18 Managing Sociophonetic Data in a Study of Regional Variation

Valerie Fridland and Tyler Kendall

1 Introduction

The data that are the focus of this data management use case stem from our Vowels in America (VIA) project, a multiyear, multipronged project funded by the National Science Foundation, the University of Nevada, Reno, and the University of Oregon.[1] The larger aim of the project was to examine the role of regionally based social and linguistic experience in shaping speakers' production and perception of vowel quality. For the project, we collected speech production data and administered a series of speech perception tests in a number of research sites in the Northern, Southern and Western United States. While the research design, in large part, was fully conceived before beginning the project in 2005, several aspects of the research, and a number of field sites, were modified or added in subsequent stages. In the following sections, we overview the project from its initial stages to its current form, highlighting, in particular, our data collection, data processing, data storage, and data sharing procedures. While the discussions here are based on our specific experiences running this particular project, we hope that our discussions of our data management practices are relevant to a wider range of projects in regional language variation, dialectology, and sociophonetics.

2 Data in regional variation and dialectological studies

Historically, the study of regional variation has been the focus of the field of dialectology. Dialectological projects have centered on the use of elicitation and questionnaire-based techniques that are designed to elicit information about regional forms (see Chambers & Trudgill 1998). Early dialectology, in an effort to chart dialect boundaries, often presented data in the form of a linguistic atlas or dictionary (e.g., Cassidy & Hall 1985–2013; Kurath 1949; McDavid & O'Cain 1980; Pederson, McDaniel, & Adams 1986–1993). However, more recent interests have moved toward gaining a better understanding of the social processes by which regional variation is achieved, often focusing on diffusion within social space as well as across geographic space. This change in research goals has been accompanied by a parallel shift in looking more deeply at gradient phonetic/phonological variables and in using sophisticated data collection, measurement, and mapping tools.

For example, recent work in dialectology has moved to applying and developing geospatial statistical methods (in what is often called dialectometry; Wieling & Nerbonne 2015) and has also moved increasingly into online data collection and data presentation (e.g., Vaux & Golder 2003; Grieve, Asnaghi, & Ruette 2013; Huang et al. 2016). Recent work in sociophonetics has also taken a large interest in regional variation and examined dialectological questions based on speech recordings and close analyses of regionally variable pronunciation features (see Kendall & Fridland 2021). Based on a sample of over seven hundred speakers from urban areas across the United States, the *Atlas of North American English* (Labov, Ash, & Boberg 2006) has led this movement toward new, large-scale regional studies based on speech recordings. Our work on the VIA project is, at its core, about documenting linguistic differences across regional varieties, but we also share an interest in capturing the social dynamicity of regional shifts and view our work as a sociophonetically driven dialectological undertaking. As such, for its methods and management tools, the VIA project combines the need for large-scale data collection techniques informed by dialectology and the computationally driven instrumental data measurement and presentation techniques informed by sociophonetics.

3 Overview of our project

Using work by Labov, Ash, and Boberg (2006), Eckert (2008), and Clarke, Elms, and Youssef (1995) as a jumping off point, the VIA project was designed to look in more detail at inter- and intraregional distinctions in US dialects, in terms of comparing production-based participation in regional shift processes across a large sample of speakers in and across each regional field site and in terms of examining correlations between the vowel shifts and understudied phonetic aspects such as duration and vowel inherent spectral change. In addition, though the underlying assumption in most sociophonetic work on regional dialects is that productive choices are made as a result of variation in receptive input, there has been limited work measuring regionally diverse vowel perception. Thus, we simultaneously administered vowel perception tasks in each regional study site, with the goal to examine how regional dialect experience mediates both production and perception, building on work by Clopper and Pisoni (2004), Evans and Iverson (2004, 2007), Hay, Warren, and Drager (2006), Niedzielski (1999), and Sumner and Samuel (2009) that suggests regional background influences speech processing. In summary, our research project had several aims:

1. To examine the extent to which there is variable participation in regional vowel shifts within and across regionally situated locales (e.g., Kendall & Fridland 2012; Fridland & Kendall 2012)
2. To determine whether degree of individuals' vowel shift participation influences vowel perception, and whether this varies by region (Kendall & Fridland 2012, 2016, 2017; Fridland & Kendall 2012, 2015, 2017)
3. To investigate how other aspects of production such as duration and vowel inherent spectral change tie into vowel shift participation (Fridland, Kendall, & Farrington 2014; Kendall & Fridland 2017; Farrington, Kendall, & Fridland 2018)
4. To examine whether the addition of talker-specific social information alters listeners' performance during perception tasks (Fridland & Kendall 2018)

We shift now to describe in greater detail our data collection, storage, and processing procedures.

4 Data collection procedures

Our first step was to determine where and how to collect data. Initially, we started with at least one university-based field site in each of three regions (per the dialect regions identified by Labov, Ash, & Boberg 2006), where we were on faculty or had strong ties with local faculty (Memphis, Tennessee; Oswego, New York; and Reno, Nevada). The goal was to collect production data from fifteen to twenty speakers in each field site, as well as have participants complete a vowel categorization task (detailed in section 4.2). To do this in remote locations, local fieldworkers were recruited within each region through our faculty associates, and they received a stipend for their work. These fieldworkers, all of whom were local students with some basic linguistics experience, were trained on the recording equipment and on study and recruitment procedures by one of the primary investigators (PIs) and then were also provided with a summary information packet prepared for the project by the PIs that reviewed the instructions so that speech sample collection would be performed uniformly across field sites. These instructions and packet reviewed the participant requirements (e.g., native English speakers, living in the field site from minimally age four, with at least one parent from that region) and the recording procedures (instructions for using and ordering speech prompt materials and a review on how to use the recording equipment), and contained written-out scripts for fieldworkers to use to read to participants to make sure all subjects were presented with identical instructions across field sites and to control the study information/background given to participants. They also provided consent materials to participants, who signed a consent form (in person or online, depending on which aspect of the study they were involved in). Fieldworkers were tasked with recruiting subjects via the friend of a friend technique (Milroy 1980) or were students recruited by local faculty members. In Reno, Nevada, and Oswego, New York (and later also in Eugene, Oregon), the study was run through a university subject pool, and students received credit for participation. Due to the long-distance nature of the data collection, we occasionally had difficulty getting people willing to complete both the production and perception tasks, which prompted us to apply for a small grant from University of Nevada, Reno, to help cover a small stipend ($40) for participants upon completion of both production and perception tasks, for those not receiving academic credits.

As our research proceeded over the next few years, we were able to expand the research to include participants in several other locations within each region. Over the course of the project, we ended up with speech production

and perception data from eight field sites in the United States: three within the Southern region (Memphis, Tennessee; Blacksburg, Virginia; and Raleigh, North Carolina), two within the Northern region (Oswego, New York, and Chicago, Illinois), and two within the West (Reno, Nevada, and Eugene, Oregon). We also obtained various California participants through our work in Oregon and Nevada and this enabled us to treat California as the third Western site (these participants were short-term residents of Oregon and Nevada, and we interpret these as best representative of their long-term hometown location). For the main vowel identification study, we were able to collect data from over 650 participants from across the three larger US regional dialect areas. A subset of over 90 of these participants also contributed production data (a reading passage and word list; see section 4.1), allowing us to compare a subset of our participants' production to their perception of target vowel pairs ($n=80$). Finally, for a follow-up study examining how social information affects vowel identification (Fridland & Kendall 2018), we also recruited participants for the perception part of the study from additional field sites in Alabama and Tennessee.

4.1 Production data

The research involved the collection of acoustic data elicited via a reading passage and word list, both of which were composed by the first author (Fridland 1998, 1999, 2001) to produce vowel variants in a variety of consonantal contexts, with a particular interest in Southern vowel features (one of the main regional dialects of focus initially). The reading passage and word list are provided in the appendix. As the project evolved, other contexts of interest emerged (e.g., lax front vowels preceding velar consonants, low back vowel merger), and a secondary word list, also included in the appendix, was created to provide additional tokens for those contexts. We chose to add a second word list rather than modify the first so that all participants would be presented the materials in the identical order, which helps to limit possible coarticulatory or prosodic differences across speakers. The reading passage and word list were printed on paper, double-spaced, with the word list printed down two columns per page.

The recordings were collected with a Tascam digital recorder and a Shure WH30XLR head-mounted microphone or with a Marantz digital recorder and a Shure SM93 lavalier microphone by a local fieldworker in a quiet setting or, where possible, in a speech lab. All speakers read the same reading passage and word list(s) with the same instructions (to read the passage over before recitation and to pause briefly between each word list recitation).

4.2 Perception data

As the project was designed to investigate the link between use of regional vowel variants and the perception of these variants, our collection of speech production data was paired with a vowel identification task. The web-based vowel identification task was designed by first author Fridland and constructed by a computer science PhD student, Sohei Okamoto, who had taken linguistics coursework with Fridland at University of Nevada, Reno. The online task was designed to first present the online consent form, give a sound-level check (e.g., clicking a button to be sure the participant's volume setting was ideal), provide written instructions and request demographic information, and, finally, provide a practice vowel identification task (using a vowel not tested in the actual task). After this preliminary material was presented, participants were reminded one last time of the instructions (in written format), then presented with the actual perception experiment. This involved presenting for each stimulus token (randomized between and within blocks) the labels for the two continuum end points, such as *bait* and *bet* (each selectable to record a participant's response), and a PLAY button, which provided the aural stimuli when pressed. Once designed, it was run on a customized University of Nevada web server managed collaboratively by Okamoto and the university's information technology team. The use of a local, customized server was partly to provide full control of functionality but also avoided any confidential data being managed on a third-party server. It also allowed for better delivery of aural stimuli than commercial software allowed for at the time of development. The authors recognize that many projects are unable to develop a customized server for speech perception experiments, but note that the number of publicly available platforms, both commercially (such as Qualtrics) and via open source (such as PsyToolkit), are growing rapidly. The custom work that was done to develop our perception experiment platform is likely not needed today.

Those interested in being part of the study were provided the study URL, a log-in, and password and were able to take the study from any Internet-accessible computer. Participants, who were asked to use earphones, first answered a series of demographic questions about

age, hearing ability, gender, and occupational and residence history. Participants were also prompted to fill in a text box with their first name, state abbreviation (which provided us quick field site information), and numerical street address (e.g., ValerieTN2760) as a means of providing each participant with a unique code so that their identifying information could be separately stored. The same code was used for their recorded production data, so they could be easily matched. As mentioned previously, in several field sites (Oswego, New York; Reno, Nevada; and Eugene, Oregon), we were able to collect a large amount of data on speech perception by recruiting through university subject pools. While this made geographically diverse large-scale data collection much easier, it did limit the age range of participants to the eighteen to thirty bracket for the most part.

The vowel identification stimuli were based on the synthesis of natural speech data provided by an adult male speaker from the Western United States with unmarked dialectal vocalic features per Clopper and Pisoni (2004). The speaker was recorded reading monosyllabic word pairs selected for vowel class and matched for consonantal environment (e.g., *date/debt, beat/bit*). The vowel identification test was designed by synthesizing consonant-vowel-consonant pairs for each vowel pair drawing from the labial, alveolar, and velar points of articulation, using the speaker's natural category means as a guide. Word pairs (minimal pairs) were selected for vowel categories with high dialectal variability across US dialects (/i/~/ɪ/, /e/~/ɛ/, /æ/~/ɑ/, /u/~/ɪ/, and /o/~/ʌ/). A pilot vowel categorization test using three consonantal environments suggested that bilabial and alveolar environments showed the greatest contrast in perception, and those two were selected for use in the final vowel identification test design. For each vowel pair studied, the continuum range was determined based on the sample speaker's production values for each of the two selected vowel categories. Based on these end points, the intermediate stimuli were created by Bartek Plichta through Akustyk, a vowel synthesis program he created. Using the web-based design just described, the experiment was set up so that one stimulus was presented per trial, and then participants were prompted to make a decision about what word they heard from two options (e.g., *bait* or *bet*). Each step in each vowel continuum was heard four times randomized over the course of the study. Generally, participants were given no background on the speaker, other than what they could interpret from the signal source alone (e.g., male adult). For those participants contributing both speech production and speech perception data, they were instructed to do the online task at least one day prior to being recorded for the speech production study to minimize the possibility that participants' speech productions would be influenced by the perception task.

The software used to create the stimuli was available publicly as a part of Plichta's Akustyk plug-in for Praat until 2014 (Plichta 2004–2014), but the website and software are no longer maintained. While the stimulus synthesis software was made open source by the designer, we were not involved in the life of the software after our stimuli were created and, like other users of Akustyk, no longer have access to it. The downside to partnerships such as this is that the researcher is, in such an arrangement, limited in the ability to control what happens to software for the long term. Yet, the benefits and time savings of not duplicating software or scripts that already exist and utilizing resources created by researchers who are more experienced in design aspects of those tools often outweighs the risk of this loss of control of the software.

5 Data storage

Data storage efforts for this project have changed over the years. In short, as the project has evolved in its various ways, our data management practices have changed as well. For the purposes of this data management use case, we especially want to acknowledge that these changes have not always been improvements but rather reflect the practicalities of a large-scale project conducted across many field sites and universities, something we address in this section.

5.1 Production data

By *speech production data*, we mean the actual audio files of recordings (in waveform audio file format). We will return to discussing derived data sets, such as acoustic measurements. Speech production recordings were initially uploaded from the digital recorders to the PIs' computers for analysis and storage, as was the typical practice at the time we began the project, and followed the guidelines for participant confidentiality set forth in the consent documents. Subsequently, speech

recordings were moved to the web server at University of Nevada that hosted the perception data, which acted as a centralized data repository for the project. These recordings were accessible to project staff through a simple, password-protected download interface. However, in 2011, we modified our human subjects protocol to allow us to store the data in a more formalized data archive environment instead, in line with field-based best practice to increase accessibility and data longevity. For more detailed discussion of data archiving and its benefits, we refer readers to Andreassen (chapter 7, this volume).

Based on earlier work developing the Sociolinguistic Archive and Analysis Project (SLAAP; Kendall 2007), the second author had developed a web-based archive for laboratory-based speech recordings at Northwestern University, the Online Speech/Corpora Archive and Analysis Resource (OSCAAR; Kendall 2010). Unlike SLAAP, whose architecture was designed around sociolinguistic interview recordings (relatively long recordings of multiparty conversational interactions), OSCAAR was designed around laboratory-based production recording collections, where recording files typically are quite short utterances (such a single sentence or word production) by a single speaker, with very many files per speaker, and where speech recordings relate to specific elicitation materials, such as word lists and reading passages. In short, OSCAAR, which housed other laboratory recordings, such as the Wildcat Corpus (Van Engen et al. 2010), was well suited to house our speech recordings. So, the recordings were stored centrally in OSCAAR for a few years. In 2013, however, OSCAAR began to be redesigned and the VIA and OSCAAR teams decided it no longer made sense for these data to reside at Northwestern University. (OSCAAR has since been further redesigned and renamed SpeechBox; it is available at https://speechbox.linguistics.northwestern.edu .) It was also the case that we had continued collecting perception and production data in the intervening years and the formal, archived collection in OSCAAR no longer represented our entire data set. The files were moved to a networked desktop server in the Language Variation and Computation Laboratory, directed by Kendall, at the University of Oregon. This computer was backed up regularly and also used to store and manage files for the Corpus of Regional African American Language (see Kendall & Farrington, chapter 14, this volume) and other projects. However, the computer's system was never designed to be a full-fledged archive, and we admit the entire data collection remained somewhat underorganized for several years.

In the summer of 2017, the team undertook the project of organizing and reformatting all of the data for the project. The data now reside in SLAAP, which provides a centralized, password-protected interface to the production recordings and metadata. While SLAAP remains designed primarily around the storage and analysis of sociolinguistic interview recordings, and so is less well suited to these data than OSCAAR, its password protected web-based interface and storage capabilities provide a more robust long-term home for the VIA data. Per the consent arrangement with subjects, the actual recordings are not publicly available, but only accessible to the PIs and those working directly on the project.

5.2 Perception data

By *speech perception data,* we mean the responses obtained from the individual participants in the perception experiments. These basically represent fixed-choice responses summarized for each stimuli step (i.e., the percentage heard as one of the two choices for the four trials for each step), along with a range of metadata (including demographic information) collected from each participant. The web server software automatically generates and stores these data for each participant. It also calculates crossover points (where a participant goes from hearing mostly one variant to mostly the other variant) for each continuum.

These speech perception data have continued to reside on the web server developed by Okamoto at University of Nevada. The server has a password-protected customized interface to the perception data that lets us query perception data for individual participants or groups of participants and even generates simple plots and summary statistics of the perception data. It also allows for the exportation of query results. In addition to the server, the authors maintain local copies of the perception data. Local copies of the perception data have also been augmented over time based on our research interests. So, for example, around 580 participants were geocoded (with latitude and longitude of their self-reported hometown) for our dialectometric analysis of spatial patterns in perception (Kendall & Fridland 2016). Copies of the perception data are also stored on SLAAP along with the production data, although the primary

storage for the perception data remains the original site through which they were collected.

6 Data processing

For analysis, the data are processed in various ways.

6.1 Production data

The speech recordings have been processed several times using different techniques. For our main set of studies (Fridland & Kendall 2012, 2015; Fridland et al. 2014; Kendall & Fridland 2012, 2017), vowel measurements were taken by hand using a customized set of Praat scripts (e.g., http://lingtools.uoregon.edu/scripts/vowel_capture_aug09.praat). These were done over a period of years, and actual measurements were done by a few different analysts, working under the guidance of the authors. Craig Fickle, an undergraduate student at the University of Oregon and the main Research Experience for Undergraduates research assistant under our National Science Foundation grants, conducted the majority of hand-measurements, with second author Kendall checking and correcting his data. These measurements focused on a subset of the words in our word lists and reading passages, with an average of 133 tokens measured per speaker and an average of 10 tokens measured for each major vowel category per speaker. Most of our hand-measurements focused on taking two measurement points per vowel, one at one-third of the vowel's duration and one at two-thirds. Figure 18.1 displays a vowel plot for one speaker, exemplifying the set of vowel measurements taken for each speaker and demonstrating a common analytic and display technique for vowel data like these (see Thomas 2011). For some recent work (Farrington, Kendall, & Fridland 2018; Gunter, Vaughn, & Kendall 2020), measurements were taken using a new Praat script to support an interest in examining dynamic properties of vowels in greater detail. While the original hand-measurements delimited vowels

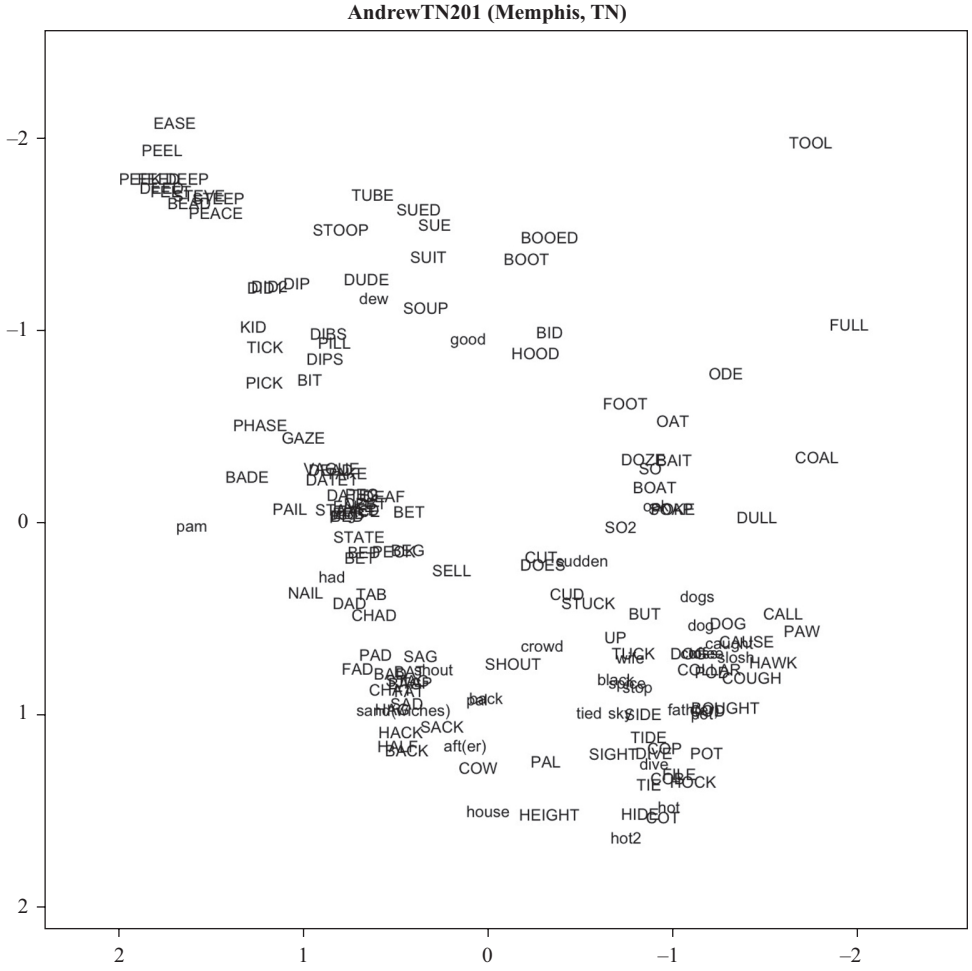


Figure 18.1
Example vowel plot showing one-third measurement point for 155 vowel tokens for participant AndrewTN201 (male from Memphis, Tennessee, between thirty-one and forty years old). Data are Lobanov normalized (Kendall & Thomas 2010).

manually, the data have been forced aligned using the Montreal Forced Aligner (McAuliffe et al. 2018) for these more recent interests. Recent approaches to vowel measurement, especially for US English, have increasingly relied on automated techniques, such as the use of the University of Pennsylvania FAVE (Forced Alignment and Vowel Extraction) Program (Rosenfelder et al. 2014). Future work with the VIA production data may follow this recent practice and use a tool such as FAVE to analyze the entirety of our collected data.

Kendall (2013) used the data stored in OSCAAR and tools built in to OSCAAR to examine speech rate variation in the reading passages. Kendall and Vaughn (2015, 2020) have used some of the speech recordings and hand-measured vowel data to explore a bootstrap measurement simulation process.

6.2 Perception data

Once retrieved from the server, perception data are processed in several ways. Many of our analyses (e.g. Kendall & Fridland 2012, 2017) examine the data through logistic (mixed-effect) regressions of the individual trials. For these analyses, the percentages stored on the server (i.e., an aggregate measure over the four instances each listener was presented for each stimulus step) were converted back to individual trial-level data points, with one categorical judgment per data row. For example, if a participant heard one step of the /ɑ/-/æ/ stimuli as 75% /æ/, this would be converted to three rows of "participant heard /æ/" and one row of "participant heard /ɑ/." We then typically presented this derived data using a vowel identification function, which shows a participant's averaged vowel identification along each step between the end points of the continuum. An example showing vowel identification functions for three participants for the *bait~bet* continuum is shown in figure 18.2. This allows us to easily visualize individual or aggregate groups of speakers in comparison to one another. Though we have not found it to be a particularly useful metric for most of our subsequent analyses, we have also compared speakers on the basis of crossover points, or the point along the continuum at which vowel perception shifts from hearing one category to the next over 50% of the time (e.g., Kendall & Fridland 2016).

6.3 Analyzing perception and production

A large part of our initial interest was to examine the relationships between perception and production at both individual and community levels. This has required combining and processing our data in various ways. And more than any other aspect of our analyses, this has been the place we have spent the most effort on exploring our data. At its most basic, combining our perception and production data involves generating new independent variables for our perception data, where information from the production data is summarized in some way and then related to the perception data. This can involve, for instance, including a participant's Euclidean distance between /e/ and /ɛ/ (a summary measure generated for each speaker—we use this as an example as we have found it to be quite useful; see Fridland & Kendall 2012; Gunter, Vaughn, & Kendall 2020; Kendall & Fridland 2012) as a new potential predictor for each perception data point. However, the number of potential production measures that can be tested in relation to perception measures is boundless, and relationships between perception and production need not be linear (see Fridland & Kendall 2012). For the most part, our approach

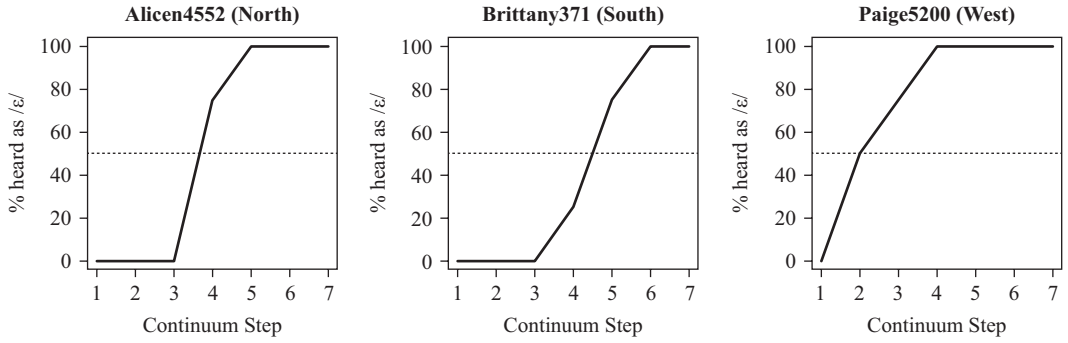


Figure 18.2
Identification functions for three participants for *bait~bet*.

to delimiting how to quantify production variables (as Euclidean distance, along a scale of F1 or F2 measures, in quartiles) has varied depending on the vowel in question and the type of vowel shift involved (backing, merger, relative movement). Much of this work is a process of trial and error, although trials are typically based in hypotheses about potential relationships between production and perception (such as that the distance between a participant's own /e/ and /ɛ/ vowel categories, in production, might relate to how that participant perceives the boundary between /e/ and /ɛ/).

7 Data sharing

Because we did not ask participants' permission to store their data in an open access format or to share it publicly, we did not set up the data archive to be publicly accessible. Thus, the audio recordings stored in SLAAP remain password-protected and not available publicly. However, we are able to share data in an anonymous format with other researchers on a permission-only basis. Likewise, about midway through the project we also changed the consent form on the online perception study to provide consent to share data more widely. Within the limitations imposed by the initial human subjects protocol, the PIs are hoping to make some of the anonymized metadata available more broadly and are currently in discussion about how best to approach this aim. Though we were not able to share our primary data publicly, our project has made available a number of products beyond publications. First, specialized software for vowel synthesis (as part of Akustyk) and phonetic data analysis and presentation scripts/code (e.g., three-dimensional plotting software; Fridland & Kendall 2009; http://lingtools.uoregon.edu/tools/durplot3d/) have been created for the project and other types of programs may be developed as we discover need. Second, speech samples and a database of perception task data have been previously and will continue to be generated during the course of the project, and, as just discussed, we plan to make some aspects of this data accessible more widely in an anonymized format. Finally, we also host public-facing information about the VIA project on Fridland's university webpage (https://packpages.unr.edu/fridland/), which seeks to make our research and findings available more widely to a non-specialist audience.

8 Lessons learned

The VIA project discussed here has been a fairly large-scale and long-term collaborative project. Our primary funding support ended in 2016, but nonetheless we continue to work with these data and the overall project continues to expand and grow. Our perception data continue to provide fuel for new insights into regional patterns of speech perception and linkages between perception and production, and our production data, in particular, form the basis for ongoing work on regional vowel patterns (Farrington, Kendall, & Fridland 2018; Fridland & Kendall 2019), as stimuli for new perception experiments (Gunter, Vaughn, & Kendall 2020), and as the basis for work on vowel analysis methods (Kendall & Vaughn 2015, 2020). We think it is fair to say that neither of the authors realized we were beginning such an expansive collaboration when we started to work together about a decade ago. As a part of this breadth and expansiveness, a number of issues have come up, several of which have been mentioned, and we believe this all provides some lessons for future data management and use for linguistic researchers. Before closing, we survey some specific points along these lines.

First, our project has relied on having a lot of "cooks in the kitchen." The breadth of our regional data collection, over 650 perception participants from across the United States, was only possible by partnering with fieldworkers local to our research field sites. And technical aspects of the project were facilitated by other collaborators. This was crucial for our project but also led to some problems. For instance, while the project's initial stimuli construction supported the development of a synthesis tool shared with the larger research community, it also meant that the PIs did have direct control or ownership over the software. When the Akustyk plug-in for Praat and its supporting website went offline, a number of researchers were dismayed, but it also meant that we too are now unable to generate new stimuli using the same methods originally used. To some degree this is not a huge issue because more tools and techniques are available now for vowel synthesis than were when we started this project, but it still represents an issue for possible replicability and it means we have been limited to expansions of our project that do not require small adjustments to our stimuli. Similarly, we have been extremely lucky and grateful to have the support of Sohei

Okamoto. Okamoto continued to help us by supporting the web server long after he completed his PhD dissertation and by providing his expertise on an ongoing basis. In part, this has been made possible because he continued at University of Nevada, Reno, in a post-doctorate position, and, also, because of his genuine interest in the project and good relationship with Fridland. However, without him we would be somewhat hindered in our ability to modify the database architecture that he set up to support the perception study and its database.

Second, while having a lot of research assistants has aided our data collection and analysis greatly, it has also created a sometimes chaotic situation for keeping up with how and where the data were stored and in what format. Each fieldworker was responsible for uploading the data to a shared server, but, depending on project and fieldworker, these were often in different folders and organized in different ways. In particular, our file-naming conventions have been irregular and not well organized. (For best practice in file naming and version control, see Mattern, chapter 5, this volume). Each fieldworker had different file-naming conventions and even differed in whether they recorded all recordings as one file or as multiple files. As well, it was often difficult to track the level of processing each sound file had been through (e.g., just raw data, Praat TextGrids, FAVE output), as we had difficulty ensuring all recordings were documented on the same master spreadsheet (several of which had been created over the course of the project, adding to the confusion). In some cases, this led to the need, at the data analysis phase of each project, to recreate the wheel. Recently, we have tried to make the data organization more transparent and clearer by having one research assistant go through and document the data and relabel everything with a more uniform practice. Our hope is that housing the data in the SLAAP archive will mean it remains more organized and centralized moving forward. However, when the data were put in OSCAAR originally, we did not expect to have to move them again. Overall, much wasted time and effort could have been saved had we been more calculated and meticulous in our data management practices from the outset.

9 Conclusion

As our ongoing project has evolved over the last ten years, our data collection, storage, and processing methods have needed to evolve as well. Partly this reflected the changing scope of our project and changing opportunities, but it also reflected a changing set of standards for how we might best manage data in linguistics. Best practices and standards have changed rapidly over the last ten years, and we have attempted to modify our procedures and data management as much as we could given the constraints of our original study design and human subjects protections. In retrospect, the main changes we would have made, based on our experience, would have been to have been more principled in our data labeling and archiving at the outset and to have set up our initial institutional review board protocol to allow for more public use of the data. However, we have created a number of public use and open-source materials during this work and also adapted our methods and data management as much as possible to allow our data to be processed and shared more effectively. We believe we've learned important lessons that inform our future research through this project and we hope this data management use case helps other researchers plan and manage their own data collections.

Appendix

Elicitation materials for production data.

Reading passage

Some mornings in the summertime, when the sky is fair and the lawn covered in dew, the good Duke Post and his wife Peg walk down to the brook by their house. There, beside the trees, is their favorite place to sit, talk and sip coffee. Her father, Don, and his dog, Bookie, often stop by to chat while their children, Betty and Kate, toss off their shoes and leap headfirst into the deep brook. It makes Peg feel like a kid again to watch them dive, shout and slosh around in the water and swing off the old black tire tied to the oak tree.

One hot, hazy, dull afternoon, she gave a call to their friends Pam and Ben Powder, inviting them over for supper. On the way, their truck got stuck in the mud and they showed up an hour late, for which they caught a good deal of teasing. But soon the crowd was having fun and the good hosts put out tunafish sandwiches, hot dogs, a big pot of bean soup and beer bread. When they were done eating, it was a sin that no one had saved room for Peg's tasty spice cake that was yet to come.

Word list 1

Dutch, Took, Cop, Dude, Shout, Bet, Collar, Cough, Town, Dock, Tide, Foot, Up, Bit, Tuck, Pail, Soap, Peel, Ghost, Dish, Deaf, Cause, Nail, Bade, Dip, Soup, Dive, Full, Seat, Beg, Bait, Poke, Same, Hock, Cow, Date, Pill, Kid, Call, Tool, Pot, Gate, Tick, Sue, Debt, Bid, Hood, Type, Gave, Boot, Tie, Cot, Dad, Bought, Paw, File, Dog, Did, Coal, Deep, Hawk, Bed, Sell, Sad, Does, Half, Deed, Peek, Dull, Pal, Bead, Date, Boat, Pod, Take, Booed, Pad, Doze, Did, But, So

Word list 2

Stayed, Fad, Vogue, Hug, Bad, Hack, Side, Dose, Stack, Tug, State, Ebb, So, Stuck, Tube, Feed, Bat, Hag, Sued, Sack, Suit, Height, Cut, Dead, Cod, Diss, Cud, Fatigue, Bet, Pick, Peace, Stoop, Bite, Pig, Hook, Bag, Sag, Dip, East, Back, Tab, Boutique, Bead, Sight, Feet, Phase, Tap, Ode, Steep, Dog, Dibs, Cob, Dips, Stag, Ease, Fake, Tease, Oat, Gaze, Step, Face, Chat, Seed, Fat, Peg, Peck, Steve, Hide, Chad, Vague, Pet, Regular, Pleasure, Bagel, Heck, Botany, Leg, Pagan, Measure, Lag, Egg, Shoot, Keg, Leggings, Treasure

Note

1. We are grateful to the National Science Foundation (grants BCS-0518264, BCS-1123460, and BCS-1122950), as well as the University of Nevada, Reno, and the University of Oregon for supporting the work described in this chapter. We especially thank Sohei Okamoto, Craig Fickle, Charlie Farrington, and Kaylynn Gunter for support with various aspects of the project. In addition, we want to acknowledge Bartek Plichta for his work helping to develop the stimuli, and the numerous faculty and students who helped recruit participants and collect data at field sites across the United States.

References

Cassidy, Frederic, and Joan Hall. 1985–2013. *The Dictionary of American Regional English*. Cambridge: Harvard University Press

Chambers, J. K., and Peter Trudgill. 1998. *Dialectology*. 2nd ed. Cambridge: Cambridge University Press.

Clarke, Sandra, Ford Elms, and Amani Youssef. 1995. The third dialect of English: Some Canadian evidence. *Language Variation and Change* 7:209–228.

Clopper, Cynthia, and David Pisoni. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics* 32 (1): 111–140.

Eckert, Penelope. 2008. Where do ethnolects stop? *International Journal of Bilingualism* 12 (1–2): 25–42.

Evans, Bronwen G., and Paul Iverson. 2004. Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *Journal of the Acoustical Society of America* 115 (1): 352–361.

Evans, Bronwen G., and Paul Iverson. 2007. Plasticity in vowel perception and production: A study of accent change in young adults. *Journal of the Acoustical Society of America* 121 (6): 3814–3826.

Farrington, Charlie, Tyler Kendall, and Valerie Fridland. 2018. Vowel dynamics in the Southern Vowel Shift. *American Speech* 93 (2): 186–222.

Fridland, Valerie. 1998. The Southern Vowel Shift: Linguistic and social factors. PhD dissertation, Michigan State University.

Fridland, Valerie. 1999. The Southern Vowel Shift in Memphis, TN. *Language Variation and Change* 11 (3): 267–285.

Fridland, Valerie. 2001. Social factors in the Southern Shift: Gender, age and class. *Journal of Sociolinguistics* 5 (2): 233–253.

Fridland, Valerie, and Tyler Kendall. 2009. Mapping production and perception in regional vowel shifts: The effects of vowel duration and formant trajectories. Presentation given at the New Ways of Analyzing Variation Conference 38, University of Ottawa, Ottawa, Canada, October 22–25.

Fridland, Valerie, and Tyler Kendall. 2012. Exploring the relationship between production and perception in the mid front vowels of U.S. English. *Lingua* 122 (7): 779–793.

Fridland, Valerie, and Tyler Kendall. 2015. Within-region diversity in the Southern Vowel Shift: Production and perception. In *Proceedings of the International Congress on Phonetics (ICPhS) 2015*. Glasgow: University of Glasgow.

Fridland, Valerie, and Tyler Kendall. 2017. Speech in the Silver State. In *Speech in the Western States*. Volume 2: *The Mountain West*, ed. Valerie Fridland, Betsy Evans, Alicia Wassink, and Tyler Kendall, 139–164. Durham, NC: Duke University Press.

Fridland, Valerie, and Tyler Kendall. 2018. Regional identity and listener perception. In *Language Regard: Methods, Variation, and Change*, ed. Betsy Evans, Erica Benson, and James Stanford, 132–152. Cambridge: Cambridge University Press.

Fridland, Valerie, and Tyler Kendall. 2019. On the uniformity of the Low-Back-Merger Shift in the U.S. West and beyond. In *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and short front vowel shifts across North America*, ed. Kara Becker, 100–119. Durham, NC: Duke University Press.

Fridland, Valerie, Tyler Kendall, and Charlie Farrington. 2014. Durational and spectral differences in American English vowels: Dialect variation within and across regions. *Journal of the Acoustical Society of America* 136 (1): 341–349.

Grieve, Jack, Costanza Asnaghi, and Tom Ruette. 2013. Site-restricted web searches for data collection in regional dialectology. *American Speech* 88:413–440.

Gunter, Kaylynn, Charlotte Vaughn, and Tyler Kendall. 2020. Perceiving Southernness: Vowel categories and acoustic cues in Southernness ratings. *Journal of the Acoustical Society of America* 147 (1): 643–656.

Hay, Jennifer, Paul Warren, and Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34:458–484.

Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59:244–255.

Kendall, Tyler. 2007. The North Carolina sociolinguistic archive and analysis project: Empowering the sociolinguistic archive. *Penn Working Papers in Linguistics* 13 (2): 15–26.

Kendall, Tyler. 2010. Developing web interfaces to spoken language data collections. In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science* 1.2. Chicago: University of Chicago. doi:10.6082/M1BK19HM.

Kendall, Tyler. 2013. *Speech Rate, Pause, and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. Basingstoke, UK: Palgrave Macmillan.

Kendall, Tyler, and Valerie Fridland. 2012. Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift. *Journal of Phonetics* 40 (2): 289–306.

Kendall, Tyler, and Valerie Fridland. 2016. Mapping the perception of linguistic form: Dialectometry with perception data. In *The Future of Dialects*, ed. Marie-Hélène Côté, Remco Knooihuizen, and John Nerbonne, 173–194. Berlin: Language Science Press.

Kendall, Tyler, and Valerie Fridland. 2017. Regional relationships among the low vowels of U.S. English: Evidence from production and perception. *Language Variation and Change* 29 (2): 245–271.

Kendall, Tyler, and Valerie Fridland. 2021. *Sociophonetics*. Cambridge: Cambridge University Press.

Kendall, Tyler, and Erik R. Thomas. 2010. *Vowels: Vowel Manipulation, Normalization, and Plotting in R*. R package, version 1.1. Software Resource. http://lingtools.uoregon.edu/norm/.

Kendall, Tyler, and Charlotte Vaughn. 2015. Measurement variability in vowel formant estimation: A simulation experiment. In *Proceedings of the International Congress on Phonetics (ICPhS) 2015*. Glasgow: University of Glasgow.

Kendall, Tyler, and Charlotte Vaughn. 2020. Exploring vowel formant estimation through simulation-based techniques. *Linguistic Vanguard* 6 (s1): 1–13.

Kurath, Hans. 1949. *A Word Geography of the Eastern United States*. Ann Arbor: University of Michigan Press.

Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2018. *Montreal Forced Aligner*. Version 1.0. http://montrealcorpustools.github.io/Montreal-Forced-Aligner/.

McDavid, Raven I., Jr., and Raymond O'Cain. 1980. *Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.

Milroy, Leslie. 1980. *Language and Social Networks*. Baltimore, MD: University Park Press.

Niedzielski, Nancy. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18:62–85.

Pederson, Lee, Susan L. McDaniel, and Carol M. Adams, eds. 1986–1993. *Linguistic Atlas of the Gulf States*. 7 vols. Athens: University of Georgia Press.

Plichta, Bartłomiej. 2004–2014. *Akustyk*. Plug-in for Praat software.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. *FAVE (Forced Alignment and Vowel Extraction)*. Program suite version 1.2.2. https://github.com/JoFrhwld/FAVE.

Sumner, Meghan, and Arthur Samuel. 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language* 60:487–501.

Thomas, Erik R. 2011. *Sociophonetics: An Introduction*. Basingstoke, UK: Palgrave Macmillan.

Van Engen, Kristin, Melissa Baese-Berk, Rachel Baker, Arim Choi, Midam Kim, and Ann R. Bradlow. 2010. The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53 (4), 510–540.

Vaux, Bert, and Scott Golder. 2003. *The Harvard Dialect Survey*. Cambridge: Harvard University Linguistics Department.

Wieling, Martin, and John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1:243–264.

This is a section of doi:10.7551/mitpress/12200.001.0001

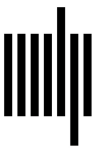
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363