Laboratory Phonology
Journal of the Association for Laboratory Phonology

Broś, K. (2025). Remote data collection in the study of ongoing sound change in Spanish – a comparative analysis. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 16(1), pp. 1–49. DOI: https://doi.org/10.16995/labphon.10557

Open Library of Humanities

Remote data collection in the study of ongoing sound change in Spanish – a comparative analysis

Karolina Broś, University of Warsaw, Warsaw, Poland, k.bros@uw.edu.pl

In this paper, I look at Canary Islands Spanish /p b/ lenition from a comparative perspective by examining the speech of the same participants produced on different occasions and under different circumstances: A field experiment consisting of reading sentences, self-recorded reproductions of the same sentences, self-recorded monologues and instant messaging app recordings of spontaneous speech. The aim of the study was to test the viability of unguided self-recordings as samples used to study stop weakening and to find out whether the use of such a method helps minimize the observer's paradox to arrive at more naturalistic speech. The results of the study support the viability of self-recordings as a data collection method. In addition, the findings also show that while data collection via self-recordings poses some challenges, it also helps gather more naturalistic productions in the spontaneous monologue condition. The data lead to the conclusion that there is an interplay between task type and recording method, with the former playing a key role in changing speech styles and pronunciation patterns of Canarian Spanish speakers.

1 Introduction

When studying speech production, gathering data remotely is not a particularly easy task. Some linguists opt for making recordings online via Zoom, for example, or Microsoft Skype or Microsoft Teams sessions (Androutsopoulos & Staehr, 2018; Lupton, 2021; Freeman & De Decker, 2021a). Many researchers in (socio)phonetics and phonology have looked especially at questions related to the reliability of such methods in measuring f_0 and formants (Byrne & Foulkes, 2007; Bulgin et al., 2010; Zhang et al., 2020; Calder et al., 2022), apart from comparing different devices to each other and to professional recorders (e.g., Leemann et al., 2020; Sanker et al., 2021; Zhang et al., 2021). Another way of getting research material is to use personal recordings made by the participants with or without researcher supervision (Freeman & De Decker, 2021b; Zhang et al., 2021; Broś, 2023). These methods are promising, as they make it easier to gather data that, it is hoped, correspond closely to the speech used on an everyday basis. Here, research focused on the issue of social media (SM) and instant messaging (IM) recordings is particularly wanting and should be explored alongside other methods that will help tackle the technical and sociolinguistic issues pertaining to phonetic and phonological analysis.

Against this background, the aim of the present study was to compare stop lenition data produced by the same speakers in different types of recording situations. Crucially, I wanted to tap into the ways in which both the recording method *and* the recording situation affect speaker productions. Hence, instead of producing various types of recordings in a single session, I collected data on three separate occasions. The study involves Spanish stop lenition, a phenomenon typical of, but not limited to, connected spontaneous speech. The dialect under analysis is Canary Islands Spanish, in which both voiced and voiceless stops undergo weakening in intervocalic position (Trujillo, 1980; Oftedal, 1985; Broś et al., 2021). More specifically, this study compares the productions of /p/ and /b/ made by eight speakers representing the Canarian dialectal community on three separate occasions: One during a field experiment involving supervised recordings made on a laptop computer, a second made by the speakers themselves during a short unsupervised recording session, and a third made by the speakers themselves via an instant messaging app. Additionally, as will be explained in Section 2, I compared a subset of data with recordings produced during my fieldwork in the course of semi-structured interviews to more reliably disentangle recording type from task type.

Looking at different types of recordings enabled me to answer two major research questions. The first question addresses whether self-recordings are a viable method that can, in principle, be expanded to include more speakers and used to collect reliable acoustic data for the study of consonantal sound changes. Second, by comparing fieldwork and experimental data with self-recorded data, I wanted to know whether self-recorded speaker productions, compared to recordings made in the presence of a researcher, are more naturalistic, less hypercorrect and absent of the observer's paradox.

The paper is structured as follows. In Section 1.1, I present background information on the studied phenomenon and the parameters used to assess the degree of stop weakening based on acoustic data. Section 1.2 briefly discusses the observer's paradox and challenges related to naturalistic data extraction. Section 1.3 presents the study's assumptions and hypotheses. In Section 2, I describe the methodology and materials used in the study. This is followed by the presentation of the results in Section 3 and discussion of key findings and issues in Section 4. Section 5 presents conclusions.

1.1 Stop lenition in Canary Islands Spanish

Canary Islands Spanish is well-known for its extensive consonantal weakening, including voicing, approximantization, debuccalization and elision (Trujillo, 1980; Oftedal, 1985; Almeida & Díaz Alayón, 1988). In the case of voiced stops, the dialect follows the pattern typical of many other varieties of Spanish, i.e., approximantization in intervocalic position, with a substantial rate of elisions, especially within words, similar to that found in Caribbean dialects (Lipski, 1994; Hualde, 2005). Apart from that, voiceless stops undergo a parallel process of weakening in the dialect, by which /p t k/ are realized as partially or fully voiced, and sometimes even approximantized segments. The prevalence of /b d g/ weakening is much greater and amounts to over 95% in intervocalic positions, while /p t k/ tend to weaken with a probability closer to 50% (Broś et al., 2021, p. 21). The latter process has been shown to be suppressed to some extent in lab conditions. For instance, a study comparing /p t k/ weakening between lab recordings of sentences repeated after a native speaker and recordings made over an instant messaging app shows a substantial increase in the frequency of voicing in the latter case (Broś, 2023). This suggests that /p t k/ weakening can be suppressed by speakers and subject to the observer's paradox (see Section 1.2). Some examples of stop weakening found in Canary Islands Spanish can be found below.

(1) Stop weakening in Canary Islands Spanish – examples of native productions

la vaca	/la baka/	[la.ˈβ̞a.ka]	'the cow'
la barrera	/la bareɾa/	[la.β̞a.ˈre.ɾa]	'the wall'
la parte	/la parte/	[la.ˈpaɾ.te] / [la.ˈbaɾ.te]	'the part'
la paciencia	/la pasiensia/	[la.pa.ˈsjen.sja] / [la.ba.ˈsjen.sja]	'the patience'

As for the parameters used to assess weakening degree and frequency in stop lenition in Spanish, researchers have used visual inspection of the spectrograms to annotate partial or full voicing, and presence or absence of a burst—in the case of /p t k/, to compare weakened and unweakened stops, as well as presence or absence of formants in both /p t k/ and /b d g/ weakening to mark approximantization (e.g., Dalcher, 2008). Furthermore, constriction and consonant duration have been used to determine the amount of weakening in several studies (e.g., Dalcher, 2008;

Hualde et al., 2011). Finally, perhaps the most robust acoustic parameter used as a proxy for consonantal weakening in Spanish is some form of relative intensity of the target segment compared to the flanking vowels. This can take the form of an intensity ratio (Ortega-Llebaria, 2004; Colantoni & Marinescu, 2010; Carrasco et al., 2012), maximum velocity (Hualde et al., 2011) or intensity difference (Hualde et al., 2010; Parrell, 2010) and has been shown to correlate well with articulatory data measuring the degree of consonantal aperture (Parrell, 2010, 2011). Since intensity difference has been used to reliably show stop lenition in Spanish in numerous works, including previous work on the same dialect (e.g., Broś et al., 2021), I adopt this measurement in the present study.

1.2 Data collection methods and the observer's paradox

The observer's paradox within linguistics was first defined by William Labov (1972, p. 209) as a situation in which we want to know how speakers talk when not supervised by a researcher, yet this knowledge can only be gained by systematic observation. Thus, getting naturalistic data from speakers is a challenge for the researcher and can be exacerbated further by the need to get a controlled sample that allows a reliable (statistical) analysis of the collected data. The latter usually takes the form of a lab study or a similar research design. However, even when gathering data in the course of fieldwork, including when we speak the language or dialect in question, recording everyday speech is often difficult. This is because speakers tend to react to the presence of the researcher by suppressing certain linguistic features, using hypercorrection or a more formal register. The social setting (or recording environment) itself can also alter the behavior of the speaker (Wagner et al., 2015). The presence of a recording device, sitting in a lab or recording studio, the speaker's awareness that the recording will be heard or analyzed by someone, and the nature of the task (e.g., reading or repeating words and phrases), can all drive (subconscious) changes in production. These issues have been addressed in the literature using such notions as (conscious) 'attention to speech' (which is minimal in the most informal styles, i.e., the so-called *vernacular*, Labov, 1978), 'audience design' (i.e., the presence and roles of interlocutors or overhearers of a speaking situation, Bell, 1984) and 'researcher as audience' (Wilson, 1987), among others. In the latter case, the researcher may play different roles, depending on the research method, and may even be a member of the audience indirectly, via the recording device, which acts as a proxy in the minds of study participants. Thus, speakers may be affected during data collection in a variety of ways and determining what is 'natural speech' may be extremely difficult without a proper comparative approach.

Technological innovation can help us at least partially remedy the situation. For example, we can gather speech samples remotely with the use of the speakers' personal devices, which brings us closer to the everyday situations experienced by participants. Furthermore, by removing ourselves from the process of speech production, i.e., resigning from supervising data collection, we may even escape the observer's paradox altogether.

Due to the COVID-19 pandemic, many researchers turned to alternative ways of collecting data, and a number of new studies on the matter were published as a result. Thanks to these initiatives within the field of phonetics and phonology, we now know that some types of remote recordings are better than others. For instance, Freeman and De Decker (2021a) show that video conferencing apps such as Zoom may be a good way of collecting data for the analysis of general vowel arrangement properties, but researchers looking at small formant differences and vowel overlap, as well as nasalization, should proceed with caution. Some distortions and excessive within-data variation can pose a problem to acoustic analyses. Zhang et al. (2020) compared recordings made over Zoom with those produced over smartphones using lossless format apps such as Recorder or Awesome Voice Recorder, showing that smartphone use is more reliable and can serve for analyzing some parts of the speech signal, e.g., prosody. They also note that irregular waveforms and filtering artefacts resulting in temporal misalignments can be produced in the course of conversion of recordings made on Zoom, and that remote recordings are, in general, more reliable when analyzing vowels than tones. In a similar study, Zhang et al. (2021) show that, unlike smartphone recordings, Zoom recordings are burdened with sudden intensity drops, and thus smartphones may be more suitable for collecting data aimed at providing some types of phonetic analysis (e.g., f_0 and formant measurements).

The potential reliability of homemade recordings for F1 and F2 analysis was also confirmed by Freeman and De Decker (2021b) in a study that simulated the use of personal devices by speakers. Importantly, they recommend that laptops be used whenever possible, as they are more reliable than iPads or smartphones. Contrary to that, however, Sanker et al. (2021) make slightly different recommendations after comparing various types of devices and several types of recording software in a simultaneous recording setup. Their analysis shows that different types of distortions are produced when recording on an iPad, smartphone or laptop without an external microphone. They also point to Facebook Messenger, Skype and Zoom as software options that affect many aspects of the speech signal.[1] In their investigation, Sanker et al. go beyond vowel productions and compare many different phonetic parameters that play a role in acoustic analysis. These include center of gravity (crucial for the analysis of fricatives), harmonics-to-noise ratio (important when looking at voicing and noise parameters in the data), spectral tilt (that can involve intensity and f_0 distortions), and sound duration, among others. The authors of the study suggest that care be taken in making sure that possible distortions are taken into account, together with such issues as file format (lossless vs. lossy), sampling rate (preferably 44.1 kHz/s), filtering options and filter-related artefacts, as well as recording comparability (preferably, limiting recordings to one session and comparing the same recording types) when conducting remote fieldwork or using self-recordings.

[1] It is worth noting that as remote recording technologies improve, these studies may need to be revisited in the future.

As for the use of smartphone recordings in phonetic research, not much has been published so far beyond the controlled methodological studies mentioned above and looking at phenomena related to vowel production and nasalization. It is worth noting that the general suitability of smartphones as recording devices comparable to traditional means was demonstrated earlier, in 2011 (De Decker & Nycz, 2011). However, researchers should proceed with caution given the findings mentioned above. Moreover, using alternative devices instead of traditional fieldwork recorders is not the same as collecting data remotely. The building block added by the pandemic concerns the latter issue.

Yet another subject for linguists to take up is unsupervised recording, on which even less research has been done to date.[2] Gittelson et al. (2021) report one such study that looked at crowd-sourced data gathered via a dedicated phone application. The participants had to engage with the app and predetermined elicitation protocols, but they used their phones independently of third persons in order to take part in the study. The project involved an investigation of the sociolinguistic aspects of nonmodal phonation in English and, apart from helping researchers reach more participants than via traditional means, it showed promising results related to inter-speaker variation. According to the authors, the ability to collect big data from a more diverse sample helped discover new factors influencing the use of non-modal phonation. An earlier study by Hall-Lew and Boyd (2017) studied sociophonetic variation in sibilant production based on a small sample of self-recordings made by four speakers of American English. By comparing interviews and controlled speech samples produced under supervision with the self-recordings, the researchers concluded that self-recordings are suitable for intra-speaker comparisons and can give us insight on the range of productions that may not have been captured otherwise. They also note, however, that such recordings may be produced with a range of styles across speakers, which should be taken into account in inter-speaker comparisons.

Finally, a recent study by Broś (2023) taps into intra-speaker differences in the production of voiceless stops in Canary Islands Spanish depending on whether the speech was supervised and recorded in a lab setting or came from unsupervised recordings taken from an IM (instant messaging) application (WhatsApp). Apart from looking at the suitability of such recordings for phonetic analysis, the most important question was whether and to what degree native speakers suppress /p t k/ lenition in lab speech. Since WhatsApp samples were examples of natural speech and participants had no expectation that their recordings would undergo third-party analysis, Broś assumed that these samples represented everyday speech spoken by representatives of the dialect. The study showed that the two tested recording types are suitable for intensity-based comparisons and phonetic annotation. Based on the voicing and other parameters of the target

[2] It is worth mentioning sociolinguistic studies that focus on speaking styles and registers, and intra-speaker variation, such as e.g., Podesva (2007; 2011a; 2011b) or Sharma (2011). The primary interest of all those studies, however, was the social dimension of speech and not phonetic or phonological generalizations.

segments, a discrepancy between the recording types was discovered, with significantly more lenition taking place when speaking via IM and, importantly, a levelling effect between speakers. While the number of weakened sounds and the degree of weakening differed substantially from speaker to speaker in lab speech, all speakers produced the same amount and type of lenitions when talking to their friends over the phone application.[3] Overall, the results of the study show what we potentially lose when collecting lab speech and how this affects our generalizations concerning the way in which a given speech community talks. Potentially, we may significantly underestimate the frequency of a given process or overlook some contexts of application of a phonetic or phonological rule. This is especially important when studying language variation and change.

In the present study, I use recorded speech produced over WhatsApp as a baseline condition that, presumably, best corresponds to the actual speech of the inhabitants of the Canary Islands. Apart from that, I test a guided self-recording method as a possible means towards reaching this baseline condition and getting more natural speech compared to data produced during a field experiment consisting in reading sentences. Finally, I use fieldwork recordings of spontaneous speech to disentangle recording type from task type. With such a multivariate comparison, we can potentially work out an optimal method of data collection to discover the speaking habits closer to a community's everyday speech behavior.

1.3 The current study: Goals and hypotheses

In this study, I use bilabial stops for two reasons. First, only /p b/ were used in the experimental condition that I take to compare recording types (supervised vs. unsupervised) and tasks (scripted sentences vs. spontaneous productions). The second reason is phonetic. In general, labials are more suitable for across-task comparisons, given that they are less susceptible to the effects of vowel coarticulation. Moreover, it has been suggested that labial stops are the most compatible with voicing compared to other stop series (Ohala 1983, p. 195) and while their constriction is the longest, their voicing durations can be easily prolonged by passive or active vocal tract expansion. In the same vein, Maddieson (1984, p. 36) states that "voicing is more readily sustained in a bilabial plosive than in any other," because the air can keep flowing into the oral cavity for a longer period before air pressure equals subglottal pressure when the closure is far away, i.e., at the lips. This facility in voicing and the challenging task of maintaining voicing in other stops, especially velars, is supported by the data from the dialect. Broś and Lipowska (2019) have shown that while all stops are voiced by Gran Canarian speakers in post-vocalic position around 45% of the time, /p/ is the most likely to be voiced, while /k/ tends to be shortened more often, and/or approximantized instead.

[3] Actually, one speaker had even more lenition than the rest in that condition, reaching almost 100% of voicing, and was the heaviest voicer of /p t k/ in the lab (around 80%, compared to the sample mean of 44%).

As the main goal of this study is to compare different recordings and establish the viability of self-recordings for studying lenition, recording type becomes a crucial variable. However, research on the differences between self-recorded speech and controlled recording sessions in the context of particular phonetic and phonological phenomena is limited. There are only a few small-scale studies that might inform us on what is to be expected of self-recorded speech. Boyd et al. (2015), for instance, have shown that self-recordings present the most advanced vocalic changes corresponding to the California Vowel Shift compared to recorded interviews. Similarly, the study by Hall-Lew and Boyd (2017), described in Section 1.2, which analyzed /s/ productions by California English speakers, showed that self-recordings can help uncover a wider range of phonetic variants produced by a given community that surpasses what is revealed by interview data. Thus, it may be assumed that self-recordings should be the closest to natural productions made by speakers for purposes other than linguistic analysis upon researcher instruction. Assuming that IM recordings are instances of natural speech *par excellence* (as suggested in previous work by Broś, 2023), I would like to test the hypothesis that self-recordings will be closer to this ideal compared to experimental data, both in the acoustic lenition marker (i.e., intensity difference) and in the frequency of lenition.

Furthermore, the recording types used in this study also differ in what type of speech (or task) is used. Presumably, scripted speech in the form of sentences provided by the experimenter may induce less naturalistic productions whose features differ compared to monologues or interviews. While Martínez-Gil (2020) argues that /b d g/ approximantization takes place even in slow and careful speech, there is some experimental work that shows task effects with differences in the degree of constriction of /b d g/ between conversational speech and reading tasks in monolingual and heritage Spanish speakers (e.g., Carrasco et al., 2012; Lozano, 2021). As for voiceless stops, Lewis (2001) found a significant difference in voiceless stop productions between a word-list reading task and conversational speech among Colombian Spanish speakers. Similarly, Hualde et al. (2011) compared scripted and unscripted data, showing that spontaneous speech results in more /p t k/ voicing among Spanish speakers from Majorca. There is also some evidence from the dialect studied here: Herrera Santana (1997) tested Gran Canarian speakers and reported that when asked to read lists of words containing intervocalic voiceless stops, speakers voiced them only rarely. Given this evidence, I assume that scripted speech (read or repeated sentences) will induce less lenition compared to spontaneous speech (monologues and interviews).

Finally, this study also assessed how a person's speech rate, which is tied to speaking styles and registers, can change depending on the type of recording used for gathering data. Since speakers use different modes of speech that are influenced by whether or not their produced sentences are scripted, and whether a third person (here, the experimenter) is directly monitoring their speech, I assume that differing speech rates will be found in the different recording samples analyzed in this study. Moreover, higher speech rates induce shorter sound durations and, indirectly, greater coarticulation, gestural overlap and/or undershoot (Cohen Priva, 2017). This may translate into more advanced lenition, as manifested by acoustic parameters (e.g., a smaller intensity difference, more voicing, etc.) and/or into a higher probability of applying lenition in a given context. Previous studies have shown that speech rate plays a role in lenition in general (Cohen Priva & Gleason, 2020), and in Spanish in particular (Soler & Romero, 1999; Hualde et al., 2011; Nadeu & Hualde, 2015; Melero-García, 2021; see also Broś et al., 2021, Section 5.5). Thus, speech rate should be included in the statistical analyses comparing tasks and recordings. I expect that higher speech rates will increase the probability and amount of voicing, whereas in the case of /b/, I expect that higher speech rates will translate into more deletions.

The hypotheses to be tested in this study are as follows.

Hypothesis 1
Speakers differ in their production of underlying /p/ and /b/ across all recording types: Self-recordings show a smaller intensity difference, i.e., more lenition, compared to the field experiment, with values closer to IM recordings.

Hypothesis 2
The frequency of occurrence of weakening depends on the recording: a) In the case of /p/, there is more voicing in self-recordings compared to the experiment and more voicing still in IM recordings; b) in the case of /b/, there is more deletion in self-recordings compared to the experiment, and even more in IM recordings.

Hypothesis 3
Speakers differ in their production of underlying /p/ and /b/ depending on task type: Spontaneous speech (monologues) will be correlated with more lenition compared to scripted speech (sentences), regardless of recording type. This will result in:

- a smaller intensity difference
- a greater frequency of voicing in /p/
- a greater frequency of deletion in /b/

In all cases, I assume a qualitative difference between underlying /p/ and /b/, i.e., a greater intensity difference in /p/ than in /b/, given that the former is typically produced as either a voiceless or a voiced stop and the latter – as an approximant (when not deleted).

Hypothesis 4
Since /p b/ weakening is a connected speech phenomenon and has been shown to be sensitive to the durational properties of language, I expect speech rate to modulate weakening in both segments. The higher the speech rate, the greater the probability of lenition (by voicing or deletion, accordingly) and the smaller the intensity difference.

2 Method

2.1 Materials and equipment

The data analyzed in this paper come from eight native speakers of Canary Islands Spanish (six males, two females) who participated in a field experiment in 2021 (Broś & Krause, 2024). Twelve participants from that study were asked to provide self-recordings so that a comparative analysis could be performed. Only eight of them completed the task. The remaining four either refused for lack of time or were unresponsive.

The self-recording consisted of two parts. First, each participant was asked to talk about their job, recent vacation or any other topic of choice for approximately one minute. As a second task, they were asked to listen to a recording with 40 sentences and repeat what they heard. A speaker of the dialect produced the sentences for this recording. A period of silence was inserted after each sentence to give the participant enough time to repeat it. The silence was longer than the sentence itself.

To ensure data comparability, I selected two independent productions of the same 20 sentences taken from the 2021 study, containing 17 target words with an intervocalic /p/ and 13 words with an intervocalic /b/.[4] Each target segment appeared at the beginning of a word, flanked by two vowels /a/. Examples are provided below.

Sample sentences from the study:

> *L**a bar**rera estaba mal colocada y el portero no veía.*
> ‘The wall was incorrectly placed, and the goalkeeper could not see’.
>
> *L**a pa**trulla ha encontrado el ladrón en l**a ba**rca.*
> ‘The patrol found the robber on the boat’.

For maximal comparability, the same sentences as those used in the self-recordings were analyzed in the experiment condition. The number of segments considered, however, is greater due to the greater number of repetitions (between five and ten) produced during the experiment.

The present study included an additional sample for six participants. Since I had access to WhatsApp recordings of these persons, and such data have been deemed good examples of

[4] The full list of sentences used in the study is provided in the Appendix. In principle, I chose ten sentences per segment, in which either /p/ or /b/ was found in stressed or unstressed position. The words were evenly distributed. Additionally, I also annotated other /p b/ segments that were found in the same phonetic environment, i.e., at the beginning of a word between two /a/ vowels. This gave me additional seven /p/ occurrences and three /b/ occurrences per participant. Thus, there were, at most, 60 observations per participant in the self-recorded sentences condition. All the analyzed segments are marked in bold in the sentence list in the Appendix.

naturalistic speech in the dialect (Broś, 2023), I decided to use them as a baseline in the analyses. These examples consisted of recordings made in a group conversation between the participants and, in some cases, recordings exchanged with me, via WhatsApp. The participants agreed to an analysis of the recordings for the purposes of the study. To ensure maximal comparability of the results, I only considered bilabials /p b/.

Samples 1 to 3, i.e., the field experiment and self-recordings, were recorded using Audacity[5] or WaveEditor for Android[6] in WAV format. Initially, I had planned for all self-recordings to be made using Audacity on a laptop computer, but this was impossible for three speakers, as they do not have laptops. Thus, I instructed them to use a smartphone app instead. After some research, I tested and then recommended WaveEditor, which allows for good quality recordings in WAV format at a high sampling rate (48 kHz/s in its default setting). Audacity was used with the default settings and a 44.1 kHz sampling frequency. The participants were asked to save the files in WAV format and send them via email or a cloud drive. In principle, I wanted to make sure samples were comparable. Since the field experiment had recordings made using Audacity on a laptop computer, the intent was to use the same tools in the present study, but in a self-recording design in which speakers independently use the software, record themselves and provide the resultant files.[7] Sample 4 consisted of recordings in OPUS format, which is a lossless file type used by WhatsApp.[8] The recordings were then converted to WAV for analysis in Praat (Boersma & Weenink, 2022).

Finally, given the intra-speaker differences found between sentence productions and spontaneous speech (monologues), I decided to include a fifth data sample to disentangle recording type from task type. This was possible on a subset of data, given that four of the participants had earlier participated in fieldwork interviews with me. Those were conducted in the same locality as the experiment, but with a professional recording device (a Zoom H4N digital recorder with a Shure SM10a headworn microphone with a 44.1 kHz sampling frequency). Since the samples were made in the presence of the experimenter who also controlled them, they constitute a good comparative sample for a 2 × 2 (recording x task) design, in which field sentences and field monologues may be compared to self-recorded sentences and self-recorded monologues.

[5] http://audacityteam.org.

[6] Sound-Base Audio, LLC.

[7] Audacity was deemed appropriate to obtain good quality recordings when using a laptop computer in a recent study by Sanker et al. (2021). Although a built-in mic should be used with caution according to the authors of that study, I believe that the setup used in my study was suitable, given my particular research questions (possible distortions reported by Sanker and colleagues were avoided by default, given the acoustic parameters I look at), and the need to ensure comparability and replicability of the results vis à vis the field experiment.

[8] https://www.whatsapp.com.

Thus, all in all, I analyzed five data samples:

a) 8 field experiment recordings of read sentences

b) 8 self-recordings of the same sentences in a repeating condition

c) 8 self-recordings of spontaneous speech (monologue)

d) 6 instant messaging app recordings made via WhatsApp

e) 4 field recordings from semi-structured interviews (monologue)

All recordings were annotated in Praat. All target segments and flanking vowels were entered in the TextGrids together with the information on voicing in the case of /p/; deletion rate in the case of /b/, stress, and word in which a given segment occurred; and the sentence it belonged to. A sample annotation is presented in the Appendix.

Following previous work on Spanish stop weakening in the same dialect (Broś et al., 2021), target segments were deemed voiced based on the visual inspection of the waveforms and spectrograms when the voicing bar and pulses were present in more than 50% of the total duration of the sound. The beginning of the stop corresponded to the end of the preceding vowel, and the end was marked at the beginning of the periodic cycle of the following vowel. Approximant pronunciations were distinguished by the lack of burst, full voicing and presence of formants on the spectrogram.[9] To delimit these segments, I followed previous work (e.g., Eddington, 2011; Hualde et al., 2011) by looking at the dip in the intensity contour from vowel to consonant and a subsequent inversion of this trend in the transition to the following vowel. When the intensity curve seemed flat and the differences in intensity between the flanking vowels were too small to reliably decide whether an approximant was produced in between, I deemed the segment deleted.

A custom-made Praat script extracted the segmental and sentence information, together with sound durations and intensity measurements in dB extracted from Praat's intensity object. This gave me a total of 2,377 observations from eight speakers in the main analysis database (recordings 1–4). Of those, the majority came from the experiment (n = 1467), followed by self-recorded sentences (n = 526), self-recorded monologues (n = 229) and IM recordings (n = 155). In the latter case, the observations come from a subset of six speakers, as mentioned before. The fifth recording gave me an additional pool of 434 observations. Given data subsetting depending on the model, the number of observations taken under analysis is stated separately for each case in Section 3.

[9] Note, however, that the first two features alone do not determine approximantization, as both voiced and voiceless stops in Canary Islands Spanish are often produced without a burst (see Broś et al., 2021 for details). Approximants have a characteristic waveform and a weak to strong formant structure depending on the degree of aperture, as reflected in several acoustic parameters, including intensity and HNR.

Table 1 lists the number of segments per participant and recording type, deleted /b/ sounds included. Note that spontaneous recordings contain fewer segments, as speakers do not necessarily use intervocalic /p b/ in their productions.[10]

Participant	Experiment	Self-recorded sentences	Self-recorded monologue	Instant messaging	Fieldwork interviews	Total
P6	181	67	24	19	0	291
P7	200	60	13	0	0	273
P10	187	63	58	0	0	308
P14	182	65	28	36	0	311
P17	181	68	16	10	118	393
P19	173	71	38	8	68	358
P20	192	69	17	20	83	381
P21	171	63	35	62	165	496
Total	**1467**	**526**	**229**	**155**	**434**	**2811**

Table 1: Number of intervocalic /p b/ segments analyzed per participant.[10]

2.2 Data analysis

The compiled database was analyzed based on the annotated parameters and intensity measurements taken from Praat. Based on the raw measurements, relative intensity difference (calculated as the minimum intensity of the target segment subtracted from the maximum intensity of the preceding vowel) was added to the database. It is worth noting that intensity difference can be measured comparing the target segment with either the preceding or the following vowel. The former method was used, e.g., by Martínez-Celdrán and Regueira (2008), Figueroa and Evans (2015) and Broś et al. (2021), while the latter was used by Hualde et al. (2011) and Carrasco et al. (2012), among others. I used the difference between the minimum intensity of the target consonant and the maximum intensity of the *preceding* vowel, as this measurement is less sensitive to intensity differences in the vocalic segment that may result from word stress. The choice of this measurement method is especially important in the case of sentence reading or repetition, as both stressed and unstressed syllable conditions are included.[11]

[10] Participant code names were kept the same as in the field experiment.

[11] For instance, *la banda*, ‘the band’, with a stressed vowel vs. *la barrera,* ‘the wall’, with an unstressed vowel; in both cases the vowel in *la* is unstressed.

I also calculated speech rate as a mean of ten measurements taken randomly from sentences produced in each recorded sample, excluding pauses. I made sure that the measurements included sentences from various parts of the recording and excluded short phrases ridden with hesitations or prolonged vowels that might falsify the speaking rate. The resulting average number of syllables per second was established for each recording and added in statistical model formulae to test the role of speaking speed in the application of lenition in the dialect.

The database was then subjected to a statistical analysis in R (R Core Team, 2020) using the packages *lme4* (Bates et al., 2018) for building models and *emmeans* (Lenth, 2019), for the calculation of simple effects. Descriptive plots and effects plots were generated using the *ggplot2* (Wickham, 2016) and *ggeffects* (Lüdecke, 2018) packages, respectively.

I ran three initial models on the main database: 1) a linear mixed-effects model, with intensity difference as a dependent variable; 2) a binomial logistic mixed-effects regression model, with voicing (binary) as a dependent variable to test the probability of voicing in /p/ depending on the speech sample; and 3) a binomial logistic mixed-effects regression model, with deletion (binary) as a dependent variable to explore the probability of deletion of /b/ depending on the speech sample. After that, I built three follow-up models on a subset of data containing four speakers, with fieldwork interview data included and instant messaging excluded to disentangle recording type from task type.

Optimal models were achieved by using maximal model structures and the backward direction of the *step()* function (linear mixed models) or by model comparison using *anova()* (logistic regressions). The exact formulae of all the models are provided in Section 3. All model assumptions were checked, including normal distribution of residuals in linear mixed models. I also used pairwise comparisons available in the *emmeans* package to correct for multiple tests.

3 Results

I present the results in the order of the models fitted for the purposes of the study. First, I present the results for intensity difference, i.e., the amount of lenition observed in the data, assuming that a smaller intensity difference corresponds to more lenition. I explore the roles played by underlying segment, speech rate and recording. I then look at the frequency of voicing in /p/, and frequency of deletion in /b/, in separate models. Finally, I add fieldwork interviews to explore the effects of task type versus recording type, in accordance with Hypothesis 3.

3.1 Intensity difference

The comparison of the main dataset (four samples) in terms of the key lenition parameter, i.e., intensity difference, shows that all recording types are comparable (see **Figure 1**). Crucially, there is a clear difference between underlying /p/ and /b/, as expected. Underlying voiced stops are quite similar, with values corresponding to weak approximants (~5 dB), while underlying voiceless stops are realized in the range of stops (15–25 dB). It must be remembered that these ranges vary depending on whether the segments are voiced or not. The former would get intensity difference closer to 15 dB, while tense voiceless pronunciations can get as high as 30 dB, or more, on the scale. These differences will be explored below.

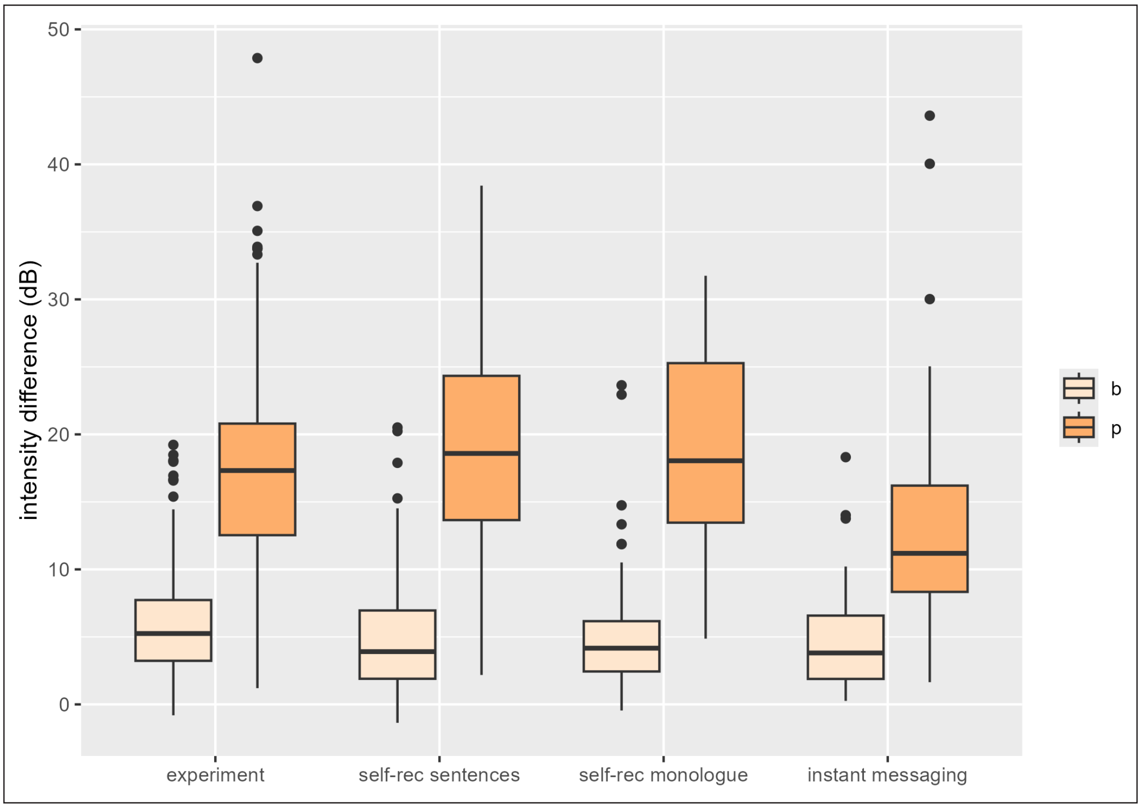

Figure 1: Comparison of intensity difference values for /p/ and /b/ across the four types of recordings.

In **Figure 1**, we can see that IM recordings have the lowest values for /p/ compared to other recordings, which may be due to the different distributions of voiced vs. voiceless /p/ realizations, or to other factors, e.g., the distance between the mouth and the microphone. Additionally, IM recordings seem to have slightly smaller between-category differences, i.e., the distance between the realizations of /p/ and /b/. Nonetheless, this is due to the smaller intensity difference in /p/. The values for /b/ are virtually the same as those of other recordings.

To explore the in-between categories, I looked at density plots. In **Figure 2**, we can see three distinct sound categories in the data: voiceless stops (with the greatest intensity difference), voiced stops (in the mid-range) and voiced approximants (the lowest values). Approximants come from underlying /b/, while voiced and voiceless stops are realizations of underlying /p/.[12] The density plots also make it immediately clear that the type of task and recording do play a role in how /p b/ are produced: While there is a visible trimodal distribution of the three types of surface segments in spontaneous speech, an overlap between voiced and voiceless stop realizations is more pronounced when sentences are read aloud or repeated after a recording, which suggests that both [b] and [p] coming from /p/ are realized with similar intensity scores. Also, note that in IM recordings, there are three visible categories that are nevertheless much closer together on the intensity difference scale. Importantly, voiced realizations of /p/ have much lower intensity scores than their counterparts in other recordings (~10dB vs ~15dB). I will return to these particular results later in this section.

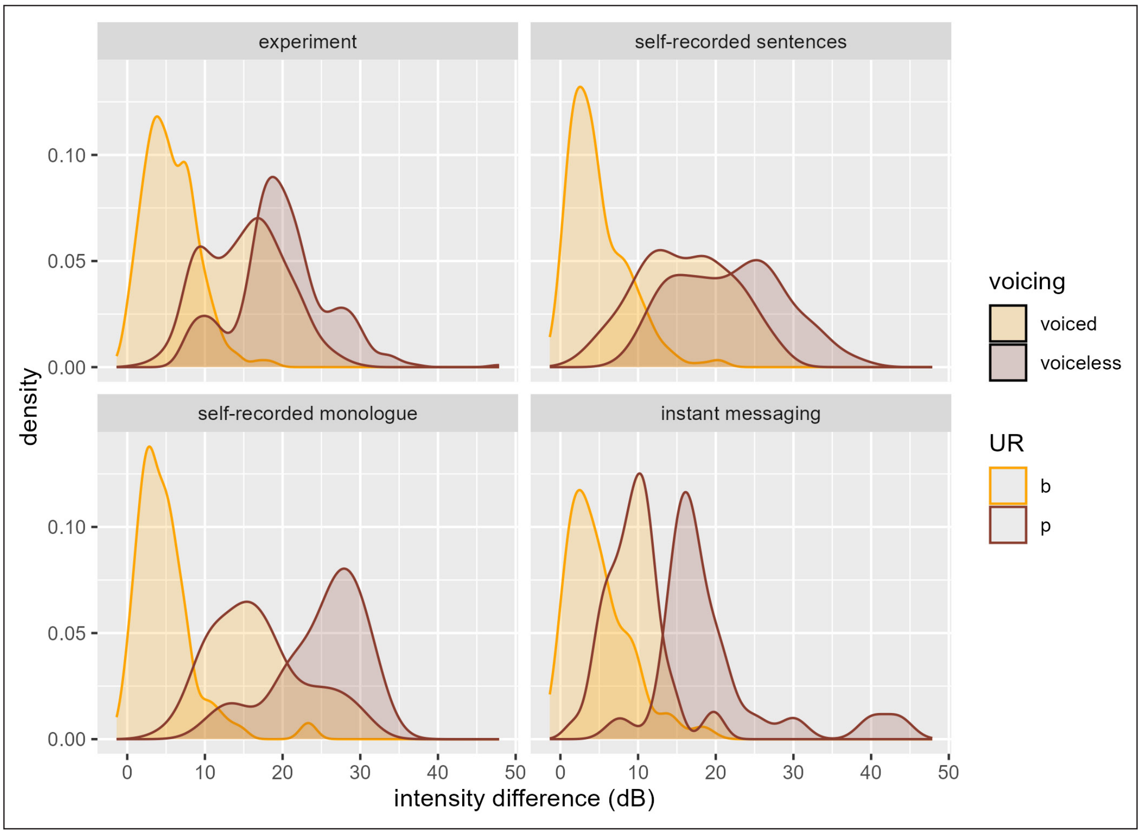

Figure 2: Density plots for intensity difference across recording types by voicing and underlying segment.

[12] This is the expected result. Since all /b/ were intervocalic, I did not expect many unweakened realizations (there were exactly seven of them). As for the /p/, while it is possible to have them realized as approximants intervocalically (see Broś et al., 2021), in this database the number of such sounds was negligible (n = 12). Rather, these sounds were usually realized either as voiceless stops, or as partially or fully voiced weakened stops, with or without a burst (for similar results, see also Broś, 2023).

As mentioned in the Introduction, speech rate is one factor that may affect the amount of voicing and the acoustic parameters. The general trend is that experiment data are the fastest while, quite surprisingly, self-recordings are the slowest (see **Figure 3**).

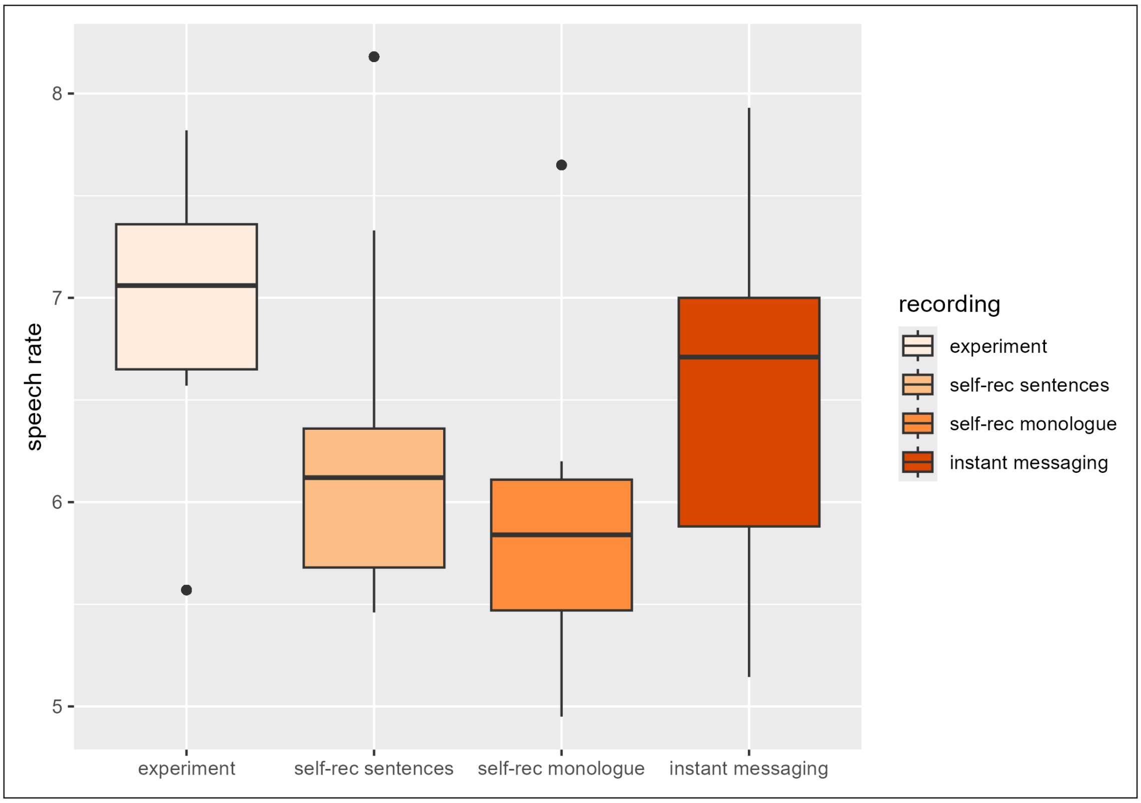

Figure 3: Speech rate depending on recording type.

3.1.1 Model 1: Predicting lenition degree, as marked by intensity difference

The abovementioned observations were tested using the following linear mixed-effects model:

> *intensity difference ~ recording type + underlying segment + speech rate + recording type : underlying segment + recording type : speech rate + (1 + underlying segment | participant) + (1 | word)*

Model 1 was run on 2,207 observations and fitted by REML, with *t*-tests using Satterthwaite's method. The reference value of recording type was experiment, and the reference value of underlying segment was /p/. The results presented in **Table 2** show significant effects of underlying segment and recording (where self-recorded sentences and self-recorded monologues, but not IM recordings, differed significantly from the experiment). Speech rate was not significant. There were also significant interactions between all levels of recording type in underlying /b/ compared to /p/ in experiment recordings, and a significant interaction between self-recordings and speech rate compared to the experiment.

In the case of recording, pairwise comparisons based on the marginal means from the model show that there was a significant difference in the values of intensity difference between the experiment and IM recordings ($\beta = 2.356$, $df = 232$, $t = 4.093$, $p < .001$), between self-recorded sentences and IM ($\beta = 2.741$, $df = 274$, $t = 4.525$, $p < .001$), and between self-recorded monologues and IM ($\beta = 3.590$, $df = 1005$, $t = 5.152$, $p < .001$), but not between self-recordings and the experiment, or self-recorded sentences and self-recorded monologues. As for the overall variance explained by the model, $R^2 = 0.543$, most of it can be attributed to the underlying segment.

	estimate	**std. error**	**df**	***t* value**	***p* value**
(Intercept)	18.962	2.419	44.764	7.838	<.001
self-recorded sentences	6.45	2.116	858.071	3.048	<.01
self-recorded monologue	–7.899	3.673	325.889	–2.15	<.05
instant messaging	3.022	4.729	564.151	0.639	.523
/b/	–11.045	1.313	11.143	–8.413	<.001
speech rate	–0.315	0.286	60.961	–1.102	.275
self-recorded sentences : /b/	–2.887	0.474	2100.838	–6.092	<.001
self-recorded monologue : /b/	–2.125	0.915	79.585	–2.323	<.05
instant messaging : /b/	2.841	1.118	116.087	2.541	<.05
self-recorded sentences : speech rate	–0.693	0.314	1035.418	–2.204	<.05
self-recorded monologue : speech rate	1.528	0.589	364.241	2.593	<.05
instant messaging : speech rate	–1.019	0.719	593.439	–1.416	.157

Table 2: Summary of results of the first model.

To explore the interactions, I looked at the contrasts generated using marginal means from the model. They show that all differences between recording types for /p/ productions are significant, except for the difference between self-recorded sentences and self-recorded monologues ($\beta = -.469$, $df = 382.09$, $t = -0.611$, $p = .999$). All differences between recording types and underlying segments were significant. As for the differences in /b/ productions across recording types, they were not significant. In other words, /b/ was produced the same way across recording types, while /p/ differed across conditions. Contrasts based on the marginal means calculated for the interaction of the two variables are presented in **Table 8**, in the Appendix. A graphical illustration of this interaction can be appreciated in **Figure 4**, while **Figure 5** shows the interaction between speech rate and recording. As expected, there is a slight tendency in the direction of a smaller intensity difference with increasing the speech rate. However, self-recorded monologues defy this, showing an opposite trend.

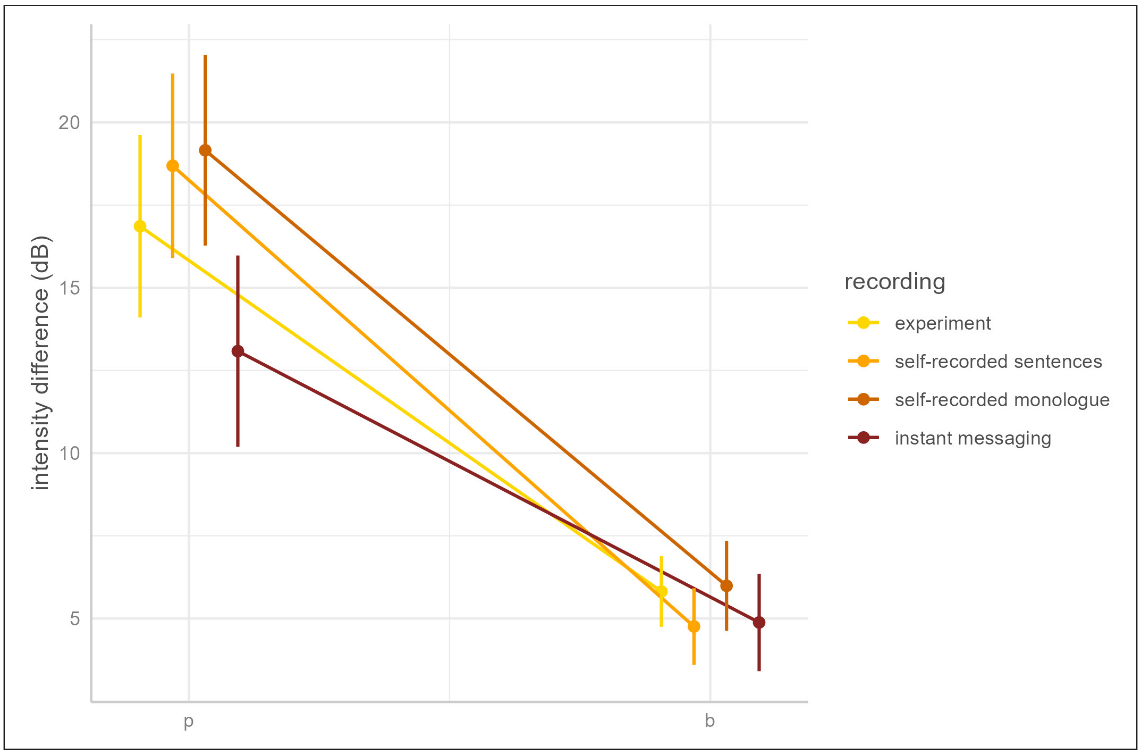

Figure 4: Effects plot for the interaction between recording type and underlying segment (/p/ or /b/).

3.2 Frequency of voicing

Descriptive statistics show that the weakening of /p/ is variable and depends on the speaking situation, as expected from the literature (see **Figure 6**). Self-recorded monologues had the greatest number of lenited /p/ segments, and IM recordings had the second highest number Sentences induced more unweakened pronunciations.

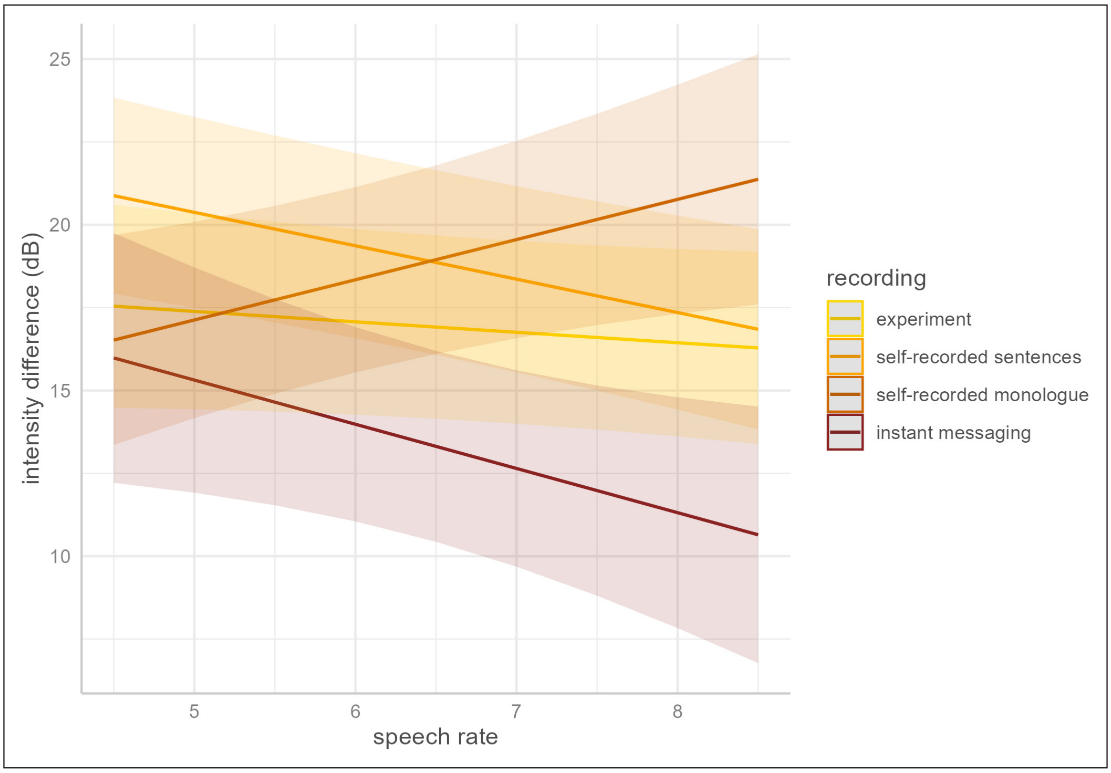

Figure 5: Effects plot for the interaction between recording type and speech rate.

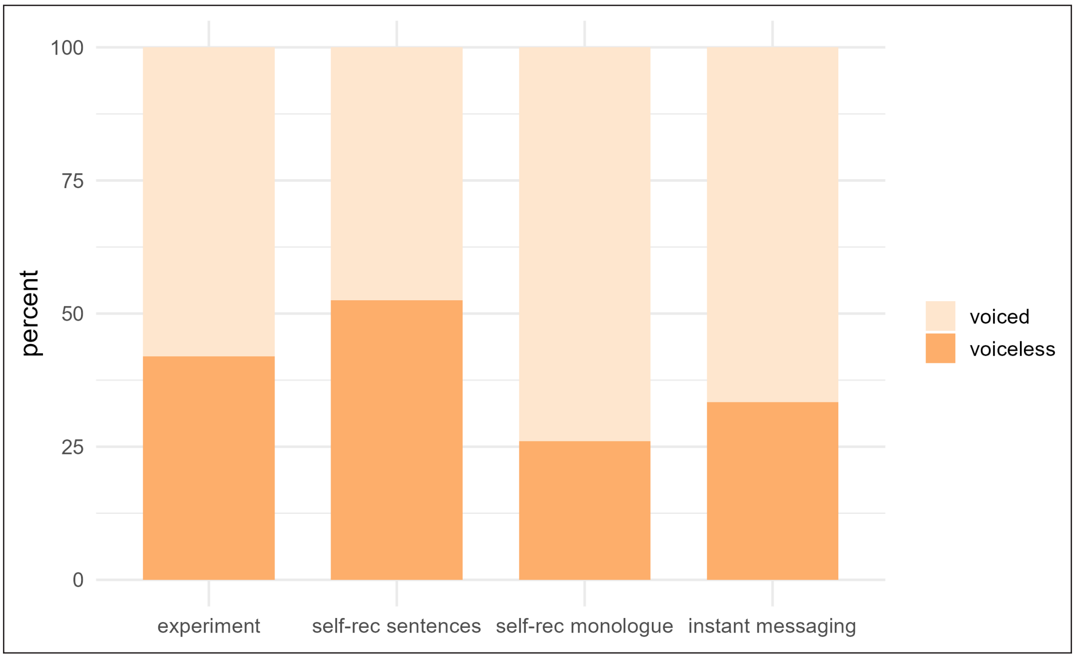

Figure 6: Percentage of voicing produced by the speakers depending on recording type. The data are taken from /p/ production only.

It is worth noting that speakers differ in the number of voiced sounds they produce. **Figure 7** shows that the general tendency is to increase the amount of voicing in spontaneous speech compared to producing scripted sentences, regardless of how much voicing is produced in the 'unnatural' experiment condition, with the exception of two participants (P14 has virtually no change, and P7 shows a reverse trend). Moreover, half of the speakers go from nearly 100% to fully 100% of voicing in spontaneous speech.

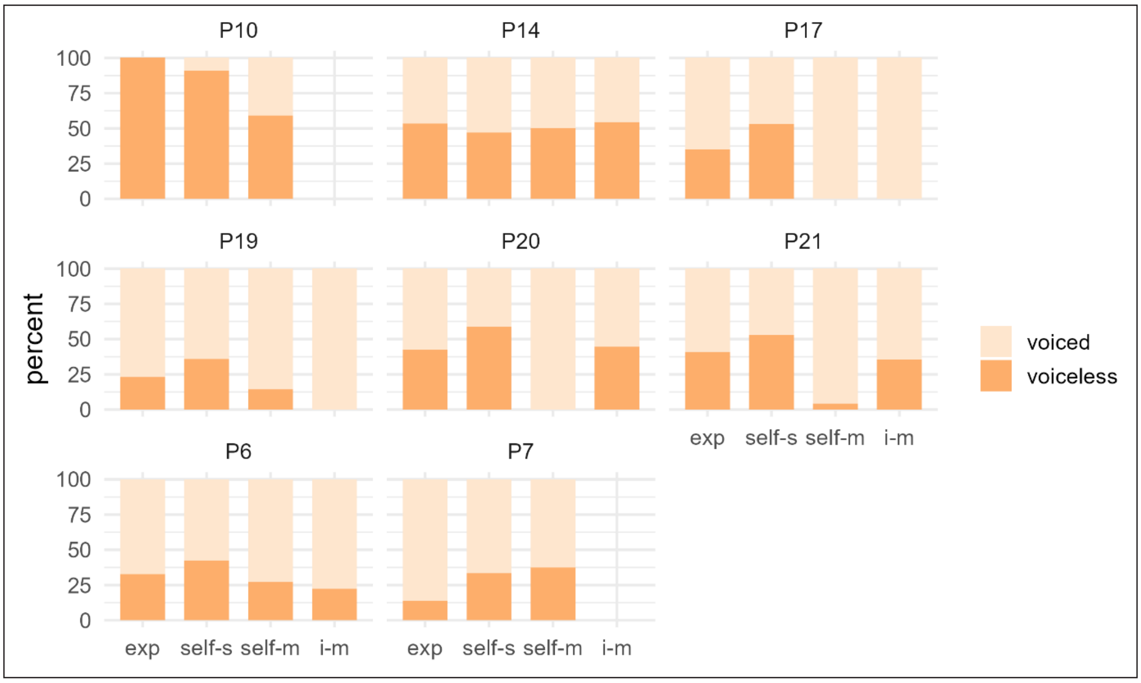

Figure 7: Comparison of voicing frequencies presented by each speaker in each recording type. The figure shows data for /p/ only. Recording types: exp = experiment, self-s = self-recorded sentences, self-m = self-recorded monologue, i-m = instant messaging. Note that only six speakers provided IM recordings, hence not all speakers have a bar corresponding to this recording type.

3.2.1 Model 2: Predicting the probability of voicing

A subset of data (N = 1225) was used to predict the probability of voicing in /p/ depending on speech rate and recording type. Model 2 was fit using *glmer()* with the binomial function and the 'bobyqa' optimizer, nAGQ set to 10:

voicing ~ recording + speech rate + (1 | participant)[13]

[13] Importantly, when speech rate is added as an interaction term in this and other logistic regressions from this analysis, the VIF inflates to impossible values (over a few hundred), which suggests multicollinearity and unreliable model results. It also leads to suspiciously high odds ratios and very large confidence interval ranges. After inspecting the data for outliers and other possibilities, I attribute these problems to small sample size and only add speech rate as a fixed factor when it improves model fit.

The results of the model fit show a significant effect of recording for the monologue condition compared to the experiment but no other significant effects (see **Table 3**). Pairwise comparisons based on the marginal means from the model confirm the significant difference in the probability of voicing between the experiment and the self-recorded monologues (β = –1.536, z = –4.501, p < .001). They also show a significant difference between self-recorded sentences and self-recorded monologues (β = –1.894, z = –5.564, p < .001), *and* between self-recorded monologues and IM recordings (β = 1.158, z = 2.821, p < .05; see **Table 10** in the Appendix). **Figure 8** shows the effect plot from the model.

	estimate	**std. error**	***z* value**	***p* value**
(Intercept)	–0.739	1.217	–0.607	.544
self-recorded sentences	–0.358	0.189	–1.896	.058
self-recorded monologue	1.537	0.341	4.502	<.001
IM	0.378	0.284	1.329	.184
speech rate	0.141	0.166	0.851	.395

Table 3: Estimates from Model 2. *Experiment* was the reference level of the recording type variable.

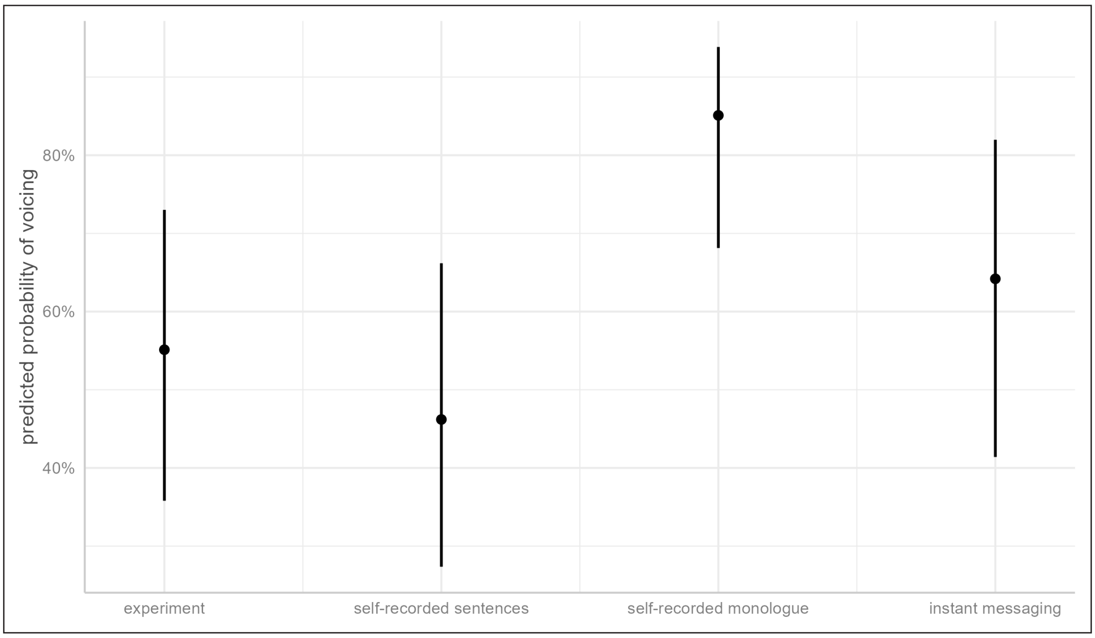

Figure 8: Predicted probabilities of voicing depending on recording type.

3.3 Frequency of /b/ deletion

While the weakening of this segment happens nearly 100% of the time, most of it takes the form of approximantization. In some cases, however, /b/ is elided and the frequency of this process tends to vary depending on the speaker and the recording situation. **Figure 9** shows that /b/ deletion is not that frequent and only goes above 25% in IM. There seems to be no effect of task type, as self-recorded monologues do not have higher rates than sentences.

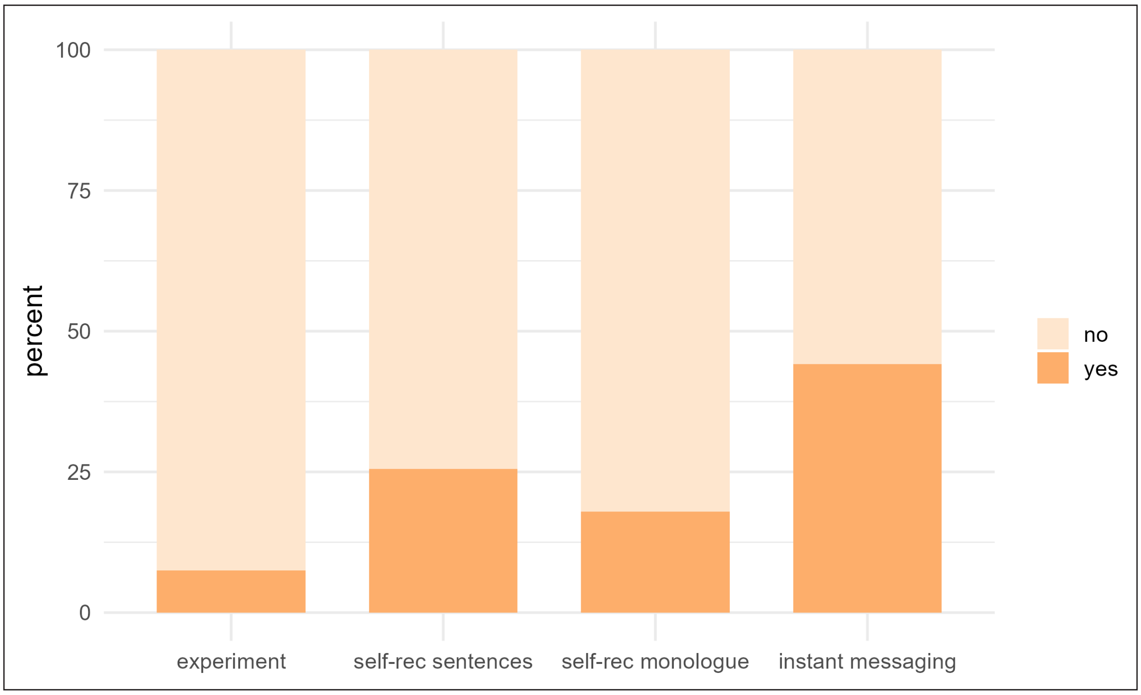

Figure 9: Rate of /b/ deletion depending on recording type. The data are taken from /b/ productions only. Deletions are marked as 'yes', while /b/ retention by approximantization is marked as 'no'.

3.3.1 Model 3: Predicting the probability of deletion

The above was tested via a binomial logistic regression analogous to Model 2, run on the /b/ database (N = 1152).

$$\textit{deletion} \sim \textit{recording} + \textit{speech rate} + \textit{(1} \mid \textit{participant)}$$

The results show a main effect of recording, with the probability of deletion being greater in all types of recordings compared to the experiment (see **Table 4**). The effect of speech rate was not significant. Pairwise comparisons show significant differences between the experiment and self-recorded monologues (β = –1.657, z = –5.511, p < .001), self-recorded sentences (β = –1.032, z = –4.544, $p < .001$) and IM recordings ($\beta = –2.611$, $z = –8.701$, $p < .001$), as well as a significant difference between self-recorded monologues and IM ($\beta = –0.955$, $z = –2.820$, $p < .05$) and between self-recorded sentences and IM ($\beta = –1.580$, $z = –5.214$, $p < .001$), but not between the self-recorded monologues and the self-recorded sentences ($\beta = 0.625$, $z = 2.176$, $p = .130$, see **Table 12** in the Appendix). The effect plot showing the probability of deletion depending on the recording is presented in **Figure 10**.

	estimate	**std. error**	***z* value**	***p* value**
(Intercept)	–4.098	1.016	–4.033	<.001
self-recorded sentences	1.657	0.301	5.511	<.001
self-recorded monologue	1.032	0.227	4.544	<.001
IM	2.611	0.3	8.701	<.001
speech rate	0.208	0.142	1.465	.143

Table 4: Estimates from Model 3. *Experiment* was the baseline level of the recording type variable.

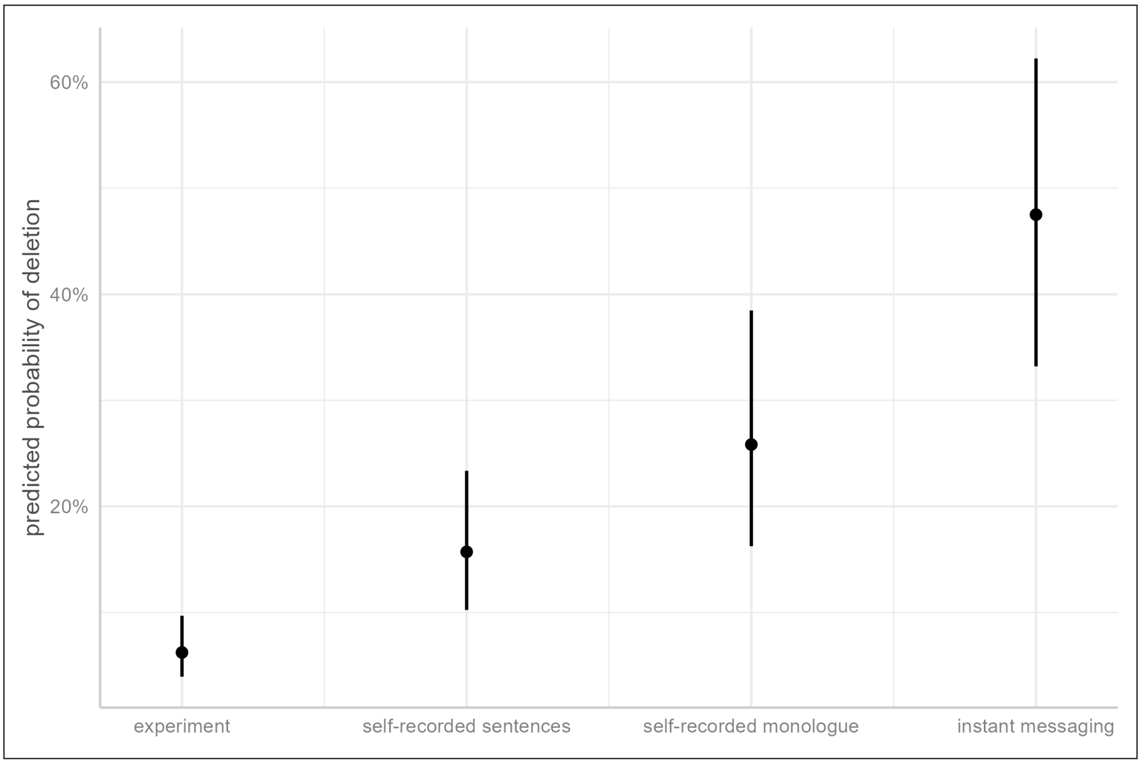

Figure 10: Predicted probabilities of /b/ deletion depending on recording type.

3.4 Disentangling recording type from task type

Given the possibility that task type may have a different effect on stop lenition than recording type, the relationship between the two was explored more closely. An additional sample of recordings from four of the participants made it possible to build a 2 × 2 × 2 model on a subset of data.[14] I compared two types of tasks (spontaneous speech and sentence reading/repetition) in a more controlled setting (with the experimenter present) to the same two tasks in an uncontrolled setting (no experimenter, self-recordings).

3.4.1 Model 4: Predicting lenition degree by segment and recording type

For Model 4, I first built a linear mixed effects model with intensity difference as a dependent variable:

intensity difference ~ recording type + task type + segment + recording type : task type + task type : segment + recording type : segment + (1 + task type + recording type | participant) + (1 | word)

The model was run on 1448 observations. The results show a significant main result of segment, with no significant results of recording type or task type (see **Table 5**). There were, however, significant interactions between recording type and segment, task type and segment, and recording type and task type. The interaction plot in **Figure 11** shows a discrepancy in the intensity difference of /p/ between the production of controlled sentences vs. monologues in recordings produced in the presence of the experimenter. However, pairwise comparisons based on *emmeans* only show significant differences between /p/ and /b/. No within-segment

	Estimate	std. error	df	*t* value	*p* value
(Intercept)	6.001	1.668	5.677	3.598	<.05
self-recording	–1.168	1.729	3.222	–0.676	.545
uncontrolled task	–0.452	2.837	3.74	–0.159	.882
/p/	7.146	1.111	139.464	6.429	<.001
self-recording : uncontrolled	–1.982	0.666	1143.281	–2.978	<.01
uncontrolled : /p/	3.09	1.171	169.441	2.638	<.01
self-recording : /p/	4.105	0.546	1262.116	7.52	<.001

Table 5: Summary of results for Model 4.

[14] I would like to thank one of the reviewers of this paper for suggesting this option. I used data from four participants because only four participated in the fieldwork interviews.

differences in either recording type or task type were reliable (see **Table 13** in the Appendix). The overall variance explained by the model was $R^2 = .384$, most of which can be attributed to the underlying segment.

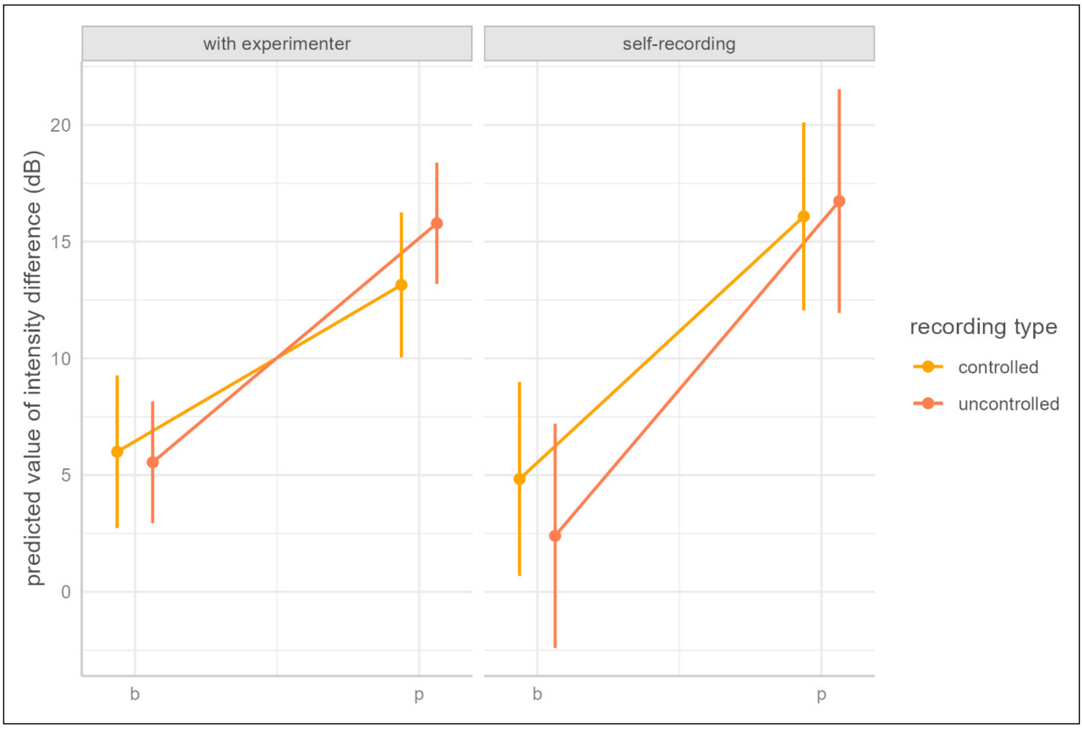

Figure 11: Interaction between segment, recording type and task type.

3.4.2 Model 5: Predicting the probability of voicing by task and recording type

For Model 5, a follow-up model was built to explore the probability of voicing across the tested conditions given the results concerning /p/. The number of observations for this model was 801.

*voicing ~ recording type * task type + (1 | participant)*

The results show a significant effect of recording type and task type, and a significant interaction between the two, by which the probability of voicing is greater in the uncontrolled task in self-recordings (see **Table 6**).

As for the interaction, the effect of task is significant only in the self-recording condition (see **Figure 12**). This is confirmed by pairwise contrasts estimated based on the marginal means from the model, which show significant differences between all the conditions, except for the difference between the controlled and uncontrolled task in the *with experimenter* condition (β = –0.456, z = –2.415, p = .074, see **Table 15** in the Appendix).

Variable	Level	Estimate	std. error	*z* value	*p* value
(Intercept)		0.629	0.242	2.598	<.01
recording type	self-recording	–0.637	0.209	–3.046	<.01
task type	uncontrolled	0.456	0.189	2.415	<.05
recording type: task type	self-recording : uncontrolled	2.428	0.655	3.709	<.001

Table 6: Estimates from Model 5. The reference levels were: *with experimenter* for recording type and *controlled* for task type.

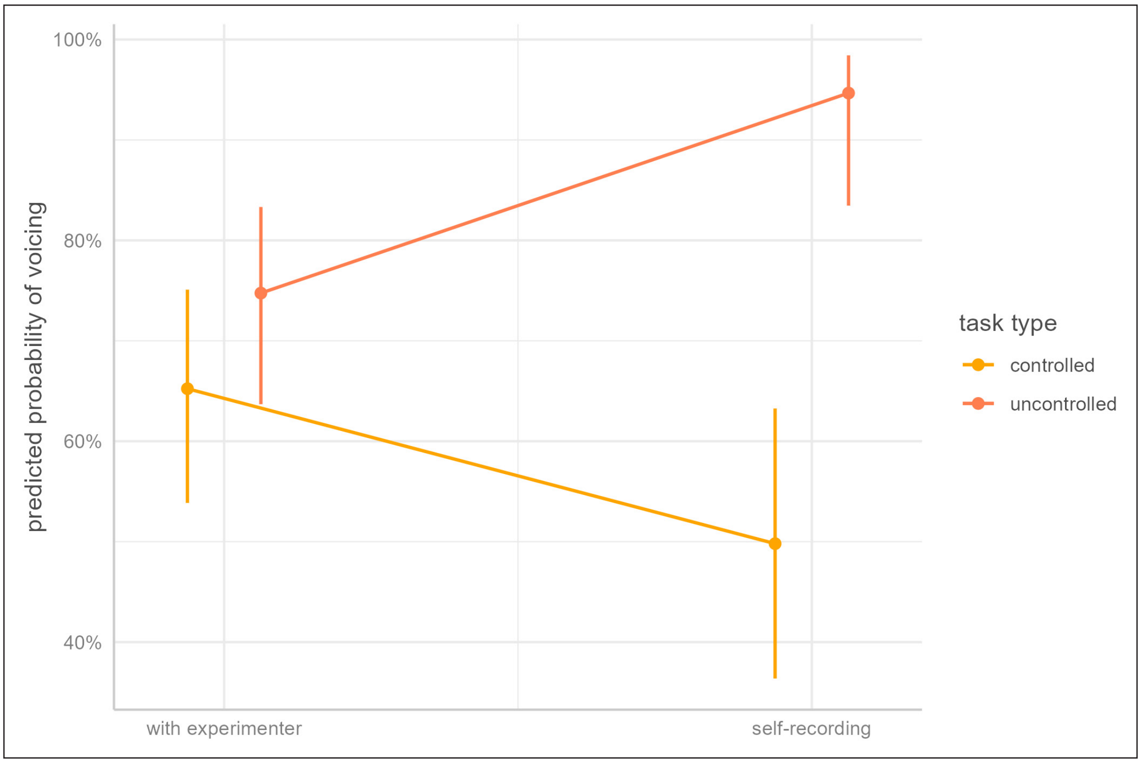

Figure 12: Interaction between recording type and task type in predicting the probability of voicing in the dialect.

3.4.3 Model 6: Predicting the probability of deletion by task and recording type

Finally, for Model 6, a follow-up model predicting the probability of deletion depending on recording type and task type was built. The number of observations for this model was 864.

$$deletion \sim recording\ type * task\ type + (1 \mid participant)$$

The results of the model show a significant main effect of task type and recording type (see **Table 7**). There was also a significant interaction between recording type and task type. This is illustrated in **Figure 13**, below, where we can see that task type is relevant in predicting deletion only in the *with experimenter* condition. The predicted probability of deletion is higher in monologues compared to sentences. The significance of this difference was also confirmed by pairwise contrasts based on the marginal means from the model, which show significant differences between the controlled and uncontrolled task in the *with experimenter* condition (β = –1.456, z = –7.169, p < .001), and between the controlled task in self-recordings and uncontrolled task *with experimenter* (β = –0.917, z = –3.787, p < .001, see **Table 17** in the Appendix).

Variable	**level**	**estimate**	**std. error**	***z* value**	***p* value**
(Intercept)		–1.865	0.265	–7.042	<.001
recording type	self-recording	0.539	0.258	2.092	<.05
task type	uncontrolled	1.457	0.203	7.169	<.001
recording type : task type	self-recording : uncontrolled	–1.321	0.423	–3.122	<.01

Table 7: Estimates from Model 6. The reference levels were *with experimenter* for recording type and *controlled* for task type.

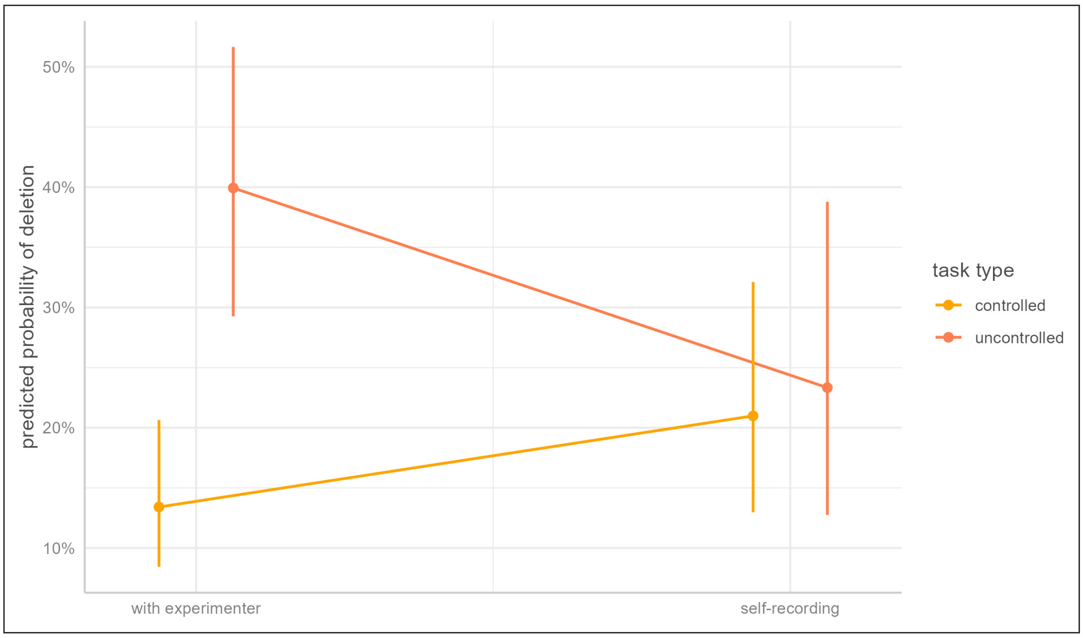

Figure 13: Interaction between recording type and task type in predicting the probability of deletion in the dialect.

4 Discussion

The results of the present study show that the differences between recordings in terms of the amount of lenition applied by a speaker are more nuanced than was predicted by Hypothesis 1. In the case of /b/, there are no reliable variations in intensity difference depending on the recording, which supports the prediction that approximantization is a categorical process that does not depend on the recording situation. As for /p/, there are significant differences between the experiment, IM, and self-recordings, but not between self-recorded sentences and self-recorded monologues, which means that /p/ productions do depend on the recording situation. However, the predicted values of intensity difference in self-recordings are not closer to IM than to the experiment. This may be due to the fact that /p/ lenition is a variable, non-phonologized process in the dialect, an interpretation that finds further support in the models fit for voicing frequency.

Hypothesis 2a) finds partial support in the data: The probability of voicing is greater in the IM and self-recordings than in the experiment. However, the probability of voicing is not reliably the highest in IM data compared to all other recordings. Against this background, the data fully confirm hypothesis 2b). The deletion of /b/ has the highest probability in IM and is higher in self-recordings than in the experiment, which is the expected pattern.

Follow-up models testing Hypothesis 3 show that the roles of recording type and task type are, to some extent, intertwined. When looking at intensity difference, both /p/ and /b/ are produced the same way regardless of whether or not the experimenter is present, and whether the speaker is producing predefined sentences or speaking freely. When we look at the frequency of voicing, however, we can see that the task matters, albeit not so much when the experimenter is present. In self-recordings, the probability of voicing is much lower when producing scripted sentences compared to spontaneous speech. In the supervised condition, the difference between the tasks is much smaller, and the overall probability of voicing is in-between the probabilities predicted for the self-recordings. These results suggest that speakers suppress the lenition of /p/ in the presence of the experimenter more than they do in self-recordings, and, at the same time, they suppress the lenition of /p/ in a controlled task compared to spontaneous speech, even without the experimenter present.

The data concerning /b/ deletion rate show that speaker productions depend on the task, as the probability of deletion is much lower when producing controlled sentences. This is expected since speakers do not plan these utterances, and their reproductions are necessarily more controlled. However, this effect is observed only in the presence of the experimenter. In self-recordings, the nature of the task does not affect the probability of deletion.

Finally, speech rate resulted significant in predicting intensity difference but not the frequency of voicing or deletion. Thus, Hypothesis 4 finds weak support in this comparative analysis. It seems that task and recording type play a greater role in stop lenition in the dialect, although the lack of definite support for speech rate as a moderator of the process may be due to the small sample size.

4.1 Comparison between recordings and the categoricity of weakening

Both data exploration and inferential statistics show that the general intensity profiles of the investigated segments are comparable. Both /p/ and /b/ are produced with similar intensity difference values within their own categories across recordings. Furthermore, /p/ has a greater range of values, as it can be either voiced or voiceless on the surface. When we look at surface sounds, there is a clear three-category division in all tested labials. What is particularly interesting is that the two subcategories of /p/, i.e., [p] and [b], have a lot of overlap in the sentence condition. In other words, when speakers are asked to either read out scripted sentences or repeat them after another speaker, they do not seem to have two categories of segments (weakened vs. unweakened), but rather the two overlap to a great extent in terms of intensity difference despite there being voicing and lack of burst in only one of them. In spontaneous speech, however, speakers do seem to have three competing surface categories, which suggests a switch in the phonological (allophonic) category between competing forms. It must be noted, however, that fieldwork recordings explored in the second part of the analysis break this pattern. Although they contain spontaneous speech, there is an overlap between voiced and voiceless realizations of /p/ (see **Figure 16** in the Appendix). This is perhaps due to the presence of the experimenter and may speak to the importance of the recording situation. The analysis of voicing reported in Section 3 showed that task type seems to be especially relevant in self-recordings, while productions made in the presence of the experimenter are more similar between tasks, which may point to the emergence of categoricity in surface forms in unsupervised spontaneous data.

4.2 Inter- and intra-speaker differences, and the observer's paradox

The data suggest that /p/ weakening is more prevalent in spontaneous speech, although not all speakers follow this pattern (see **Figure 7**). This may be due to many factors, such as individual differences in speaker awareness or reactions to the speaking or recording situation. During the field experiment, I noted that while some speakers do not change their speaking patterns much when being recorded and performing the unnatural task of reading sentences, others get nervous, make more mistakes and have to repeat several of the sentences, or use a lot of hypercorrections. Also, there might be an effect of familiarity with the experimenter: Some participants may have felt less inclined to 'perform well' compared to others. Finally, in the self-recordings, two of the speakers were a bit nervous (voice trembling at times), and one of them (P6) seemed to have a prepared speech for the monologue, which is probably why she had an exceptionally high speaking rate in this condition.

If we accept that spontaneous speech is closer to natural productions made every day in the community, then the increased frequency of voicing in the monologue is evidence that we indeed gathered appropriate data that show more naturalistic speech closer to that ideal. It should not escape our attention, however, that there exists a discrepancy between self-recorded sentences and the self-recorded monologue in almost all speakers, even though their productions were recorded in the same session. Also, although these recordings have overall the slowest speech rates, the greatest number of /p/ weakening is produced precisely in them. This observation is also confirmed by the category distinctions illustrated on the density plots (**Figure 2**), where we see a discrepancy in changing surface category from [p] to [b] between the monologue and repeated sentences, and a clear distinction between spontaneous productions (monologue + IM) and sentences in general. Similar tendencies can be seen in the prevalence of /b/ deletion depending on the recording. The probability of elision is higher in self-recorded monologues, and higher still in IM. It is also higher in spontaneous speech from fieldwork interviews compared to the field experiment (see **Figure 13**).

These results lead to the conclusion that perhaps the experimenter (or observer) is not the most important element to remove from speech production recordings to capture natural speech. The data show that the task, or the way language is used by the speakers, plays an even more important role in influencing their speech production in the context of the observer's paradox. Scripted sentences help control the investigated conditions and elicit exactly what we need as researchers but lead to a substantial change of style on the part of the speaker, even when sentences are repeated from memory, and even when the speaker is the one making recordings and controlling the situation.

As noted in the Introduction, recording environment can influence study participants' speech patterns, possibly suppressing certain sound processes. However, the data presented in this paper show that, apparently, entering a 'recording situation', in which the participant is aware that their speech production will be subject to post-hoc evaluation, is not sufficient to explain differences in speaker productions. An analysis of fieldwork recordings of spontaneous speech demonstrates an interplay between recording type (situation) and task type. Having no flexibility over what is being produced diminishes the probability of applying a weakening process (be it /p/ voicing or /b/ deletion), although it does not necessarily affect the way particular variants are produced acoustically (intensity difference). Thus, it is possible that speech planning is organized differently depending on the type of task and input speech. Our cognitive and motor mechanisms responsible for changes in pronunciation, such as coarticulation, gestural undershoot and overlap, and other issues, may be highly dependent on whether we are merely repeating sentences written by someone else or spontaneously creating our own. This unsurprising result finds support in previous work in sociophonetics (e.g., Lewis, 2001; Hualde et al., 2011; Lozano, 2021). Repeated and redundant words or phrases are also different from the point of view of speech production (Clark & Wasow, 1998; Aylett & Turk, 2004). There has also been some work on the differences in speech planning under lexical competition or in error making conditions, focused on lab vs. spontaneous speech (e.g., Pouplier & Goldstein, 2010; Alderete et al., 2021). Different speech planning mechanisms and the resultant motor arrangements, and perhaps differences in other language-external factors modulating speech, such as social factors, attitudes, speaking habits, and so on, may result in different weakening patterns and influence researchers' generalizations concerning sound change.

In the present case, were we to look at scripted sentences only, we might consider /p/ weakening in the dialect as much less prevalent and much more continuous or unstable than it is in natural productions. We might also underestimate the frequency of /b/ deletion. The self-recordings employed in this study helped us estimate the true scale of the studied processes and uncover additional factors influencing speaker productions. They point to the differences in the frequency of process application across different tasks and, hence, speech styles, as well as the fact that some weakening parameters remain virtually unaffected, regardless of the task or recording type. In our case, intensity difference, a marker of degree of weakening, proved not to be that sensitive to the data gathering process. More research is needed to address these questions more thoroughly, and new research paradigms should be sought to test different task types and induce more controlled or research question-aligned, yet spontaneous, productions. The use of turn-taking and speaker interactions is perhaps a good direction to follow (e.g., Lozano, 2021, successfully used conversation dyads in her study design). Interactive tools and applications could also be explored. If we want to make use of self-recordings, reading maps, giving instructions or other types of semi-structured speech could be elicited in a way similar to research on intonation. Additionally, care should be taken to disentangle the experimenter/interlocutor from a perceived interlocutor. Since we can see a difference between self-recorded monologues and instant messaging in this study, it may be that when participants know a researcher will analyze their speech recordings, they suppress their natural productions to some extent, albeit not as much during a study made with the experimenter present.

The present study offers sufficient evidence that self-recordings can be used successfully in studying some types of sound change, such as stop weakening, and that self-recorded monologues are quite close to the natural productions we might hear over an IM app, which should be elicited at least as a comparative sample. The study itself touches on some important questions concerning the viability of remote data collection (see Section 4.3, below) on the one hand, and the importance of using authentic productions, and comparing lab and spontaneous speech on the other, which is crucial for us to see whether the conclusions taken from controlled studies can be supported with more naturalistic data. The latter endeavor is not new, and the debate over lab vs. uncontrolled speech is still far from over. As Wagner et al. (2015, p. 10) put it in their introduction to a special issue on the topic in the *Journal of Phonetics*, "the best way to approach research questions is by embracing a certain methodological pluralism.” Depending on the phenomenon we want to study, various methodological settings can be considered. Care should also be taken to elicit as much speech as possible to obtain enough data for a comparative sample when necessary. We have seen that not many tokens of sounds of interest have been found in short monologues and, moreover, total numbers of sounds analyzed differ from speaker to speaker. Perhaps using certain topics or keywords would be a good option if we can predict that certain groups of sounds will be more frequent in a sample using a specific grammatical or semantic category.

4.3 Challenges and recommendations

Many advantages of remote data collection have been described in this paper, but researchers should also take the challenges into account and address them. One challenge is finding an adequate number of participants that can provide data according to the researcher’s instructions. In remote recordings, the load of performing the study is transferred from the researcher to the prospective participant. Instead of coming to a lab or place of encounter, the participant has to prepare the equipment and setup, as well as download software that will enable quality recording. They also have to make sure that the recording is saved in a correct format and send it to the researcher, which is sometimes complicated, as larger files cannot be easily sent by email. Thus, some participants may be initially interested in sending in their data, but never find time for it. Moreover, remote data collection cannot be performed on the go; place and time are important factors, as is ambient noise, which can also discourage prospective participants, leaving them unwilling to complete the task. Compared to an online study or survey that asks participants to simply click on the correct answer, a production study conducted in the way described here is a nuisance. In the case of the present study, several people promised to perform the task but never did it, despite being reminded for two months after the recordings were due to be shared with me.

To encourage participants to complete the study, incentives, such as remuneration, prizes, or other perks, should be considered to ensure a representative sample. Or, use third parties to assist in data collection and plan ahead to have more time before the data is in. Also, it might be useful to invite twice as many people as necessary to participate. Getting a larger sample will help address such issues as bad audio quality resulting from the participants’ non-compliance with the instructions, and other, independent factors. For example, reverberation in the room, uncontrolled background noise, and so on. Rejecting some participants from the sample will not affect statistical analysis if more recordings are made than absolutely necessary to ensure a good signal-to-noise ratio and study effect size. In this study, the sample size was small overall, especially in the IM condition. This is because segments other than /p b/ were excluded from the dataset for better data comparability. Larger sample sizes should be used in future investigations.

For some participants, following instructions correctly can pose a challenge. In our case, each participant got an email with a full package containing recorded sentences to repeat and instructions with precise information on how to proceed. Screenshots were provided illustrating where to click in Audacity or mobile app and how to save the data in the correct format. Still, three of the participants had to re-save the files and/or resend them because they were provided in an incorrect format. Thus, even though most people are familiar with the use of smartphones and computers, certain tasks will not be intuitive for them. It is therefore worth checking on the participants and repeating crucial information several times.

Another challenge is the lack of control of certain recording parameters. In our study we saw a difference in overall intensity depending on recording type. Self-made recordings were slightly louder than the ones from the field experiment, which may be due to a smaller distance of the microphone from the mouth in the former case. On the other hand, IM recordings, which are typically made with the mouth close to the smartphone, had lower intensity values compared to all other recordings. Hence, we cannot reliably say that the distance between the mouth and the device is indeed responsible for the discrepancies. In any case, researchers using remote data collection methods should proceed with caution, as we cannot be absolutely sure how sound is processed via the different IM and social media apps. Furthermore, we have no way of determining whether or not the speaker maintains a consistent distance from the microphone for the duration of the recording. Acoustic measurements such as intensity are particularly sensitive to such changes. On a positive note, the within-participant differences involving different underlying segments, voiced vs. voiceless realizations of the /p/ or stressed vs. unstressed syllables were virtually the same across the different recording types in this study. Using relative rather than absolute values of key parameters was certainly helpful here and should be the way to proceed whenever possible.

Finally, since some of the recordings analyzed in this study were made using a phone rather than a laptop with Audacity, there may be some concern about the differences in quality between the recordings. Upon careful file examination, it does not appear that there were substantial differences between the two types of devices used. I ran an additional statistical model to see if there was a difference between laptop and phone recordings. There was no statistical effect of device in the intensity measurements ($F = 0.019, p = .895$, see **Figure 17**). Any differences found in the data can be attributed to the difference between underlying segments rather than the device used. Thus, we may conclude that both laptops and smartphones are suitable for the purposes of analyzing stop lenition in Spanish.

While unguided remote data collection can pose a challenge, it is nevertheless a useful research method for phoneticians, phonologists and sociolinguists. There are obvious advantages of gathering production data this way: It is less resource-intensive, it does not require travelling arrangements and captures more data in a similar amount of time. Additionally, it eliminates researcher bias and helps us get samples of casual, everyday speech. However, as noted in Section 4.2, we should be aware of the possibility that the absence of a third party does not necessarily eliminate the problem of unnatural data. Drawing on the results of this study, I see three parts of the equation in gathering naturalistic speech productions: the setting (recording situation and speaker's awareness of it), the observer (researcher, third party present) and the material/task used. In most studies, it is impossible to eliminate all the three obstacles to natural speech production, hence cross-comparisons of different types of recordings can help mitigate the observer's paradox. We have seen in this comparative study that speaker behavior changes quite substantially in terms of frequency of applied weakening when they are speaking freely, without preconceived words or phrases or frame sentences. At the same time, the only recording type that does not imply an observer but does imply an audience (i.e., the recipient of the message) at the time of recording is a casual voice recording sent via an IM or social media app. By contrast, self-recordings made for a researcher have an implied observer and may still induce a change of style. Thus, crowd-sourced data consisting of social media messages might be helpful in uncovering the true range of productions and their frequency in a given speech community, while controlled recordings can be used to explore acoustic detail in a more reliable manner.

5 Conclusions

The aim of the present study was to provide a comparative analysis of several types of speaker productions and test the viability of unsupervised self-recordings as a data collection method that allows for a thorough study of stop lenition. The results confirm that the method used to gather data is suitable for the phonetic analysis of this process. Moreover, the method proved crucial in uncovering factors influencing speech production in the dialect. More specifically, it helped demonstrate that not only the way speakers are recorded, but also the way they are asked to speak, hinders naturalistic productions. Thus, an interplay between recording type and task type seems to be at work. Importantly, sentence reading/repeating tasks render similar results across recording sessions using different methods. The data is more dispersed, and phonetic categories blend together more, compared to spontaneous speech, which shows more optionality with categorical sound changes. However, the effect of task seems to be different in self-recordings compared to data gathered in the presence of the experimenter, the latter being sufficient to diminish the frequency of process application. Based on this study, I recommend that to minimize the observer's paradox and elicit more authentic speech to tap into the phonetics-phonology interface, researchers should rethink the elicitation model or collect comparative samples of (self-recorded) spontaneous productions before making generalizations concerning a given linguistic process.

Appendix

1. List of sentences used in the experiment and self-recording conditions

La **b**arrera estaba mal colocada y el portero no veía.

La **b**anda de música empezó el concierto con la **b**amba.

La **b**ase científica del Covid es indudable.

La **b**araja española tiene cuarenta cartas con cuatro palos.

La **b**amba es un baile latinoamericano muy conocido.

La **v**acuna contra el Covid debería ser obligatoria **p**ara todos.

La **b**atata del potaje no era muy dulce.

La **b**asura acumulada en el Océano Atlántico es una **p**asada.

La **v**aca de Juan cuesta mucha **p**asta.

La **b**aba de caracol se usa **p**ara producir cosméticos.

La **p**arte más difícil de ser padre es tener que aprenderlo.

La **p**ágina web del gobierno ha **p**arado de funcionar.

La **p**aella valenciana es la más auténtica de todas las paellas.

La **p**atrulla ha encontrado el ladrón en la **b**arca.

Se llama **p**anza de burro cuando está nublado en verano.

La **p**andilla de mi barrio es bastante conflictiva.

La **p**asta de dientes que compramos no sirve para niños.

La **p**alanca de cambios de mi coche ha **p**arado de funcionar.

La **p**aciencia de esa mujer me tenía **b**astante impresionado.

La **p**aga mensual es más baja de lo que pensaba **P**aco.

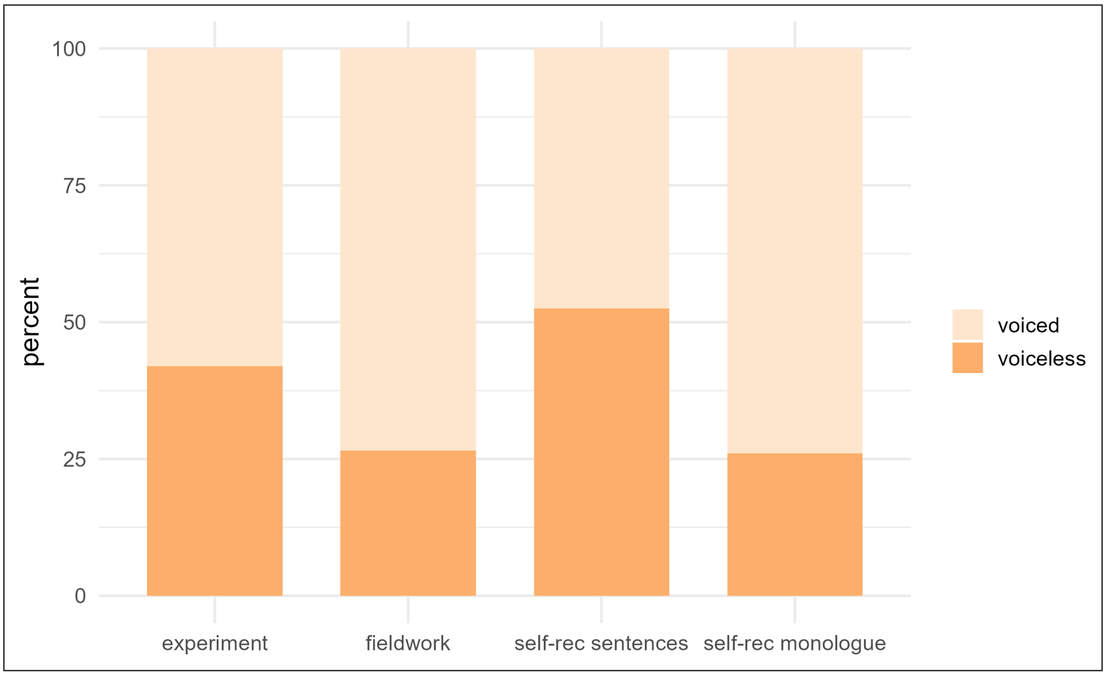

Figure 14: Voicing frequencies depending on recording type showing fieldwork spontaneous speech instead of IM as a comparison.

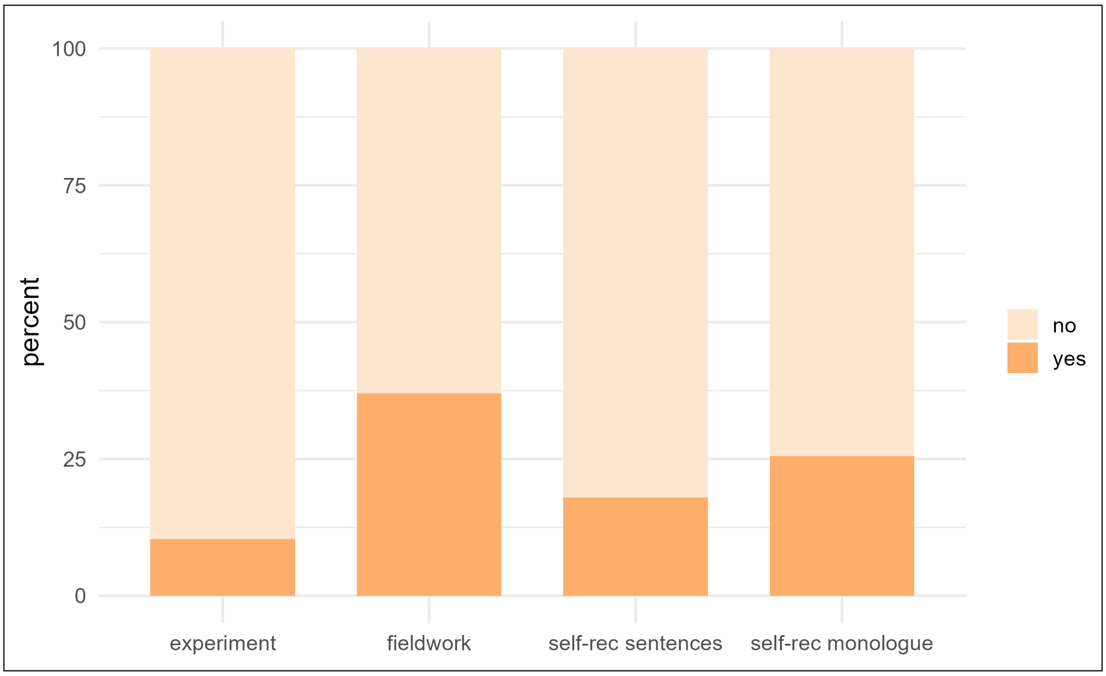

Figure 15: Deletion of /b/ depending on recording type showing fieldwork spontaneous speech instead of IM as a comparison.

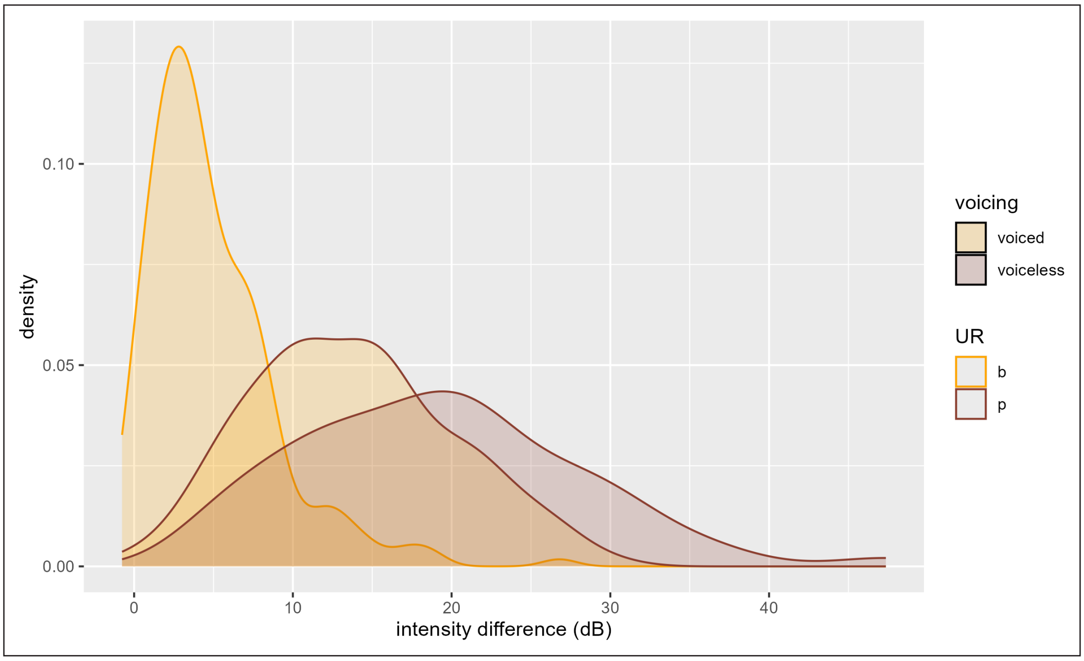

Figure 16: Density plot showing intensity difference for each of the underlying and surface categories in fieldwork data.

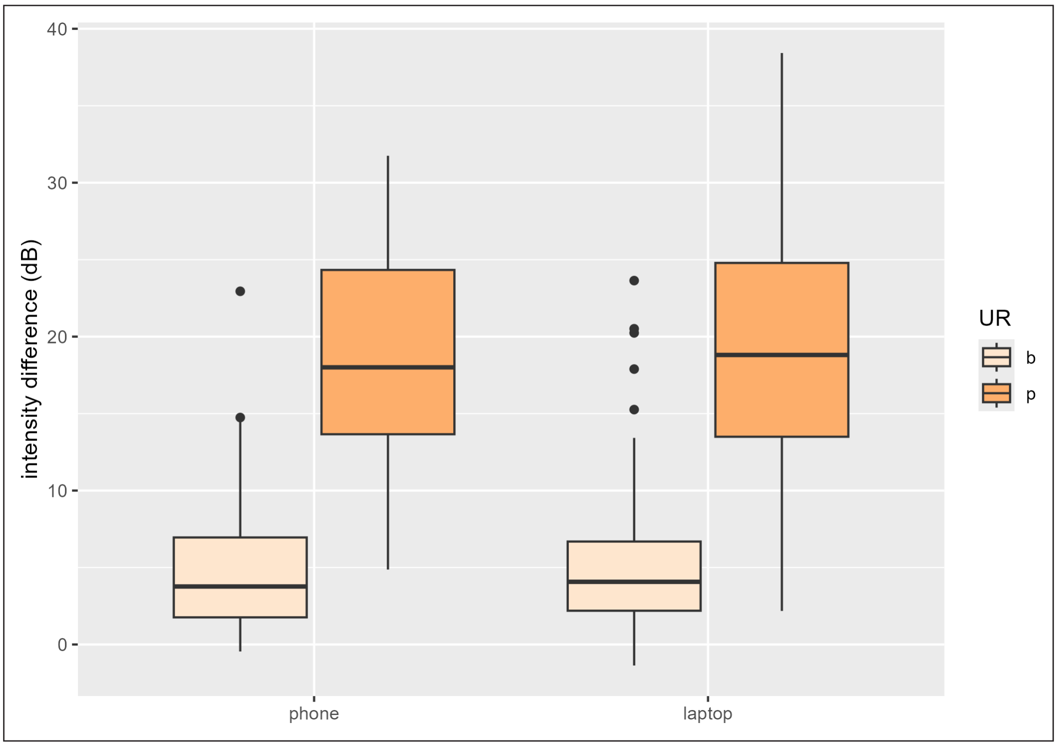

Figure 17: Intensity comparison of self-recordings made using an Android phone and a laptop computer.

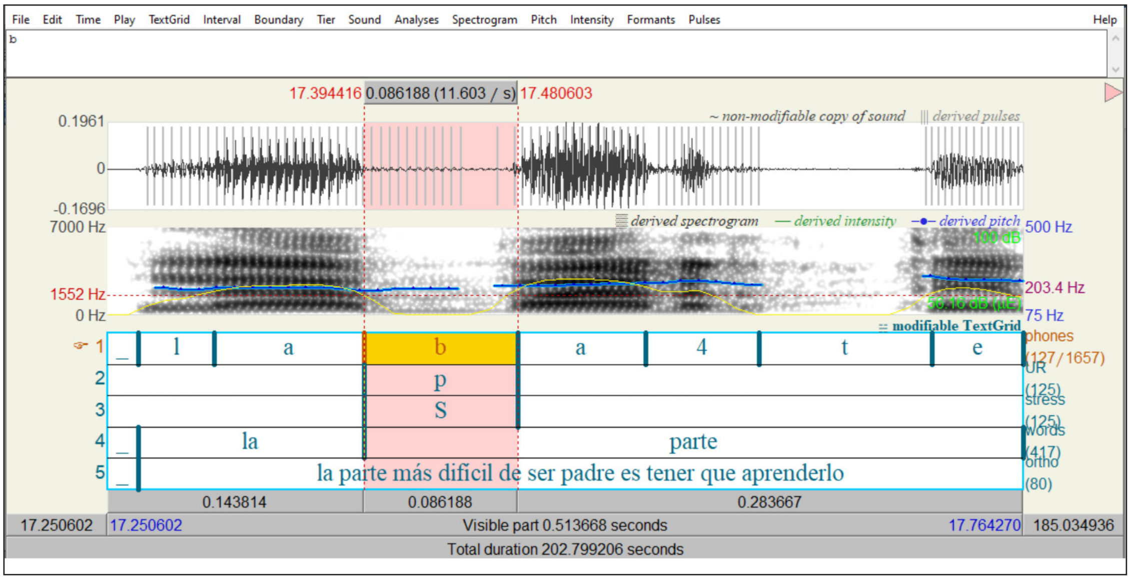

Figure 18: Sample annotation of the data showing la parte, 'the part' [la.baɾ.te]. The first tier shows surface productions, the second tier shows the phonemic representation (here: /p/), while the third tier shows stress, marked as "S" when the key syllable is stressed and "U" when unstressed. The remaining tiers show the word and sentence level.

Contrast	estimate	SE	df	*t* ratio	*p* value
experiment /p/ – self-recorded sentences /p/	–1.829	0.357	1477.171	–5.122	<.001
experiment /p/ – self-recorded monologue /p/	–2.297	0.722	321.981	–3.182	<.05
experiment /p/ – IM /b/	3.776	0.746	281.966	5.061	<.001
experiment /p/ – experiment /b/	11.045	1.315	11.348	8.399	<.001
experiment /p/ – self-recorded sentences /b/	12.104	1.339	12.173	9.040	<.001
experiment /p/ – self-recorded monologue /b/	10.873	1.389	14.245	7.826	<.001
self-recorded sentences /p/ – self-recorded monologue /p/	–0.469	0.767	382.096	–0.611	.999
self-recorded sentences /p/ – IM /b/	5.605	0.782	328.937	7.169	<.001

(Contd.)

Contrast	estimate	SE	df	*t* ratio	*p* value
self-recorded sentences /p/ – experiment /b/	12.874	1.338	12.104	9.622	<.001
self-recorded sentences /p/ – self-recorded sentences /b/	13.932	1.348	12.537	10.337	<.001
self-recorded sentences /p/ – self-recorded monologue /b/	12.702	1.410	15.046	9.006	<.001
self-recorded sentences /p/ – IM /b/	13.809	1.441	16.203	9.582	<.001
self-recorded monologue /p/ – IM /p/	6.073	0.849	1094.908	7.153	<.001
self-recorded monologue /p/ – experiment /b/	13.343	1.385	14.044	9.636	<.001
self-recorded monologue /p/ – self-recorded sentences /b/	14.401	1.409	15.028	10.220	<.001
self-recorded monologue /p/ – self-recorded monologue /b/	13.171	1.341	12.479	9.818	<.001
self-recorded monologue /p/ – IM /b/	14.278	1.481	18.250	9.644	<.001
IM /p/ – experiment /b/	7.269	1.391	14.304	5.226	<.01
IM /p/ – self-recorded sentences /b/	8.328	1.413	15.208	5.895	<.01
IM /p/ – self-recorded monologue /b/	7.097	1.456	17.242	4.875	<.01
IM /p/ – IM /b/	8.204	1.485	18.384	5.527	<.01
experiment /b/ – self-recorded sentences /b/	1.059	0.367	1449.175	2.883	.077
experiment /b/ – self-recorded monologue /b/	–0.172	0.802	194.167	–0.215	1.000
experiment /b/ – IM /b/	0.935	0.863	215.188	1.084	.960

(Contd.)

Contrast	estimate	SE	df	*t* ratio	*p* value
self-recorded sentences /b/ – self-recorded monologue /b/	–1.231	0.841	236.872	–1.464	.826
self-recorded sentences /b/ – IM /b/	–0.123	0.897	253.600	–0.137	1.000
self-recorded monologue /b/ – IM /b/	1.107	0.956	966.923	1.158	.943

Table 8: Model 1: Pairwise contrasts based on the marginal means calculated for the interaction between recording type and underlying segment (SE = standard error, df = degrees of freedom). Here and elsewhere, simple effects were tested via *t*-tests, *emmeans()* function, with familywise error contained using Tukey's method. The most relevant, i.e., within-segment category contrasts, are shaded in grey.

	estimate	lower.CL	upper.CL
(Intercept)	0.017	0.002	0.122
self-recorded sentences	2.805	1.798	4.377
self-recorded monologue	5.241	2.908	9.448
IM	13.614	7.560	24.516
speech rate	1.231	0.932	1.628

Table 9: Odds ratios based on the estimates from Model 2. *Experiment* was the baseline level of the recording type variable.

Contrast	estimate	SE	Df	*z* ratio	*p* value
experiment – self-recorded sentences	0.358	0.189	Inf	1.896	.230
experiment – self-recorded monologues	–1.537	0.341	Inf	–4.502	<.001
experiment – IM	–0.378	0.284	Inf	–1.329	.544
self-recorded sentences – self-recorded monologues	–1.895	0.341	Inf	–5.565	<.001
self-recorded sentences – IM	–0.736	0.299	Inf	–2.464	.066
self-recorded monologues – IM	1.159	0.411	Inf	2.822	<.05

Table 10: Model 2: Pairwise contrasts based on the marginal means calculated for recording (SE = standard error, df = degrees of freedom).

	Estimate	**lower.CL**	**upper.CL**
(Intercept)	0.017	0.002	0.122
self-recorded sentences	2.805	1.798	4.377
self-recorded monologue	5.241	2.908	9.448
IM	13.614	7.560	24.516
speech rate	1.231	0.932	1.628

Table 11: Odds ratios based on the estimates from Model 3. *Experiment* was the baseline level of the recording type variable.

Contrast	**estimate**	**SE**	**df**	***z* ratio**	***p* value**
experiment – self-recorded monologues	–1.657	0.301	Inf	–5.511	< .001
experiment – self-recorded sentences	–1.032	0.227	Inf	–4.544	< .001
experiment – IM	–2.611	0.300	Inf	–8.701	< .001
self-recorded monologues – self-recorded sentences	0.625	0.287	Inf	2.176	.130
self-recorded monologues – IM	–0.955	0.338	Inf	–2.820	< .05
self-recorded sentences – IM	–1.580	0.303	Inf	–5.214	< .001

Table 12: Model 3: Pairwise contrasts based on the marginal means calculated for recording (SE = standard error, df = degrees of freedom).

Contrast	**estimate**	**SE**	**df**	***t* ratio**	***p* value**
with experimenter /b/ – self-recording /b/	2.159	1.735	3.262	1.244	.642
with experimenter /b/ – with experimenter /p/	–8.691	0.637	227.937	–13.652	< .001
with experimenter /b/ – self-recording /p/	–10.637	1.831	4.033	–5.810	< .05
self-recording /b/ – with experimenter /p/	–10.850	1.827	4.007	–5.939	< .05
self-recording /b/ – self-recording /p/	–12.796	0.763	400.536	–16.779	< .001

(Contd.)

Contrast	estimate	SE	df	*t* ratio	*p* value
with experimenter /p/ – self-recording /p/	–1.946	1.730	3.225	–1.125	.701
with experimenter controlled – self-recording controlled	–0.885	1.706	3.054	–0.519	.949
with experimenter controlled – with experimenter uncontrolled	–1.093	2.747	3.283	–0.398	.975
with experimenter controlled – self-recording uncontrolled	0.005	3.581	3.196	0.001	1.000
self-recording controlled – with experimenter uncontrolled	–0.209	2.841	3.243	–0.073	1.000
self-recording controlled – self-recording uncontrolled	0.889	2.779	3.446	0.320	.987
with experimenter uncontrolled – self-recording uncontrolled	1.098	1.780	3.601	0.617	.921
/b/ controlled – /p/ controlled	–9.198	1.118	200.952	–8.229	< .001
/b/ controlled – /b/ uncontrolled	1.443	2.837	3.735	0.509	.952
/b/ controlled – /p/ uncontrolled	–10.845	2.840	3.751	–3.819	.066
/p/ controlled – /b/ uncontrolled	10.641	2.796	3.524	3.805	.075
/p/ controlled – /p/ uncontrolled	–1.647	2.773	3.411	–0.594	.928
/b/ uncontrolled – /p/ uncontrolled	–12.288	0.531	604.248	–23.128	< .001

Table 13: Model 4: Pairwise contrasts based on the marginal means calculated for each of the three interactions from the model (SE = standard error, df = degrees of freedom).

	estimate	lower.CL	upper.CL
(Intercept)	0.532	0.331	0.856
self-recording	1.890	1.255	2.849
uncontrolled task (monologue)	0.633	0.437	0.917
self-recording * uncontrolled task	0.088	0.024	0.318

Table 14: Odds ratios based on the estimates from Model 5 (voicing probability).

Contrast	**estimate**	**SE**	**df**	***z* ratio**	***p* value**
with experimenter controlled – (self-recording controlled)	0.637	0.209	Inf	3.046	<.05
with experimenter controlled – with experimenter uncontrolled	–0.456	0.189	Inf	–2.415	.074
with experimenter controlled – (self-recording uncontrolled)	–2.247	0.612	Inf	–3.672	<.01
(self-recording controlled) – with experimenter uncontrolled	–1.093	0.238	Inf	–4.602	<.001
(self-recording controlled) – (self-recording uncontrolled)	–2.884	0.629	Inf	–4.586	<.001
with experimenter uncontrolled – (self-recording uncontrolled)	–1.791	0.620	Inf	–2.889	<.05

Table 15: Model 5: Pairwise contrasts based on the marginal means calculated for the interaction between recording type and task type (SE = standard error, df = degrees of freedom).

	estimate	**lower.CL**	**upper.CL**
(Intercept)	0.155	0.092	0.260
self-recording	1.715	1.035	2.842
uncontrolled (monologue)	4.292	2.882	6.392
self-recording * uncontrolled task	0.267	0.117	0.612

Table 16: Odds ratios based on the estimates from Model 6 (deletion probability).

Contrast	**estimate**	**SE**	**df**	***z* ratio**	***p* value**
with experimenter controlled – (self-recording controlled)	–0.539	0.258	Inf	–2.092	.156
with experimenter controlled – with experimenter uncontrolled	–1.457	0.203	Inf	–7.169	<.001
with experimenter controlled – (self-recording uncontrolled)	–0.675	0.346	Inf	–1.955	.205
(self-recording controlled) – with experimenter uncontrolled	–0.918	0.242	Inf	–3.788	<.01
(self-recording controlled) – (self-recording uncontrolled)	–0.136	0.368	Inf	–0.370	.983
with experimenter uncontrolled – (self-recording uncontrolled)	0.781	0.337	Inf	2.318	.094

Table 17: Model 6: Pairwise contrasts based on the marginal means calculated for the interaction between recording type and task type (SE = standard error, df = degrees of freedom).

Data accessibility statement

Full results of the models and additional graphs and tables are provided in the Appendix.

The aggregate data and scripts used to provide the above analysis are available on the Open Science Framework, Project name: *Remote Data Collection – lenition in Spanish*, DOI 10.17605/OSF.IO/ZYFJR.

Ethical statement

The experiment was performed in accordance with the recommendations of the ethical committee of the University of Warsaw. Participants gave their informed consent to the use of their data. The same applies to the self-recordings, which were personally sent to me by the participants, and to the IM recordings that were made available to me for analysis.

Competing interests

The author has no competing interests to declare.

References

Alderete, J., Baese-Berk, M., Leung, K., & Goldrick, M. (2021). Cascading activation in phonological planning and articulation: Evidence from spontaneous speech errors. *Cognition, 210*, 104577. https://doi.org/10.1016/j.cognition.2020.104577

Almeida, M., & Díaz Alayón, C. (1988). *El español de Canarias.* Santa Cruz de Tenerife.

Androutsopoulos, J., & Staehr, A. (2018). Moving methods online. In A. Creese & A. Blackledge (Eds.), *The Routledge handbook of language and superdiversity* (pp. 118–132). Routledge.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech, 47*, 31–56.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2018). lme4: linear mixed-effects models using 'Eigen' and S4. R package (version 1.1-17). cran.r-project.org/web/packages/lme4

Bell, A. (1984). Language style as audience design. *Language in Society, 13*, 145–204.

Boyd, Z., Elliott, Z., Fruehwald, J., Hall-Lew, L., & Lawrence, D. (2015). An evaluation of different sociolinguistic elicitation methods. *Proceedings of the 18th International Congress of Phonetic Sciences*, ed. by the Scottish Consortium for ICPhS 2015. The University of Glasgow.

Broś, K. (2023). Using social media as a source of analysable material in phonetics and phonology – lenition in Spanish. *Linguistics Vanguard.* https://doi.org/10.1515/lingvan-2021-0153

Broś, K., & Krause, P. A. (2024). Stop lenition in Canary Islands Spanish–a motion capture study. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 15*(1), 1–50. https://doi.org/10.16995/labphon.9934

Broś, K., & Lipowska, K. (2019). Gran Canarian Spanish non-continuant voicing: Gradiency, sex differences and perception. *Phonetica, 76*, 100–125.

Broś, K., Żygis, M., Sikorski, A., & Wołłejko, J. (2021). Phonological contrasts and gradient effects in ongoing lenition in the Spanish of Gran Canaria. *Phonology, 38*(1), 1–40. https://doi.org/10.1017/S0952675721000038

Bulgin, J., De Decker, P., & Nycz, J. (2010). Reliability of formant measurements from lossy compressed audio. *British Association of Academic Phoneticians Colloquium*. University of West Minister.

Byrne, C., & Foulkes, P. (2007). The 'mobile phone effect' on vowel formants. *International Journal of Speech Language and Law*, *11*(1), 83–102. https://doi.org/10.1558/ijsll.v11i1.83

Calder, J., Wheeler, R., Adams, S., Amarelo, D., Arnold-Murray, K., Bai, J., Church, M., Daniels, J., Gomez, S., Henry, J., Jia, Y., Johnson-Morris, B., Lee, K., Miller, K., Powell, D., Ramsey-Smith, C., Rayl, S., Rosenau, S., & Salvador, N. (2022). Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for vocalic analysis. *Linguistics Vanguard*, 20200148. https://doi.org/10.1515/lingvan-2020-0148

Carrasco, P., Hualde, J. I., & Simonet, M. (2012). Dialectal differences in Spanish voiced obstruent allophony: Costa Rican versus Iberian Spanish. *Phonetica, 69,* 149–179.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37*(3)*,* 201–242. https://doi.org/10.1006/cogp.1998.0693.

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language, 93*, 569–597.

Cohen Priva, U., & Gleason, E. (2020). The causal structure of lenition: A case for the causal precedence of durational shortening. *Language, 96*, 413–448.

Colantoni, L., & Marinescu, I. (2010). The scope of stop weakening in Argentine Spanish. In M. Ortega-Llebaria (Ed.), *Selected proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology* (pp. 100–114). Cascadilla.

Dalcher, C. V. (2008). Consonant weakening in Florentine Italian: A cross-disciplinary approach to gradient and variable sound change. *Language Variation and Change*, *20*(2), 275–316.

De Decker, P., & Nycz, J. (2011). For the Record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics, 17*(2), Article 7.

Eddington, D. (2011). What are the contextual phonetic variants of /β, ð, ɣ/ in colloquial Spanish? *Probus, 23*, 1–19.

Figueroa Candia, M. A., & Evans, B. G. (2015). Evaluation of segmentation approaches and constriction degree correlates for spirant approximant consonants. Poster presented at *the 18th International Congress of Phonetic Sciences*. Glasgow.

Freeman, V., & De Decker, P. (2021a). Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *The Journal of the Acoustical Society of America, 149*(2)*,* 1211. https://doi.org/10.1121/10.0003529

Freeman, V., & De Decker, P. (2021b). Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices. *Language & Linguistics Compass*, e12435. https://doi.org/10.1111/lnc3.12435

Gittelson, B., Leemann, A., & Tomaschek, F. (2021). Using crowdsourced speech data to study socially constrained variation in nonmodal phonation. *Frontiers in Artificial Intelligence*, *3*, 565682. https://doi.org/10.3389/frai.2020.565682

Hall-Lew, L., & Boyd, Z. (2017). Phonetic variation and self-recorded data. *University of Pennsylvania Working Papers in Linguistics, 23*(2), Article 11. https://repository.upenn.edu/pwpl/vol23/iss2/11

Herrera Santana, J. (1997). Estudio acústico de /p, t, c, k/ y /b, d, y, g/ en Gran Canaria. In M. Almeida & J. Dorta (Eds.), *Contribuciones al estudio de la lingüística hispánica (Homenaje al profesor Ramón Trujillo)* (pp. 73–86). Montesinos.

Hualde, J. I. (2005). *The sounds of Spanish*. Cambridge University Press.

Hualde, J. I., Simonet, M., & Nadeu, M. (2011). Consonant lenition and phonological recategorization. *Laboratory Phonology, 2*(2), 301–329. https://doi.org/10.1515/labphon.2011.011

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, W. (1978). *Field methods used by the research project on linguistic change and variation*. University of Pennsylvania Press.

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., & Messerli, J. (2020). Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*, *6*(3), 20200061. https://doi.org/10.1515/lingvan-2020-0061

Lenth, R. V. (2019). emmeans: Estimated marginal means, aka least-squares means. R package. https://cran.r-project.org/web/packages/emmeans/index.html

Lewis, A. M. (2001). *Weakening of intervocalic /ptk/ in two Spanish dialects: Toward the quantification of lenition processes*. [Doctoral dissertation, University of Illinois, Urbana-Champaign].

Lipski, J. M. (1994). Spanish stops, spirants, and glides: From consonantal to [vocalic]. In M. Mazzola (Ed.), *Issues and theory in Romance languages* (pp. 67–86). Georgetown University Press.

Lozano, C. J. (2021). *Prosodically driven continuity lenition: A phonological account of spirantization in Colombian heritage Spanish*. [Doctoral dissertation, University of California, Davis].

Lüdecke, D. (2018). Ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.

Lupton, D. (Ed.) (2021). Doing fieldwork in a pandemic (crowd-sourced document), revised version. Retrieved December 2024. https://docs.google.com/document/d/1clGjGABB2h2qbduTgfqribHmog9B6P0NvMgVuiHZCl8/edit?tab=t.0#heading=h.ze8ug1cqk5lo

Maddieson, I. (1984). *Patterns of sounds*. Cambridge University Press.

Martínez-Celdrán, E., & Regueira, X. L. (2008). Spirant approximants in Galician. *Journal of the International Phonetic Association, 38*(01). http://doi.org/10.1017/S0025100308003265

Martínez-Gil, F. (2020). Spirantization and the phonology of Spanish voiced obstruents. In S. Colina & F. Martínez-Gil (Eds.), *The Routledge handbook of Spanish phonology* (pp. 34–83). Routledge.

Melero-García, F. (2021). Lenition of syllable-initial /p t k/ in a variety of Andalusian Spanish: Effects of linguistic factors and speech rate. *Estudios de Fonética Experimental, 30,* 169–187.

Nadeu, M., & Hualde, J. I. (2015). Biomechanically conditioned variation at the origin of diachronic intervocalic voicing. *Language and Speech, 58*(3), 351–370. https://doi.org/10.1177/00238309145547273

Oftedal, M. (1985). *Lenition in Celtic and in Insular Spanish: The secondary voicing of stops in Gran Canaria.* Universitetsforlaget.

Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In P. MacNeilage (Ed.), *The production of speech* (pp. 189–216). Springer-Verlag.

Ortega-Llebaria, M. (2004). Interplay between phonetic and inventory constraints in the degree of spirantization of voiced stops: Comparing intervocalic /b/ and intervocalic /g/ in Spanish and English. In T. L. Face (Ed.), *Laboratory approaches to Spanish phonology* (pp. 237–53). Mouton de Gruyter.

Parrell, B. (2010). Articulation from acoustics: Estimating constriction degree from the acoustic signal. *The Journal of the Acoustical Society of America, 128*(4), 2289–2289.

Parrell, B. (2011). Dynamical account of how /b, d, g/ differ from /p, t, k/ in Spanish: Evidence from labials. *Laboratory Phonology, 2*, 423–449.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics, 11*, 478–504.

Podesva, R. J. (2011a). Salience and the social meaning of declarative contours: Three case studies of gay professionals. *Journal of English Linguistics, 39*, 233–264.

Podesva, R. J. (2011b). The California vowel shift and gay identity. *American Speech, 86*, 32–51.

Pouplier, M., & Goldstein, L. (2010). Intention in articulation: Articulatory timing in alternating consonant sequences and its implications for models of speech production. *Language and Cognitive Processes*, *25*(5), 616–649. https://doi.org/10.1080/01690960903395380

R Core Team. 2020. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Retrieved May 1, 2023, from https://www.R-project.org/

Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., & Bowern, C. (2021). (Don't) try this at home! The effects of recording devices and software on phonetic analysis. *Language, 97*(4), e360–e382. https://doi.org/10.1353/lan.2021.0075

Sharma, D. (2011). Style repertoire and social change in British Asian English. *Journal of Sociolinguistics, 15,* 464–492.

Soler, A., & Romero, J. (1999). The role of duration in stop lenition in Spanish. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences*. Vol. 1., pp. 483–486.

Trujillo, R. (1980). Sonorización de sordas en Canarias. *Anuario Letras, 18,* 247–254.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48,* 1–12.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer.

Wilson, J. (1987). The sociolinguistic paradox: Data as a methodological product. *Language and Communication, 7*(2), 161–177.

Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2020). Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings. *The Journal of the Acoustical Society of America, 148*(4), 2717.

Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America, 14*9(6), 3910–3916.