Vowel variability in elicited versus spontaneous speech: evidence from Mixtec

Christian DiCanio[a,*], Hosung Nam[a], Jonathan D. Amith[b,c], Rey Castillo García[d], D. H. Whalen[a,e]

[a]*Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511*
[b]*Department of Anthropology, Gettysburg College, 300 N. Washington Street, Gettysburg, PA 17325*
[c]*Department of Anthropology, National Museum of Natural History, Smithsonian Institution*
[d]*Secretaría de Educación Pública, Avenida de la Juventud, Chilpancingo, Guerrero, C.P. 39090, Mexico*
[e]*Speech, Language, Hearing Sciences Program, Graduate Center, City University of New York, 365 5th Ave, New York, NY 10016*

Abstract

This study investigates the influence of speech style, duration, contextual factors, and sex on vowel dispersion and variability in Yoloxóchitl Mixtec, an endangered language spoken in Mexico. Oral vowels were examined from recordings of elicited citation words and spontaneous narrative speech matched across seven speakers. Results show spontaneous speech to contain shorter vowel durations and stronger effects of contextual assimilation than elicited speech. The vowel space is less disperse and there is greater intra-vowel variability in spontaneous speech than in elicited speech. Furthermore, male speakers show smaller differences in vowel dispersion and duration across styles than female speakers do. These phonetic differences across speech styles are not entirely reducible to durational differences; rather, speakers also seem to adjust their articulatory/acoustic precision in accordance with style. Despite the stylistic differences, we find robust acoustic differences between vowels in spontaneous speech, maintaining the overall vowel space pattern. While style and durational changes produce noticeable differences in vowel acoustics, one can closely approximate the phonetics of a vowel system of an endangered language from narrative speech. Elicited speech is likelier to give the most extreme formants used by the language than is spontaneous speech, but the usefulness of phonetic data from spontaneous speech has still been demonstrated.

*Corresponding author

Email addresses: dicanio@haskins.yale.edu (Christian DiCanio), nam@haskins.yale.edu (Hosung Nam), jonamith@gmail.com (Jonathan D. Amith), castagr@hotmail.com (Rey Castillo García), whalen@haskins.yale.edu (D. H. Whalen)

Keywords: style, vowels, dispersion, variability, endangered languages, forced alignment, mixtec

1. Introduction

At a phonetic level, speech sounds are extraordinarily variable across languages and speakers. It is clear that some of this variability is conditioned by social and dialectal factors (see Foulkes and Docherty (2006); Labov (2001); Labov et al. (2006); Munson et al. (2006), and the references therein), and that language-specific phonological constraints also play a role. For instance, the number of vowels in a language may inversely correlate with the degree of permitted coarticulatory variability (Manuel, 1990, 1999). Even closely-related languages, like Cree and Chickasaw, permit different degrees of variability in the production of similar vowels (Gordon et al., 1997; Johnson and Martin, 2001). While these studies carefully investigate certain sources of phonetic variability, a common thread is their use of speech data restricted to a certain style. In order to test sociophonetic and phonological factors, style is controlled. By *style*, we refer to the spontaneity of the speech data (spontaneous vs. elicited), its embedding (citation vs. carrier sentences), care (careful, casual), rate (fast, slow), and its discourse context (narrative, oral history, conversation, etc.). While the strategy of controlling for style is completely natural for doing phonetic research, a better understanding of its effect on speech production would allow us to both make clearer predictions regarding how individuals vary and permit a more direct comparison of stylistically-divergent phonetic data.

Elicited speech is scripted and often reflects a careful style. By contrast, spontaneous and conversational speech are typified by increased vowel reduction (Aylett and Turk, 2006; Barry and Andreeva, 2001; Gahl et al., 2012; Harmegnies and Poch-Olivé, 1992; Meunier and Espesser, 2011; Smiljanić and Bradlow, 2005). Reduction is also typical of speech in read prose (compared with read wordlists) (Keating and Huffman, 1984) and faster speech rates (Van Son and Pols, 1992). A shared explanation for vowel reduction across these studies is the idea that vowels of shorter duration will fail to reach their target; they will undergo vowel *undershoot* (Lindblom, 1963, 1990; Moon and Lindblom, 1994). Yet, it remains unclear to what extent style imposes additional changes on speech production targets *independent* of duration, as the effects of style and duration are rarely disentangled. Moreover, if individuals vary in a consistent manner across styles, then one could more confidently compare data across styles. If they do not vary in a consistent manner, then our conclusions from careful speech data in a language should not be used to theorize about spontaneous speech data.

We address these questions in the current study where we examine vowel production in elicited citation words and spontaneous speech for native speakers of Yoloxóchitl Mixtec, an endangered Oto-Manguean language spoken in Guerrero, Mexico (Castillo García,

2007). We compared elicited vowel data across 10 native speakers to spontaneous speech data (narratives) from seven of the same speakers. We tested intra-category vowel variability, vowel dispersion, and durational differences in vowel production. These measures were examined in relation to style, vowel identity, speaker sex, and coarticulatory factors. Greater vowel reduction and centralization was observed in spontaneous speech, but strong effects of duration and context were also observed. These findings are discussed in relation to both the literature on vowel production and endangered language corpus data.

1.1. Background: Yoloxóchitl Mixtec Phonology

Yoloxóchitl Mixtec (ISO-639 xty), henceforth YM, is an endangered Mixtecan language spoken in and around the town of Yoloxóchitl in Guerrero, Mexico. There are approximately 2,500 speakers, all of whom speak the Yoloxóchitl variant of Mixtec. Nearby communities have already undergone a shift towards speaking Spanish, specifically in the younger generation. However, this has not yet occurred in Yoloxóchitl. The segmental phonology of the language is described in Castillo García (2007) and Amith and Castillo García (no date). Words may also contrast in glottalization, though this is best understood as a feature of bimoraic feet rather than a segment in the consonant inventory, e.g. /$\mathrm{ja^1a^1}$/ '*slow*' and /$\mathrm{ka^3ni^4}$/ '*drool (n.)*' vs. /$\mathrm{ja^1ʔa^1}$/ '*soot*' and /$\mathrm{ka^3ʔni^4}$/ '*fever (n.), kill (v.)*'. The phonetic properties of the consonant system are examined in DiCanio et al. (submitted). The consonant and vowel inventories are shown in Tables 1 and 2.

Table 1: Yoloxochitl Mixtec Consonant Inventory

	Bilabial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Labialized Velar
Plosive	p	t				k	$\mathrm{k^w}$
Nasal	m		n				
Post-stopped nasal	$\mathrm{m^b}$		$\mathrm{n^d}$			$\mathrm{ŋ^g}$	
Tap			ɾ				
Affricate				tʃ			
Fricative	β	s		ʃ			
Approximant			l		j		

There are five vowels in YM, each of which contrasts in nasality. Vowels are also obligatorily nasalized following a nasal consonant, e.g. /$\mathrm{na^3a^4}$/ '*dark*' is produced as [$\mathrm{nã^3ã^4}$]. Content words are minimally bimoraic, consisting of either a monosyllable with

Table 2: Yoloxochitl Mixtec Vowel Inventory

	Front	Central	Back
Close	i, ĩ		u, ũ
Close-mid	e, ẽ		o, õ
Open		a, ã	

a long vowel or a disyllable with two shorter vowels, e.g. /$\mathrm{ta^1a^3}$/ '*man*' vs. /$\mathrm{ke^1ta^3}$/ '*to enter*'. While most words have this structure, maximal words are trimoraic, e.g. /$\mathrm{ki^3ʃa^3a^4}$/ '*to begin*', /$\mathrm{ku^3n^di^3ka^2}$/ '*to be inside*'. While not the subject of the current study, the language is notable for its complex tonal inventory (DiCanio et al., 2014). Tone is assigned at the moraic level. Up to five tones may occur on the penultimate mora (1, 3, 4, 13, 14) and they may be combined with up to 8 tones on the ultimate mora (1, 2, 3, 4, 13, 24, 42, 32) (DiCanio et al., 2014). The tones here reflect values on the Chao tone scale, where /1/ is low and /4/ is high. Tones transcribed with a single value reflect phonological level tones and those transcribed with multiple values reflect contours. Many of these tonal combinations occur in both monosyllabic and disyllabic words, though the latter shows a greater number of possibilities. For instance, a total of 20 tonal melodies are contrastive on monosyllabic words and more in disyllabic words. The current paper presents the first instrumental study of Mixtec vowel acoustics. We will focus on the formant patterns of oral vowels, ignoring in this first analysis the more detailed questions of vowel/tone interactions and the acoustics of nasalized vowels.

1.2. Background

1.2.1. Linguistic and stylistic factors in vowel production

Research on the production of vowels has focused on their variation from contextual assimilation and from contexts where they become centralized, or *reduced*. Crucial to both themes is the idea that there exists an ideal vowel target for an individual speaker and that coarticulation and centralization reflect deviations from it.[1] The phonetic parameters which correlate with these deviations are the time window available for the production of the vowel and the demands on articulation made by adjacent consonants and vowels. Given a sufficiently short duration, the speech articulators may fail to reach an ideal vowel target, resulting in vowel *undershoot* (Lindblom, 1963, 1983, 1990; Meunier and Espesser, 2011). The more typical, reduced vowels approach a schwa-like vowel closer to the center

[1]Note, however, that it is also possible that a particular vowel have context-dependent targets (Stevens and House, 1963).

of one's vowel space (Moon and Lindblom, 1994). In a related way, vowels produced in contexts with adjacent sounds requiring very different articulations also diverge from an ideal target (Hillenbrand et al., 2001; Moon and Lindblom, 1994; Ohde and Sharf, 1975; Öhman, 1966, 1967; Recasens, 1984). For instance, a back rounded vowel in the context of alveolar or palatal consonants may be produced with substantial fronting and a higher F2 position than in other contexts (Flemming, 2003).

The effect of these two phonetic parameters on vowel production is well-established. Yet, to what degree the *non-phonological* factors contribute to changes in duration or contextual assimilation is an ongoing topic of investigation. Such factors include lexical category, redundancy, speech rate, speech style, and differences inherent to specific languages or individuals.[2] With respect to lexical category, function words are typically shorter than content words. Investigating vowel reduction using a corpus of conversational French, Meunier and Espesser (2011) found a relationship between vowel duration and centralization of the vowel space. One of the contributors to the durational differences across the data was the content/function word distinction. Vowels in monosyllabic content words had an average duration of approximately 70 ms whereas those in monosyllabic function words had an average duration of approximately 57 ms. This latter value corresponded to a lowering of 60 Hz in F1 for /a/, though this was the only vowel the authors were able to investigate for the content/function distinction.

Vowel reduction has also been observed to correlate with differences in predictability or redundancy (Jurafsky et al., 2001; Aylett and Turk, 2004, 2006). That is, words with high probability tend to be produced with more centralized vowels than words of low probability. Investigating a corpus of English citation speech, Aylett and Turk (2006) find a strong relationship between lexical redundancy and syllable duration. This relationship is offset by prosodic prominence, where syllables with greater prominence were longer (see also de Jong (1995)). These duration differences correlated with a significant, but small degree of vowel centralization. These effects were stronger for F1 than for F2 and substantial vowel-specific differences were observed. The relative weakness of the spectral effects here may reflect the fact that the corpus consisted entirely of citation speech when compared with the much stronger effect of duration on spectral differences shown in Meunier and Espesser (2011) for conversational speech. Note that the syllable durations reported in Aylett and Turk (2006, 3054) lie between 120 - 400 ms, whereas the mean *vowel* duration found in Meunier & Espesser is 73 ms. If one assumes that vowels are generally longer than consonants (Kewley-Port et al., 2007; Perkell et al., 2004; Pisoni, 1973), then this latter value would lie at the low end of the predicted vowel durations from the syllable duration values. Differences in style (and language, see section 1.2.2) may be

[2]Neighborhood phonological density also plays a role in phonetic reduction (Gahl et al., 2012).

responsible for differences in the degree of vowel reduction between these studies.

While it is possible to observe stylistic differences *across* studies which analyze different types of data, rather few studies explicitly focus on speech style. In a study on vowel variation in five Japanese speakers, Keating and Huffman (1984) investigate the degree of acoustic overlap produced in elicited and read speech. They observe greater overlap among vowel categories in read prose than in elicited tokens, though the variability was asymmetric. The high vowel /ɯ/ was produced with substantially greater F2 variability in prose, while the vowel /a/ is produced with substantially greater F1 variability. The authors argue that the result of such reduction in prose is to "preserve the skewing of the vowel allophones towards the high front region of the space" (p.201, ibid).

In an extensive study on Dutch speech style, Koopmans-van Beinum (1980) investigated vowel production for ten speakers for vowels produced in isolation, in isolated words, in read speech, in a retold short story, and in free conversation. A strong correlation was found between vowel duration and the Euclidean distance between a vowel in an F1xF2 space and the global centroid of the speaker's vowel space. These effects were strongest for the more peripheral vowels /i, u, a/ and weakest for the open-mid and front rounded vowels /y, ø, œ, ɛ, ɔ/, which contain more centralized F1 and F2 values (ibid, p.66). Vowel duration was longest for vowels and words produced in isolation, shorter for words produced in read speech and retold speech, and shortest in conversational speech. This stylistic effect was also mediated by distinctive vowel length; greater durational differences with changes in speech style were found for long vowels than for short vowels.

Moon and Lindblom (1994) investigate the relationship between duration and speech style in English by comparing front vowel production in citation forms to those produced in clear speech. Clear speech is typified by an expansion of the vowel space and tighter clustering of vowels within categories (Chen, 1980; Chen et al., 1983). Vowel duration was examined in relation to formant displacement, which was defined as the Euclidean distance between the formant values for a given vowel in the fixed context /wVl/ and its formant values in a more neutral /hVd/ context. The authors found citation form vowels to be shorter and more prone to contextual assimilation than clear speech vowels. While there was a strong relationship between duration and formant displacement across all speakers, duration was not sufficient to account for all the observed patterns of vowel undershoot. In particular, the relationship between duration and variability in formant displacement was weaker for clear speech than for citation forms.

These studies highlight robust methodological differences in the estimation of vowel undershoot (or formant displacement). Meunier and Espesser (2011) analyzed raw formant values without any comparison to a vowel or speaker-specific centroid, while Koopmans-van Beinum (1980) analyzed formant values in relation to a speaker-specific centroid. Both Moon and Lindblom (1994) and Gahl et al. (2012) examined formant values in re-

lation to a citation-target value. These latter two methods provide a basis for measuring formant displacement both in terms of intra-categorical variability and inter-categorical variability (classic dispersion), a factor used in the current study. These two measures allow us to determine whether undershoot across speech styles is produced more by shifting specific vowel centroids towards the center or by simply increasing the variability within categories. Duration is predicted to influence both these measures, but these different measures are rarely distinguished in studies on vowel reduction. If we find greater undershoot in spontaneous speech to be determined primarily by increases to inter-categorical variability, this means that style induces more fundamental changes to the basic structure of a vowel space. If we find it to be determined by intra-categorical variability, it means that spontaneous speech style simply increases variability within a category, but one can more directly compare mean formant values found in spontaneous speech corpora with those found in elicited speech. These considerations have ramifications for research using spontaneous speech corpora from endangered languages.

1.2.2. Endangered languages and language-specific differences in vowel production

The studies above highlight differences in the directionality and degree of reduction across languages. For instance, reduction in Japanese asymmetrically involves a fronting of the vowel space, while it appears to be more symmetrical in Dutch. One source for these differences is inventory size. That is, languages with larger vowel inventories may restrict vowel undershoot so as to maintain perceptual distinctiveness within a more crowded vowel space. Manuel (1990, 1999) found that Shona and Ndebele (Bantu), which have 5-vowel systems, exhibit greater anticipatory vowel-to-vowel coarticulation than Sotho, which has a 7-vowel system. Mok (2010) finds that English undergoes greater vowel-to-vowel coarticulation than Thai does, regardless of the interval duration between successive syllables. While not explicitly discussed in her work, Thai has a larger vowel inventory than English, which may have also contributed to this difference. [3]

Despite some work on cross-linguistic differences in vowel production, our knowledge of variability in vowel production is very limited. Investigating short vowel reduction in Creek (Johnson and Martin, 2001, 96) note that "*...though there may be a set of general causes for short vowel centralization - coarticulation with neighboring consonants and/or differences in degree of overall vocal effort, for example - these motivating factors are managed differently by speakers of different languages.*" While the dimensions on which different languages vary may eventually be shown to be more universal predictors, the

[3] Another source of vowel reduction variability is speech rate. Speech rate is known to differ substantially across individuals, languages, and dialects of the same language (Kendall, 2009; Pellegrino et al., 2011; Verhoeven et al., 2004), though it has not been investigated as a factor in cross-linguistic vowel reduction.

language-specific characteristics, even among languages with similar inventories, remain robust. Just how these dimensions are managed in YM is an additional motivation for the current study.

A final motivation lies in assessing the validity of endangered language documentation corpora for descriptive phonetic fieldwork. In relation to using corpora, Ladefoged (2003) states:

> From a phonetician's point of view, there is *no point* in making lengthy recordings of folk tales, or songs that people want to sing. Such recordings can seldom be used for an analysis of the major phonetic characteristics of a language, except in a qualitative way. You need sounds that have all been produced in the same way so that their features can be compared. (p.9, Ladefoged (2003), emphasis ours)

Corpora from endangered languages are often collected with high quality audio recordings, but it remains to be shown how such data could be used in phonetic analysis. Given that these recordings usually consist of spontaneous speech data and folk tales, there are clear stylistic differences between the documentation corpora on the one hand and the more careful speech recommended by Ladefoged on the other. In this study, we evaluate Ladefoged's claim by comparing careful, elicited speech with speech in personal narratives and folk tales.

2. Methods: elicited and spontaneous vowel production

2.1. Speech materials

The elicited data set for the current study comes from a corpus of 261 isolated words (see also DiCanio et al. (2013)). Speakers were asked to repeat each word six times for a total of 15,660 word tokens (261 words x 6 x 10 speakers). However, many speakers reproduced more than six repetitions for each word and the analyzed number of tokens totalled 17,880. The words were originally elicited to explore differences in the production of tone in words of different sizes. Of the 261 words, 169 were disyllabic, 89 were monosyllabic, and three were trisyllabic. While some words contained additional tonal morphemes (tone marks verb aspect, negation, and person), the majority (177/261) were monomorphemic.[4] The entire corpus was hand-labeled by Leandro DiDomenico, a linguistics graduate student at Université Lyon 2 and hand-corrected by the first author. This data set consisted

[4]Note, however, that the morphological complexity of verbs is problematic here since there is no infinitival form for most verbs. Verb stems either carry the tonal melody of the potential aspect or the completive. Thus, most verbs are inherently polymorphemic.

of a total of 27,973 vowels. Of these, 5,806 were nasalized and were excluded from the analysis, leaving a total of 22,167 analyzed vowels. Nasalized vowels were excluded on lexical grounds, i.e. phonologically nasal vowels.

The spontaneous vowels occurred in a set of 8 spontaneous speech recordings produced by 7 speakers (we used two samples for one speaker), taken from a large-scale language documentation corpus collected and transcribed by Amith and Castillo García. Spontaneous speech data was analyzed from four males and three females. The recordings consisted of personal narratives and stories. The speech extracts were chosen based on quality of recordings and comparability across speakers. For a few speakers, only one spontaneous extract was available whereas for others, more than an hour was available. Rather than compare a substantial amount of data from a single speaker to rather less from another, we chose to limit our spontaneous speech to 25 minutes per speaker. Yet, note that within a 10 minute extract, there are at least several hundred instances of each vowel, comparable to the number of tokens in the elicited corpus. The average speech extract duration was 12 minutes, but individual's extracts varied between 6 and 22 minutes. A total of 95 minutes of spontaneous speech was analyzed. The spontaneous data was segmented using the University of Pennsylvania forced alignment system (Yuan and Liberman, 2008, 2009). Vowel boundaries were hand-corrected by the first two authors. This data set consisted of a total of 23,050 vowels. Of these 6,831 were nasalized and were excluded from the analysis leaving a total of 16,219 analyzed vowels.

These two corpora reflect distinct speech styles and were originally collected to address two issues in language documentation: phonetics and phonology (elicited corpus) and morphosyntax and semantics (spontaneous speech). The distribution of vowels was found to be similar across data sets. Table 3 provides the number of each vowel found in each corpus. The ranking of vowels in terms of frequency is identical across the corpora, though /a/ is less common in spontaneous speech while /i/ is more common.

Table 3: Oral vowel frequency across corpora

Elicited			Spontaneous		
Vowel	Number	Percentage	Vowel	Number	Percentage
/i/	3,960	17.9%	/i/	4,854	29.9%
/e/	1,063	4.8%	/e/	1,209	7.5%
/a/	11,519	52.0%	/a/	6,047	37.3%
/o/	2,202	9.9%	/o/	1,729	10.7%
/u/	3,420	15.4%	/u/	2,380	14.7%

2.2. Speakers and data collection

For the elicited data, 10 speakers were selected from the Yoloxóchitl Mixtec community: four females and six males. Their mean age was 41.6 years old; four were 51–71 years old, while six were 22–38 years old. All participants were fluent native speakers of Yoloxóchitl Mixtec and were born in Yoloxóchitl. No participant reported having a history of speech or hearing disorders. The wordlist was randomized and elicited in two parts. As there is limited literacy and no standard orthography in YM, the wordlist was elicited by having the speaker repeat Castillo García's production of the word. Castillo García is a native speaker of YM. For the spontaneous data, there were, as mentioned, 7 speakers, all of whom had also provided elicited data. For the recording session, Castillo García began the discourse in YM, asking the speaker to provide a personal narrative or story. The 8 stories chosen here were all personal narratives. Most of these speakers were bilingual in Spanish, as is typical in most of Mexico. However, Spanish is not used in Yoloxóchitl between Mixtec speakers and thus YM remains the dominant language for this community. Both the elicited and spontaneous data were recorded in a quiet room in the nearby town of San Luis Acatlán on a Marantz PMD671 portable audio recorder with a Shure SM10A head-mounted microphone.

2.3. Measures and coding

All vowel segments were extracted from the corpora with the aid of a script for Praat (Boersma and Weenink, 2013) written by the first author. This script generated 8 acoustic measures: vowel duration, F1, F2, F3, center of gravity, standard deviation, skewness, and kurtosis. Of these, only duration, F1, F2 values were are analyzed in the present paper. For all spectral measures, the mean value from each of three equal intervals over the duration of the vowel was extracted. The decision to extract dynamic data over three points (rather than more) was motivated by the lower limit of vowel duration in the spontaneous speech corpus: many vowels were shorter than 60-70 ms, which cannot be easily subdivided beyond three time points. All speech data was first resampled at 16 kHz (from 48 kHz) to meet the requirements of the forced alignment program, and formant measurement was done using the LPC covariance method with a prediction order of 16 and a window size of 15 ms.

While it is typical to use a window size of 25 ms for formant tracking, we used a smaller that average window size here as this allowed us to extract dynamic measures from vowels as short as 40 ms. Only vowel midpoint values were extracted from vowels shorter than 40 ms, not dynamic data. Vowels shorter than 40 ms. comprised just 3.3% of the total data and came almost entirely from the spontaneous speech corpus. Thus, excluding dynamic data from these vowels does not overly bias the data toward longer duration vowels. We also examined whether using this smaller window size for formant estimation

resulted in erroneous formant values for longer vowels for a given speaker by comparing window sizes of 15 and 25 ms We found a maximum difference of approximately 5 Hz for F1 values and 20 Hz. for F2 between data estimated with different window sizes. However, most formant values were identical across the two window sizes. The covariance method was used as it more consistently estimated formant values across time than the burg method in a small speech sample we tested (see also Shadle et al. (2013)).

In the data set we used for analysis, each vowel token was coded for style (spontaneous vs. elicited); speaker; vowel; the preceding segment identity and its place of articulation; and the following segment identity and its place of articulation. Given our use of dynamic measurements, the data was also coded for time (with three levels). As a comparison point for individual vowel formant values, two sets of centroids were calculated: one representing the mean formant values for the entire vowel space for a given speaker in a given style (10 speakers (elicited) + 7 speakers (spontaneous) = 17 centroids), and another representing the mean formant values for individual *vowels* for a given speaker in a given style (5 vowels * (10 speakers (elicited) + 7 speakers (spontaneous)) = 85 centroids). The former was used for measuring *vowel dispersion*, while the latter was used for measuring *intra-categorical variability* in vowel production.

2.4. Notes on normalization

Studies on vowel production differ dramatically in their use of normalized formant values. Non-normalized formant values have been used in several studies (Moon and Lindblom, 1994; Meunier and Espesser, 2011; Aylett and Turk, 2006), with a particular aim of *"avoiding any theoretical assumptions concerning the relationship between human vowel production and perception"* (Aylett and Turk, 2006, 3050). Investigating phonetic variation across vowel inventories in a large corpus of data, Becker-Kristal (2010) argues that log scale distances between vowels and centroids tend to overestimate the degree of centralization in lower frequencies, and he instead uses a linear scale in his study. This argument stands in contrast to practice in sociolinguistics of applying a log mean normalization (Labov, 2006; Labov et al., 2006). We chose to analyze z-score normalized, linear formant values here, following the typical procedure for z-score normalization (Johnson, 2008). Mel and bark scales have also been used in other studies; though they are more linear than logarithmic in the frequency region covered by F1 and F2. Note, however, that since the dependent variables considered in this study are the distance between speaker-specific vowel productions and their centroids, our data carry an additional by-speaker or by-vowel normalization.

3. Results: elicited and spontaneous vowel production

The results of this study were evaluated using linear mixed effects models with random effects (Baayen, 2008). These models are particularly appropriate to the current data as mixed models allow for a combination of continuous and discrete predictors, permit the inclusion of by-subject and by-item random effects, and do not require design balance; an important requirement in spontaneous corpus data. Four series of linear mixed effects models were used. Each series consisted of two separate models constructed for the first two formants. The first series treated vowel dispersion as the response variable. Vowel dispersion was calculated as the distance between the z-score normalized value of a given formant and the mean for that formant in a speaker's vowel space for a particular style (spontaneous vs. elicited). This measure was extracted from the middle third of each vowel duration. Condition (spontaneous vs. elicited), vowel (i, e, a, o, u), duration (continuous), and speaker sex were treated as fixed effects. Random intercepts were included for Speaker and Word and a random Condition by Speaker slope was also included. Given the imbalance in word types across the spontaneous speech data, models with more fully-specified random effects structures for Word failed to converge.

The second series used intra-vowel variability as the response variable. This measure reflects the difference between the z-score of a formant value from the middle third of the vowel duration and the formant mean for that *vowel category* for a speaker in a particular style. The third and fourth series were identical to the first two models, respectively, except for the inclusion of three additional fixed effects: preceding consonant place of articulation, following consonant place of articulation, and time (with three levels: initial, medial, final). Adjustments made for these models are described in section 3.2.1. The effects of predictor variables were evaluated by using an analysis of variance on the linear mixed effects model. This model relies on the Satterthwaite method to approximate for degrees of freedom via the *lmerTest* (Kuznetsova et al., 2013) and reports both an F statistic and p values, but only a lower bound on degrees a freedom. There is currently no way to approximate the upper bound on degrees of freedom for linear mixed effects models (Baayen, 2008). All statistics were calculated using R (R Development Core Team, 2013).

3.1. Dispersion

Table 4 shows a summary of descriptive statistics for the data. Condition was a significant predictor in the model for both F1 ($F[1] = 11.8$, $p < .01$) and F2 values ($F[1] = 52.8$, $p < .001$). The mean dispersion of F2 was greater than that of F1. The average dispersion for F1 in elicited speech was 117.8 Hz, compared to 87.1 Hz in spontaneous speech. Meanwhile, the average F2 dispersion in elicited speech was 545.3 Hz, compared to 391.2 Hz in spontaneous speech.

Table 4: Summary descriptive statistics for data (means). All formant values are given in Hertz and all duration values are given in milliseconds. Formant values here reflect average values across the middle third duration for each vowel.

		Elicited						**Spontaneous**					
	Vowel	F1	s.d.	F2	s.d.	Duration	s.d.	F1	s.d.	F2	s.d.	Duration	s.d.
Female	i	433.0	54.3	2751.9	211.7	237	113.1	455.5	108.1	2523.0	305.3	81	59
	e	528.3	73.5	2558.2	133.6	276	143.9	560.7	93.7	2283.0	288.4	95	52
	a	895.6	85.1	1703.2	121.2	203	95.1	768.3	137.6	1799.8	233.3	94	60
	o	582.3	74.9	1232.4	406.8	262	146.8	566.9	73.3	1431.5	333.3	90	52
	u	483.6	82.1	1349.5	527.4	212	98.5	499.5	134.5	1589.4	452.3	77	64
Male	i	349.2	45.7	2223.3	205.4	199	101	372.8	69.8	1939.5	220.8	96	78
	e	447.4	58.1	2037.5	171.3	235	123	443.4	56.9	1728.9	204.8	100	60
	a	692.5	84.7	1418.9	133.2	176	86	543.6	95.8	1405.8	201.7	98	68
	o	472.1	63.9	961.0	184.4	214	120	434.2	62.1	1093.9	204.8	99	69
	u	388.8	71.7	983.0	238.1	177	80	375.4	82.1	1112.0	308.8	85	54

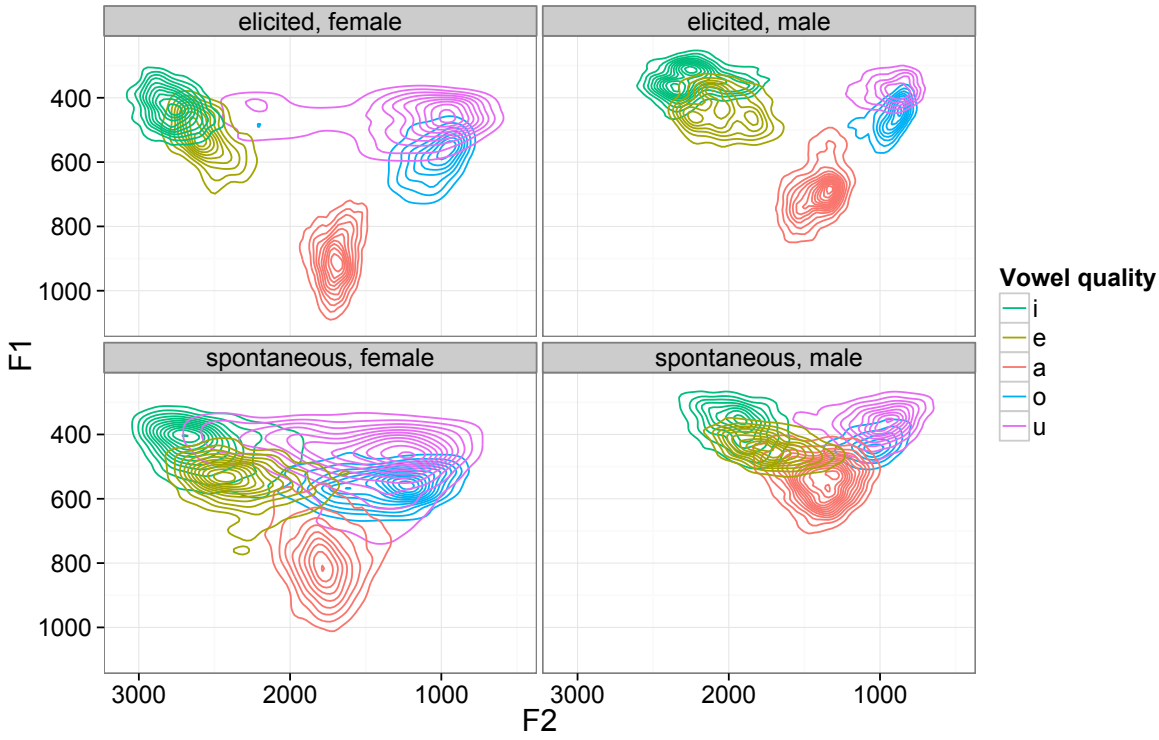

Figure 1: Vowel dispersion by style and sex. Formant values are drawn as two-dimensional contours using kernel density estimation, a smoothing algorithm which assumes randomly-distributed values around the median.

Figure 1 shows vowel dispersion by style and sex. In this figure, we observe a marked change in vowel space across styles. Within the elicited vowel space, some overlap occurs in the F1 dimension between the high close /i, u/ and the mid close vowels /e, o/. With the exception of certain fronted /u/ productions among females, front, central, and back vowels all seem to occupy distinct F2 regions. Within the spontaneous token vowel space, there is a noticeable expansion of the peripheral vowels into the center of the vowel space, causing greater overlap in F2 values across vowels. This change in overlap was greater than the changes in F1 value, which were mainly restricted to the vowel /a/. While dispersion varies significantly by vowel, significant interactions of vowel and style were also observed for both the F1 model ($F[4] = 14.5$, $p < .001$) and the F2 model ($F[4] = 27.7$, $p < .001$).

While Figure 1 shows clear differences in non-normalized formant values, sex was not a significant predictor of the degree of dispersion of F1 or F2 in the data, nor were there

any interactions between sex and style. However, a significant interaction between sex and vowel was found both for the F1 model ($F[4] = 4.4$, $p < .01$) and the F2 model ($F[4] = 20.9$, $p < .001$). These interactions correspond to less dispersion among males for the vowel /a/ (it is more centralized than it is for females) and greater dispersion among females for the vowel /u/ (it is more fronted than it is for males). Note that the former effect is more difficult to ascertain from the figure, but the latter pattern is clear. For females, F2 values for /u/ in spontaneous speech may extend near the median values for the distribution for /i/, whereas for males; they almost never overlap the F2 values for /i/.

Independent from style, duration had a strong significant effect on vowel dispersion in the F1 model ($F[1] = 32.4$, $p < .001$) and especially in the F2 model ($F[1] = 944.1$, $p < .001$). An interaction between Condition and Duration was found as well, both for the F1 model ($F[1] = 10.7$, $p < .001$) and the F2 model ($F[1] = 29.2$, $p < .001$). The effect of duration on formant dispersion depended on style. The average vowel duration in elicited speech was more than twice as long as that in spontaneous speech (see Table 4). As duration increased, both F1 and F2 dispersion decreased. It is possible to visualize this effect by dividing vowel duration values in spontaneous speech into distinct ranges of duration, shown in Figure 2. In this figure, we observe that, as duration increases, the average formant values for males and females in spontaneous speech approach those values found in elicited speech. While dispersion varies with duration in both elicited and spontaneous speech, the dispersion differences within spontaneous speech are greater.

A strongly significant vowel by duration interaction was found both for the F1 model ($F[4] = 292.6$, $p < .001$) and for the F2 model ($F[4] = 206.5$, $p < .001$). In the elicited speech, mid vowels /e, o/ were significantly longer (46 ms, or roughly 23% longer) than peripheral vowels /i, a, u/. The increased durational differences between mid vowels across styles led to greater changes in dispersion. A significant duration by sex interaction was also found, though only for the F2 model ($F[1] = 11.7$, $p < .001$). Similar to the vowel by duration interaction, there was a wider variation in duration values for vowels produced by female speakers than by male speakers. This led to greater F2 dispersion across tokens produced by females. Finally, a significant three-way interaction of style * duration * sex was observed for the F2 model ($F[1] = 52.7$, $p < .001$). This interaction reflects a greater degree in dispersion across styles observed for female listeners, but only where duration values significantly differed with style.

Since duration varied by style, one potential confound in the model presented here is that style influences duration indirectly and that speech style is not an *independent* predictor of vowel dispersion. To address this concern, an additional linear mixed effects model was constructed which excluded Condition (speech style) as a predictor and compared with the model here using analysis of variance. The results show that the model containing Condition as a predictor to be significantly better than the one excluding it, both for F1

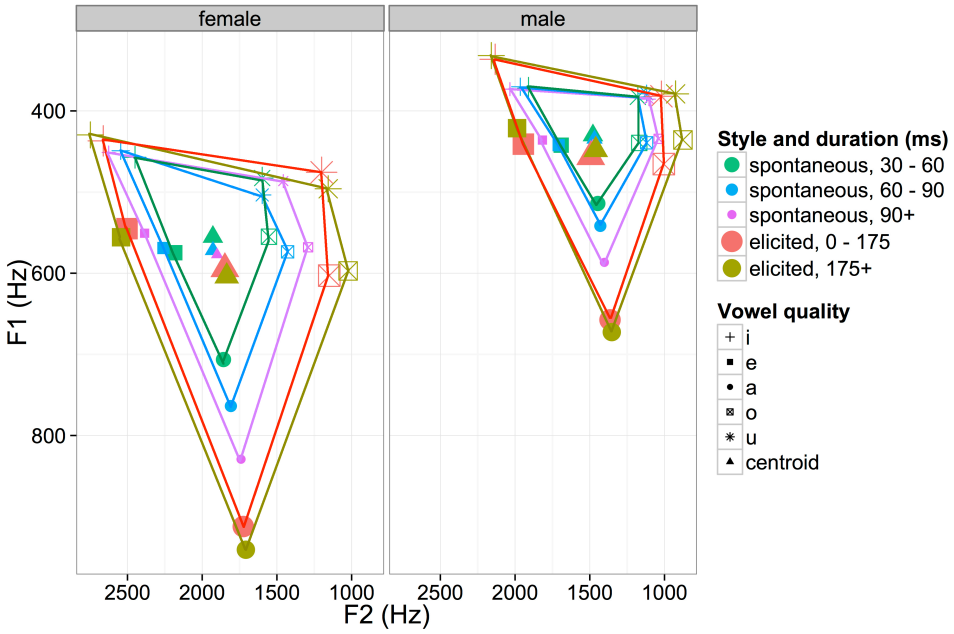

Figure 2: Vowel space as a function of duration, style, and sex. Spontaneous speech data are divided into bins of 30-60 ms, 60-90 ms, and 90 ms and more. Elicited speech data are divided into bins of 0-175 ms. and 175 ms and more.

dispersion (χ^2[45] = 248, p < .001; AIC = -68397 vs. AIC = -68605) and F2 dispersion (χ^2[45] = 641, p < .001; AIC = 393 vs. AIC = 994). The lower AIC values here reflect the better model with style as a factor. This demonstrates that the independence of style and duration in predicting formant dispersion.

To summarize, strong effects of style, vowel, and duration were found for vowel dispersion. Vowels in spontaneous speech were typified by formant values closer to the vowel centroid than vowels in elicited speech, though this effect was stronger for F2 than for F1. While sex was not a significant predictor, it interacted with vowel quality. Males produced more retracted /u/ tokens and raised /a/ tokens relative to females. Mid vowels (/e, o/), which happened to be longer in elicited speech than other vowels, underwent greater changes in dispersion than other vowels across styles. As females produced longer vowels in elicited speech than males, their speech was also typified by a greater overall influence of duration on dispersion.

3.2. Intra-categorical variability

Measures of dispersion may correspond closely with measures of intra-vowel variability when movement away from a vowel target mean corresponds with movement towards a vowel space centroid. These two dependent variables were correlated roughly equally for each formant in the data analyzed here, F1-cor($D(i)_{wc}$, $D(i)_{bc}$) = 0.499, F2-cor($D(i)_{wc}$, $D(i)_{bc}$) = 0.510. As a result of this correlation, certain predictors will have similar effects for each of these measures. We keep this in mind in our interpretation of these results.

Figure 3 shows intra-vowel variability by style, vowel, and sex. A significant effect of style on intra-vowel variability was found for both the F1 model ($F[1] = 11.7$, $p < .01$) and the F2 model ($F[1] = 18.2$, $p < .001$). Like the vowel dispersion data, the effect size was greater for F2 than for F1, suggesting that the production of spontaneous speech involves both greater contraction of the entire vowel space along the F2 dimension as well as increased variability within individual vowels. A significant effect of duration was found, though only for the F1 model ($F[1] = 7.8$, $p < .01$). Shorter vowels tended to be produced further from the vowel target mean F1 than longer vowels. The lack of any effect of duration on F2 intra-categorical variability result differs starkly from the results for vowel dispersion where the effect of duration was very strong.

A significant effect of vowel was found both for the F1 model ($F[4] = 21.8$, $p < .001$) and the F2 model ($F[4] = 60.9$, $p < .001$). The vowel /a/ varied more along the F1 dimension than did other vowels while back vowels /o, u/ varied more than other vowels along the F2 dimension. A significant vowel by style interaction was found for the F1 model ($F[4] = 9.3$, $p < .001$) and the F2 model ($F[4] = 16.1$, $p < .001$) as well. Since the back vowels showed greater intra-categorical variability in general, there was a significantly smaller influence of style on their variability. The same was true for the vowel /a/. While sex was nearly significant as a predictor of intra-categorical vowel variability in the F1 model ($F[4] = 6.0$, $p = 0.055$) and in the F2 model ($F[4] = 4.9$, $p = .059$), stronger interactions of vowel and sex were observed for F1 ($F[4] = 3.6$, $p < .01$) and for F2 ($F[4] = 21.8$, $p < .001$). In particular, back vowels produced by females showed greater variability along the F2 dimension than they did for males.

A significant duration by style interaction was found for the F1 model ($F[1] = 9.7$, $p < .001$) and the F2 model ($F[1] = 4.2$, $p < .05$). The effect of duration on vowel variability in the F1 dimension was stronger within the spontaneous speech than within the elicited speech. This effect has to do with the differences in the distribution of duration across styles. Whereas the spontaneous speech contained many vowel tokens that were particularly short (under 50 ms), such tokens were absent from the elicited speech. It is probable that more vowel productions reached a target F1 value given the longer durational window in the elicited speech and fewer productions reached a target F1 value in the spontaneous speech, resulting in greater variability. Jaw opening, which is inversely correlated with F1,

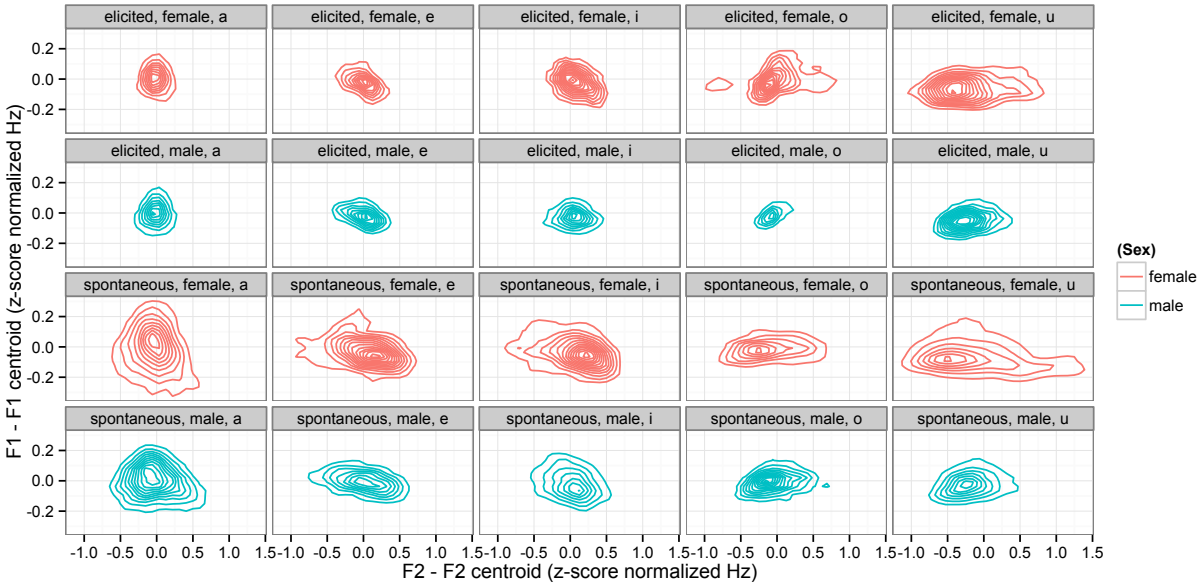

Figure 3: Intra-vowel variability as a function of style, vowel, and speaker sex. Formant variability is drawn as a two-dimensional contour using kernel density estimation, a smoothing algorithm which assumes randomly-distributed values around the median.

was greater in elicited speech than in spontaneous speech.

Similar to the dispersion data in Section 3.1, one potential confound in the model presented here is that style influences duration indirectly and that speech style is not an *independent* predictor of vowel dispersion. An additional linear mixed effects model was constructed which excluded style as a predictor and compared with the model here using analysis of variance. The results show that the model containing style as a predictor to be significantly better than the one excluding it, both for F1 variability ($\chi^2[45] = 265$, $p < .001$; AIC = -77286 vs. AIC = -77510) and F2 variability ($\chi^2[45] = 180$, $p < .001$; AIC = -10216 vs. AIC = -10356). The lower AIC values here reflect the better model with style as a factor. This demonstrates that the independence of style and duration in predicting formant variability.

Some of the effects here mirror those found for vowel dispersion. Vowels produced in spontaneous speech are produced with both greater within-vowel variability in F1 and F2 and with less dispersion in the overall F1xF2 space. Sex interacted with vowel quality in both models. The fronter production of back vowels /o, u/ in females resulted in both an increase in within-vowel variability and less dispersion overall. However, these two

measures diverged with respect to the role of duration as a predictor. While duration plays a strong role in overall dispersion, it plays less of a role in predicting within-vowel variability for F2. One possible source of such variability is that centralized vowels are further away from the articulatory limits of the vowel space and thus have more freedom to vary than do the elicited vowels. Another possibility may lie in the contextual effects of the adjacent consonants. We explore these factors below.

3.2.1. Effects of place of articulation on vowel variability

The elicited and spontaneous speech corpora were not controlled by vowel with respect to place of articulation for the adjacent consonant. There were some language-specific reasons for this. First, mid vowels /e, o/ are noticeably less frequent than peripheral vowels /i, a, u/ (see Table 3). As a result, these vowels did not occur with certain preceding or following consonants. Second, there is a phonological restriction in YM and common across Mixtecan languages where labial consonants never precede a back vowel (Longacre, 1957; Silverman, 2002). These gaps restrict both the vowels and possible places of articulation that we can compare within a model examining contextual effects on vowel production. To address this, the effects the place articulation of preceding and following consonants on vowel production were examined only for peripheral vowels /i, a, u/. The preceding place of articulation included four levels: alveolar/dental (/t, n, $\mathrm{n^d}$, ɾ, s, l/), post-alveolar/palatal (/tʃ, ʃ, j/), velar (/k, $\mathrm{ŋ^g}$/), and glottal (/ʔ, h/).

Restricting our data analysis to these four levels results in a smaller data set than analyzed previously (38,383 vowels above vs. 24,715 vowels here, or roughly 64% of the data). F1 and F2 variability were tested with four fixed effects: Condition, Vowel quality, Time (3 levels), and preceding place of articulation; two random intercepts (Speaker and Word), and a random Condition by Speaker slope. These models were evaluated identically to those we previously analyzed.

Figure 4 shows the influence of preceding place of articulation on F1 and F2 values by vowel and style. The preceding place of articulation showed a significant effect on vowel variability, but the effect was substantially greater for the F2 model ($F[3] = 227.4$, $p < .001$) than for the F1 model ($F[3] = 16.6$, $p < .001$). The preceding place of articulation had only a small influence on F1 targets for the three examined vowels. This was mainly restricted to both greater variability (raising) of F1 in the production of the vowel /a/. This asymmetry in influence is verified by the significant interaction of place of articulation by vowel for F1 ($F[6] = 10.4$, $p < .001$). A strong influence of glottal and alveolar place of articulation were responsible for the much stronger effect of place on variability in F2 values. Vowels produced following an alveolar consonant showed less variability than those that followed other places of articulation. A strong place by vowel interaction was also found for the F2 model ($F[6] = 73.3$, $p < .001$). The effects of place of articulation

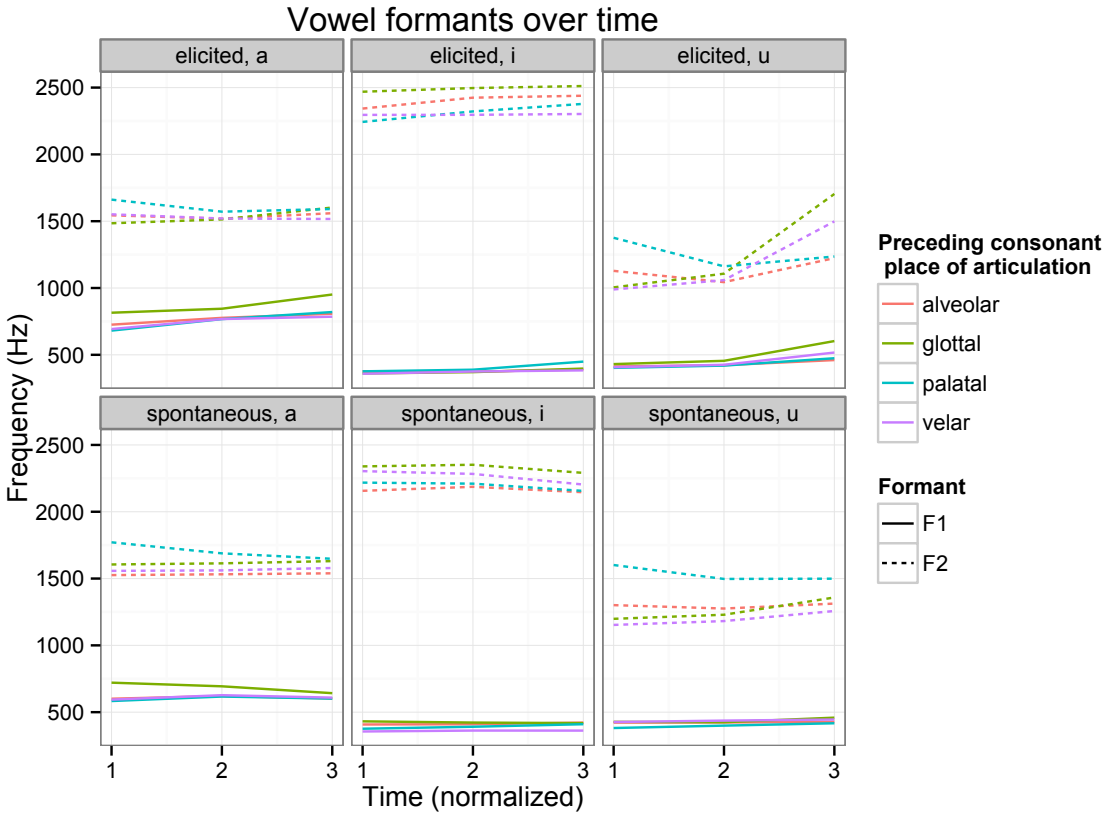

Figure 4: Formant values by vowel as a function of time showing the effects of preceding place of articulation and style.

on F2 variability were much stronger for high vowels /i, u/ than for the low vowel /a/ and stronger still in back vowels than in front vowels.

There was a significant interaction between style and place of articulation for both the F1 model ($F[3] = 9.5$, $p < .001$) and the F2 model ($F[3] = 3.8$, $p < .01$). Place of articulation exhibited a stronger influence on vowel variability in spontaneous speech than in elicited speech. Yet, while we have considered a number of predictor variables along with style in these models, we have not yet considered the role of duration in predicting vowel variability. If the effects of place are stronger in spontaneous speech than in elicited speech, it may be that the inherently shorter duration of vowels in spontaneous speech is entirely responsible. In essence, we are examining to what degree we can control for the more random factors which vary between the styles (place, gender, speaker-specific

formant differences, vowel variability, temporal dynamics) and test only the influence of style.

To test whether style was a significant predictor independent from these other effects, we evaluated its role via a model comparison using analysis of variance. We constructed a pair of models each for F1 and F2, the first which included five fixed effects (style, vowel, duration, place of articulation, and time), and the second which excluded only style. Random intercepts were applied for word and for speaker, along with a random slope of style by speaker. We found a significant difference between the models of F1 variability (χ^2[48] = 1093, p < .001) and between the models of F2 variability (χ^2[48] = 964, p < .001). The AIC value for the fully-specified AIC model for F1 was -96814, compared to a value of -95817 for the model excluding style as a predictor. The AIC value for the fully-specified AIC model for F2 was -6584.6, compared to a value of -5716.1 for the model excluding style as a predictor. The model containing style as a predictor was significantly better than the model excluding it.

One confound with such a comparison though is the inherent differences in what proportion of vowels in spontaneous and elicited speech may undergo greater coarticulatory effects. As vowels in spontaneous speech are shorter, a larger percentage of such vowels is closer to the C-V transition than for vowels in elicited speech. As a result, vowels in spontaneous speech may inherently undergo greater influence from adjacent vowels due to their durational characteristics, not simply as an effect of style. One way to *partially* control for this is to restrict the data set to those vowel durations which happen to overlap between both narrative and elicited speech. We re-ran the statistical model comparisons in the previous paragraph with a subset of the data where vowel duration was set to be between 80 - 120 ms. This consisted of just 4,502 tokens, or 18.2% of the vowels considered previously. We found a a significant different between the model of F1 variability excluding Condition as a factor and the one containing it (χ^2[37] = 186, p < .001; AIC = -21304 vs. AIC = -21476). The same result was found between the comparable models of F2 variability (χ^2[37] = 160, p < .001; AIC = -2536 vs. AIC = -2682). Even when duration is carefully controlled, the effect of speech style remains. In sum, though there are strong influences of vowel quality, duration, gender, and place of articulation on the degree of vowel variability in the data, the effect of style remains robust; spontaneous speech is qualitatively distinct from elicited speech for YM speakers.

3.3. Individual differences

So far, we have modelled speaker differences as random effects. To what extent do speakers actually vary in the degree of vowel variability as a function of style. Examining the random effects structure in Figure 5, we observe speaker differences in how much style played a role in determining their vowel productions. Since we do not have spontaneous

speech data from three speakers, their values are excluded here. In general, speakers were quite similar in their degree of F1 variability, with most clustering near the intercept value. Two speakers (CTC, VRR) exhibited less overall F1 variability and speaker VRR showed a much smaller influence of style on vowel variability. Speaker GNS exhibited greater F2 variability overall and a stronger effect of style on their productions than the other speakers did. Meanwhile, speakers CTC and VRR exhibited less overall F2 variability and a weaker influence of style.

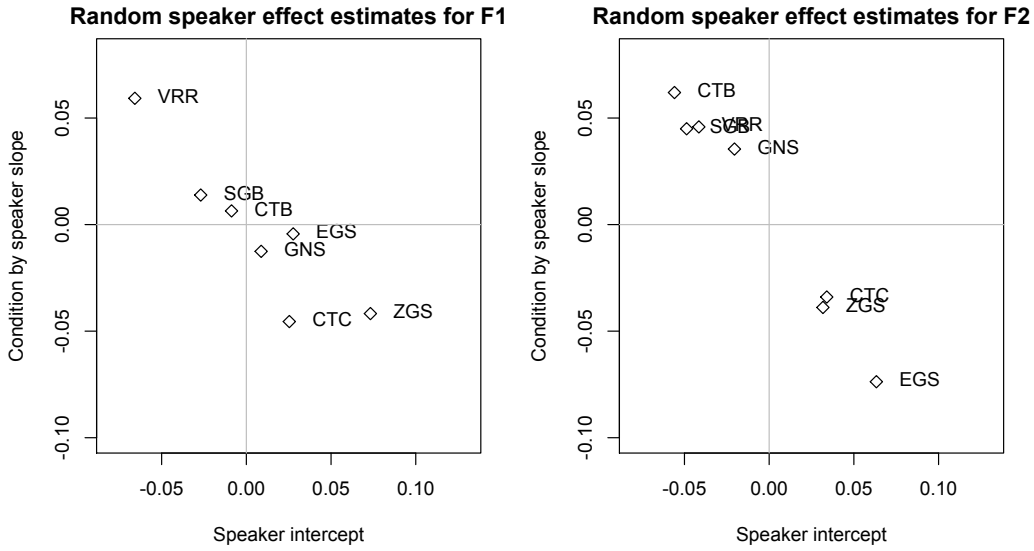

Figure 5: Random effect estimates on F1 and F2 variability. Positive intercept values (x axis) indicate less overall formant variability, while negative values indicate greater overall variability in the formant. Positive slope values (y axis) indicate greater overall effect of style on formant variability.

With a few exceptions, speakers are very similar in their overall degree of vowel variability across style. The stronger random effects for speakers GNS, CTC, and VRR, all of whom produced spontaneous and elicited speech, correspond strongly with the tightness of these speaker's vowel spaces. Speaker GNS is female and has the most disperse vowel space among all speakers. Speakers CTC and VRR are male and have the most compact vowel spaces among all the speakers. Given these inherent differences, one predicts a greater influence of style on vowel production for speaker GNS and a weaker one for speakers CTC and VRR.

4. Discussion

4.1. Speech style and vowel undershoot

The spontaneous speech style resulted in vowels that were both less disperse and showed greater variability than those produced in elicited speech. Spontaneous speech in YM is typified by faster overall speech rate (as is probably true of all languages), resulting in vowel durations which were, on average, less than half as long as those produced in elicited speech. Yet the durational differences across speech styles do not account for the entirety of the stylistic differences, a finding which agrees with previous work by Moon and Lindblom (1994). Stylistic effects were robust even when duration and other factors were considered by comparing statistical models with and without style as a factor.

4.1.1. Phonetic factors

Speech style interacts in nuanced ways with the other phonetic factors considered in this study. Vowel duration played a strong role in determining the degree of vowel dispersion in the YM data, though its effect varied by style. It was a stronger limiting factor for vowel dispersion in the spontaneous speech corpus, which consisted of a wider range of (shorter) vowel duration values, than in the elicited speech corpus, which consisted of a smaller range of longer values. Such findings suggest that, given a sufficiently large durational window, speakers are more likely to reach an ideal vowel target.

The findings here not only support the theory of *vowel undershoot* (Lindblom, 1990; Moon and Lindblom, 1994), but they have general implications for understanding the effect of duration in studies on vowel production. Recall that vowel duration played a large role in dispersion in the French spontaneous corpus examined by Meunier and Espesser (2011), but a weaker role in the English clear speech corpus examined by Aylett and Turk (2006). As in the elicited data for YM, the increased vowel duration in this latter study may have resulted in a greater percentage of vowel productions which were sufficiently long enough to reach an ideal target. Given a greater proportion of longer targets, the overall variation with duration would vary less, resulting in a *smaller effect* of duration on spectral dispersion. The effect of duration on style is non-linear in this fashion, and in this way, vowel dispersion is sensitive to corpus style.

One can analyze the patterns of vowel reduction from the perspective of articulatory phonology (AP). In AP, speech gestures are discrete constricting actions on the vocal tract. They are defined as critically-damped second order systems and specified by an activation onset and offset time and by a set of dynamic parameters (target, stiffness, and damping). If a gesture's activation is truncated during a faster speech rate (such as that induced by spontaneous speech), it is likely not to be able to reach the target (Saltzman and Munhall, 1989). Such incompletely achieved targets would lead to less disperse formants in vowel space.

AP not only explains why vowels become less disperse but also why they become more variable when they are shortened. Once a gesture's activation is long enough to reach the target, modulating the duration would not change the final position of the gesture. This prediction is borne out in the current data, as the longer elicited vowels were less sensitive to changes in vowel duration than the spontaneous vowels were. On the other hand, if a gesture's activation is too short to reach the target, even a small change in the gestural duration would result in the final position of the gesture, leading to greater variability in articulation/acoustics. This account requires an assumption that the stiffness of the gesture, i.e. how quickly a gesture reaches the target, is not varied drastically so as to rescale the activation duration. As duration alone does not explain the pattern of vowel undershoot across speech styles, changes in *gestural stiffness* may be responsible.

Phonetic variability between the duration of different vowels and in the effect of place of articulation were responsible for differences in the degree of undershoot among vowels. For instance, the effect of style on vowel dispersion was strongest for mid vowels, which were (surprisingly) significantly longer than the peripheral vowels in the elicited speech. As a result, these vowels underwent greater durational changes across the styles than the other vowels did; only the male /o/, however, showed noticeably smaller within-vowel variability for the elicited style (Figure 3). This pattern remains a puzzle, but it may be due largely to the small number of tokens of mid vowels in the spontaneous corpus.

The preceding place of articulation had a stronger effect on the F2 variability in back vowels than in front vowels. Note that contrasts in place of articulation in the front of the oral cavity are more common cross-linguistically than contrasts further back in the oral cavity (Ladefoged and Maddieson, 1996). As a result of this asymmetry, one expects to observe a greater number of contrasts which will cause tongue advancement of back vowels than contrasts which will cause tongue retraction of front vowels. YM follows this cross-linguistic trend. In the models including place of articulation as a fixed effect, two contexts were responsible for F2 fronting of back vowels (alveolar, palatal), while only one was responsible for F2 retraction (velar). Asymmetries in the location of place contrasts result in greater contextual assimilation for back vowels.

The effects of place of articulation interacted with style and duration in similar ways; contextual assimilation was stronger in spontaneous speech and when duration was shorter than in elicited speech or when duration was longer. However, some phonetic factors remained unexplored. For instance, there is a relationship between tone and vowel duration (DiCanio et al., 2012). Similarly, word-final syllables in YM are longer than non-final syllables and may host a greater number of tonal contrasts (ibid). This observation suggests that final syllables have greater prosodic prominence than non-final syllables. Prosodic prominence results in hyperarticulation of phonological contrasts (de Jong, 1995; Keating et al., 2000). Yet, spontaneous speech vowels were not coded for either tone nor prosodic

prominence in the current study. These additional factors may also help explain some of the durational and spectral differences among vowels.

4.1.2. Sex and individual differences in vowel production

Languages differ in the degree in which different phonetic factors play a role in speech production, but they also differ in how individual speakers may vary. Within the YM data, we observed three findings pertaining specifically to speaker sex. First, the close back vowel /u/ was produced with a more advanced articulation (higher F2) by female speakers than by male speakers. Second, male speakers produced a higher /a/ vowel (lower F1) than female speakers. This finding jibes well with previous work showing that sex differences in vowel production occur mostly in the F1 dimension (Henton, 1992). The first effect resulted in less overall dispersion of /u/ for female speakers while the second resulted in a smaller overall vowel space in male speakers. Third, female speakers produced longer elicited vowels than male speakers did, but males produced slightly longer spontaneous vowels than females did. The average duration of female elicited vowels was 238 ms, compared to 200 ms for males. The average duration of female spontaneous vowels was 87 ms, compared to 96 ms for males. Females' elicited vowels were 2.74 times longer than their spontaneous vowels, whereas the same ratio for males was 2.08:1. As a result of this difference, females as a group showed stronger effects of duration and style on vowel variability than males did.

This third finding is intriguing with respect to previous findings on sex and speech rate. While there is little cross-linguistic study on the topic, the common observation is that males speak faster than females (Byrd, 1994; Henton, 1992; Kramer, 1978; Labov, 1966). For instance, males were found to speak 6.2% faster than females in the TIMIT corpus (Byrd, 1994). If we evaluate vowel duration as a measure of speech rate (instead of the more typical syllables/second measure), we find that males speak 19% faster than females in elicited YM speech. While greater than the difference reported for English, this finding matches the direction of the previous findings. Yet, females speak 10% *faster* than males in spontaneous speech. While it is necessary to do a more detailed analysis of speech rate in YM, these findings suggest that speech rate differences between sexes may depend on speech style.

Individuals were found to vary somewhat in degree of vowel variability as a function of style. However, speaker variability correlated with the degree in which style influenced each speaker. The female outlier shown in Figure 5 shows both greater individual vowel variability overall and a strong influence of style on her vowel productions. The two male outliers show both smaller individual vowel variability and a weak influence of style on their productions. Even though speakers' formant values were normalized, individual differences were found in the degree of permitted vowel variability.

4.2. Endangered language corpora and comparability across corpus types

Taken as a whole, the results of this study suggest that one cannot approximate stylistic differences in vowel production by applying a single transformation to one's data. What might this mean for descriptive phonetic work which uses spontaneous speech recordings? If the goal of such work is to characterize the typical, or average production of a particular sound contrast, then the researcher must understand the relationship between the potential undershoot of the contrast and its duration. As spontaneous speech recordings provide a larger range of exemplars which vary more substantially in duration, we believe it is possible to examine this natural range of productions and provide an ecologically valid glimpse of the sound system of the language. For instance, Figure 2 demonstrates that it is possible to bin spontaneous speech vowels into groups of different duration ranges and produce a vowel space which more closely reflects one found in careful, elicited speech as the duration increases.[5] However, to truly capture the vowel space of a given language, one must consider both elicited and spontaneous speech data.

Though spontaneous speech in YM is qualitatively different from elicited speech, we disagree with Ladefoged's claim that folk tales are of limited use for phonetic descriptions of a language. Without controlling for context, we found robust differences between vowels. Moreover, by their nature, vowels in elicited tokens occurred in a more limited range of contexts. This contrasts with the fuller range of contexts found in spontaneous speech, where each vowel co-occurred with all possible consonant places of articulation. Spontaneous speech data permits the investigation of large contextual variability in speech production. Finally, tools such as mixed effects modeling allow one to confidently examine contextual assimilation when data is unbalanced (Baayen, 2008; Bates, 2005). While the large-scale phonetic analysis of spontaneous speech may not have been possible with older statistical tools, the more recent availability of these tools permits it.

The inclusion of phonetic context in speaker normalization procedures might also allow us to derive what Ladefoged considers an ideal vowel space for a language. That is, not only would vocal tract length differences be taken into account by comparing F_0, F1 and F2, but known context effects of neighboring segments could be added as well. Such procedures might need to be language-specific, given differences in amount of coarticulation across languages (Manuel, 1990, 1999), and perhaps even with differences in average vocal tract morphology (Proffitt et al., 1975). Current algorithms for citation speech alone are not completely accurate (Becker-Kristal, 2010; Flynn, 2011), and issues of genuine differences across speakers remain despite normalization. Vowel spaces derived from spontaneous speech may approximate vowel spaces observed in elicited speech, but an

[5]One might also question whether a phonetic description of a language should demand a citation form standard, given the novelty of elicitation for speakers of many endangered/undescribed languages.

analysis of speech data from a variety of styles would seem to be necessary for estimating the the "true" vowel space of a language.

5. Conclusions

Vowel variability and dispersion differ between spontaneous speech and elicited speech styles in YM. Such effects are not reducible to changes in the distribution of vowel duration. The fully-crossed linear mixed effects models which contained speech style as a predictor for formant dispersion, formant variability, and the influence of place of articulation were significantly better than those which excluded speech style. The influence of speech style varied with inherent durational differences found between vowels and with differences in speech rate between males and females. Controlling for these predictors, one can more closely approximate elicited speech data with folkloric texts and personal narratives, thereby allowing a clear comparison across speech styles. The relationships among the vowels are the same even as the formant values change with duration and style.

Best practices in the documentation of endangered languages involves the collection of high-quality audio recordings containing culturally-relevant, spontaneous speech samples. The necessity of high quality in these recordings is motivated by the desire for verification of the language's phonology and for the future potential investigations of the language's phonetics. While we find that one can approximate more careful speech using spontaneous speech data, there are still stylistic differences between the two. For traditional *phonological* descriptions based on more careful, elicited productions, our findings argue that spontaneous speech data is useful for the purposes of verification only where researchers are aware of the phonetic and non-phonetic factors which influence speech production variability. However, for the *phonetic* description of a language, spontaneous speech data is useful and may even allow the researcher to investigate contextual influences not typically present in elicited speech.

Despite the strength of the factors we considered here, additional phonetic factors, like tone and prosody, may play a role in the degree of vowel dispersion and variability. Future work on YM phonetics will focus specifically on the production of tone in spontaneous speech data. Such an investigation will involve tonal segmentation, which will allow us to evaluate the realization of tone in spontaneous speech, the effect of tonal and segmental environment on the realization of tone, and, pertinent to this study, the role of tone in vowel production. Our findings also raise a larger issue regarding the influence of contextual assimilation on vowel dispersion and variability. Studies which investigate the effect of consonant place of articulation on vowel production frequently use single words spoken in isolation (Hillenbrand et al., 2001). Our findings show that these effects are attenuated in elicited speech, even when factors such as duration are included in the model.

While contextual effects on vowel production come out most clearly from those studies which involve maximally phonetically-divergent consonant contexts, e.g. /wVl/ in Moon and Lindblom (1994), the strength of contextual assimilation is also sensitive to stylistic differences in speech production.

While the relationship between contextual variability and speech style has been investigated for YM, it is unclear how it compares to work in other languages. If one takes seriously the claim that different " motivating factors are managed differently by speakers of different languages" (Johnson and Martin, 2001, 96), then one anticipates that both the degree of contextual assimilation and its dependence on style to vary across languages. The generality of these relationships depends on active future work on a diverse sample of different languages.

Further study of endangered language data requires not simply the existence of speech corpora in the language, but it also depends on the amenability of those corpora to automatic methods of phonetic analysis. Large data sets are not feasibly analyzed by hand, but can be aided by tools such as forced alignment, even when such tools are not trained on the target language (DiCanio et al., 2013). Those phoneticians interested in exploring more of the world's languages should take advantage of the increase in documentation, even though the challenges to analysis are not insignificant. Having more information about the state of languages today will be of immense value, especially as languages fall silent, preventing later data collection. The rewards of analyzing endangered language data are sufficient to justify this effort.

Acknowledgements

The YM corpus was elicited by Castillo García, Amith, and DiCanio with support from Hans Rausing Endangered Language Programme Grant MDP0201 and NSF grant 0966462. The authors would like to thank Leandro DiDomenico for his help with transcription labeling of the elicited speech corpus. This work was supported by NSF grant 0966411 to Haskins Laboratories (Whalen, PI).

References

Amith, J. D. and Castillo García, R. (no date). *A dictionary of Yoloxóchitl Mixtec*.

Aylett, M. and 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(1):31–56.

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, 119(5):3048–3058.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, 353 pages.

Barry, W. and Andreeva, B. (2001). Cross-language similarities and differences in spontaneous speech patterns. *Journal of the International Phonetic Association*, 31(1):51–66.

Bates, D. M. (2005). Fitting Linear Mixed Models in R. *R News*, 5:27–30.

Becker-Kristal, R. (2010). *Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus*. PhD thesis, UCLA.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. www.praat.org.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, 15:39–54.

Castillo García, R. (2007). Descripción fonológica, segmental, y tonal del Mixteco de Yoloxóchitl, Guerrero. Master's thesis, Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS), México, D.F.

Chen, F. (1980). Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level. Master's thesis, MIT, Cambridge, MA.

Chen, F., Zue, V., Picheny, M., Durlach, N., and Braida, L. (1983). Speaking Clearly: Acoustic characteristics and intelligibility of stop consonants. *Working Papers: Speech Communication Group, MIT*, 2:1–8.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, 97(1):491–504.

DiCanio, C., Amith, J., and Castillo García, R. (2012). Phonetic alignment in Yoloxóchitl Mixtec tone. Talk Presented at the Society for the Study of the Indigenous Languages of the Americas Annual Meeting.

DiCanio, C., Amith, J. D., and Castillo García, R. (2014). The phonetics of moraic alignment in Yoloxóchitl Mixtec. In *Proceedings of the 4th Tonal Aspects of Language Symposium*. Nijmegen, the Netherlands.

DiCanio, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., and Castillo García, R. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *Journal of the Acoustical Society of America*, 134(3):2235–2246.

DiCanio, C., Zhang, C., Whalen, D. H., Amith, J. D., and Castillo García, R. (submitted). Phonetic structure in Yoloxóchitl Mixtec consonants.

Flemming, E. (2003). The relationship between coronal place and vowel backness. *Phonology*, 20:335–373.

Flynn, N. (2011). Comparing vowel formant normalisation procedures. *York Papers in Linguistics*, Series 2(11):1–28.

Foulkes, P. and Docherty, G. J. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, 34:409–438.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66:789–806.

Gordon, M., Munro, P., and Ladefoged, P. (1997). The phonetic structures of Chickasaw. *UCLA Working Papers in Phonetics*, 95:41–67.

Harmegnies, B. and Poch-Olivé, D. (1992). A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. *Speech Communication*, 11:429–437.

Henton, C. (1992). The abnormality of male speech. In Wolf, G., editor, *New Departures in Linguistics*. Garland Publishing, New York.

Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, 109(2):748–763.

Johnson, K. (2008). *Quantitative Methods in Linguistics*. Blackwell Publishing.

Johnson, K. and Martin, J. (2001). Acoustic vowel reduction in Creek: Effects of distinctive length and position in the word. *Phonetica*, 58:81–102.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J. L. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Keating, P., Cho, T., Fougeron, C., and Hsu, C.-S. (2000). Domain-initial articulatory strengthening in four languages. In Local, J., Ogden, R., and Temple, R., editors, *Papers in laboratory phonology 6*, chapter 10. Cambridge University Press.

Keating, P. A. and Huffman, M. K. (1984). Vowel variation in Japanese. *Phonetica*, 41:191–207.

Kendall, T. S. (2009). *Speech rate, Pause, and Linguistic Variation: An Examination through the sociolinguistic archive and analysis project*. PhD thesis, Duke University.

Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *Journal of the Acoustical Society of America*, 122(4):2365–2375.

Koopmans-van Beinum, F. (1980). *Vowel contrast reduction, an acoustic and perceptual study of Dutch vowels in various speech conditions*. PhD thesis, University of Amsterdam, The Netherlands., Academische Pers B. V., Amsterdam.

Kramer, C. (1978). Perceptions of female and male speech. *Language and Speech*, 20:151–161.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2013). *lmerTest (R package)*.

Labov, W. (1966). Hypercorrection by the lower middle class as a factor in linguistic change. In Bright, W., editor, *Sociolinguistics*, pages 84–113. Mouton and Co., The Hague.

Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Blackwell.

Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, 34:500–515.

Labov, W., Ash, S., and Boberg, C. (2006). *Atlas of North American English: Phonetics, Phonology and Sound Change*. Mouton de Gruyter.

Ladefoged, P. (2003). *Phonetic Data Analysis*. Blackwell.

Ladefoged, P. and Maddieson, I. (1996). *Sounds of the World's Languages*. Oxford: Blackwell, 425 pages.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35:1773–1781.

Lindblom, B. (1983). The economy of speech gestures. In MacNeilage, P. F., editor, *The Production of Speech*, pages 217–243. Springer-Verlag.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the h&h theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers.

Longacre, R. E. (1957). Proto-Mixtecan. In *Indiana University Research Center in Anthropology, Folklore, and Linguistics*, volume 5. Indiana University Research Center in Anthropology, Folklore, and Linguistics, Bloomington.

Manuel, S. Y. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America*, 88(3):1286–1298.

Manuel, S. Y. (1999). Cross-language studies: relating language-particular coarticulation patterns to other language-particular facts. In Hardcastle, W. J. and Hewlett, N., editors, *Coarticulation: Theory, Data, and Techniques*, chapter 8, pages 179–198. Cambridge University Press.

Meunier, C. and Espesser, R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, 39:271–278.

Mok, P. P. (2010). Language-specific realizations of syllable structure and vowel-to-vowel coarticulation. *Journal of the Acoustical Society of America*, 128(3):1346–1356.

Moon, S.-J. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 96(1):40–55.

Munson, B., McDonald, E. C., DeBoe, N. L., and White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, 34:202–240.

Ohde, R. N. and Sharf, D. J. (1975). Coarticulatory effects of voiced stops on the reduction of acoustic vowel targets. *Journal of the Acoustical Society of America*, 58:923.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39:151–168.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41:310–320.

Pellegrino, F., Coupé, C., and Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 87(3):539–558.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. ., Stockmann, E., Tiede, M., and Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America*, 116:2338–2344.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, 13(2):253–260.

Proffitt, W., McGlone, R., and Barrett, M. (1975). Lip and tongue pressures related to dental arch and oral cavity size in Australian Aborigines. *Journal of Dental Research*, 54:1161–1172.

R Development Core Team, Vienna, A. (2013). R: A language and environment for statistical computing [computer program], version 3.0.2. http://www.R-project.org, R Foundation for Statistical Computing.

Recasens, D. (1984). Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America*, 76(6):1624–1635.

Saltzman, E. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382.

Shadle, C. H., Nam, H., and Whalen, D. H. (2013). Measurement of formants in synthetic vowels. *Journal of the Acoustical Society of America*, 134:4068.

Silverman, D. (2002). The diachrony of labiality in Trique and the functional relevance of gradience and variation. In Goldstein, L., Whalen, D. H., and Best, C. T., editors, *Laboratory Phonology 8: Varieties of Phonological Competence*. Mouton de Gruyter.

Smiljanić, R. and Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *Journal of the Acoustical Society of America*, 118(3):1677–1688.

Stevens, K. N. and House, A. S. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research*, 6(2):111–128.

Van Son, R. and Pols, L. C. W. (1992). Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, 92(1):121–127.

Verhoeven, J., De Pauw, G., and Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Duth in Belgium and the Netherlands. *Language and Speech*, 47(3):297–308.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics - 2008*.

Yuan, J. and Liberman, M. (2009). Investigating /l/ variation in English through forced alignment. In *Interspeech - 2009*, pages 2215–2218.