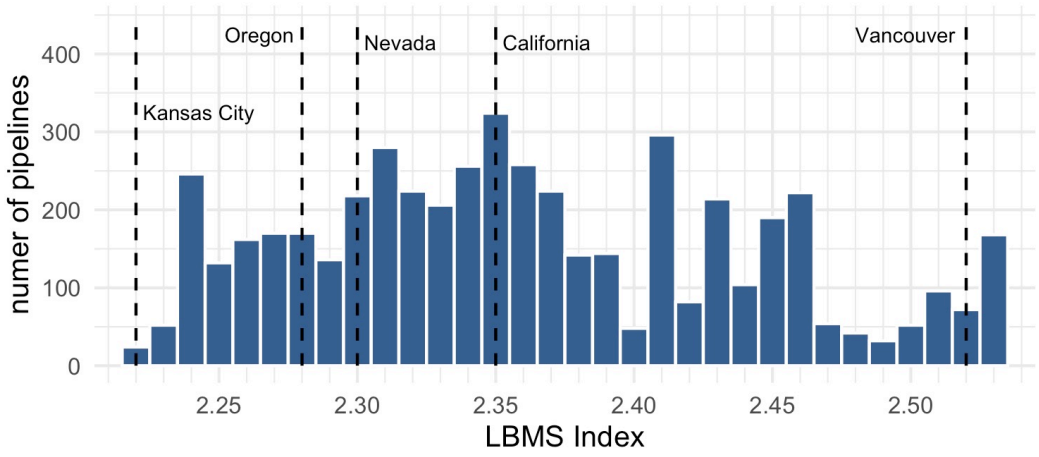

Figure 5: Distribution of mean LBMS Indices across the entire sample.

6 Discussion

To summarize, the same data was processed 5,040 different times by adjusting the order of operations, producing hundreds of unique outcomes per speaker per phenomenon in most cases. In Experiment 1, differences in Pillai scores for the low back merger across permutations was small. But for prenasal-raising, the differences between permutations were huge. For measuring shifting, Experiment 2 compared vowel measurements to so-called "benchmarks" and Experiment 3 calculated the LBMS Index. The distribution of those values was quite large—large enough that some people were considered shifted in some permutations and not others. In all three cases, the order of operations mattered. The differences between these permutations were on the same magnitude as reported sociolinguistic differences.

One important implication for these results is that comparison across studies becomes problematic since order of operations is typically not reported. In the case of Jen, a naïve analysis may conclude that her LBMS Index is 2.34, but when the order of operations is unknown, that number is more or less a random draw from the range of possible values that Jen's data could have produced. A similar issue arises when trying to compare one dataset to another. It is theoretically possible that the LBMS Index reported in one study happened to use the pipeline that systematically produces the highest LBMS Indices while another study happened to use the pipeline that produces the lowest LBMS Indices. Without detailed documentation of these steps—and familiarity with how changes in the order affects the overall results—interpreting numbers from other studies is difficult.

7 Recommendations and Conclusions

Since there is no precedent for an order of operations, I conclude this study by offering a recommended order. Please note that this is an opinionated order and is largely based on reasoning rather than on the results of these experiments. Broadly, the order is centered on the timing of normalization: bad data should be removed before normalizing and then good data should be removed after.

1) The first step in the analysis should be to appropriately identify allophones so that later steps in the pipeline (such as outlier detection) happen by *allophone* rather than by broad phonemic categories. For American English data, this typically means that prenasal, prelateral, prerhotic, postcoronal, prevelar, and prevoiceless allophones should be identified, depending on the vowel and variety in question. This subgrouping should happen even if such allophones are not part of the research question and will be excluded later in the pipeline. As a part of this step, stopwords and reduced vowels should also be tagged appropriately

since they arguably belong to a different distribution than vowels in stressed, content words. At this point, no data is excluded yet; it is just reclassified.

2) Once all the data has been categorized, bad data should be removed first so that subsequent steps are unaffected by non-data. This typically involves identifying and handling outliers, which should be done by *allophone* rather than by phoneme. Stopwords and unstressed vowels should also be treated as independent categories as well.[6] This step ensures that the remaining observations moving forward are actual linguistic data, representing formants from all vowels and allophones included in the sample, regardless of whether they're the object of study, so that normalization is based on as complete a vowel space as possible.
3) At this point, I believe it is best to normalize the data. Again, it is beyond the scope of this paper to recommend a specific normalization technique, but at least based on the Lobanov and *ANAE* procedures that were used in this paper, the location of normalization within the data analysis pipeline is the most important factor on the overall results.
4) Finally, it is now appropriate to subset the data and remove any data that is not part of the subsequent analysis. This may include removing diphthongs, presonorant tokens and other allophones, tokens from stop words, unstressed vowels, and isolating midpoints from their trajectories.[7]

Regardless of whether this recommended order is implemented, it is important that sociophoneticians begin reporting the order of operations as part of their methods sections. Relatedly, if specific numbers from a previous study are being compared to and if the order of operations is not reported in that study, this oversight should be acknowledged and the results should be interpreted with care.

More broadly, sociophoneticians and quantitative linguists must be mindful of how their data is being processed. Be aware of what kind of normalization procedure or outlier detection method is used, what assumptions they make about the data, and what cognitive process they aim to model. Massaging data for the sake of massaging data might make sense when looking at just numbers, but since these numbers represent actual language data, every step of our analysis should theoretically be grounded some cognitive process.

Sociophoneticians and researchers generally should stay flexible and adapt to new tools and techniques. A particular method may be easy to comprehend, simple to implement, or standard in the field, but it is important to not get too attached to any one technique. If recent studies find that some method is no good for whatever reason, researchers should be willing to abandon their habitual methods. Comparability with previous work may be important, and using a combination of traditional and innovative techniques may be necessary temporarily to bridge the gap between old and new work. Just as a teacher constantly adapts their lesson plans incorporate new developments in the field, a researcher should likewise modify their analysis to reflect recent findings in linguistic methodology.

Finally, I encourage more conversation and publications about methods in sociophonetic analysis. To reiterate, studies that explicitly compare similar techniques are good for the field because they uncover issues in current methodologies. Quantitatively-minded linguists should not hesitate to propose new methods—even for steps in the pipeline that appear to be resolved. While some such new techniques may not stand the test of time, these failed attempts may help discover others that become the new standard for a time.

This paper explored one small part of sociophonetic data analysis, the order of operations, and

---

[6] At this point, I tentatively argue all unstressed vowels should be treated as a single group for the purposes of outlier detection. One could argue that they should be grouped by their underlying phoneme, but sometimes that is not possible to deduce. Separating types of unstressed vowels (as in *Rosa's* and *roses* for people who make the distinction) can complicate the analysis since not all people will likely have the same distribution. Additional meta-analysis of sociophonetic methods can further inform this step of the pipeline, particularly when it comes to the handling of stopwords.

[7] It is important to note that if trajectory data is never considered, either because a custom Praat script never extracted it or because the columns in FAVE's output that contain that trajectory information were ignored, then this step will automatically be first in a researcher's pipeline and is contrary to the recommended order provided here. I believe it is necessary to work with vowel trajectories at least a little bit when detecting outliers and normalizing, even if they are not part of the analysis.

found that small changes in the ordering of otherwise identical functions can result in a different interpretation of the data. A recommended order was provided to help mitigate this issue. It is my hope that this paper inspires others to consider their methods more carefully so that we as a field may learn from our mistakes. Even though this paper says nothing about how language works, more informed data analysis can allow researchers to conduct more informed linguistic analysis.

References

Adank, Patti, Roel Smits, and Roeland Van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America* 116(5). 3099–3107.

Barreda, Santiago. 2021a. Fast Track: fast, (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard* 7(1).

Barreda, Santiago. 2021b. Perceptual validation of vowel normalization methods for variationist research. *Language Variation and Change* 1–27. https://doi.org/10.1017/S0954394521000016.

Becker, Kara (ed.). 2019a. *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and short front vowel shifts across North America* (Publication of the American Dialect Society 104). Durham, NC: Duke University Press.

Becker, Kara. 2019b. Introduction. In Kara Becker (ed.), *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and short front vowel shifts across North America* (Publication of the American Dialect Society 104). Durham, NC: Duke University Press.

Dinkin, Aaron. 2018. Revisiting the Inland North Fringe. Presentation presented at the New Ways of Analyzing Variation 47, New York City, NY.

Evanini, Keelan. 2009. *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*. Philadelphia, PA: University of Pennsylvania Ph.D. dissertation.

Faber, Alice, and Marianna Di Paolo. 1995. The discriminability of nearly merged sounds. *Language Variation and Change* 7(1). 35–78.

Fridland, Valerie, and Tyler Kendall. 2019. On the Uniformity of the Low-Back-Merger Shift in the U.S. West and Beyond. In Kara Becker (ed.), *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and short front vowel shifts across North America* (Publication of the American Dialect Society), vol. 104. Durham, NC: Duke University Press.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and linguistics compass* 3(1). 359–383.

Johnson, Daniel Ezra. 2014. Progression in regression: Why natural language data calls for mixed-effects models. Self-published manuscript.

Johnson, Daniel Ezra. 2015. Quantifying vowel overlap with Bhattacharyya's affinity. Presented at the New Ways of Analyzing Variation (NWAV44), Toronto.

Kelley, Matthew C. and Benjamin V. Tucker. 2020. A comparison of four vowel overlap measures. *The Journal of the Acoustical Society of America* 147(1). 137–145. https://doi.org/10.1121/10.0000494.

Kendall, Tyler, and Charlotte Vaughn. 2020. Exploring vowel formant estimation through simulation-based techniques. *Linguistics Vanguard* 6(s1). https://doi.org/10.1515/lingvan-2018-0060. https://www.degruyter.com/view/journals/lingvan/6/s1/article-20180060.xml (4 November, 2020).

Labov, William. 1994. *Principles of linguistic change. Vol. 1: Internal features* (Language in Society). Oxford: Wiley-Blackwell.

Labov, William, Sharon Ash, and Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Walter de Gruyter.

Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89(1). 30–65. https://doi.org/10.1353/lan.2013.0015.

Lenzo, Kevin. 2013. *The CMU Pronouncing Dictionary*. Carnegie Mellon University. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Lobanov, Boris M. 1971. Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America* 49(2B). 606–608. https://doi.org/10.1121/1.1912396.

Mahalanobis, Prasanta Chandra. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* 2(1). 49–55.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association*.

Nearey, Terrance Michael. 1978. *Phonetic feature systems for vowels*. University of Alberta Dissertation.

Nycz, Jennifer, and Lauren Hall-Lew. 2013. Best practices in measuring vowel merger. *Proceedings of Meetings on Acoustics* 20(1). 060008. https://doi.org/10.1121/1.4894063.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. *FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2*.

Sankoff, David. 1975. *VARBRUL 2*.

Sankoff, David, Sali A. Tagliamonte, and Eric Smith. 2005. *Goldvarb X: A variable rule application for Macintosh and Windows*. Department of Linguistics, University of Toronto.

Stanley, Joseph A. 2020. *Vowel dynamics of the Elsewhere Shift: A sociophonetic analysis of English in Cowlitz County, Washington*. Athens, Georgia: University of Georgia Ph.D. Dissertation.

Strelluf, Christopher. 2019. Structural and Social Correlations with the Low-Back-Merger Shift in a U.S. Midland Community. In Kara Becker (ed.), *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and short front vowel shifts across North America* (Publication of the American Dialect Society), vol. 104. Durham, NC: Duke University Press.

Swan, Julia. 2019. The Low-Back-Merger Shift in Seattle, Washington, and Vancouver, British Columbia. In Kara Becker (ed.), *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and short front vowel shifts across North America* (Publication of the American Dialect Society), vol. 104. Durham, NC: Duke University Press.

Tagliamonte, Sali A., and Alexandra D'Arcy. 2017. Individuals, communities and the sociolinguistic canon. Presented at the New Ways of Analyzing Variation (NWAV) 46, Madison, WI.

Department of Linguistics
Brigham Young University
4064 JFSB, Provo, UT 84602
joey_stanley @byu.edu