Interpreting the Order of Operations in a Sociophonetic Analysis

Joseph A. Stanley
Brigham Young University
joey_stanley@byu.edu
ORCID: 0000-0002-9185-0048

Abstract

Sociophonetic data analysis involves a pipeline of processing steps to convert a raw spreadsheet of acoustic measurements to interpretable results. While most studies report the steps used in their pipeline, very few explicitly report their order in which those steps were applied. This study analyzes a dataset containing vowel formant data from 53 speakers by processing it 5,040 unique ways, each representing a different permutation of seven processing steps. To analyze the effect that an order has on the overall results, pairs of pipelines that differed only by swapping two adjacent steps, were compared. The most important steps in the pipeline were when normalization happened, how outliers were detected, and when good data was excluded. This study illustrates the what happens when these steps are rearranged relative to each other in order to justify and recommend the following order of operations: classifying allophones, removing outliers, normalizing, and then subsetting.

Keywords: sociophonetics, data processing, quantitative methods

1 Order of operations

As an increasingly quantitative field, sociophoneticians and other linguists who work with numeric data are constantly bombarded by new methods, techniques, tools, and statistics. Many of these tools are meant to facilitate tasks that were once labor-intensive, giving the current generation of researchers access to larger datasets and opening doors to new research questions. For example, automatic formant extraction has made it possible to process and analyze an order of magnitude more data than previous techniques (Labov, Rosenfelder & Fruehwald 2013: 35) and it is not uncommon to hear of datasets containing hundreds of thousands, if not millions, of vowel tokens (e.g. Olsen et al. 2017; Brand et al. 2021; Kendall & Farrington 2021).

Regardless of whether a project uses fully automatic or completely manual methods, the typical goal of data processing is to convert audio into a spreadsheet of acoustic measurements to allow for statistical analysis. To accomplish this task, one must send the data through a pipeline of processing steps that includes aligning text to speech, extracting acoustic measurements like vowel formants, and filtering the full dataset to include only the tokens of interest. Since modern quantitative research values transparency and reproducibility in methodology sections, most sociophonetics papers today fortunately provide detail on what processing procedures were done to the data.

However, Stanley (2022) finds that it is not enough to simply report the processing steps in a methods section because the *order* in which those steps applies matters. Stanley observed that the overall interpretation of a dataset can be different if, for example, normalization occurs *before* removing outliers as opposed to occurring *after* removing outliers. To test this more rigorously, he processed the same spreadsheet 5,040 times, each with a different pipeline of data analysis steps. The component parts of those analyses were identical from one pipeline to the next, but it was the order in which those steps were applied that varied. He found that what the results show and what the research may ultimately conclude about a particular dataset can change somewhat dramatically depending on the order of operations. He ends the study with a recommended order that sociophonetic researchers can adopt (Figure 1), as well as a call to action, urging researchers to report their analyses with greater detail.

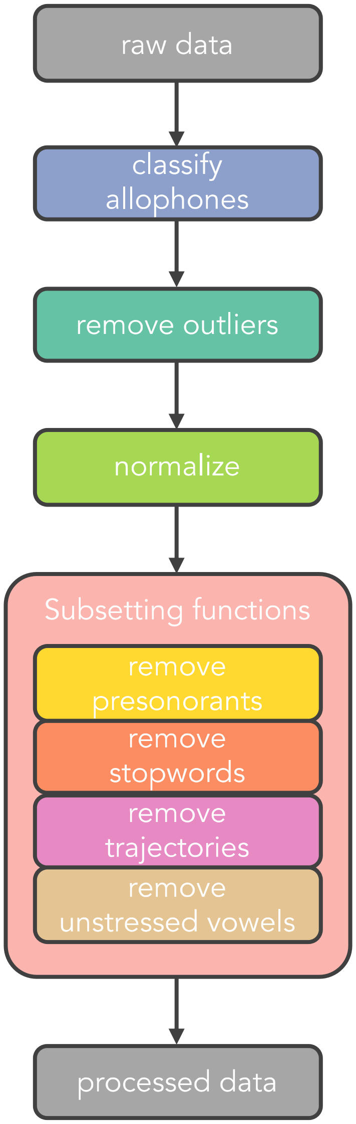

Figure 1: The order of operations recommended by Stanley (2022) and justified in this paper.

Reporting the order of operations in methods sections is unusual (see Brand et al. 2021 fig. 2, for a rare example) but very good for transparency. However it does little good if researchers are not familiar with how the order affects the overall results. A detailed methods section may explain that normalization happened before outliers were removed, but it is not currently clear what effect that order had on the results. How should a reader evaluate the results of one study that normalized the data before removing outliers against another study that transposed those two steps?

This paper addresses this gap and explores in more detail the effect that some orderings are likely to have on the results of a study. It is obviously not feasible to describe every possible permutation, but instead, some of the most important patterns will be illuminated and explained. The goal for this paper is to arm readers with some working knowledge of how to interpret others' order of operations[1] and to justify the order presented in Stanley (2022).

2 Method

2.1 Data

The data for this study is the same as what was used in Stanley (2022). To begin, vowel formant measurements were automatically extracted from interviews with 53 Western American English speakers. Such raw data is typically not analyzed directly and usually goes through a pipeline of processing steps that ultimately trims it down to the subset that will be used for analysis. For example, this subset may contain only midpoints of stressed vowels in certain phonetic environments from content words, with outliers removed and the data normalized to facilitate comparison across speakers.

For this study, seven steps were selected, based on their apparent ubiquity in sociophonetic studies on North American English, to ultimately accomplish this processing: defining allophones, removing outliers, normalization, removing presonorant tokens, removing stopwords,[2] isolating midpoints from vowel trajectories, and removing unstressed vowels. These seven steps can be rearranged into 5,040 permutations or orders of operations, so the original raw data was therefore processed 5,040 times, each representing

[1] The information presented in this study can, unfortunately, be misused. If a particular result is desired, such as a higher LBMS Index or lower Pillai score, a simple "fix" would be to process the data using a different pipeline. What is presented in the next sections are essentially "cheat codes" to do just that. Hopefully, as more researchers, reviewers, and editors become familiar with order of operations, they will be in a better position to identify poorly conducted and reported methods.

[2] Stopwords are words that are very frequent and are often subject to vowel reduction. While there is no established list of what is or is not a stopword, such lists typically contain pronouns, prepositions, other grammatical particles, and perhaps some very frequent content words.

a different hypothetical analysis of the same data. These permutations collectively generated 5,040 spreadsheets, each representing a hypothetical analysis conducted on the original data.

For each of the resulting spreadsheets, the following three metrics were then extracted for each speaker:

- Pillai scores (Pillai 1955; Nycz & Hall-Lew 2013) were calculated between /ɑ/ and /ɔ/ to quantify the Low Back Merger (*i.e.* the merger of *cot* and *caught*). Pillai scores were also calculated between preobstruent and prenasal allophones of /æ/ to assess the degree of prenasal raising in words like *ban*, *sand*, and *hand*.
- The normalized F1 and F2 measurements of /ɛ/ and /æ/ were compared to the "benchmarks" derived from the *Atlas of North American English* (ANAE; Labov, Ash & Boberg 2006) as a way to determine whether a particular speaker has shifted those vowels.
- The Low-Back-Merger Shift (LBMS; Becker 2019a) was calculated using the LBMS Index (Becker 2019b). The shift involves the lowering and centralizing of /ɪ/, /ɛ/, and /æ/ and is a more gradient measure than the binary comparison to ANAE benchmarks. The LBMS Index calculates the average Euclidean distance between those three vowels and /i/ in the Lobonov-normalized space.

These three measures were chosen simply because they are commonly used in current North American sociophonetic research, particularly among those who study the spread of Low-Back-Merger shift across time and space.

In the end, a final spreadsheet was compiled and contained these five metrics (Pillai scores for the Low Back Merger and prenasal raising, whether /ɛ/ or /æ/ were shifted past the "benchmarks," and the LBMS Index) for each of the 53 speakers, for each of the 5,040 permutations.

2.2 Analysis and processing

While Stanley (2022) describes the results from this spreadsheet in general terms, the current study dives deeper by uncovering patterns in the results. Consider, for example, a hypothetical speaker with the pseudonym "Justin" whose raw formant data is freely available in an online repository. Researcher A processes Justin's data using these steps presented on the left side of Figure 2. Researcher B, who doing an independent analysis, uses a similar pipeline, only the first two steps, removing stopwords and removing outliers, were swapped, as seen in the right side of Figure 2. Researcher A concludes that Justin's low back vowels had a Pillai score of 0.051. Researcher B finds it to be 0.055. This is admittedly a small difference, but it is a difference nonetheless. Stanley (2022) shows that the magnitude of this difference is typical across the 53 speakers in this dataset. However, the question that remains is whether it is always the case that the first order of operations produces a lower Pillai score than the second.

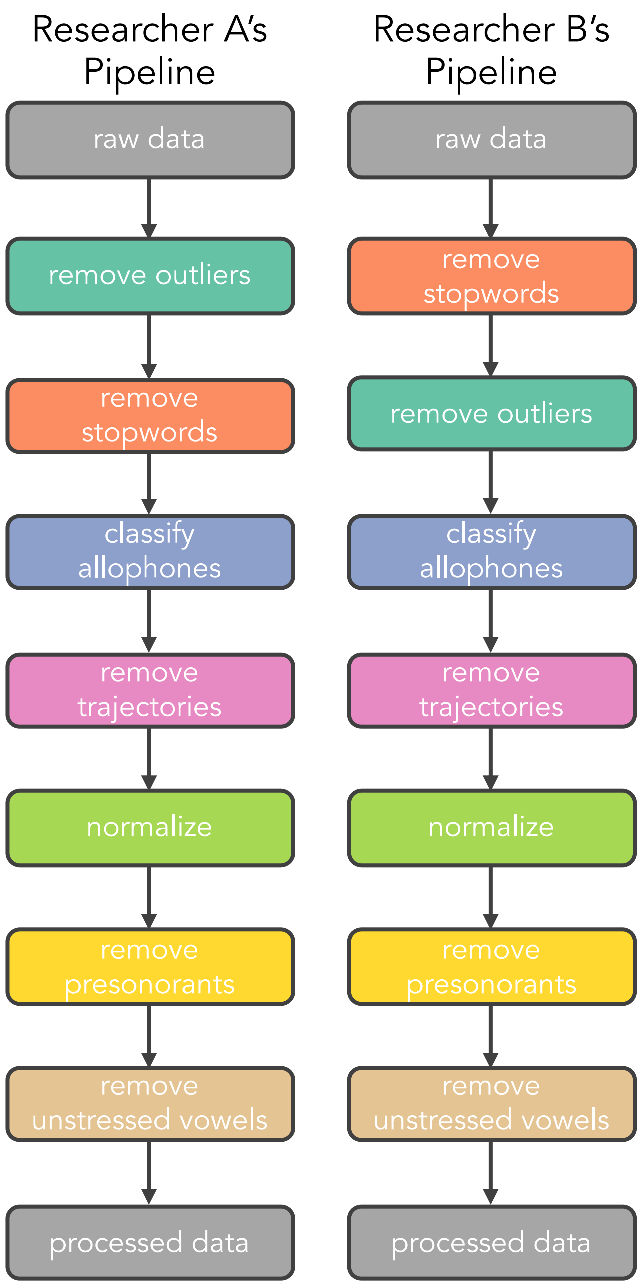

Figure 2: Two similar pipelines used by hypothetical researchers. Note that these do not reflect the order of operations recommended in this paper.

After attempting different approaches, I found that the best way to systematically explore the permutations was to compare all pairs of permutations where the only difference between them was that two adjacent steps were switched. I focus on steps that are adjacent in the recommended order of operations presented in Figure 1. By focusing on small changes like swapping adjacent steps, general trends vis-à-vis those steps can still be observed. These comparisons were done using the near() function in the dplyr package (Wickham et al. 2018) in the R programming language (R Core Team 2021), which allows for a very small amount of tolerance (on the order of eight decimal places) so that numbers that were not truly identical, but for all intents and purposes are the same, could be treated as identical.

With these comparisons in hand, I asked two broad questions: what permutations resulted in identical output and for permutations that did change the output, what reliable patterns were there? The following sections answer these questions. Section 3 identifies functions that, when swapped, do not change the overall result. Section 4 addresses the first step in the recommended pipeline, allophone classification, and considers its relationship to the second step, outlier removal. Section 5 briefly describes outlier removal and why it should happen before normalization. Section 6 then focuses on normalization and its relationship to the final step, the subsetting functions, described in section 3.

3 Permutations that make no difference

The first category of results includes identifying which permutations resulted in identical output. For some of the seven functions used here, it makes no difference if they get applied in different orders so long as they are adjacent to each other in the pipeline. For example, removing presonorant tokens and removing stopwords have no effect on each other because there are no tokens that are defined as presonorant only after stopwords have been removed, and vice versa. The set of observations that are excluded in these steps is fixed: regardless of the normalization procedure, whether outliers have been removed, or how the vowels are classified, the exact same set of observations will be excluded each time. What other pairs like these are there?

Upon examination of the 5,040 permutations, it appears that most of the procedures that involved subsetting the data are independent of each other and can be swapped with no effect on the overall results. Specifically, these functions are isolating midpoints from their trajectories, removing presonorant tokens, removing tokens from stopwords, and removing tokens with lexical stress. Regardless of whether these functions all happen at once or in adjacent steps in the pipeline, there is no effect on the overall results.

For the remainder of this paper, these four functions will be collapsed down into a single "subsetting" function. This is justified theoretically because they accomplish the same purpose of removing data that is good but of no interest to the researcher's current project. It is also justified from a programming standpoint since they can easily be accomplished using the same function. As shown in Figure 1 and described in Section 6, it is recommended that these subsetting steps all happen at the end of the data analysis pipeline, after defining vowel classes, removing outliers, and normalization have occurred.

We now turn to the permutations that do have an effect on the overall result and the more complicated matter of identifying what effect they have.

4 Steps 1 and 2: Allophone classification and outlier removal

While the previous section described subsetting functions, which remove good but uninteresting data, this section focuses on the effect of outlier removal, which excludes data considered to be bad, and its relationship to allophone classification. This step is different from the other subsetting functions discussed above because what is considered an outlier changes depending on the steps that have already occurred.

Identifying and (potentially) excluding outliers is an important step in any data analysis project. Ideally, all extreme values and outliers should be hand-checked, though this is infeasible when analyzing large datasets. One compromise is to do a blanket exclusion of all extreme observations under the assumption that they are true outliers. This is sometimes done by taking the *z*-scores for each formant, for each vowel, for each speaker and remove observations that are more than, say, two standard deviations from the mean. An alternative is to use Mahalanobis distances, which can be thought of as finding the centroid for each vowel for each speaker, and fitting an ellipse centered around that point, rotated and stretched to best match the distribution of the data. How much data is encompassed by the ellipse is determined by the researcher, and observations that fall outside of it are excluded.

To examine the effect of outlier exclusion and its relationship to classifying allophones, consider prenasal raising, which is sometimes measured using Pillai scores. The vowel in words like *bat* and *ban* are underlyingly /æ/, and what the Pillai scores quantify is the amount of overlap between the two allophones. A higher Pillai score suggests less overlap and, consequently, a more raised /æN/. As shown in the left panel of Figure 3, if outlier detection happens by phoneme, that vowel's centroid will be somewhere between the two allophones. For speakers with little overlap in their allophones (or in smaller, cleaner datasets), this centroid may be in an area where no actual tokens occur. Based on Mahalanobis distances from that single center point, only the most raised tokens of /æN/ and the most lowered tokens of preobstruent /æ/ will be considered outliers. Because only these tokens are removed, and low observations of /æN/ that are in the "territory" of preobstruent /æ/ and high tokens of preobstruent /æ/ that are in the "territory" of /æN/ are retained, the two clusters are artificially drawn together, resulting in greater overlap and a lower Pillai score.

However, if allophones have already been defined, outliers can be determined and excluded by allophone. As shown in the right panel of Figure 3, this creates two centroids, appropriately centered among the cluster of tokens for the speaker's allophones. Now, some of the extreme observations that were considered outliers in the other pipeline are retained in the analysis and some of the observations that were in the territory of the other allophone have been removed. The result is that the amount of overlap between the two is smaller since they are more distinct from one another, yielding a higher Pillai score.

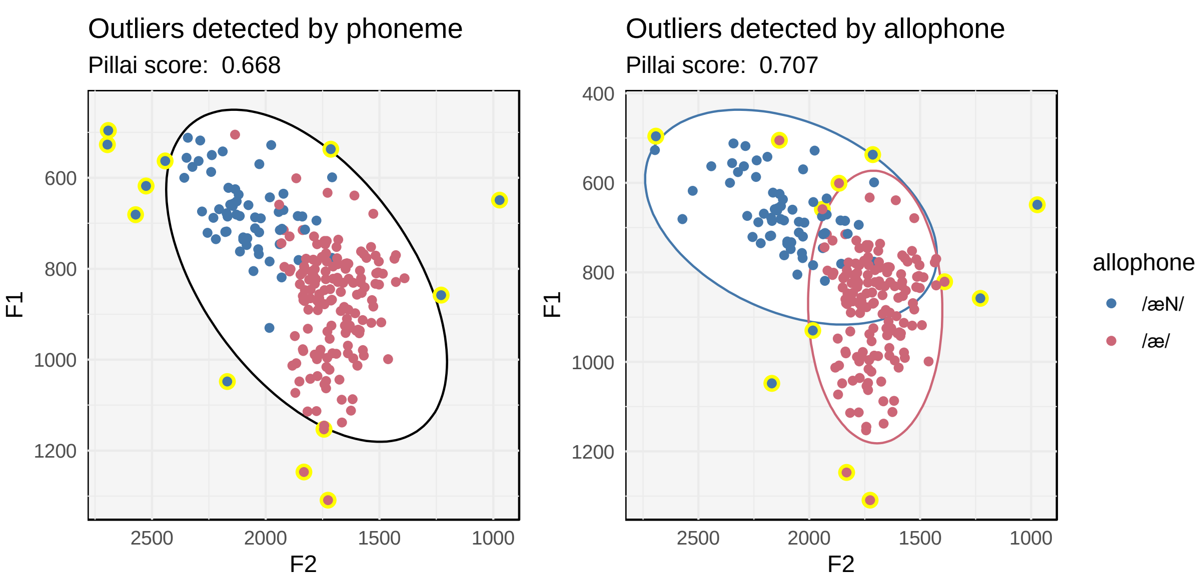

Figure 3: F1 and F2 measurements of ban (blue) and bat (red) from "Sabrina," a female speaker born in Idaho in 1977. Ellipses, which were determined by Mahalanobis distances, show the inclusion region(s); any observations outside of the region are highlighted in yellow and would presumably be excluded from further

analysis. Note that, for the purposes of providing a clean visual, the subsetting functions have already been applied, though that is not the order of operations recommended in this paper.

So, a simple swap of processing steps – defining allophones and outlier removal – can influence the results of the study. A researcher who removes outliers by *phoneme* will find less prenasal-raising (or any allophonic split) than a researcher who removes outliers by *allophone*. For a single speaker this difference may be inconsequential, but when such differences are consistent across many speakers, it can influence how an allophonic split is interpreted in a sample, perhaps leading the researcher to determine whether that split is present in the speech community. Figure 3 makes it apparent that outliers should be detected at the allophonic level, so defining allophones should come before outliers are detected.

5 Steps 2 and 3: Outlier removal and normalization

Figure 1 suggests that outliers be removed before normalization. The goal of normalization is to map acoustic data to the perceptual vowel space, so that an [æ] produced by a smaller person and a perceptually identical [æ] produced by a larger person line up even though they are acoustically different (Barreda & Nearey 2018); essentially it aims to mimic the human ear by eliminating physiological differences while maintaining sociolinguistic differences. One method is the Lobonov (1971) technique which transforms formant measurements into *z*-scores independently for each speaker and each formant. Another technique is the one described in the ANAE (Labov, Ash & Boberg 2006: 39–40) which, for each speaker, finds the mean log-transformed formant frequency, compares it to the mean log-transformed formant frequency for all speakers in the sample, and uses that single parameter to adjust all formants for that speaker.

Rather than intensely scrutinize what happens when these two steps are swapped, I instead offer a logical explanation. Both normalization procedures used here rely on an average vowel formant frequency, whether that be per formant (as in Lobanov) or with all formants pooled together (as in ANAE). Mathematically, the mean is sensitive to outliers and the presence of a single extreme observation can shift the mean away from its “true” value in the direction of the outlier. It makes sense then that outliers be removed before normalizing so that their presence does not interfere with the resulting normalized values of the remaining "good" data.

6 Steps 3 and 4: Normalization and subsetting

Finally, we explore last pair of functions in Figure 1, the effect that vowel normalization has on a dataset that includes good but irrelevant observations (such as the ones removed in the subsetting step), and on one that does not. Examining the effect of normalization is more complicated than other steps in this study because two different procedures are used. Comparison to ANAE "benchmarks" of course necessitates the use of ANAE normalization so that the values are comparable. When calculating the LBMS Index, Becker (2019b) recommends using Lobonov (*z*-score) normalization. The Pillai score analysis here used ANAE normalization, though Stanley (2021) shows that Pillai scores are unaffected by either of these two normalization procedures.

In the recommended order of operations, normalization happens before subsetting has occurred. How does this compare to normalizing after subsetting? Figure 4 illustrates this comparison with both normalization procedures applied to two speakers' data. Within each of the four panels, the data is shown in two ways. For each observation, arrows start at the point in the normalized F1-F2 space where that observation is located when normalization occurs before subsetting (which is the recommended order). Arrows point to the location where those same observations are located in the normalized F1-F2 space when normalization occurs after subsetting the data. That is, each vowel token is represented by an arrow and the arrows collectively show the effect on the full vowel space when normalization occurs after subsetting.

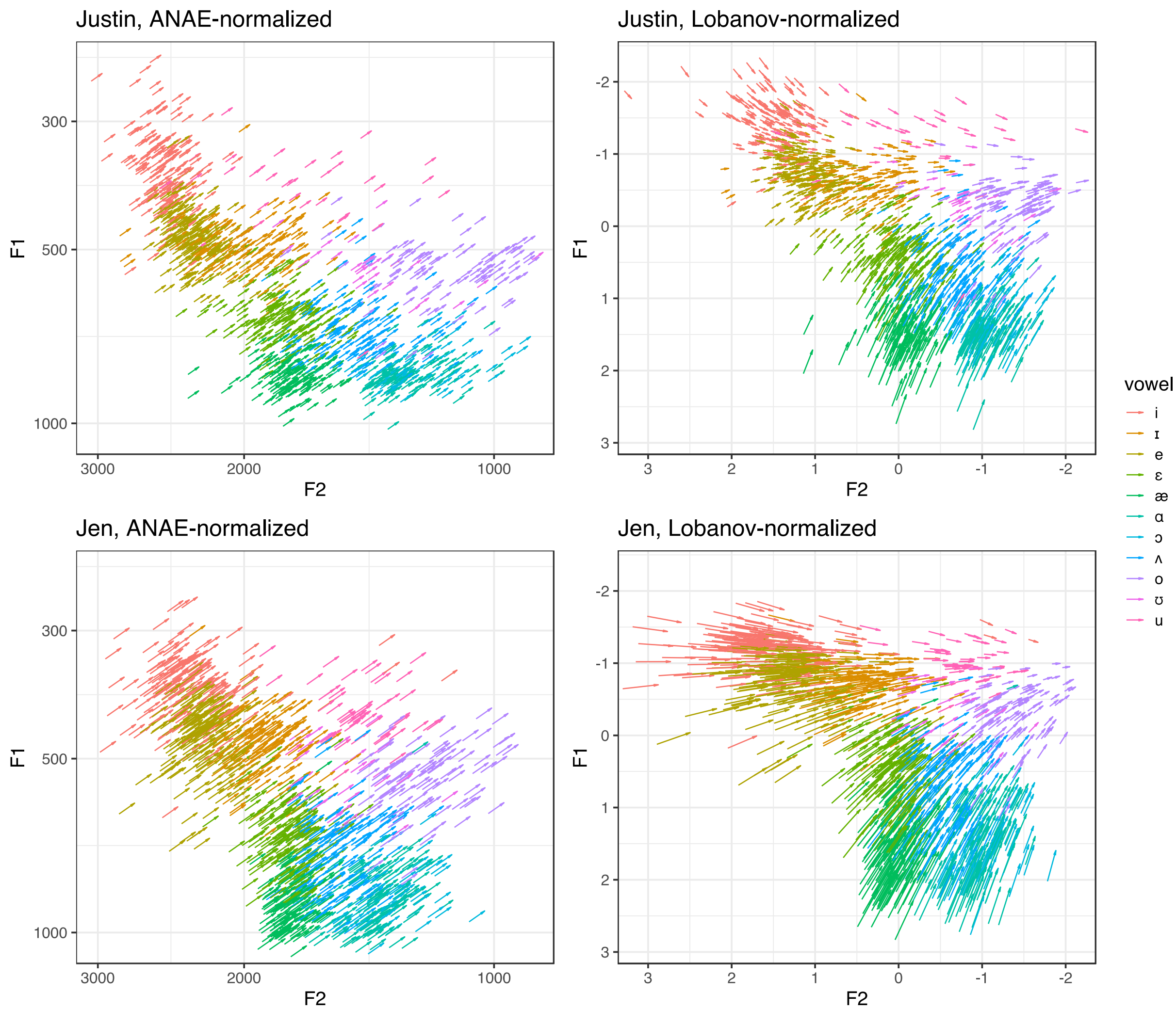

Figure 4: Vowel plots from two representative speakers showing how data is shifted when normalizing before subsetting the data to after subsetting, for two methods of normalization. Note that the panels on the left are on a logarithmic scale. Also note that data that was removed in the subsetting step was removed from the plot so that its effect on the remaining "good" data is clearer.

In the left two panels, which show the effect of the ANAE normalization procedure, all the arrows point in the same direction towards the upper right, and within speakers the arrow length is equal for all observations. Since there is just one parameter per speaker and it applies equally to both F1 and F2, the effect of normalization is the same on all vowels for both formants. This diagonal shift has to do with finding the average formant value for all tokens for both formants – essentially locating the center of the vowel space – with higher average formants resulting in a lower/fronter global position in the vowel space. A higher average is compensated by the normalization procedure by lowering all formants more so than a vowel space with a lower average.

In terms of order of operations, subsetting before normalizing causes the ANAE procedure to lower both formants for all observations, resulting in a higher/backer position in the vowel space. This means that the data that is removed in the subsetting step is, on average, higher/backer than the remaining data. It is difficult to determine precisely why this shift is consistently upward for all speakers since the tokens that are removed (unstressed vowels, stopwords, irrelevant allophones) are not typically thought of as being higher/backer than the data that is retained. But one possible reason is that because trajectories, particularly those of elsewhere allophones, are often U-shaped (Nearey 2013 Fig. 11; Stanley 2020 Fig 4.4), midpoints are, on average, lower in the vowel space than their trajectories. Therefore, only retaining midpoints (e.g. the data represented by the tail end of the arrows) brings the average formant value up, meaning a lower position in the F1-F2 space. This lower position means the normalization procedure must compensate by pulling all formants down, which results in a higher/backer position in the vowel space (e.g. the data represented by the arrowheads).

The right two panels of Figure 4 are more complicated. This time, instead of adjusting all points equally, the Lobonov normalization procedure has a varying types and amounts of influence, depending on the speaker and the formant. For the two speakers shown there, we see that higher formants lower and lower formants raise, resulting in a vertically compressed vowel space. All their vowels back somewhat, though the degree of backing is conditioned by F2, with front vowels backing more than back vowels. The pattern is similar but not identical across speakers; in Figure 4 we see that the two speakers' data was affected in slightly different ways. Indeed, each of the 53 speakers in this sample had a different instantiation of this pattern, suggesting that Lobonov exerts an influence that is less predictable than what the ANAE procedure does.

As far as the reason for the compression/backing pattern, this can be explained by thinking about the kind of data that is removed in the subsetting step and how that excluded data might affect the mean and dispersion of the vowel space. Because trajectories are U-shaped, vowel data that includes trajectories (e.g., the version of data represented by the arrowheads) will be, on average, higher in the F1-F2 space than vowel data that only includes midpoints (e.g. the version of the data represented by the tail end of the arrows). So to compensate for this difference in height, the normalization procedure brings the higher portion of the vowel space down, more than it would for the lower part of the vowel space. This is represented visually in Figure 4 by the lowered version of the vowel space in the right panel. The compression in the vowel space is the result of the reduced vowels typical of unstressed tokens and stopwords. The vowel space before subsetting is considered less disperse and has a lower standard deviation because of the dense clustering of tokens in the center of the vowel space. Again, the normalization procedure compensates for this by doing the opposite; in this case, the denser vowel space is spread out more.

Disregarding the difference between the normalization procedures for the moment,[3] Figure 4 should cause some concern when considering why these particular normalization procedures are used. For the ANAE procedure, all vowels for a speaker appear higher and backer if normalization happens after subsetting the data. Recall that the ANAE procedure is often used to see whether certain vowels are, on average, past the "thresholds" that have been derived from ANAE's maps. Subsetting the data before normalization may result in a speaker's /ɛ/ to be higher than the threshold of 650Hz, while subsetting the data after normalization could result in that speaker's /ɛ/ lower than the threshold. In Stanley (2020), I show that this is precisely what happened for 39 of the 53 speakers in this dataset. Since the order of operations used in the ANAE is not known, it is impossible to make an accurate comparison.

For the Lobonov procedure it is more complicated because normalizing after subsetting the data resulted in a vertically compressed vowel space and backer vowels. Artificial vowel space compression like this means vowels appear closer together in the normalized vowel space, and since the LBMS Index (among many other measures not mentioned in this study) quantifies distances between vowels in the Euclidean space, those distances will be shorter. In other words, normalizing after subsetting the data will typically result in a lower LBMS

[3] It is out of the scope of this paper to evaluate the pros and cons of these two normalization procedures, though it appears that the error introduced by the ANAE procedure is smaller and more predictable than that by the Lobonov method. I refer interested readers to Barreda & Nearey (2018) for a more in-depth treatment.

Index and those speakers will appear less shifted than if normalization happened before subsetting the data.

So which pipeline is preferred? Unlike the subsetting, removing outliers, and allophone classification steps, the reason for why normalization should occur where it does in Stanley's (2022) recommended order is not as clear. While it is true that swapping the normalization and subsetting steps produces different results, it is difficult to determine which one of those results is better. For now, I can only recommend that sociophonetic pipelines normalize before subsetting. The reason is that it may be important to include *all* data from any one speaker – including stopwords, irrelevant allophones, unstressed vowels, and vowel trajectories – so that the resulting normalized dataset represents that speaker's full vowel space in the normalized, perceptual vowel space.

One may argue that subsetting should happen before normalization so that comparisons across studies are more meaningful. It is true that a study that collects interview data will exclude many more tokens than another study that only elicits wordlist data. An argument can be made that by subsetting before normalization, what remains from the interview study more closely matches what is even collected in the wordlist study, and therefore the effect of normalization is more comparable between the two. However, a follow-up study of the interview data that happens to focus on something that was previously excluded, like vowel trajectories, will end up with a different input into the normalization step than the first study. There will be a difference between the normalized data in the first study and the normalized data in the second study. In other words, a single token of [æ] will have different normalized F1-F2 measurements between the two studies. This makes no sense since the token has not changed whatsoever. For this reason, I believe it is important to normalize before subsetting so that any analysis of that particular token of /æ/ will be based on the same normalized F1-F2 measurements.[4]

---

[4] This order may cause issues with imbalanced across speakers (some speakers producing far fewer vowel tokens while others producing many more) or within speakers (an over- or underrepresentation of a particular vowel, which pulls the mean in its direction) (cf. Brand et al. 2021). Again, it is outside the scope of this paper to evaluate different normalization procedures, but these problems may be indicative of an imperfect

7 Conclusions

The previous sections have highlighted just a few of the consequences that may occur when a dataset is processed with different orders of operations. Classifying vowels into allophones, stopwords, and unstressed vowels should happen first so that outlier detection can do a better job at finding truly deviant tokens. Outliers should be removed before normalization so that bad data does not have an effect on where in the normalized F1-F2 space the good data is located. Subsetting functions like isolating midpoints from their trajectories, removing uninteresting allophones, removing stopwords, and removing tokens with lexical stress do not interact and should be done more or less simultaneously and should happen after normalization occurs so that the vowels of interest end up in the same position in the normalized vowel space regardless of whether vowel trajectories, unstressed vowels, etc. are studied. The order of operations is important and should be included in methods sections and preregistrations[5] of studies (Nosek et al. 2018). This study has shown why the order is important and how to interpret the order of operations that are presented in future papers.

An important takeaway of this study is that the full range of vowel pronunciations for a given speaker should be collected, if possible. If a script is set up to only extract acoustic measurements from midpoints, or if a filter is set in place to skip over stopwords and uninteresting allophones, that essentially forces part of the subsetting step to occur first in the pipeline. The consequences of subsetting that data before normalizing is a shift in the vowel space in the direction of the arrows in Figure 4. If those filters were removed from that script and new acoustic measurements were extracted from the same audio, the normalized data will be in the opposite direction of the arrows shown in Figure 4, even though they represent the exact same vowel tokens.

Quantitative sociophonetic analysis has come a long way in the past several decades. An explosion of methods, techniques, functions, and procedures have been proposed and used. The dust has settled somewhat and some of these procedures have become standard because

---

procedure. Additional work that simulates these imbalances (e.g. Barreda & Nearey 2018) may be fruitful in the quest for a normalization procedure that truly models the human ear.

[5] I thank the anonymous reviewer who recommended that I add this suggestion.

they have been shown to be empirically superior in some way; we are becoming more equipped to properly analyze sociophonetic data. The dust has continued to settle enough for this study to point out the importance of order of operations. Future scholars should continue to scrutinize our linguistic methods so that we can better analyze sociolinguistic variation.

References

Barreda, Santiago & Terrance M. Nearey. 2018. A regression approach to vowel normalization for missing and unbalanced data. *The Journal of the Acoustical Society of America* 144(1). 500–520. https://doi.org/10.1121/1.5047742.

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.

Brand, James, Jen Hay, Lynn Clark, Kevin Watson & Márton Sóskuthy. 2021. Systematic co-variation of monophthongs across speakers of New Zealand English. *Journal of Phonetics* 88. 101096. https://doi.org/10.1016/j.wocn.2021.101096.

Kendall, Tyler & Charlie Farrington. 2021. The Corpus of Regional African American Language. Eugene, Oregon: The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal.

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

Labov, William, Ingrid Rosenfelder & 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.

Lobonov, B. M. 1971. Classification of Russian vowels spoken by different listeners. *The Journal of the Acoustical Society of America* 49. 606–608.

Nearey, Terrance M. 2013. Vowel inherent spectral change in the vowels of North American English. In Geoffrey Stewart Morrison & Peter F. Assmann (eds.), *Vowel inherent spectral change*, 49–85. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-14209-3_4.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven & David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11). 2600–2606. https://doi.org/10.1073/pnas.1708274114.

Nycz, Jennifer & 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.

Olsen, Rachel M., Michael L. Olsen, Joseph A. Stanley, Margaret E. L. Renwick & William A. Kretzschmar Jr. 2017. Methods for transcription and forced alignment of a legacy speech corpus. *Proceedings of Meetings on Acoustics* 30(1). 060001. https://doi.org/10.1121/2.0000559.

Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics* 26(1). 117–121. https://doi.org/doi:10.1214/aoms/1177728599.

R Core Team. 2021. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

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

Stanley, Joseph A. 2021. Pillai scores don’t change after normalization. https://joeystanley.com/blog/pillai-scores-dont-change-after-normalization. (2 November, 2021).

Stanley, Joseph A. 2022. Order of operations in sociophonetic analysis. In *University of Pennsylvania Working Papers in Linguistics*. 28(2), Article 17. Available at: https://repository.upenn.edu/pwpl/vol28/iss2/17.

Wickham, Hadley, Romain François, Lionel Henry & Kirill Müller. 2018. dplyr: A grammar of data manipulation. https://CRAN.R-project.org/package=dplyr.