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
17 Managing Legacy Data in a Sociophonetic Study of Vowel Variation and Change

James Grama

1 Introduction: The focus of this data management use case

This data management use case provides a description of the workflow of a sociophonetic study of language variation and change using legacy data (Grama 2015). This study investigated change in the vowel system of Pidgin (known to linguists as Hawai‘i Creole), an English-lexified creole spoken in Hawai‘i. In that study, two existing corpora were used to conduct a trend study that compared Pidgin speakers at two time points: the 1970s and the 2000s. Analysis was conducted on acoustic vowel measurements taken from speech elicited using sociolinguistic-style interviews (cf. Labov 1972). This use case serves as a meta-analysis of the methods applied in the Grama study and is intended for researchers interested in using naturalistic legacy data to identify longitudinal change. Throughout the chapter, methodological considerations are made that concern sociophonetic studies, the management of legacy data (especially when that data was not originally designed to address sociophonetic research questions), and further issues that are relevant to variationist research.

The chapter is organized as follows: first, the language setting for the current study is described, along with a discussion of my positionality to Pidgin (section 2). Then, I discuss the way archived interviews were selected for analysis, as well as characteristics of the archived corpora used in Grama (2015) (section 3). This is followed by a description of the way vowel data was transcribed (section 4.1), force aligned (section 4.2), manually checked (section 4.3), extracted (section 4.4), cleaned (section 4.5), and normalized (section 4.6). Issues concerning data transparency (cf. Gawne & Styles, chapter 2, this volume), figure interpretability (section 5.1), and storing and sharing of data (section 5.2) are then briefly discussed. Throughout, the Grama study is evaluated in the context of best practices in sociophonetics, and deviations from optimal procedure are noted, where relevant.

I now turn to a brief discussion of the social setting of this study, as the social landscape of any variety is paramount to interpreting the results and understanding why the study took the shape it did.

2 A brief history of Hawai‘i and the development of Pidgin

Hawai‘i’s complex history of contact began from its discovery. The islands were originally settled by Polynesian seafarers between 1190 and 1293 CE (Wilmshurst et al. 2011:1816; Walworth 2014:258), and the following two centuries saw intercultural movement among neighboring Polynesian cultures (Collerson & Weisler 2007). It is generally accepted that sustained contact between Hawai‘i and other eastern Polynesians declined sharply in the fifteenth century (Drager 2012b:62). In 1778, Hawai‘i was irrevocably altered by James Cook’s arrival. European contact opened the floodgates for foreign influx at an unprecedented scale; the islands were quickly exploited for their sandalwood and used both as a strategic launching point during the whaling trade and a stopover point in the fur trade of the early 1800s (Reinecke 1969:24). Foreign presence was debilitating for Hawaiians and their language (‘ōlelo Hawai‘i).

The research discussed in this paper was made possible by funding from the Russell J. and Dorothy S. Bilinski Dissertation Fellowship Award and the University of Hawai‘i at Mānoa. Many thanks are due to the editors of this handbook, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, to two anonymous reviewers, and to Melody Ann Ross for providing helpful feedback on earlier drafts. All remaining errors are my own. Greatest thanks go to the participants who graciously lent their voices to these collections.
Throughout the 1800s, the number of native Hawaiians dropped precipitously due largely to foreign disease (Stannard 1990:330). An increase in Christian missionaries in 1820 further contributed to the decline of ʻōlelo Hawai‘i, as English was elevated as the language of the church, economic advancement, and social capital (Drager 2012b:63). The overt prestige of English was further entrenched by the steady growth of US influence in schools. A wealthy, English-speaking minority affected policies that gradually forced ʻōlelo Hawai‘i schools to switch to English as the primary language of instruction. All of this was designed to prepare the youth of Hawai‘i for “participation in an American-type community” (Stueber 1964:144). In July 1887, a group of wealthy US businessmen coerced King Kalākaua under threat of force to sign a new constitution that stripped the Hawaiian monarchy of its authority and disenfranchised native Hawaiians. A mere five years later, the Kingdom of Hawai‘i was overthrown by a wealthy, white minority, and Hawai‘i was illegally annexed by the US in 1898.

The rapidly changing social climate of Hawai‘i in the nineteenth century was spurred on by the establishment of sugarcane plantations in 1835. The plantations lured laborers worldwide, initially, Cantonese and Portuguese, then Japanese, laborers from the Philippines, and later, from Germany, Korea, Puerto Rico, Spain, and islands throughout the Pacific. While the plantation foremen originally spoke ʻōlelo Hawai‘i, English soon took its place, reflecting a changing social climate that was increasingly dominated by an English-speaking minority. Because many languages were used on the plantations with no shared first language among the workers, an English-based pidgin arose to facilitate communication. Children raised in this context commonly spoke both the language(s) of their parents and a creolized version of English—Hawai‘i Creole, known locally as Pidgin (Kawamoto 1993). As English speakers solidified a position of overt economic power, Pidgin increasingly took on the role of lingua franca among Hawai‘i Locals, and within three generations (by the 1930s), Pidgin was a distinct language from English (Roberts 2004).

Pidgin has become closely linked with a Local identity, due in part to its development alongside Hawai‘i’s changing demography. This identity encompasses many cultural backgrounds and ethnicities, the extent of which is beyond the scope of this use case (for detailed accounts, see Fujikane 1997; Rohrer 1997; Ohnuma 2002), but a crucial opposition places Local at conceptual odds with Haole. For many, the category “Local” comprises a non-white person, born in Hawai‘i, with a familial connection to the plantations and Pidgin, while the category “Haole” comprises a white person whose presence on the islands is the result of historical exploitation via colonialism. “Haole” is a construct characterized by external forces (e.g., whiteness, the military, tourism, English dominance), while “Local” evinces solidarity against that force (Ohnuma 2002). Attitudes toward Pidgin and English reflect this opposition. Pidgin is marginalized by many as “broken English,” while English is upheld as “proper” (Drager & Grama 2014). Moreover, English is often cited as a language of economic mobility, and despite Pidgin’s use in publicly visible domains (e.g., politics, the news), it is nevertheless viewed by many as a language whose use should be restricted to the home and close personal relationships (see Marlow & Giles 2008, 2010).

2.1 A note on researcher positionality
The history of colonialism in Hawai‘i cannot be ignored when discussing the present study. In keeping with the observation that data cannot be divorced from its source (Holton, Leonard, & Pulsifer, chapter 4, this volume), the researcher’s positionality to the variety under study is key. In the context of Hawai‘i, it matters greatly that I am a white, non-native Pidgin speaker from California. My very presence as a researcher in Hawai‘i was the result of settler privilege. Because of this, the likelihood that I could elicit representative Pidgin speech data is low, given the long history of language hegemony (see Marlow & Giles 2008, 2010). Therefore, it was both prudent and methodologically necessary to make as much use of existing data as possible in lieu of collecting new data.1

Having established context, I now move on to describe the goals of the original study, as well as how legacy data were used to address those goals.

3 Using legacy data to achieve research goals
Building on phonological work on Pidgin (e.g., Bickerton 1976; Sakoda & Siegel 2008), the focus of Grama (2015) was to acoustically characterize changes in the vowel system of Pidgin over time, with considerations to both internal and external factors.2 While phonological
descriptions of Pidgin highlight that inter- and intra-speaker variation is clearly present in the language, relatively few studies have applied variationist methods to acoustic data to describe this variation (for exceptions in other creoles, see Kraus 2017; Lesho 2014; Rosenfelder 2009; Sabino 1996, 2012; Veatch 1991; Wassink 1999, 2001, 2006).

To characterize changes over time in Pidgin, Grama (2015) employed both real- and apparent-time data taken from archival recordings from two existing corpora on Kaipuleohone, the University of Hawai‘i at Mānoa’s digital ethnographic archive (Albarillo & Thieberger 2009; Berez 2013); one collection was recorded in the 1970s (with speakers born 1896–1946), and one was recorded in the 2000s (with speakers born 1947–1988). These data were appropriate for answering questions of language change over time because they represented two independent samplings of the Hawai‘i speech community approximately thirty years apart, with speakers born over a ninety-year time frame. This allowed for real-time comparisons to be made across corpora and apparent-time comparisons to be made within each corpus. That the data were already extant was particularly beneficial, given that many of the recordings were conducted between native Pidgin speakers, and that my position as a migrant, white, non-native Pidgin speaker in Hawai‘i limited my ability to reliably collect new data. These corpora—the Bickerton Collection and the Influences and Variation in Hawai‘i Creole English—are discussed in the next section.

3.1 The corpora

3.1.1 The 1970s corpus (the Bickerton Collection) The Bickerton Collection includes materials elicited from Pidgin speakers and from speakers of a range of other languages. Recordings were made between 1970 and 1980 and include people born between the mid-1890s to the late 1940s. Speakers were represented across these birthdates, but they tended to be either in their mid-to-late sixties, or in their early thirties to mid-forties; this created a natural break in the corpus, where about half the speakers were older than fifty years of age and half were younger. The collection was amassed to describe the linguistic structure of Pidgin, with special attention paid to morpho-syntactic and phonological alternations (for studies based on this corpus, see Bickerton 1976; Odo 1975, 1977). Interviews were conducted by both Pidgin and non-Pidgin speakers; the apparent dominant language of the interviewer became an important criterion for interview exclusion from Grama (2015) (see section 3.1.3). In all, the collection includes 168 recordings, which vary widely in duration. While many of the recordings are in the style of traditional sociolinguistic interviews, a number of them are recordings of radio and television programs or advertisements. Throughout the remainder of the chapter, this collection will be referred to as the 1970s corpus.

3.1.2 The 2000s corpus (Influences and Variation in Hawai‘i Creole English) The Influences and Variation in Hawai‘i Creole English collection comprises sociolinguistic-style interviews with Pidgin speakers. Recordings were made in the early-to-mid-2000s and include people born between the mid-1940s to the mid-1980s. Speakers were represented across these birthdates, but they tended to be either in their late forties, or in their early twenties, again creating a natural break in the corpus; here, approximately half the speakers were older than thirty years of age and half were younger. The corpus was amassed by Jeff Siegel and a number of research assistants to examine variation in Pidgin and the role played by external influences on language change (for studies based on this corpus, see Sakoda & Siegel 2003, 2008; Siegel 2007). Interviews were conducted by Pidgin speakers, with Pidgin speakers, and interviewees were typically friends, family members, colleagues, or otherwise previously known to the interviewer. In all, the collection includes 117 recordings, ranging from around six minutes to one hour and forty minutes. Throughout the remainder of the chapter, this collection will be referred to as the 2000s corpus.

3.1.3 Interview selection Selecting recordings from these corpora that were appropriate for a sociophonetic study of language change over time proved labor-intensive, largely because the original intent of each corpus was not to facilitate sociophonetic research. Recordings were prioritized that met five constraints. First, sufficient information had to be included in the metadata or in the recording itself to indicate that interviewees were born and raised in Hawai‘i and not people who emigrated to Hawai‘i later in life. This constraint was implemented to be as certain as possible that speakers included in the study were native Pidgin speakers, and
not second-language learners, as later acquisition can impact phonological realizations (see, e.g., Flege, Shirru, & MacKay 2003). Second, interviews were prioritized where the interviewer was also a Pidgin speaker. Given the history of linguistic hegemony in Hawai‘i, people are less likely to use Pidgin if their interlocutor does not also speak Pidgin. This is especially true in more formal domains (e.g., in a recorded conversation). Third, recordings with one interviewee were desirable because recordings with multiple interlocutors yield overlapping speech and uneven turn-taking, which makes preparing data for forced alignment more labor-intensive. Fourth, recordings needed to be of high-enough quality to undergo acoustic analysis. Quality issues rendered many recordings unusable; wind, excessive background noise, static, feedback, and quiet interviewees yielded recordings that were unlikely to produce reliable formant measurements. Fifth, recordings had to feature enough speech to map a speaker’s vowel space. Through conducting the original study, it became clear that interviews lasting around twenty minutes, or two thousand words, reliably produced enough vowel tokens across vowel category to accurately map the vowel space. Interviews that met each of these five constraints were candidates for inclusion in Grama (2015).

There was also a desire for a balanced number of speakers to ensure equal representation across demographic category. These categories consisted of corpus (1970s vs. 2000s), wherein real-time, longitudinal change could be tested; relative age within the corpus (relatively older vs. relatively younger speakers), so that change in apparent time could be tested, where relatively older speakers represent the language as it was in the past, and relatively younger speakers represent the language as it is spoken at the time of recording (Bailey 2004); and binary sex (female, male), to assess how females and males differed (if at all) in their participation in the identified changes. Pruning the two corpora based on these constraints resulted in a total of thirty-two total recordings. The distribution of speakers across the tested demographic categories included in Grama (2015) is reported in table 17.1.

Despite these restrictions, inherent differences exist between the two collections, particularly with respect to interview styles. In the 2000s corpus, interviewers and interviewees tended to be close friends or family members. The familiarity between speakers resulted in very conversational interactions. By contrast, 1970s interviewers were less likely to be previously known to interviewees and tended to feature more monologic speech styles. Additionally, differences arose that reflected Hawai‘i’s changing social climate over a thirty-year period. Young 2000s speakers as a group, for example, completed more formal schooling than 1970s speakers, who typically did not achieve more than a high school education. By contrast, 1970s speakers were more likely to have worked in positions that required manual labor than 2000s speakers did. Appendix A provides a more detailed breakdown of demographic information for each speaker.

## Data processing

The following section describes how data were processed in Grama (2015), including how interviews were transcribed and force aligned; how the resulting alignments were manually checked; and how vowel formants were extracted, cleaned, and normalized.

### 4.1 Transcription

After selection, interviews were orthographically transcribed at the utterance level. Research concerned with the acoustic properties of vowels ultimately relies on transcription of some kind. While there have been significant strides in automated speech recognition, fully automated transcription is far from accurate enough to replace human transcribers. Thus, manual transcription remains a necessary bottleneck to acquiring large amounts of high-quality vowel data (cf. Seifert et al.)
automatically generate corpus annotations and because I had prior experience using the system. Transcriber files were uploaded along with their accompanying .wav files to SOLIS, and files were tagged with available metadata. Each uploaded transcript was automatically checked for words that were not in the English CELEX dictionary (Baayen, Piepenbroock, & Gulikers 1995), with which LaBB-CAT is designed to interface. Unfamiliar items in the dictionary were added using the grapheme-to-phoneme (G2P) mapping system that CELEX employs for British English.

Errors in the transcript (e.g., misspellings, unrecognized characters) were corrected in SOLIS with LaBB-CAT’s transcription editing protocols. Forced alignment (Schiel, Draxler, & Harrington 2011). Choosing a forced aligner is therefore crucial, as they are now standard practice of workflows in many production-based sociophonetic studies.

In Grama (2015), corpus storage and forced alignment were performed using a server build of LaBB-CAT housed at the Sociolinguistics Server (SOLIS) at the University of Hawai‘i at Mānoa (Drager 2012a). This choice was made both because of LaBB-CAT’s ability to automatically generate corpus annotations and because I had prior experience using the system. Transcriber files were uploaded along with their accompanying .wav files to SOLIS, and files were tagged with available metadata. Each uploaded transcript was automatically checked for words that were not in the English CELEX dictionary (Baayen, Piepenbroock, & Gulikers 1995), with which LaBB-CAT is designed to interface. Unfamiliar items in the dictionary were added using the grapheme-to-phoneme (G2P) mapping system that CELEX employs for British English. Errors in the transcript (e.g., misspellings, unrecognized characters) were corrected in SOLIS with LaBB-CAT’s transcription editing protocols. Forced alignment

### Table 17.2

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Speaker pseudonym</th>
<th>Time transcribed (hr:min:sec)</th>
<th>Word count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old 1970s</td>
<td>Male</td>
<td>Joseph</td>
<td>00:21:16</td>
<td>2,744</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kawiia</td>
<td>00:19:42</td>
<td>2,358</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kimo</td>
<td>00:30:30</td>
<td>1,508</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manny</td>
<td>00:32:06</td>
<td>2,831</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Kaimana</td>
<td>00:18:45</td>
<td>2,324</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Keiko</td>
<td>00:15:13</td>
<td>1,689</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kaimana</td>
<td>00:18:45</td>
<td>2,324</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mali</td>
<td>00:27:56</td>
<td>2,766</td>
</tr>
<tr>
<td>Young 1970s</td>
<td>Male</td>
<td>Danny</td>
<td>00:19:36</td>
<td>1,735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eddie</td>
<td>00:19:23</td>
<td>2,245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Glen</td>
<td>00:19:15</td>
<td>1,777</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Victor</td>
<td>00:14:42</td>
<td>2,045</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Delia Jane</td>
<td>00:18:17</td>
<td>2,099</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eddie</td>
<td>00:19:23</td>
<td>2,245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mona Lisa</td>
<td>00:18:40</td>
<td>1,858</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Teresa</td>
<td>00:26:58</td>
<td>1,930</td>
</tr>
<tr>
<td>Old 2000s</td>
<td>Male</td>
<td>Grant</td>
<td>00:17:06</td>
<td>1,976</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Keoni</td>
<td>00:14:11</td>
<td>1,952</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kevin</td>
<td>00:18:36</td>
<td>1,910</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Palani</td>
<td>00:30:05</td>
<td>2,063</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Carla</td>
<td>00:13:32</td>
<td>1,927</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kahea</td>
<td>00:15:09</td>
<td>2,201</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lani</td>
<td>00:14:05</td>
<td>1,727</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pua</td>
<td>00:28:50</td>
<td>1,707</td>
</tr>
<tr>
<td>Young 2000s</td>
<td>Male</td>
<td>Alike</td>
<td>00:11:55</td>
<td>2,142</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eric</td>
<td>00:27:37</td>
<td>2,018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kaleo</td>
<td>00:17:23</td>
<td>2,230</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Myko</td>
<td>00:21:48</td>
<td>2,038</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Lena</td>
<td>00:27:30</td>
<td>1,941</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mina</td>
<td>00:33:42</td>
<td>1,840</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sarah</td>
<td>00:20:24</td>
<td>1,984</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Starla</td>
<td>00:24:09</td>
<td>1,966</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>11:16:29</td>
<td>66,100</td>
</tr>
</tbody>
</table>
was produced using the Hidden Markov Model Toolkit, HTK (Young et al. 2009), via the default train/align procedure (see Fromont & Watson 2016), which produced automated boundaries around phonemes according to lexical data from CELEX.

Force-aligned data in LaBB-CAT can be accessed in two main ways. First, the user can access the transcripts tab to interface directly with the transcript and toggle available annotation layers on and off. These layers are programmed by the user and generated automatically. Information can then be extracted as a time-aligned file in several formats, for example, as a Praat TextGrid (Boersma & Weenink 2019). While the entire interview file can be exported as one Praat TextGrid (which is preferable if one-to-one correspondence between file and speaker is desired), Praat often has issues with larger files, making them difficult to work with if processing power is at a premium. The second way to extract the force-aligned data is via LaBB-CAT’s search function. By executing a search across the segments layer, LaBB-CAT returns all instances that fit the specified criteria. Individual segments can then be extracted, a process which yields TextGrids and .wav files for all aligned intervals that correspond to interval breaks in Transcriber or annotations in ELAN. For example, the user could specify one or more speakers in LaBB-CAT and query a specific vowel. Figure 17.1 shows an example of such a search executed for Myko, a young 2000s male.

This process was performed independently for each vowel category and individual speaker. The pairs of .wav files and TextGrids (between 25 and 350 files per vowel per speaker) were then extracted from SOLIS and stored on my personal computer in a folder organized by speaker by vowel. In accordance with best practices for data management, these files will be archived in the future in one of three locations: SOLIS, Kaipuleohone, or the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).
4.3 Checking data

After forced alignment and extraction, vowel alignments were manually checked for accuracy and coded following the protocol described here. There is some debate in sociophonetics as to whether, and under what circumstances, manual checking of force-aligned vowel data is strictly necessary. For many, this comes down to what the research question is, trust in the force-aligned output, or more practical considerations (e.g., time). There are reasonable arguments for and against manually checking vowel alignments for accuracy. While some work demonstrates that force-aligned output produces inconsistent boundaries between vowels and sonorants (Strelluf 2016; Gonzalez, Grama, & Travis 2020), other work suggests that, given enough data, correcting force-aligned output only marginally improves formant measurements (Labov, Rosenfelder, & Fruehwald 2013:37–38; Gonzalez & Docherty 2018). However, much of the work that evaluates taking formant measurements from uncorrected data assesses the observed improvement in static vowel measures (e.g., single F1/F2 measurements for each vowel). Complications can arise if more measurements are taken from single vowels, as forced aligners are not as good at segmentation as humans are, despite doing the work much faster (Fromont & Watson 2016). In Grama (2015), there was the added complication of legacy data, where recordings were not performed in ideal scenarios, or where audio degradation impacted the quality of recordings prior to digitization. With such data, there is a higher risk that completely automated methods could yield spurious measurements. A middle ground is to check a random sample of vowel tokens, particularly those that occur in phonological contexts that disproportionately impact alignment (see Gonzalez, Grama, and Travis 2020). In my view, claims about the behavior of vowel trajectories necessitate more accurate alignments, given these concerns.

Given that the original study utilized legacy data and investigated vowels both statically and dynamically, it was prudent to take a more conservative approach and manually check all the force-aligned output. Each vowel was checked by hand and boundaries were corrected if necessary; then, the vowel was tagged with its appropriate lexical set (cf. Wells 1982). As is typical of sociophonetic studies, only vowels in stressed content words were prepared for analysis. Boundary correction was done using a strict set of criteria, detailed in sections 4.3.1–4.3.3, and boundaries were placed at, or as close to a zero crossing in the waveform as possible. If the cues discussed in the following sections were not available, the token was excluded.

4.3.1 Obstruents For vowels preceded by stops, the burst and aspiration were included in the consonant segment, not the vowel segment. Vowel onsets were marked where the waveform indicated periodicity and the spectrogram showed clear formant structure (e.g., rising F1). A perceptible decrease in amplitude served as an additional cue for vowels adjacent to voiced stops. Boundaries for vowels adjacent to voiceless fricatives were placed where formant structure was clear, and where the cessation of aperiodic energy coincided with a change in amplitude.

4.3.2 Sonorants For vowels bordering nasals and laterals, boundaries were placed where decreased amplitude coincided with formant dampening or a lowered F1. Boundaries between vowels and pre-vocalic /r/ were marked at maximum F3, or F2, if F3 was unavailable. Where post-vocalic /r/ was present, boundaries were placed where a dip in F3 indicated oral closure consistent with /r/ articulation (see, e.g., Johnson 2012:140). Along with amplitude, boundaries for /w/ were evaluated based on where F2 and F3 began to diverge, and boundaries for /j/ were evaluated based on where F2 and F3 began to diverge.

4.3.3 Word-initial and word-final vowels Vowel-initial words at the beginning of an utterance were often bordered by silence or glottal closure. The starting point of the vowel was therefore marked as the first relatively high-amplitude vocal pulse evident in the waveform. Vowel-final words at the end of an utterance were marked at the last high-amplitude vocal pulse evident in the waveform. In both cases, Praat’s ability to track the formant structure factored into boundary placement.

4.4 Extracting formant values

Vowel checking and formant extraction was done for each speaker and vowel independently. Before extraction, a 15% subsample of each speaker’s vowel category was checked to ensure that settings in Praat accurately captured formant behavior. While it is commonplace to use a standard set of values for vowel extraction
(e.g., five formants under 5,500 hertz for females), this was not possible in every case given the aforementioned issues regarding recording quality. Therefore, manual assignment of formant settings during formant extraction was necessary. The specific extraction values supplied to the Praat script can be found in Grama (2015:296–307).

The Praat script used for this study was based on a script written by Mietta Lennes, which had been modified by Abby Walker and Katie Drager. This script extracted vowel identity, the word in which the vowel appeared, the preceding and following phonological segments, the duration of the vowel, the fundamental frequency, and readings of the first three formants—F1, F2 and F3—from seven equidistant points, starting from 20% of the duration through the vowel and terminating at 80%. Speech rate was assessed using de Jong and Wempe’s (2009) Praat script, which calculates speech rate as a function of the number of amplitude peaks over the duration of an interval. This process yielded two different types of data for each vowel: midpoint data, which were useful for assessing the overall picture of the vowel space and the target position of monophthongs, and transition data, which were useful for quantifying formant contours over the vowel’s duration.

### 4.5 Cleaning the data
Speaker demographic information, along with formant and speech rate measurements were compiled into a single .csv file using R (R Core Team 2018) in wide format (i.e., where each row in the data frame corresponds to a single token). This yielded 11,551 vowel tokens over fourteen vowel categories across thirty-two speakers. Formant measurements were checked in R over the duration of the vowel to ensure that accurate readings were taken by the Praat script. Radical deviations from expected patterns (e.g., an F2 in /u/ of 1,200 hertz at 30% of the vowel, followed by an F2 of 500 hertz at 40%) were treated as spurious measurements, and resulted in the exclusion of the vowel token. Each speaker’s vowel space was plotted and checked for outliers. Tokens with formant measurements that fell outside the range of plausible adult physiological limits (e.g., an F1 of 2,000 hertz or an F2 of 300 hertz) were removed. Tokens that were phonologically unlikely (e.g., /i/ in the low back area of the vowel space) were evaluated in Praat and removed if they were judged to be measurement errors. Finally, a script was written to remove tokens whose formant measurements fell more than three standard deviations outside of a speaker’s vowel distribution, calculated within speaker, within vowel (cf. Hughes 2014, discussed in Foulkes et al. 2018). Figure 17.2 presents an example of a speaker’s /i/ and /æ/ before and after filtering.

A total of 353 vowel tokens were removed following these processes, yielding 11,198 tokens for analysis. A breakdown of the number of vowels per vowel class can be seen in table 17.3.

### 4.6 Normalizing vowel formants
Even when thoroughly cleaned, raw data are not usually the focus of direct analysis at the group level (cf. Han, chapter 6, this volume). Normalization is often a necessary step to interpreting patterns in vowel data. Vowel normalization seeks to neutralize differences between
Work that directly compares the efficacy of normalization techniques (e.g., Adank, Smits, & Van Hout 2004; Flynn & Foulkes 2011) tends to find that methods that are vowel-extrinsic, formant-intrinsic, and speaker-intrinsic are best at reducing variation that arises as a result of physiology, while preserving sociolinguistic variation. Vowel-intrinsic methods perform comparably worse by these same metrics. However, any vowel normalization technique shows improvement over raw hertz comparisons (Flynn & Foulkes 2011:686). A popular normalization choice is the Lobanov method (Lobanov 1971; but see Barreda & Nearey 2018), as it is vowel-extrinsic and produces interpretable plots.

The vowel data in this study were normalized using the Lobanov method, which requires that relatively equal samples of all (monophthongal) vowel categories be included to avoid artificial skewing. Lobanov converts raw hertz values to normalized \( z \)-scores by subtracting a speaker’s mean formant frequency \( \mu_i \) from a raw measurement \( F_i \), and then diving this by the standard deviation for that speaker’s formant \( \sigma_i \), as in equation (17.1).

\[
F_N^i = \frac{F_i - \mu_i}{\sigma_i}
\]

(17.1)

Because Lobanov produces values that are centered on \((0, 0)\), it is not uncommon to scale these values back to hertz (see, e.g., Labov, Rosenfelder, & Fruehwald 2013:36), though this should only be performed after all values have been normalized. While easily performed manually or via script, normalization can also be performed using the vowels package (Kendall & Thomas 2018) in R, as can the Bark difference (Traunmüller 1997), ANAE (Labov, Ash, & Boberg 2006), Nearey (1977), and Watt and Fabricius (2002) methods.

5 Reporting, storing, and sharing the data

5.1 Reporting data

Transparent, clear reporting of data is an important aspect of any sociophonetic study. To this end, I include the raw formant values measured along the vowel’s duration for all vowels across speakers (Grama 2015:308–337), as well as the normalized formant values across vowel, age, and sex (338–345). Moreover, clear figures and statistical analyses are key to achieving interpretable and reproducible findings. R is a powerful option for graphics creation.
and fitting statistical models. For graphing, ggplot2 (Wickham et al. 2019) is a popular and highly customizable option; for fitting statistical models, lme4 (Bates et al. 2019) allows the user to fit mixed models, and both lmerTest (Kuznetsova, Brockhoff, & Bojesen 2019) and pbkrtest (Halekoh & Højsgaard 2017) allow the user to derive p-values and interpretable model outputs from lme4.10

The original study strove to achieve clarity and statistical accountability. For example, the two-dimensional kernel density plots in the left panel of figure 17.3 show the distribution and overall shape of the data, giving the reader an understanding of the distribution of the overall vowel categories; this aids the interpretability of the plot. However, not every plot was equally well conceived. The right panel in figure 17.3 shows a localized polynomial regression with standard errors for the F1 of /ɪ/ over time. The regression indicates a sex difference; however, the figure lacks the data points on which the regression is based. Therefore, it is unclear how closely the regression fits the data and whether the data points vary consistently across the observed effect. A solution to this would be to include the individual data points in the plot in the right panel.

5.2 Storing and sharing the data
SOLIS served as the repository for transcribed interviews and the boundaries produced by forced alignment. Using this server allowed me to store raw data used in the original study in a password-protected online repository as I progressed through the collections. While this minimized the risk of data loss, the rest of my data (e.g., individual formant measurements, R scripts, and statistical models) is on personal repositories. And while formant values are reported in Grama (2015), no officially archived data sheet exists. This is a key area where the original study was not in line with best practices (see Andreassen, chap. 7, this volume). One way to solve this issue would be to make further use of SOLIS. Because SOLIS uses LaBB-CAT architecture, it allows for detailed annotation stores. Formant values and other information could be uploaded to the server to allow the precise study to be replicated by anyone with access. In addition, manually checked Praat TextGrids could be uploaded to SOLIS to enrich the files currently housed on SOLIS. This would also constitute a step toward sharing the data, as access is, as of writing, restricted to SOLIS users or achieved through personal requests to me.

6 Conclusion
This use case serves as a guide for those wishing to undertake similar longitudinal studies of language variation and change, especially where legacy data are required. As a short-hand reference, a schematized version of the methodological steps detailed in this chapter is available in appendix B. As a final caveat, this methodology was employed between 2014 and 2015, meaning that there are aspects of it that may not stand the test of time. Nevertheless, the steps for processing the data discussed here, many of which are maximally conservative, should remain relevant in the future.
## Appendix A: Detailed speaker demographic information in Grama (2015)

Speaker demographic and social information in Grama (2015), including the collection file name, the corpus, and the speaker's relative age, assigned pseudonym, binary sex, age, date of birth, island of residence, listed ethnicity, highest education achieved, and occupation information.

<table>
<thead>
<tr>
<th>Collection file name</th>
<th>Corpus and relative age</th>
<th>Speaker pseudonym</th>
<th>Sex</th>
<th>Age</th>
<th>Date of birth</th>
<th>Island</th>
<th>Ethnicity</th>
<th>Highest education</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB1-072</td>
<td>Old 1970s</td>
<td>Kawika</td>
<td>M</td>
<td>79</td>
<td>1896</td>
<td>Kaua'i</td>
<td>Hawaiian</td>
<td>na</td>
<td>Retired motel owner</td>
</tr>
<tr>
<td>DB1-122</td>
<td>Old 1970s</td>
<td>Joseph</td>
<td>M</td>
<td>69</td>
<td>1906</td>
<td>Hawai'i</td>
<td>Portuguese</td>
<td>No high school</td>
<td>Retired plantation worker</td>
</tr>
<tr>
<td>CS1-JA-03</td>
<td>Old 1970s</td>
<td>Manny</td>
<td>M</td>
<td>58</td>
<td>1922</td>
<td>Hawai'i</td>
<td>Filipino</td>
<td>High school</td>
<td>Farmer, real estate</td>
</tr>
<tr>
<td>DB1-164</td>
<td>Old 1970s</td>
<td>Kimo</td>
<td>M</td>
<td>54</td>
<td>1921</td>
<td>O'ahu</td>
<td>Part Hawaiian</td>
<td>High school</td>
<td>Retired roofer, plantation</td>
</tr>
<tr>
<td>DB1-073</td>
<td>Old 1970s</td>
<td>Miki</td>
<td>F</td>
<td>68</td>
<td>1907</td>
<td>Kaua'i</td>
<td>Japanese</td>
<td>High school</td>
<td>Retired barber</td>
</tr>
<tr>
<td>DB1-066</td>
<td>Old 1970s</td>
<td>Malia</td>
<td>F</td>
<td>64</td>
<td>1911</td>
<td>Kaua'i</td>
<td>Hawaiian</td>
<td>High school</td>
<td>Housewife</td>
</tr>
<tr>
<td>DB1-162</td>
<td>Old 1970s</td>
<td>Kaimana</td>
<td>F</td>
<td>57</td>
<td>1918</td>
<td>O'ahu</td>
<td>Hawaiian, Haole</td>
<td>High school</td>
<td>Retired</td>
</tr>
<tr>
<td>DB1-056</td>
<td>Old 1970s</td>
<td>Keiko</td>
<td>F</td>
<td>55</td>
<td>1918</td>
<td>Kaua'i</td>
<td>Japanese</td>
<td>High school</td>
<td>Home management</td>
</tr>
<tr>
<td>DB1-165</td>
<td>Young 1970s</td>
<td>Eddie</td>
<td>M</td>
<td>39</td>
<td>1936</td>
<td>Hawai'i</td>
<td>Part Hawaiian</td>
<td>High school</td>
<td>Construction worker</td>
</tr>
<tr>
<td>DB1-059</td>
<td>Young 1970s</td>
<td>Victor</td>
<td>M</td>
<td>37</td>
<td>1938</td>
<td>Kaua'i</td>
<td>Portuguese</td>
<td>High school</td>
<td>NA</td>
</tr>
<tr>
<td>CS1-EA-03</td>
<td>Young 1970s</td>
<td>Danny</td>
<td>M</td>
<td>30</td>
<td>1942</td>
<td>O'ahu</td>
<td>Filipino</td>
<td>High school</td>
<td>Floorer</td>
</tr>
<tr>
<td>CS1-GN-02</td>
<td>Young 1970s</td>
<td>Glen</td>
<td>M</td>
<td>25</td>
<td>1944</td>
<td>Hawai'i</td>
<td>Japanese</td>
<td>High school</td>
<td>Contract laborer</td>
</tr>
<tr>
<td>DB1-074</td>
<td>Young 1970s</td>
<td>Mona Lisa</td>
<td>F</td>
<td>48</td>
<td>1927</td>
<td>Kaua'i</td>
<td>Filipino</td>
<td>High school</td>
<td>NA</td>
</tr>
<tr>
<td>DB1-075</td>
<td>Young 1970s</td>
<td>Leilani</td>
<td>F</td>
<td>42</td>
<td>1933</td>
<td>Kaua'i</td>
<td>Hawaiian</td>
<td>High school</td>
<td>Housewife, retired</td>
</tr>
<tr>
<td>DB1-120</td>
<td>Young 1970s</td>
<td>Delia Jane</td>
<td>F</td>
<td>35</td>
<td>1940</td>
<td>Hawai'i</td>
<td>Filipino</td>
<td>High school</td>
<td>Adult education instructor</td>
</tr>
<tr>
<td>DB1-065</td>
<td>Young 1970s</td>
<td>Teresa</td>
<td>F</td>
<td>35</td>
<td>1940</td>
<td>Kaua'i</td>
<td>Filipino</td>
<td>College</td>
<td>Air national guard</td>
</tr>
<tr>
<td>CS2-053</td>
<td>Old 2010s</td>
<td>Grant</td>
<td>M</td>
<td>56</td>
<td>1951</td>
<td>O'ahu</td>
<td>Japanese</td>
<td>College</td>
<td>Government worker</td>
</tr>
<tr>
<td>CS2-029</td>
<td>Old 2010s</td>
<td>Kevin</td>
<td>M</td>
<td>52</td>
<td>1955</td>
<td>Hawai'i</td>
<td>Hawaiian</td>
<td>NA</td>
<td>Unemployed, ex-military/farmer</td>
</tr>
<tr>
<td>CS2-017</td>
<td>Old 2010s</td>
<td>Palani</td>
<td>M</td>
<td>44</td>
<td>1963</td>
<td>Hawai'i</td>
<td>Part Hawaiian</td>
<td>NA</td>
<td>Shop owner</td>
</tr>
<tr>
<td>CS2-030</td>
<td>Old 2010s</td>
<td>Keoni</td>
<td>M</td>
<td>40</td>
<td>1967</td>
<td>Hawai'i</td>
<td>Part Hawaiian</td>
<td>High school</td>
<td>NA</td>
</tr>
<tr>
<td>CS2-037</td>
<td>Old 2010s</td>
<td>Pua</td>
<td>F</td>
<td>58</td>
<td>1949</td>
<td>O'ahu</td>
<td>Part Hawaiian</td>
<td>High school</td>
<td>NA</td>
</tr>
<tr>
<td>CS2-040</td>
<td>Old 2010s</td>
<td>Lani</td>
<td>F</td>
<td>49</td>
<td>1958</td>
<td>O'ahu</td>
<td>Part Hawaiian</td>
<td>High school</td>
<td>Housewife</td>
</tr>
<tr>
<td>CS2-027</td>
<td>Old 2010s</td>
<td>Carla</td>
<td>F</td>
<td>46</td>
<td>1961</td>
<td>Hawai'i</td>
<td>Portuguese</td>
<td>High school</td>
<td>Unemployed</td>
</tr>
<tr>
<td>CS2-011</td>
<td>Old 2010s</td>
<td>Keaha</td>
<td>F</td>
<td>42</td>
<td>1965</td>
<td>Kaua'i</td>
<td>Part Hawaiian</td>
<td>High school</td>
<td>Ranch worker</td>
</tr>
<tr>
<td>CS2-052</td>
<td>Young 2010s</td>
<td>Kaleo</td>
<td>M</td>
<td>22</td>
<td>1985</td>
<td>Maui</td>
<td>Hawaiian, Korean, Haole</td>
<td>College</td>
<td>Student</td>
</tr>
<tr>
<td>CS2-056</td>
<td>Young 2010s</td>
<td>Myko</td>
<td>M</td>
<td>22</td>
<td>1985</td>
<td>Kaua'i</td>
<td>Hawaiian</td>
<td>College</td>
<td>Student</td>
</tr>
<tr>
<td>CS2-046</td>
<td>Young 2010s</td>
<td>Eric</td>
<td>M</td>
<td>21</td>
<td>1986</td>
<td>Hawai'i</td>
<td>Chinese, Filipino</td>
<td>College</td>
<td>Student</td>
</tr>
<tr>
<td>CS2-051</td>
<td>Young 2010s</td>
<td>Aila</td>
<td>M</td>
<td>21</td>
<td>1986</td>
<td>Hawai'i</td>
<td>Japanese</td>
<td>College</td>
<td>Student</td>
</tr>
<tr>
<td>CS2-048</td>
<td>Young 2010s</td>
<td>Sarah</td>
<td>F</td>
<td>24</td>
<td>1983</td>
<td>O'ahu</td>
<td>Chinese</td>
<td>College</td>
<td>MA Student</td>
</tr>
<tr>
<td>CS2-019</td>
<td>Young 2010s</td>
<td>Starla</td>
<td>F</td>
<td>23</td>
<td>1984</td>
<td>Hawai'i</td>
<td>Chinese, Japanese</td>
<td>High school</td>
<td>NA</td>
</tr>
<tr>
<td>CS2-055</td>
<td>Young 2010s</td>
<td>Mina</td>
<td>F</td>
<td>21</td>
<td>1986</td>
<td>Kaua'i</td>
<td>Japanese, Haole, Chinese, Hawaiian</td>
<td>College</td>
<td>Student</td>
</tr>
<tr>
<td>RK01</td>
<td>Young 2010s</td>
<td>Lena</td>
<td>F</td>
<td>19</td>
<td>1988</td>
<td>Kaua'i</td>
<td>Filipino, Japanese</td>
<td>College</td>
<td>Student</td>
</tr>
</tbody>
</table>

F = female; M = male; NA = not available.
2. One aim of Grama (2015) was to quantify vowel variation along the creole continuum (cf. DeCamp 1971; Sato 1993), a focus that falls outside the scope of the present chapter.

3. Metadata were missing or inconsistent across recordings; it is a researcher’s duty to report these gaps, while being as vigilant as possible in our own studies to record metadata as thoroughly and representatively as possible (cf. Mattern, chapter 5, this volume). In some cases, accompanying corpus notes were mined for missing information to account for gaps in the metadata. When information crucial to the test variables was absent, the interview was excluded.

4. Recordings were excluded if they were shorter than ten minutes on these grounds.

**Appendix B: Data workflow for Grama (2015)**

The chart schematizes the workflow for Grama (2015) from interview selection to analysis. It is meant as a short-hand reference for those wishing to replicate the methodology of this study. The transcription software and forced aligner listed could easily be altered to accommodate researcher preference.

**Notes**

1. This was especially true because time and financial considerations made hiring research assistants impractical during the completion of the original study.
5. Space does not allow for a fuller description of the range of factors that might govern the selection of a forced aligner; however, aligner choice should be based on the researcher’s needs, as well as aligner performance and suitability. For a more thorough discussion of these factors, I direct the reader to Fromont and Watson (2016), McAuliffe et al. (2017), Mackenzie and Turton (2020), and Gonzalez, Grama, and Travis (2020).

6. The steps described here would be identical to those were the reader to use a local build of LaBB-CAT.

7. At the time, CELEX seemed preferable to the CMU Pronouncing Dictionary (the latter of which is optimized for US varieties), given the variable absence of post-vocalic /r/ in Pidgin. However, recent work demonstrates that automated alignment is not strongly affected by variety if the variety under study is not markedly different from the training variety (MacKenzie & Turton 2020), or if vowel coercion is performed (Gonzalez, Grama, & Travis 2020).

8. The original script is collect_formant_data_from_files.praat, available at https://github.com/lennes.

9. Some phoneticians (e.g., Rosner & Pickering 1994) argue that normalization is done to model cognitive processes that underpin vowel perception, though there is some debate in the phonetics literature as to whether listeners normalize vowels at all (see, e.g., Pisoni 1997). This chapter does not weigh in on this debate.


References


Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1980001c015900_9780262366076.pdf by guest on 16 January 2022


Grama, James. 2015. Variation and change in Hawai‘i Creole vowels. PhD dissertation, University of Hawai‘i at Mānoa.


