Praat Project:

Distinguishability of Out-of-Context Whispered Consonants

Jemmin Chang
December 3, 2014

Submitted to

Professor Tom Werner
80-282 Phonetics & Phonology I
Carnegie Mellon University

Introduction

In Assignment 3, question 7, we attempted to identify the harmonics on a spectrogram produced from a recording of whispered speech. We noted that the harmonics were very unclear and seemingly nonexistent. This observation suggested the possibility that whispered speech (i.e., breathy voice) limits or prohibits entirely vibration of the vocal folds, effectively preventing voiced sounds. Regarding voiced consonants, we hypothesized that consonants voiced during normal speech would be produced as their voiceless counterparts (same place and manner of articulation) in whispered speech. Two questions arise from this hypothesis. First, how are voiced consonants distinguished from voiceless consonants in whispered speech? Second, how are vowels articulated in whispered speech, without vibration of the vocal folds?

We attempt to answer these questions through two modes of analysis: observation of experimental results and explanation of these results by careful comparison of spectrograms of whispered speech. The language of analysis is General American English. We conclude that 1) vowels are unaffected (i.e. remain voiced) in whispered speech and that 2) voiced consonants are generally pronounced as their voiceless counterparts in whispered speech, significantly decreasing the distinguishability between consonant minimal pairs with respect to voicing.

Experimental Methodology

Motivation and Expected Results

If our hypothesis is correct, then voiced consonants are pronounced without voicing in whispered speech, making them indistinguishable from their voiceless counterparts. We tested this hypothesis by asking participants to identify words in whispered speech and marking whether they correctly identify the sounds as voiced or voiceless. By our hypothesis, we expect all consonant sounds to be identified as voiceless.

Design

In order to isolate and test the *physical* distinguishability of voiced and voiceless sounds in whispered speech, it was crucial to minimize contextual influences on participants' identification of the sounds they hear. To this purpose, we used as test words only common words of General American English. Also, the strings of words that participants listened to were unrelated and presented in a random order.

For each minimal pair (with respect to voicing) of consonant sounds, we found two or three pairs of words that differ only in these consonants. Whenever possible, we used words of the form CV or VC to minimize the influence of other consonants on the participants' identification of the words. We constructed a list of such word pairs for each pair of consonant sounds in both the initial and final positions (for some consonant pairs, a word pair could not be found, so these were excluded). The complete list of words used is shown in the code listing in Appendix A.

Using the Python script listed in Appendix A, we generated 4 strings of 22 words each. (These are also listed in Appendix A.) Each string contains one word for each distinct consonant sound in initial and final positions. We recorded each of the 4 strings in a different speaker's whispering voice (2 male speakers and 2 female speakers), at a pace of about 1 word every 1.5 seconds – just enough time for the listening participants to write down what they hear without thinking too much about it.

Then, we had 20 participants each listen to one recording (cycling through the four recordings such that 5 participants listened to each recording) and write down what they heard. We checked the participants' answers against the string of words that was recorded and marked whether the participant correctly identified the consonant sounds of interest. We ignored any variations in vowels or consonant sounds that were not being tested (e.g. the [o] and [r] in [kor]).

Results

The results largely confirmed our hypothesis. Many of the voiced consonants were consistently misidentified as their voiceless counterparts, whereas the voiceless consonants were more often identified correctly. However, the voiceless consonants were also misidentified as their voiced counterparts somewhat frequently; we might explain this by supposing a cognitive process by which a listener “overcompensates” for the voicelessness of whispered speech by assuming that voicing has been removed.

We present the aggregate data, broken down by sound type, in Table 1. The complete spreadsheet of results is included in Appendix B. We analyze the results for each minimal pair of consonant sounds and consider the spectrograms of their recordings in the Spectrogram Analysis section.

	Initial			Init-Total	Final		Fin-Total	**IF-Total**
	Stop	Fricative	Affricate		Stop	Fricative		
Voiceless	43 / 60 (71.7%)	35 / 40 (87.5%)	10 / 20 (50%)	88 / 120 (73.3%)	49 / 60 (81.7%)	32 / 40 (80%)	81 / 100 (81%)	**169 / 220 (76.8%)**
Voiced	33 / 60 (55%)	10 / 40 (25%)	10 / 20 (50%)	53 / 120 (44.2%)	37 / 60 (61.7%)	30 / 40 (75%)	67 / 100 (67%)	**120 / 220 (54.6%)**

Table 1: Correct identification counts (and percentages) by sound type

Limitations

The results of this experiment must be considered in the context of a few inevitable limitations.

In order to get the required number of participants, a convenience sample was taken (namely, my friends). A random sample of the population would of course be ideal, but unrealistic for this project. All the participants were native speakers of General American English. However, they were all Carnegie Mellon University undergraduates, and a disproportionate number (relative to the general population of GA English speakers) were of Korean heritage. What effect this sample bias may have played is unknown.

Although I normalized the volume level across each of the 4 recordings, they still varied slightly in pace and clarity. In particular, the speaker for recording 2 spoke with less deliberate enunciation than the other 3 speakers; as a result, the average number of correct responses for this recording was noticeably less than the average for the other 3 recordings. Ideally, a larger number of recordings would have normalized the effect of this variation in recording clarity.

Occasionally, participants misidentified a consonant sound with respect to more than just the voicing (e.g. wrong place or manner of articulation). However, this occurred infrequently enough (< 10% of responses) that we assume the effect of these mistakes to be negligible. In these cases we marked the participant's response as incorrect.

Spectrogram Analysis

In this section, we examine spectrograms of the various word pairs and attempt to reason about the articulatory processes that account for the more unexpected experimental results that we observed.

First, we list the pairs that yielded results close to what we expected based on our hypothesis. These were initial [t]/[d], initial [k]/[g], initial [f]/[v], initial [s]/[z], final [p]/[b], and final [k]/[g]. For these pairs, in our experiment, the voiced consonant was frequently misidentified while the voiceless consonant was usually identified correctly. The corresponding spectrograms corroborate these experimental results, showing that there is very little distinguishable physical difference between these minimal word pairs. For example, Figure 1 shows the spectrograms for (a) [lɑk] and (b) [lɑg]. The spectrograms for the remaining pairs are shown in Appendix C.

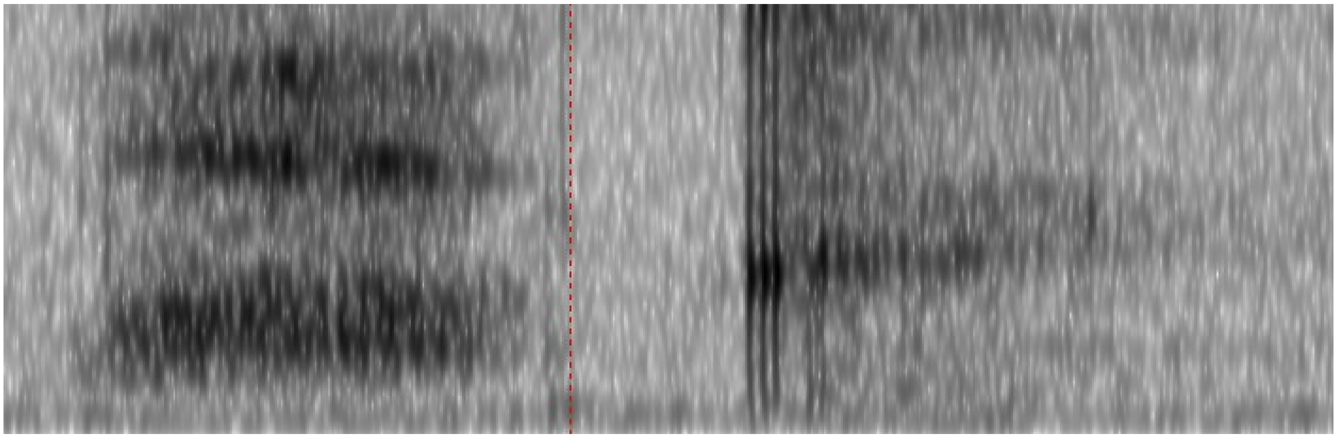


Figure 1a: Spectrogram for [lɑk] (recording 4)

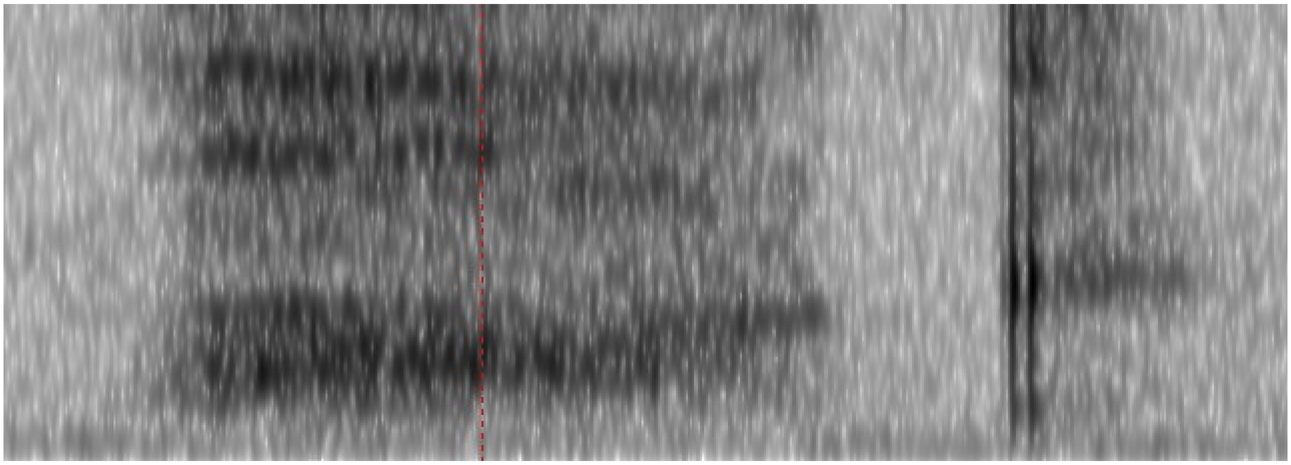


Figure 1b: Spectrogram for [lɑg] (recording 3)

Unexpected Results

The first unusual result we notice is with the initial [p]/[b] pair. Both [p] and [b] were correctly identified 13 / 20 times. We examine the spectrograms for the words [pi] and [bi] in Figure 2.

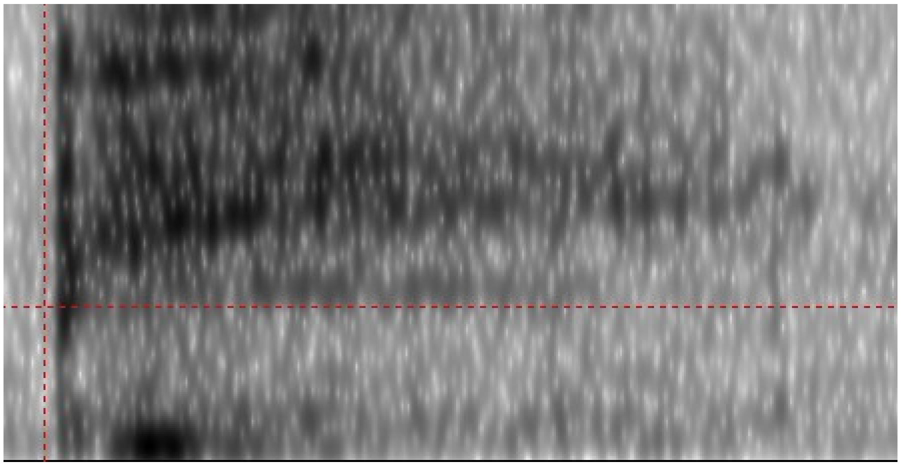


Figure 2a: Spectrogram for [pi] (recording 2)

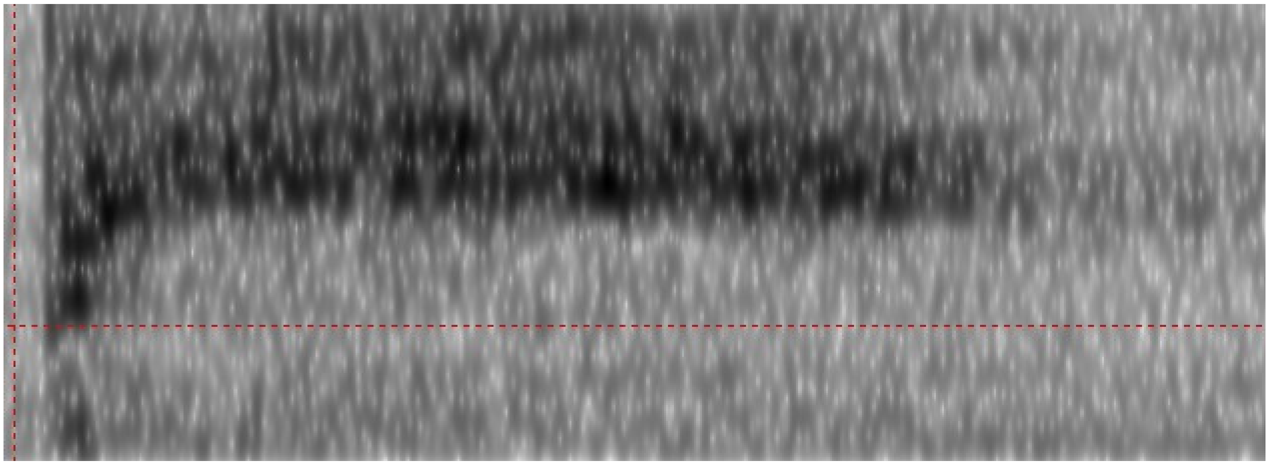


Figure 2b: Spectrogram for [bi] (recording 4)

The spectrograms look quite similar. However, we notice that in [pi] there is a more defined onset (seen as a dark vertical bar before the vowel), whereas the onset in [bi] is less defined. This is likely a result of the aspiration of the initial [p], whereas [b] is unaspirated. The "soft onset" of [bi], as opposed to the clear aspirated onset of [p], may be the distinguishing characteristic which allowed initial [b] to be identified correctly more often than other voiced consonants.

We also see unexpected results with the pair initial [tʃ]/[dʒ], with both consonants being identified correctly 10 / 20 times. Examining the spectrograms in Figure 3, however, we see little distinction between them. In this case, we can only conclude that the results were skewed significantly by the relative commonality of the word pairs. It's reasonable to say that "jar" is a more common GA English word than "char," "cheer" more common than "jeer," and "joke" more common than "choke." We hypothesize that participants tended to identify the recorded words as the more common of each of these pairs simply due to their past experience and vocabulary. We would need to construct word pairs with more equal occurrence in everyday vocabulary to get conclusive results for this consonant pair. It's quite difficult to come up with minimal pairs of equally common words beginning with these consonants, though, and the pairs used were clearly inadequate.

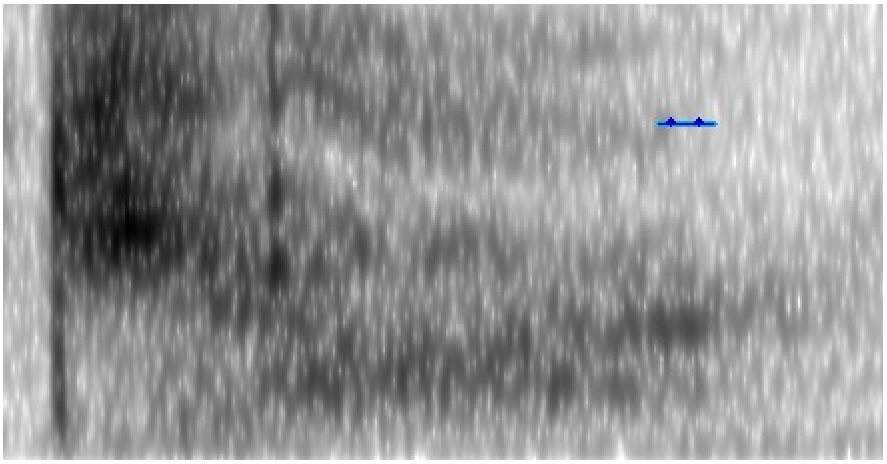


Figure 3a: Spectrogram for [tʃɑr] (recording 2)

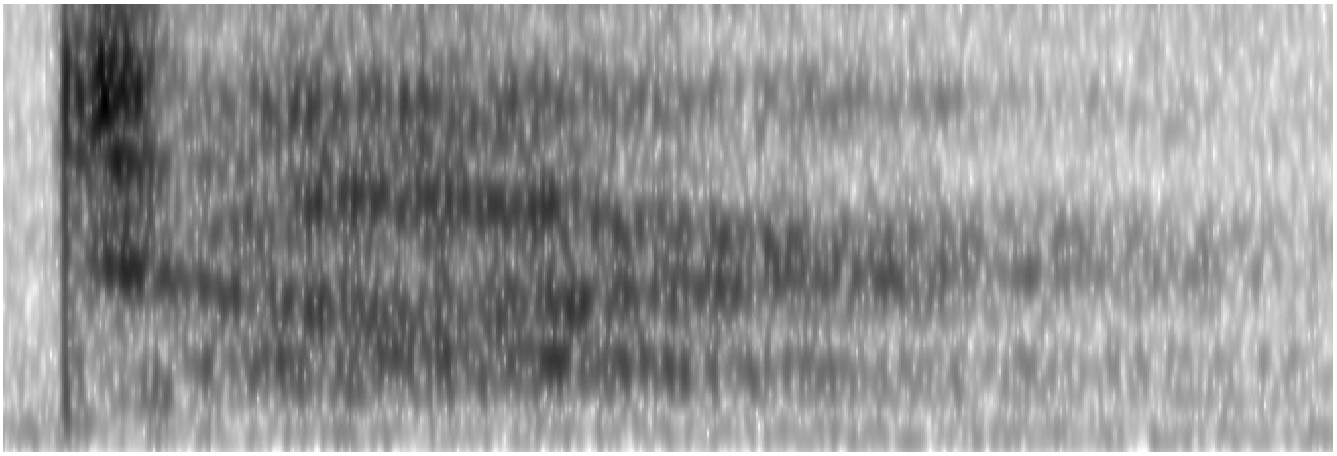


Figure 3b: Spectrogram for [dʒɑr] (recording 4)

Perhaps the most striking result is that for the final pair [t]/[d]. For this pair, the voiced consonant was identified correctly *more often* than the voiceless consonant – indeed almost always identified correctly (18 / 20 times). Comparing the spectrograms for [rɑt] and [rɑd] in Figure 4, we expect to see a clear distinction which would allow the participants' to reliably identify the final [d].

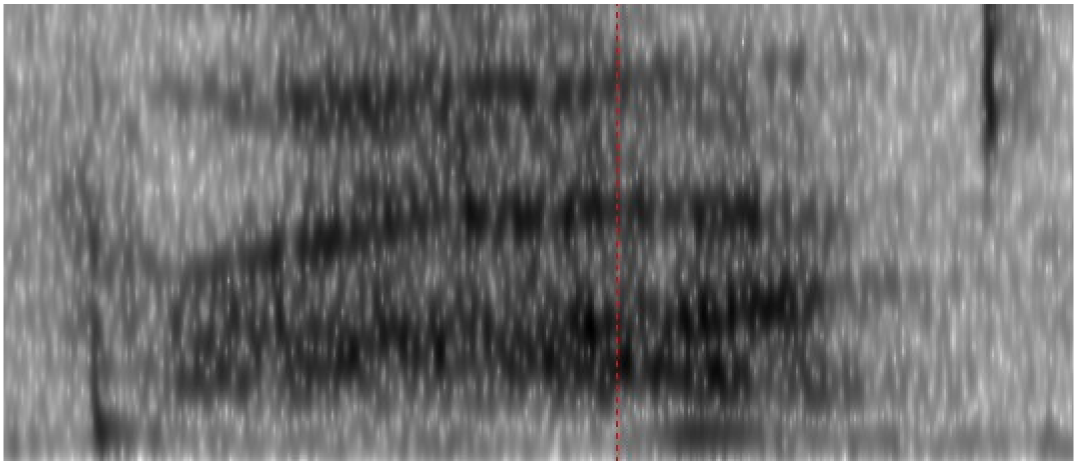


Figure 4a: Spectrogram for [rɑt] (recording 4)

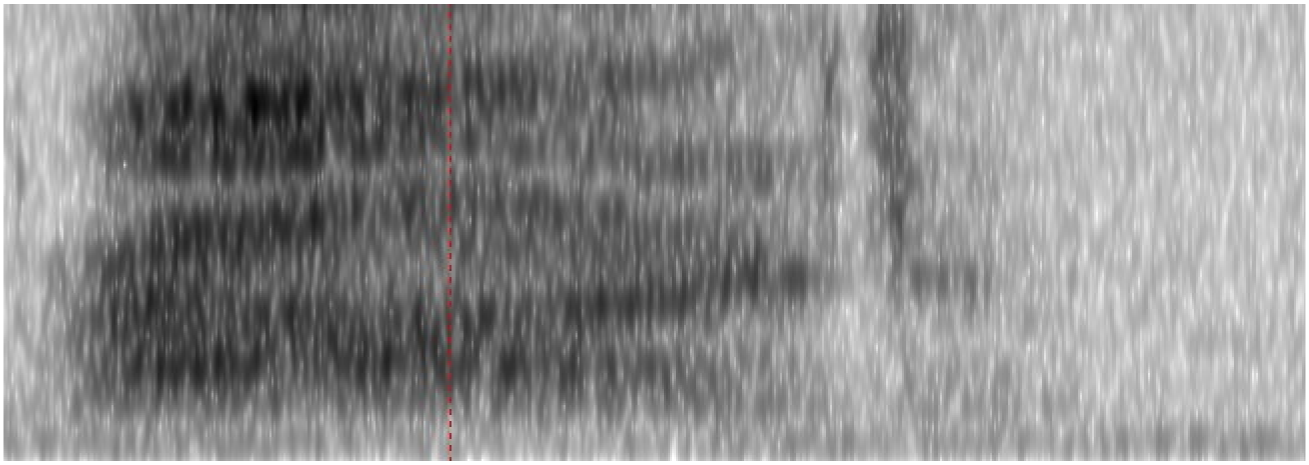


Figure 4b: Spectrogram for [rɑd] (recording 3)

Indeed we do. In [rɑt], there is a clear horizontal gap between the vowel and the final consonant. But in [rɑd] there is a very short (arguably no) gap. We might guess that this is because a vowel – a voiced sound – leads more easily into the voiced stop [d] than the voiceless stop [t]. This analysis begs another question: why was [t] misidentified more often than [d]? We must remember that each participant only heard one word from each of the minimal pairs, so no participant would have heard both [rɑt] and [rɑd] and been able to compare them. Upon hearing [rɑd], the smoothness of the close transition from vowel to [d] probably made it clear to the listener that the final consonant could not have been [t]. When hearing [rɑt], though, the large gap heard does not absolutely *preclude* the possibility of the final consonant being [d]; thus, in combination with the "overcompensation" effect described earlier, listeners might occasionally misidentify the [t] as a [d].

The last unusual result we observe is in the final pairs [f]/[v] and [s]/[z]; namely, that the voiced consonants [v] and [z] are correctly identified more often than we expect based on our hypothesis. Figures 5 and 6 show spectrograms for the relevant word pairs.

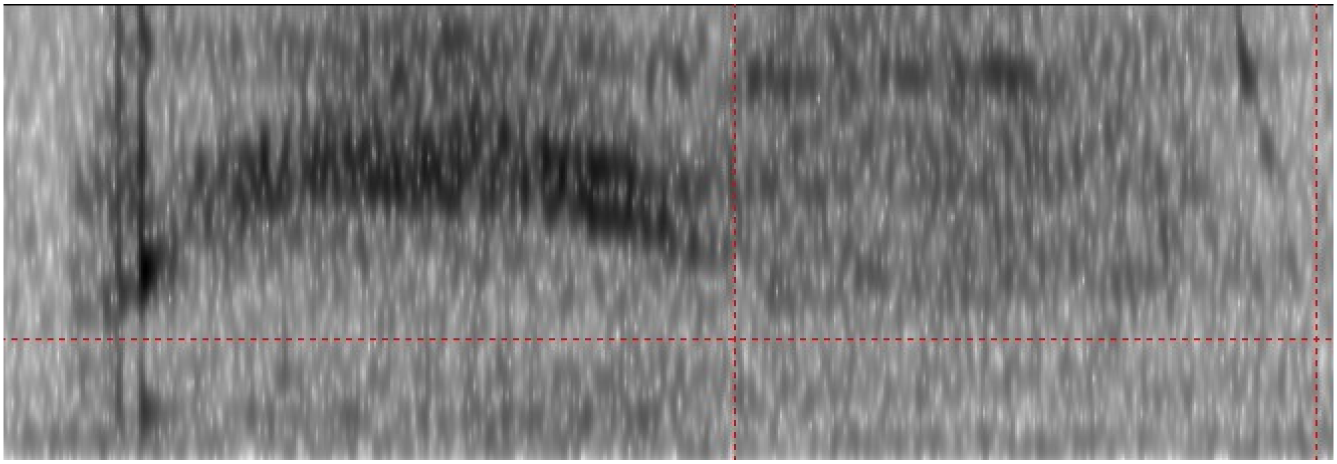


Figure 5a: Spectrogram for [lif] (recording 4)

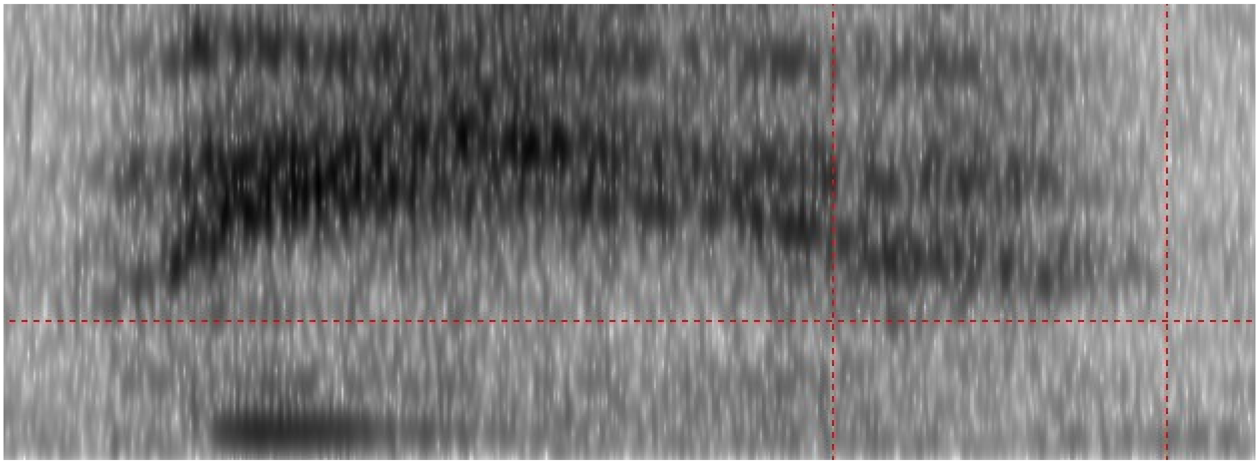


Figure 5b: Spectrogram for [liv] (recording 3)

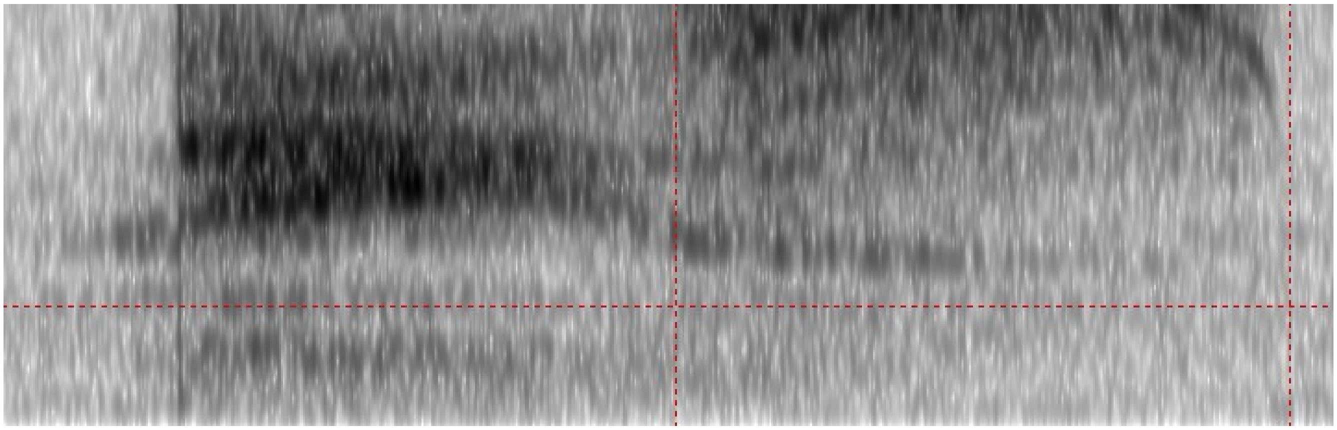


Figure 6a: Spectrogram for [nis] (recording 1)

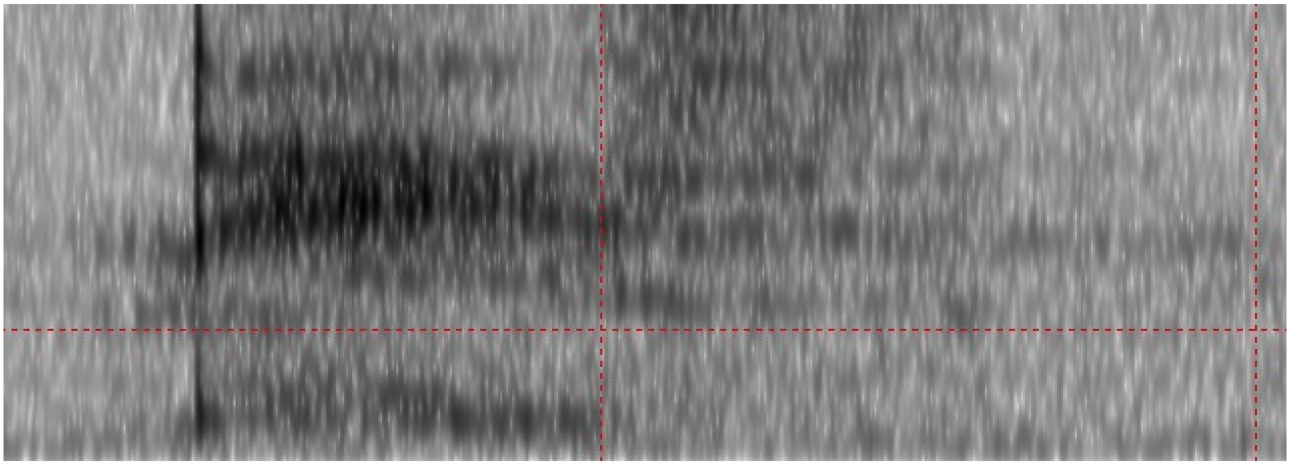


Figure 6b: Spectrogram for [niz] (recording 2)

In both of these pairs, we see something interesting occur with the final consonant (selected in both figures). With the voiceless final consonant, the spectrograms look normal – we see a region of turbulent sound as we expect with fricatives. However, with the voiced consonants, there seems to be a trace of formants extending from the preceding vowel, as we might expect to see with approximants in normal speech. This observation appears to contradict our hypothesis that all consonants are pronounced unvoiced in whispered speech, and this apparent voicing explains the much higher rate of correct identification of these voiced fricatives.

Conclusion

After conducting this experiment as well as examining the spectrograms, we conclude that our hypothesis is generally correct: whispered speech limits voicing of consonants. In response to our original questions, we also note that – importantly – whispering does not affect the voicing of vowels. We recognize that while some pairs of consonants are almost completely indistinguishable in whispered speech, both being heard as voiceless consonants, other pairs have particular acoustic features which help distinguish the voiced consonants even when voicing is removed.

With the limited scope of our experiment, it is difficult to make robust conclusions about what environments these distinctions appear in; however, our results indicate that the position of the consonants (initial or final) affects their distinguishability. We also hypothesize that the particular vowel preceding or following a consonant can affect its distinguishability. Lastly, we note that, as shown in the last set of spectrograms, whispering may not completely eliminate voicing for fricatives in certain environments. Further experimentation and careful analysis of spectrograms would be useful for determining in what environments and to what extent whispering devoices voiced consonants.

Appendix A: Experimental Materials

Listing 1: Code used to generate 4 randomly-ordered strings of words, each containing one word for each consonant sound in initial and final positions.

```
import random

initials = [['pay', 'pee', 'pie'],
            ['bay', 'be', 'buy'],
            ['tie', 'toe', 'ton'],
            ['die', 'doe', 'done'],
            ['could', 'came', 'core'],
            ['good', 'game', 'gore'],
            ['fail', 'fine'],
            ['veil', 'vine'],
            ['sue', 'seal'],
            ['zoo', 'zeal'],
            ['choke', 'cheer', 'char'],
            ['joke', 'jeer', 'jar'],
           ]

finals =   [['rope', 'lap', 'rip'],
            ['robe', 'lab', 'rib'],
            ['at', 'neat', 'rot'],
            ['add', 'need', 'rod'],
            ['rack', 'lock'],
            ['rag', 'log'],
            ['half', 'leaf'],
            ['have', 'leave'],
            ['race', 'niece'],
            ['raise', 'knees'],
           ]

all_words = initials + finals

for i in xrange(4):
  word_set = []
  for row in all_words:
    word_set.append(row[i%len(row)])
    i += 1
  random.shuffle(word_set)
  print "Whispered:"
  print " ".join(word_set)
  print
```

Listing 2: Randomly-ordered strings of 22 words. Generated by code in Listing 1.

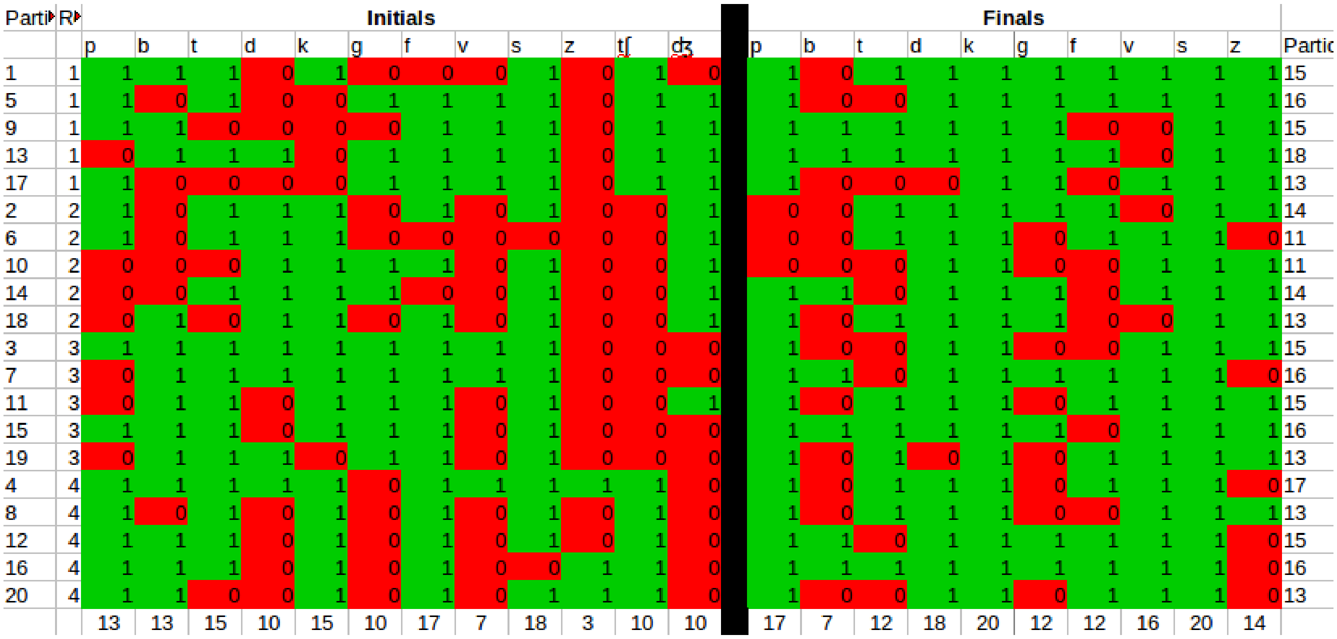
#1: be rack half leave ton log pay gore came die fail rot cheer add sue lab zeal jar niece rope vine raise

#2: leaf at good doe veil rag lap race char pee rib fine tie seal have knees zoo need core lock joke buy

#3: rack jeer zeal neat bay choke vine done toe fail rip robe sue rod game leave pie knees race could half log

#4: ton seal gore cheer rag lab niece lock zoo veil die raise jar came rope add leaf rot be fine pay have

Appendix B: Complete Experimental Results



Parti	R	Initials													Finals										
		p	b	t	d	k	g	f	v	s	z	tʃ	dʒ		p	b	t	d	k	g	f	v	s	z	Partic
1	1	1	1	1	0	1	0	0	0	1	0	1	0		1	0	1	1	1	1	1	1	1	1	15
5	1	1	0	1	0	0	1	1	1	1	0	1	1		1	0	0	1	1	1	1	1	1	1	16
9	1	1	1	0	0	0	0	1	1	1	0	1	1		1	1	1	1	1	1	0	0	1	1	15
13	1	0	1	1	1	0	1	1	1	1	0	1	1		1	1	1	1	1	1	1	0	1	1	18
17	1	1	0	0	0	0	1	1	1	1	0	1	1		1	0	0	0	1	1	0	1	1	1	13
2	2	1	0	1	1	1	0	1	0	1	0	0	1		0	0	1	1	1	1	1	0	1	1	14
6	2	1	0	1	1	1	0	0	0	0	0	0	1		0	0	1	1	1	0	1	1	1	0	11
10	2	0	0	0	1	1	1	1	0	1	0	0	1		0	0	0	1	1	0	0	1	1	1	11
14	2	0	0	1	1	1	1	0	0	1	0	0	1		1	1	0	1	1	1	0	1	1	1	14
18	2	0	1	0	1	1	0	1	0	1	0	0	1		1	0	1	1	1	1	0	0	1	1	13
3	3	1	1	1	1	1	1	1	1	1	0	0	0		1	0	0	1	1	0	0	1	1	1	15
7	3	0	1	1	1	1	1	1	1	1	0	0	0		1	1	0	1	1	1	1	1	1	0	16
11	3	0	1	1	0	1	1	1	0	1	0	0	1		1	0	1	1	1	0	1	1	1	1	15
15	3	1	1	1	0	1	1	1	0	1	0	0	0		1	1	1	1	1	1	0	1	1	1	16
19	3	0	1	1	1	0	1	1	0	1	0	0	0		1	0	1	0	1	0	1	1	1	1	13
4	4	1	1	1	1	1	0	1	1	1	1	1	0		1	0	1	1	1	0	1	1	1	0	17
8	4	1	0	1	0	1	0	1	0	1	0	1	0		1	0	1	1	1	0	0	1	1	1	13
12	4	1	1	1	0	1	0	1	0	1	0	1	0		1	1	0	1	1	1	1	1	1	0	15
16	4	1	1	1	0	1	0	1	0	0	1	1	0		1	1	1	1	1	1	1	1	1	0	16
20	4	1	1	0	0	1	0	1	0	1	1	1	0		1	0	0	1	1	0	1	1	1	0	13
		13	13	15	10	15	10	17	7	18	3	10	10		17	7	12	18	20	12	12	16	20	14	

Table 2: Spreadsheet of individual experimental results. 1 indicates participant identified the target consonant correctly; 0 indicates incorrect identification. Grouped by recording number in order, with participant totals on the far right and totals per sound in the bottom row.

Appendix C: Additional Spectrograms

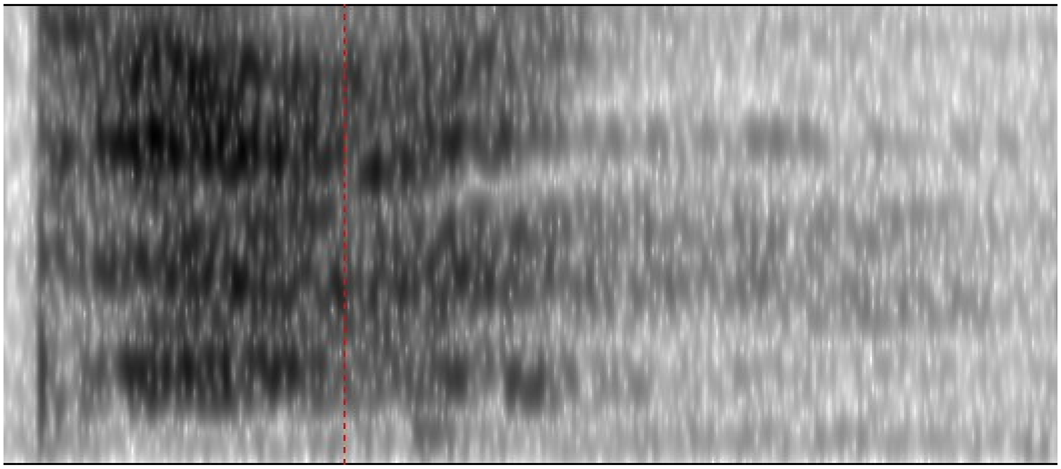


Figure 7a: Spectrogram for [tən] (recording 4)

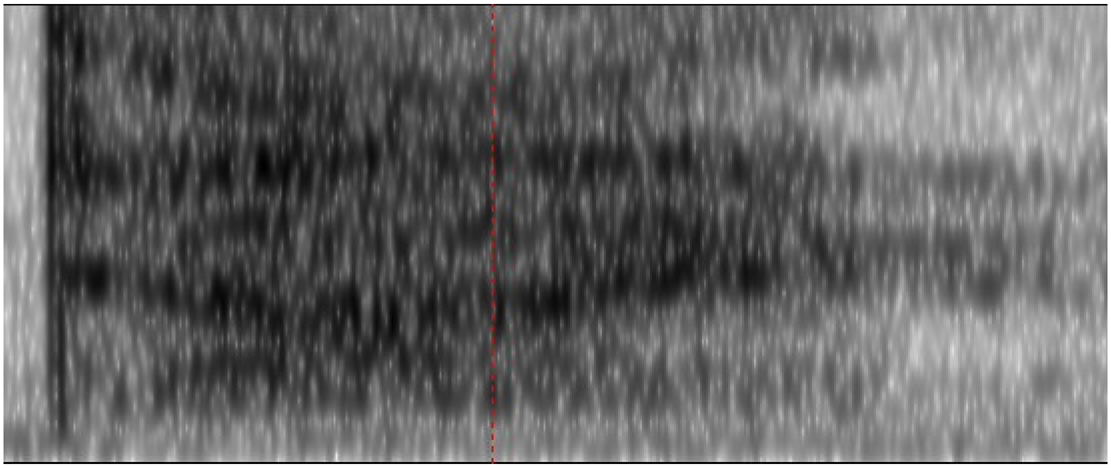


Figure 7b: Spectrogram for [dən] (recording 3)

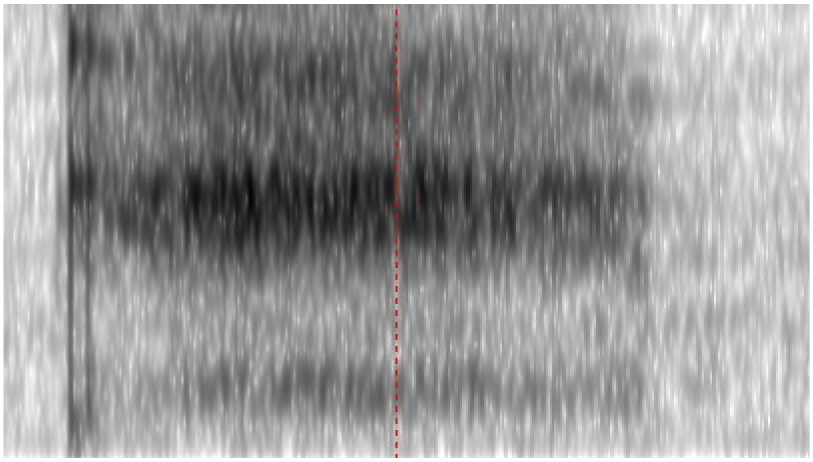


Figure 8a: Spectrogram for [keɪm] (recording 1)

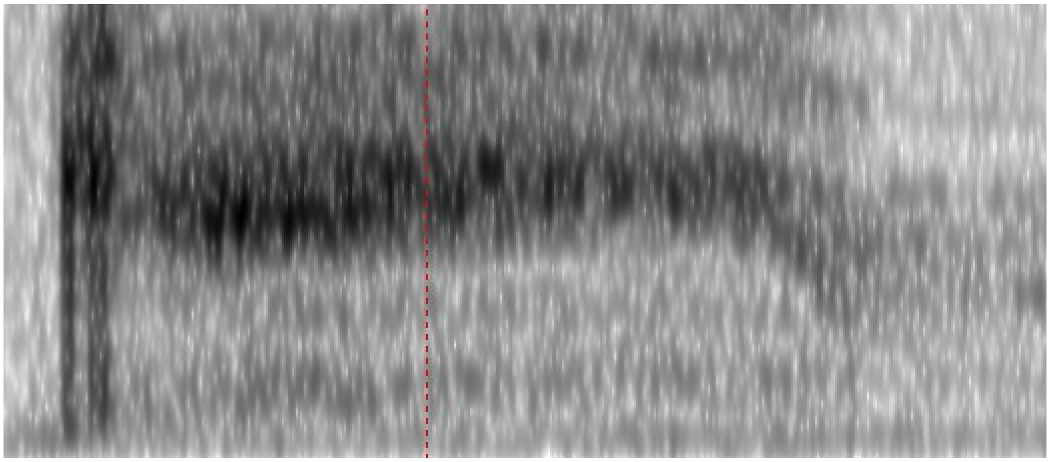


Figure 8b: Spectrogram for [geɪm] (recording 3)

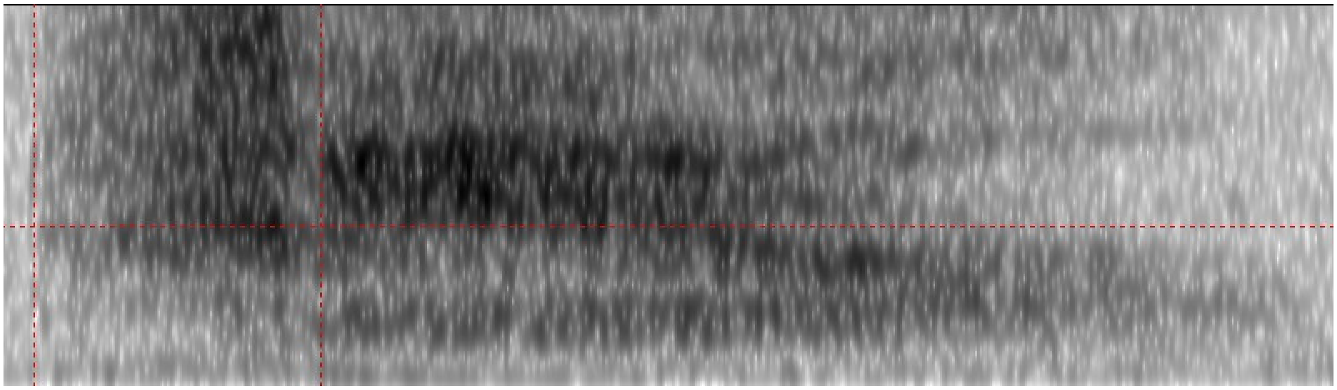


Figure 9a: Spectrogram for [feɪl] (recording 3)

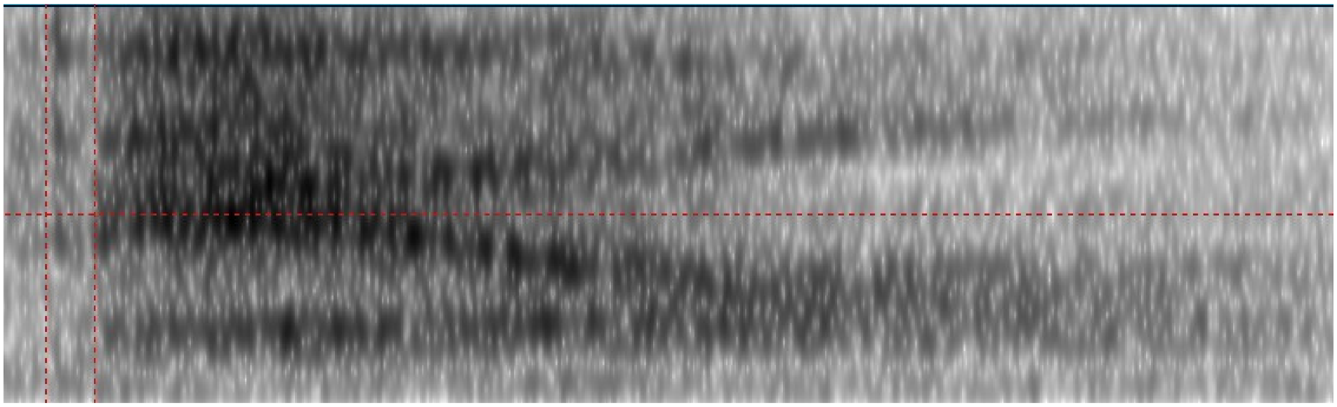


Figure 9b: Spectrogram for [veɪl] (recording 4). Note the much shorter initial consonant compared to 9a – this may have contributed to the higher recognizability of initial [v] compared to e.g. initial [z]

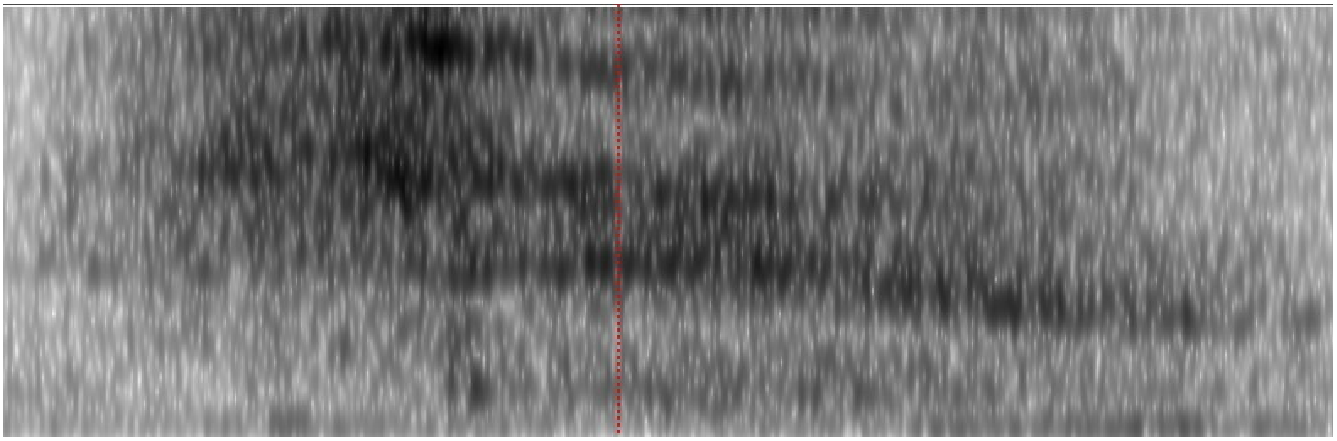


Figure 10a: Spectrogram for [su] (recording 3)

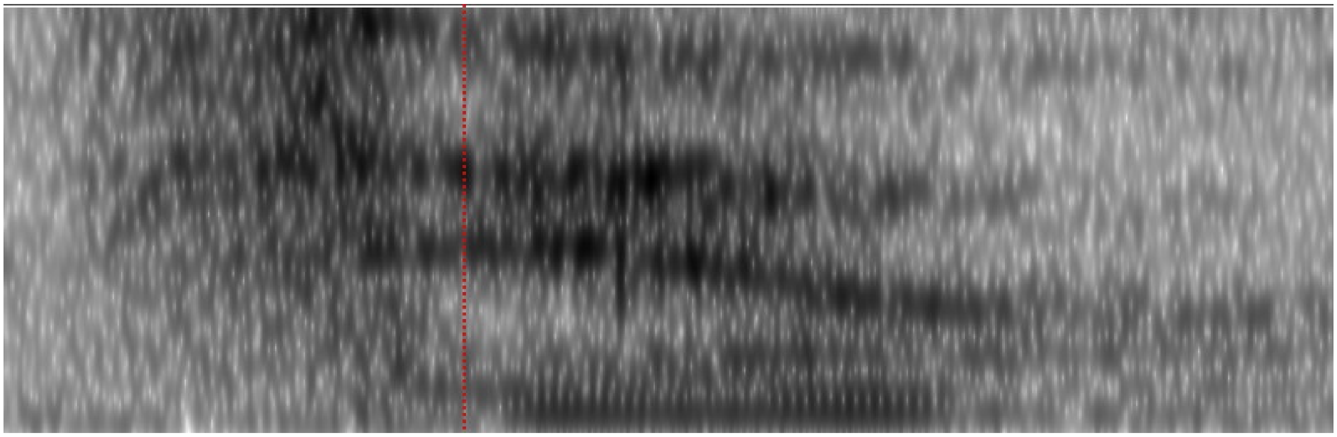


Figure 10b: Spectrogram for [zu] (recording 4)

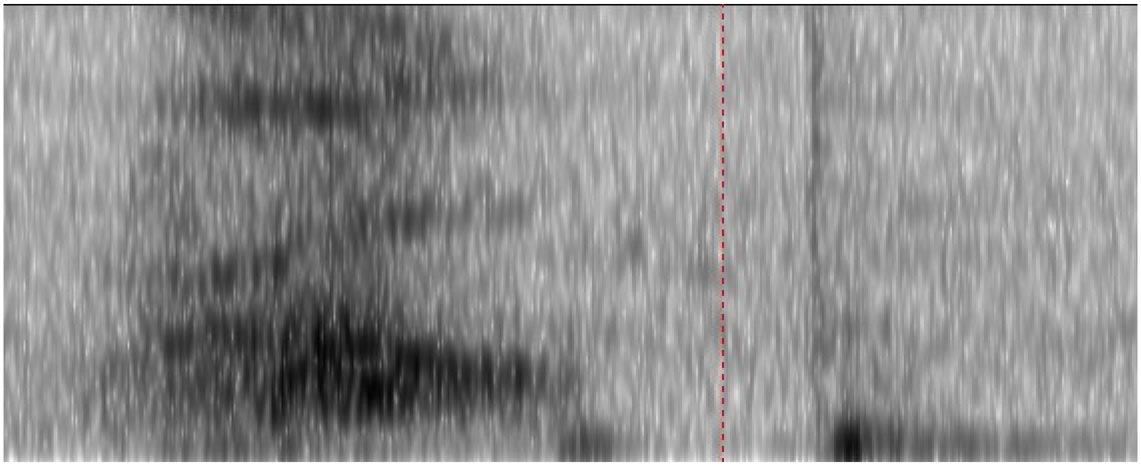


Figure 11a: Spectrogram for [rop] (recording 1)

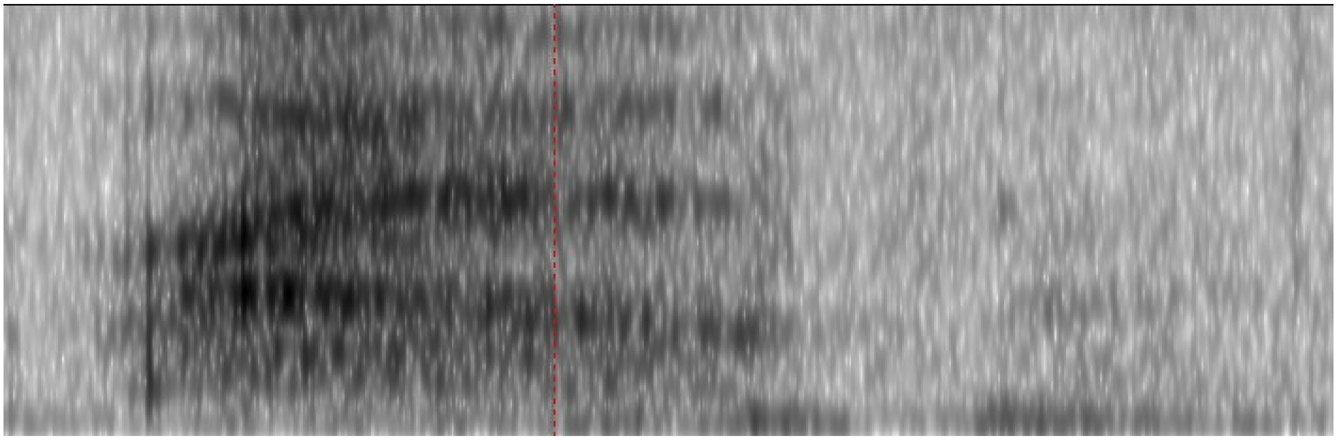


Figure 11b: Spectrogram for [rob] (recording 3)