Journal of Phonetics 92 (2022) 101147

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Extreme stop allophony in Mixtec spontaneous speech: Data, word prosody, and modelling

Christian DiCanio [a,b,*], Wei-Rong Chen [b], Joshua Benn [a], Jonathan D. Amith [c], Rey Castillo García [d]

[a] Department of Linguistics, University at Buffalo, Buffalo NY 14260, USA
[b] Haskins Laboratories, 300 George Street, New Haven CT 06511, USA
[c] Department of Anthropology, Gettysburg College, 300 N Washington St, Gettysburg, PA 17325, USA
[d] Secretaría de educación pública (SEP), Guerrero, Mexico

ARTICLE INFO

Article history:
Received 24 August 2020
Received in revised form 15 March 2022
Accepted 26 March 2022

Keywords:
Corpus phonetics
Speech reduction
Endangered languages
Mixtecan
Prosody

ABSTRACT

Word-level prosody plays an important role in processes of consonant lenition. Typically, consonants in word-initial position are strengthened while those in word-medial position are lenited (Keating, Cho, Fougeron, & Hsu, 2003). In this paper we examine the relationship between word-prosodic position and obstruent lenition in a spontaneous speech corpus of Yoloxóchitl Mixtec, an endangered Mixtecan language spoken in Mexico. The language exhibits a surprising amount of lenition in the realization of otherwise voiceless unaspirated stops and voiceless fricatives in careful speech. In Experiment 1, we examine the relationships between word position, consonant duration, and passive voicing and find that word-*medial* pre-tonic position is the locus of both consonant lengthening and less passive voicing. Non-pre-tonic consonants are produced with more voicing and shorter duration. We also find that the functional status of the morpheme plays a role in voicing lenition. In Experiment 2, we examine manner lenition and find a similar pattern – word-medial pre-tonic stops are more often realized with complete closure relative to non-pre-tonic stops, which are more often realized with incomplete closure. In Experiment 3, we model these lenition patterns using a series of deep neural networks and find that, even with limited training data, we can achieve reasonably high accuracy in the automatic categorization of lenition patterns. The results of this research both complement recent work on the phonetics of lenition in the world's languages (Katz and Fricke, 2018; White et al., 2020) and provide computational tools for modeling and predicting patterns of extreme lenition.

1. Introduction

Speech requires speakers to carefully control the timing of different articulatory gestures while simultaneously conveying information to listeners at a sufficient rate. In running speech these constraints compete with each other and speakers may *lenite* certain speech sounds. Distinct from predictable allophonic rules, like voiceless stop aspiration in English (Lisker & Abramson, 1964; Kingston, Diehl, Kirk, & Castleman, 2008), are the variable phonetic patterns which occur in reduced, running speech. For instance, voiceless stops can be produced with a variable amount of voicing, called voicing *bleed*, when preceded by a voiced segment, e.g. in English (Davidson, 2018; Westbury & Keating, 1986), in French & Spanish (Torreira & Ernestus, 2011), and in Triqui (DiCanio, 2012). These patterns are variable in each of these languages and ubiquitous in connected speech.

The rates and types of variable lenition vary by speech context (Hualde, Simonet, & Nadeu, 2011; Lewis, 2001; Torreira & Ernestus, 2011; Warner & Tucker, 2011) and by language (Shih, Möbius, & Narasimhan, 1999) (see Section 1.1 for a discussion of patterns across languages). This is true even for what is described as the same phonological category across languages, e.g. voiceless unaspirated stops (ibid). This broadly suggests that lenition does not simply result from general human constraints on articulatory-motor control, but rather that it is sensitive to differences between languages. Thus, the primary theoretical question is just which specific linguistic factors contribute to differences in the rate of lenition. One known factor is prosody – segments which occur in stressed syllables are less likely to be lenited than those which occur in unstressed syllables (Bouavichith & Davidson, 2013; Lavoie, 2001). Given that languages show great diversity in terms of

* Corresponding author.

https://doi.org/10.1016/j.wocn.2022.101147
0095-4470/

prosodic structure (c.f. Gordon (2016)), one anticipates that stress-related differences contribute to differences in consonantal reduction.

The current study examines how word position and word prosody influences patterns of consonantal lenition in a corpus of spontaneous speech in Yoloxóchitl Mixtec, an endangered Otomanguean language spoken in Mexico (García, 2007; Palancar, Amith, & Castillo García, 2016). By "prosody", we refer to word-level stress differences. Yoloxóchitl Mixtec has a small obstruent inventory (/p, t, k, k^w, s, ʃ∼h, tʃ/) and does not contrast voicing within obstruents; all obstruents are voiceless in citation or in controlled experimental contexts (DiCanio, Zhang, Whalen, & Castillo García, 2020). However, these obstruents undergo extensive patterns of lenition in running speech, where they may be fully or partially voiced and the stops may also be realized as frictionless continuants.

In Experiment 1, we examine voicing lenition in obstruents in relation to word size and prosodic structure. We find that the duration of voicing bleed during the obstruent correlates with the overall obstruent duration, which itself correlates with stress. Fixed stress occurs in stem-final syllables in Yoloxóchitl Mixtec (DiCanio, Benn, & Castillo García, 2018; DiCanio et al., 2020). Obstruents in the onset of stem-final syllables are both phonetically longer and less likely to be voiced than obstruents that occur in non-final syllables, including word-initial position. In Experiment 2, we categorize each of the stops into a number of discrete allophonic groupings in order to examine the degree of spirantization or lenition (full closure with no voicing, full closure with partial voicing, fricative realization, approximant realization, etc). Approximantization (c.f. Bouavichith & Davidson (2013); Ladd & Scobbie (2003)) is pervasive among stops in the language, occurring in 26% of all stem-final syllables and 45% of all stem non-final syllables. We describe the rates for each lenition type. We then trained several deep neural networks to predict allophonic grouping from the acoustic signal. Two-way DNN models trained on the stop/non-stop ([±continuant]) categories showed 95–98% accuracy in detecting spirantized/non-spirantized allophones. Three-way and four-way models showed 74–88% accuracy in detecting additional allophonic realizations (voiceless vs. voiced stop vs. fricative) but accuracy in detecting fricative vs. approximant realizations of the stops remained low (54–57%). The findings here both argue for a structural prosodic motivation for patterns of lenition in an under-resourced and endangered language; and provide positive evidence that computational tools can categorize surface acoustic–phonetic variation related to lenition in a non-controlled spontaneous speech corpus.

1.1. Background: Obstruent reduction in speech production

Obstruent reduction refers to the process where, relative to carefully-produced speech, an obstruent is produced with reduced spatial excursion of articulators and reduced constriction degree.[1] Obstruents produced with this type of articulatory undershoot are also shorter in duration relative to carefully-produced variants (Lavoie, 2001; Parrell, 2014; Parrell & Narayanan, 2018) and voiceless obstruents may undergo an additional process of *passive voicing* or voicing *bleed* (Beckman, Jessen, & Ringen, 2013; Davidson, 2018; DiCanio, 2012; Jansen, 2004; Schwarz, Sonderegger, & Goad, 2019; Stevens, 2000; Westbury, 1983; Westbury & Keating, 1986). As a consequence of this reduction, the target obstruent may be realized with rather different acoustic cues than in carefully-produced speech and be less perceptually distinct from adjacent speech sounds.

Reduction (or variable lenition) is ubiquitous in human speech and typical even in carefully-produced speech contexts (Lavoie, 2001; Warner & Tucker, 2011; Warner, 2019). As a result, listeners are regularly exposed to speech with different acoustic cues than one observes in idealized contexts. At the same time, reduction frequently coincides with weak prosodic positions (unstressed syllables; word-internal, phrase-medial, and intervocalic contexts) whereas obstruent fortition frequently coincides with strong prosodic positions (stressed syllables; word-initial, phrase-initial, and phrase-final contexts) (Bouavichith & Davidson, 2013; Cho & Keating, 2001; Fougeron & Keating, 1997; Katz, 2016; Keating, Cho, Fougeron, & Hsu, 2004; Katz & Fricke, 2018). Though reduction often reduces distinctness of a segment relative to its neighbors, the relationship between reduction and prosodic boundaries nevertheless aids the listener in lexical segmentation. Listeners are able use the degree of reduction to make decisions about word boundaries (Katz & Fricke, 2018) just as they are able to use more general prosodic cues for the purpose of word segmentation (Saffran, Newport, & Aslin, 1996; White, Mattys, Stefansdottir, & Jones, 2015).

1.1.1. Prosodic and non-prosodic contexts for obstruent reduction

Which prosodic positions are sensitive to patterns of reduction? A general finding throughout the literature is that stops are less likely to be reduced in phrase-initial and word-initial position than in phrase-medial or word-medial position in a variety of languages, such as Bardi, English, French, Italian, Spanish, Korean, Hungarian, and Taiwanese (Cho & Keating, 2001; Davidson, 2018; Fougeron & Keating, 1997; Kakadelis, 2018; Katz, 2016; Katz & Fricke, 2018; Keating et al., 2004; Lavoie, 2001; Lewis, 2001; White, Benavides-Varela, & Mády, 2020; Keating et al., 2003). Phrase and word-level delimitation are in fact considered to be primary goals in processes of lenition (Cho, McQueen, & Cox, 2007; Katz, 2016; Katz & Fricke, 2018); word-medial segments are lenited whereas word-initial ones are not. This process would appear to be listener-driven; if listeners rely on word-initial segments for lexical parsing, such segments ought to be hyperarticulated for the listener. Models of spoken word recognition, such as Shortlist B even encode the relative importance of word-initial segments directly (Norris & McQueen, 2008).

Katz and Fricke (2018) distinguish between two separate phenomena which are often discussed as leniting processes, *voicing lenition* and *spirantization*. Though these two processes often co-occur (Gurevich, 2011), research on reduction and lenition often focuses on just one of them. Katz and Fricke find that only spirantization (or, more often, approximantization) aids in word segmentation in an artificial language learning task. In their study, voicing lenition is the discrete, allophonic process whereby a voiceless obstruent is produced

[1] The terms *speech reduction* and *lenition* are often used interchangeably in the speech production literature, the former being more common in phonetics and the latter being more common in describing discrete phonological patterns or processes of historical sound change (c.f. Gurevich (2004)).

with full voicing in intervocalic position. However, this process has its phonetic precursors in gradient phonetic processes of voicing *bleed*, where the voicing from a preceding segment may spread into an obstruent (Hualde et al., 2011). Similarly, categorical spirantization has its phonetic precursors in processes of articulatory undershoot.

Yet, there are languages whose patterns might challenge the view that fortition/lenition is driven by the necessity to maintain clear word onsets. In languages with patterns of initial consonant mutation and/or processes of prefixation, knowing the initial consonant does not aid the listener in ascertaining lexical identity (Ussishkin et al., 2017). It is an open question as to whether prefix-heavy languages still undergo the same type of word-initial strengthening observed for languages like English, Spanish, and Korean, where suffixation is more common than prefixation. This question is relevant to the current study since the language under investigation, Yoloxóchitl Mixtec, is prefixal with fixed, stem-final stress.

Another context where obstruent reduction is found is in unstressed syllables. In a study analyzing a spontaneous speech corpus of English, Bouavichith and Davidson (2013) find that voiced stops (/b, d, g/) are produced as approximants [β̞, ɾ, ɣ̞] between 10–21% of the time (depending on the place of articulation) in the onset of stressed syllables but 47–75% of the time in the onset of unstressed syllables. Durational differences also accompanied stress - obstruents in the onset of stressed syllables were longer (39–66 ms) than those in the onset of unstressed syllables (21–43 ms). Yet, the pattern for voiceless obstruents in English is somewhat different; stress does not influence the degree of passive voicing (a type of lenition) in a spontaneous speech corpus (Davidson, 2018), though it is unclear whether voiceless obstruents are spirantized more often in unstressed contexts. In a study of spontaneous speech in Spanish and French, incomplete closure of voiceless stops (/p, t, k/) was found to be more common in unstressed syllables than in stressed syllables (Torreira & Ernestus, 2011). Closure duration was also found to be significantly shorter in stops in unstressed syllables. Similar findings on Central Colombian and Bilbao (Castillian) Spanish are reported in work by Lewis (2001), where voiceless unaspirated stops are shorter and have more closure voicing in the onsets of unstressed syllables than in stressed syllables. Stress induces a similar pattern of weakening/shortening ofc /s/ in a corpus of Madrileño Spanish speakers (Torreira & Ernestus, 2012).

Despite its close connection to prosodic structure, predicting the context where lenition will occur in a given language from a phonological representation also remains a challenge. For instance, not all voiced obstruents in a given language may lenite in a similar fashion or even in the same way in the same prosodic context (Jun, 1995). Moreover, languages differ substantially in both the magnitude and type of lenition which may occur. When we take this cross-linguistic variability seriously, there are still largely unanswered empirical questions about the relationship between lenition and other linguistic factors.

For instance, reduction also frequently coincides with changes in speech rate and speech style/register. Grosso modo, obstruents produced at a faster speech rate will be lenited relative to a slower speech rate (Cheng & Xu, 2015; Dalby, 1984; DiCanio, 2012). In a study of Itunyoso Triqui lenis (singleton) and fortis (geminate) stops, DiCanio (2012) finds variable lenition among singleton stops and affricates which correlates both with speech rate differences across speakers and inherent duration of the obstruent. The singleton postalveolar affricate had the shortest closure duration among all obstruents and was most likely to undergo lenition to [ʃ]. Words and segments also have reduced duration in spontaneous speech relative to careful speech (Baker & Bradlow, 2009; DiCanio, Nam, Amith, Castillo García, & Whalen, 2015; DiCanio & Whalen, 2015; Laan, 1997; Moon & Lindblom, 1994; Warner & Tucker, 2011) and reduced duration in high frequency lexical items relative to low frequency items (Gahl, 2008; Gahl, Yao, & Johnson, 2012). For instance, in a study on voiceless stops in Bilbao (Castillian) and Central Columbian Spanish, Lewis (2001) finds greater rates of reduction (voicing during closure, shortening of closure duration) in conversational speech than in read passages or wordlist recordings. Obstruent reduction is more common in both spontaneous speech and in high frequency lexical items.

1.1.2. Is reduction just durationally-induced?

One of the principal articulatory/acoustic components tying all of these effects together is duration. Cross-linguistically, duration is one of the strongest correlates of stress (Gordon, 2016), though the prosodic unit onto which stress manifests can vary (foot, onset, vowel, rime). In languages possessing a geminate-singleton contrast, singleton obstruents may undergo processes of variable spirantization or passive voicing whereas geminates will not (DiCanio, 2012; Hualde & Nadeu, 2011; Stevens & Hajek, 2004). The findings showing greater rates of reduction at faster speech rates also suggests that duration is the primary factor predicting obstruent reduction and articulatory undershoot.

The relationship between duration and reduction is a design feature in articulatory phonology (Browman & Goldstein, 1990; Byrd & Saltzman, 2003; Byrd & Tan, 1996; Parrell, 2014; Parrell & Narayanan, 2018). If articulatory gestures have target durations in individual languages, one anticipates that obstruents shortened by prosodic or paralinguistic factors will undergo gestural undershoot in production. In a study examining lenition in English and Spanish using MRI, Parrell and Narayanan (2018) find that the patterns of Spanish coronal spirantization ([d - ð̞]) and English coronal flapping ([d/t - ɾ]) are gradient and predictable by duration. As obstruent duration decreases, speakers are continuously less likely to achieve full tongue tip to palatal contact – even though there may still be gestural evidence for tongue movement. In a study examining non-nasal velar lenition in Iwaidja (Iwaidjan: Australia), Shaw et al. (2020) find a close correlation between degree of constriction for the velar ([ɯ] vs. [a]) as measured via ultrasound and acoustics and the overall duration of the tongue body movement.

Studies examining variation in reduction in spontaneous speech additionally confirm these findings. Madrileño Spanish voiceless stops are shown to have shorter closure duration than Continental French stops and it is the Madrileño Spanish stops which undergo greater spirantization (Torreira & Ernestus, 2011). Speakers of Central Columbian Spanish have longer closure duration for voiceless stops than Bilbao Spanish

speakers do and voiceless stops are more often lenited and (partially) voiced the latter group (Lewis, 2001).[2] In a recent study of Campidanese Sardinian, Katz and Pitzanti (2019) find that duration accounts for most of the patterns of lenition among obstruents. Though, categorical features related to the prosodic hierarchy are still useful for characterizing the lenition patterns – increased duration, drops in intensity, and more abrupt releases occur at successively higher positions in the prosodic hierarchy above the word. In a study of lenition patterns in English using the Buckeye corpus (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005), Priva et al. (2020) find that durational changes were specifically causal in predicting patterns of lenition - moreso than other linguistic variables like stress position and the informational content of the target word/phrase.

In her dissertation examining languages lacking a voicing contrast Kakadelis (2018) finds much higher rates of spirantization and passive voicing in languages like Bardi (Western Nyulnyulan) than in languages like Arapaho (Algonquian) and Sierra Norte de Puebla Nahuatl (Uto-Aztecan). In Bardi, stops /p, t, k/ have an average duration of 45–48 ms where 35% of stops are realized without closure and voicing often persists throughout the entire closure (87–95% of closure). In Nahuatl, stops /p, t, k/ have an average duration of 67–84 ms, 10.7% of stops are spirantized, and voicing persists through 63–77% of the closure. In Arapaho, stops /b, t, k/ have an average duration of 79–138 ms, only 10% of stops are spirantized, and (for /t, k/) voicing persists through 36–45% of the closure. There is a close relationship between stop duration and patterns of lenition. If a language has shorter stops (due to inherent duration targets, speech rate, word size, etc), it will have more lenited stops.[3] Moreover, even within each of the languages in Kakadelis' study, the stops with the shortest average duration underwent spirantization at a greater rate than stops with longer durations. If a stop tends to have a shorter target duration in a given language, that stop will be more likely to undergo processes of lenition.

The findings in Kakadelis (2018) are additionally relevant because they suggest that patterns of lenition can vary even in languages lacking a phonological contrast in voicing within the oral stop series. Languages possessing an obstruent voicing or aspiration contrast may limit the degree of voicing bleed in either voiceless or aspirated stops. That is, the necessity to maintain a voicing/aspiration contrast in the language may limit the degree of voicing in leniting contexts. This idea finds support in cross-linguistic, typological studies demonstrating that patterns of categorical voicing lenition rarely result in contrast neutralization (Gurevich, 2004; Gurevich, 2011). Another, so far unexamined hypothesis is the idea that in languages possessing a contrast in continuancy in the obstruent series (stop vs. fricative), stops might be less prone to spirantization. Both hypotheses are based on the more general hypothesis that phonological contrast preservation is an active force that influences surface phonetic variation within a given language (c.f. Keyser & Stevens (2006); Lindblom (1990)). Yet, as noted above, reduction is more common in voiced stops in English than in voiceless stops. At first glance this would seem to be related to contrast, but note that voiced stops have shorter closure duration than voiceless stops in English and listeners can use this cue in perception (Lisker, 1957; Lisker, 1986). Thus, the presence of a contrast might only constrain the distribution of durational values for a given stop.

1.1.3. Theoretical motivations

The current study examines patterns of variable lenition in a spontaneous speech corpus of Yoloxóchitl Mixtec with three scientific questions in mind. First, prosodic boundaries have a strong influence on consonant duration. This predicts that obstruents in word-initial position should be less likely to undergo variable lenition than obstruents in word-medial position. Second, stress also influences consonant duration. This predicts that obstruents in the onset of stressed syllables[4] should be less likely to undergo variable lenition than obstruents in the onset of unstressed syllables. Third, inherent durational differences for obstruents typically correlate with variable patterns of lenition. This predicts that duration will be closely correlated with rates of spirantization and voicing lenition. While experiments 1 and 2 investigate these scientific questions, experiment 3 models the discrete allophonic variants examined in experiment 2 using deep neural networks. The motivation for experiment 3 is to determine whether surface phonetic variants observed in a reasonably small corpus of spontaneous speech can be accurately predicted in a computational model.

1.2. Background: Yoloxóchitl Mixtec phonology

Yoloxóchitl Mixtec [jolo$^{\text{ˈ}}$sotʃit̰$^{\text{ˈ}}$mistɛk] is an Oto-Manguean (Mixtecan:Mixtec) language spoken in the towns of Yoloxóchitl, Cuanacaxtitlán, Buena Vista, and Arroyo Cumiapa in Guerrero, Mexico (García, 2007). The name "Mixtec" does not refer to the language itself, but to an ethnolinguistic grouping and language family comprising approximately twelve pan-dialectal regions and between 50–60 language varieties (c.f. DiCanio et al. (2020); Josserand (1983)). For the most part, languages spoken across pan-dialectal regions are not mutually intelligible. For instance, there is reasonably good mutual intelligibility across most of the Guerrero Mixtec languages (García, 2007), but only approximately 30% mutual intelligibility between the Guerrero and Southern Baja pan-dialectal regions (Lewis, Simons, & Fennig, 2015). There are approximately 4,000 speakers remaining, though many younger speakers are more dominant in Spanish than in Yoloxóchitl Mixtec.

Yoloxóchitl Mixtec possesses a complex lexical tone system consisting of nine distinct tones which are moraically-aligned (García, 2007; DiCanio, Amith, & Castillo García, 2014; DiCanio et al., 2018; Palancar et al., 2016). Lexical roots are minimally bimoraic (CV$^{\text{ː}}$or CVCV) but longer words are possible with both prefixation (marking negation, tense, and aspect) and with enclitic morphology, which marks person on verbs or possessors on nouns (Palancar et al., 2016). Certain words may consist of a single mora, but these are entirely functional

[2] An anonymous reviewer suggests that different dialects of Spanish may simply have different constriction targets for different stops. While we believe this is possible and worthy of study, across published studies it appears that such differences usually correlate strongly with dialectal differences in speech rate. This, in turn, directly influences closure duration.

[3] For a discussion of cross-linguistic differences in speech rate, please see Pellegrino, Coupé, and Marsico (2011); Verhoeven, De Pauw, and Kloots (2004).

[4] There are no codas in Yoloxóchitl Mixtec.

particles, i.e. adverbials like /ka^{1}/ 'still' and /ha^{143}/; or pre-nominal classifiers like/ja^{1}/ 'that (one).' Syllables are obligatorily open (CV) and glottalization is contrastive on lexical roots, occurring between moras in disyllabic roots (CV^{ʔ}CV) and monosyllabic roots (CV^{ʔ}V), e.g. [^{n}de^{3}e^{3}] 'flipped' vs. [^{n}de^{3ʔ}e^{3}] 'ground bean.'[5]

There are five phonemic vowel qualities and vowel nasalization is also contrastive, e.g. /i, e, a, o, u, ĩ, ë, ã, õ, ũ/. Previous research examining vowel production in a corpus of elicited and spontaneous speech found significant effects of speech style and duration on vowel production (DiCanio et al., 2015). Vowels are reduced in spontaneous speech relative to elicited speech but these contexts also significantly influenced overall vowel duration. The consonant inventory is relatively small, consisting of just fourteen contrastive consonants/p, t̪, k, kʷ, ᵑg, ʧ, m, n, ɾ, s̪, ʃ∼h, β, j, l/. Prenasalized stops (/^{m}b, ^{n}d/) are allophones of nasal consonants which surface before oral vowels (DiCanio et al., 2020). The fricatives [ʃ] and [h] are in free variation, but the distribution of each allophone is rather different. Out of 733 examples of [h] in the corpus (see Section 2.1), 98.6% (723/733) occurred in word-initial position. Out of 768 examples of [ʃ] in the corpus, 38.5% (296/768) occurred in word-initial position and 61.5% (472/768) occurred in word-medial position. The glottal fricative is almost never found in word-medial position and therefore it rarely occurs in the onset of the stressed syllable. This is unexpected given standard assumptions about where we expect debuccalization to occur (in word-medial position).

The four stops/p, t̪, k, kʷ/ are all voiceless unaspirated in careful speech and elicited recordings, with a positive VOT range between 11–32 ms (DiCanio et al., 2020). García (2007) describes a variable process of velar lenition, where velar and labialized velar stops will be realized as frictionless continuants [ɣ, ɣ̞ʷ]. This lenition was not observed in the recordings from eight speakers in DiCanio et al. (2020) but these speakers produced carrier sentences with target words; the speech was elicited. In spontaneous speech, patterns of stop lenition are noticeable. Voiceless fricatives are also variably debuccalized. The post-alveolar fricative freely varies with [h], though the post-alveolar articulation is much more common (García, 2007). The dental fricative /s̪/ occasionally also is produced as [h], though impressionistically this seems rarer than post-alveolar fricative debuccalization.

In addition to lexical tone, there is evidence for root-final, fixed stress in Yoloxóchitl Mixtec. Whereas five lexical tones surface on non-final syllables of roots, nine surface on root-final syllables. Root-final syllables are also consistently longer than non-final syllables (DiCanio et al., 2018). One of the interesting manifestations of stress in the language is the asymmetry in onset consonant duration. Onset consonants in final stressed syllables are longer than onset consonants in non-final, unstressed syllables. This has been observed in words in different focus conditions (DiCanio et al., 2018) and in words elicited in carrier sentences with sentential focus (DiCanio et al., 2020). In the former case, the target focused constituents were all utterance-initial and stop consonants were excluded from the analysis since one can not accurately measure voiceless stop closure duration in this position. In the latter case, stop consonants were included, but the recording conditions involved elicited and more careful speech productions.

[5] Numbers indicate tone, where /1/ is low and /4/ is high. Tone is orthogonal to glottalization and most tonal melodies can freely occur on glottalized roots.

2. Experiment 1: Lenition in Mixtec spontaneous speech

2.1. Speakers and Materials

Yoloxóchitl Mixtec has been the focus of a major language documentation project which has produced over 200 h of carefully transcribed texts (Amith & García, 2019; Amith & García, 2021). These texts consist almost entirely of spontaneous speech narratives and conversations spoken by native speakers. From this corpus, 85.5 min of spontaneous speech was selected, as produced by three female speakers and three male speakers. Individual sound files were between 318 and 1368 s in duration. Though the length of each recording varied by speaker, we chose approximately the same length of recordings from men and from women (44.7 and 41.7 min, respectively). The mean age of the speakers at the time of recording was 36.5 years (37.7 female, 35.3 male). One speaker from each gender group was older (49 and 60 years old) and the remaining four speakers were younger (22–31 years old). All speakers were fluent, native speakers of Yoloxóchitl Mixtec and used the language in their daily life. They also each resided in the town of Yoloxóchitl where their Mixtec variety is a language of daily communication.

Time-aligned morphophonemic transcription of this corpus was first done in ELAN (Wittenburg, Brugman, Russel, Klassman, & and Sloetjes, 2006) by the last author and then exported to Praat (Boersma & Weenink, 2016). Forced alignment was completed using an adapted Mixtec lexicon with the Penn aligner (P2FA) (Yuan & Liberman, 2009; DiCanio et al., 2013). Segments mis-identified by P2FA were recoded as native Yoloxóchitl Mixtec segments using an existing phonological transducer built for the language by Jonathan Amith, Jason Lilley, Rey Castillo García and Christian DiCanio. Acoustic boundaries for all segments in this corpus were then additionally corrected by hand. Given that many of the tokens were reduced and produced as approximants with continuous voicing, a simple acoustic rubric was used for segmenting these tokens – we chose the acoustic midpoint in the intensity transitions between adjacent vowels and the target consonant. Since Yoloxóchitl Mixtec has strictly CV syllable structure, adjacent vowels were always present. When obstruents were realized with abrupt acoustic boundaries in relation to adjacent vowels (loss of formant structure, abrupt shifts in intensity), the boundaries for duration measurement were easy to estimate and our hand-corrections of these boundaries often were close/identical to the boundary position estimated by the forced aligner.

Owing to the nature of spoken Mixtec, some of these words were loanwords from Spanish (mostly proper names). Out of a total of 8,683 obstruents in the corpus, 760 (8.8%) came from Spanish loanwords. After these were excluded, a total of 7,923 obstruents were analyzed. Most of the obstruents were included in this analysis (/t̪, k, kʷ, ʧ, s, ʃ, h/). The stop consonant /p/ is very rare in native Mixtec words (a pan-Mixtecan pattern) so it was excluded here. For statistical analyses exam-

ining syllable position, all polysyllabic words were grouped together and monosyllabic words were excluded. Monosyllabic words were analyzed separately (see below).

2.2. Methods

2.2.1. Continuous measures of voicing lenition

For the first analysis, both consonant duration and proportion of voicing during the calculated obstruent duration were extracted and analyzed. The proportion of voicing during consonant constriction was calculated using the normalized low frequency energy ratio (NLFER), a component of the pitch-tracking algorithm YAAPT (Kasi & Zahorian, 2002; Zahorian & Hu, 2008). The sum of absolute values of spectral samples over the low frequency regions is taken (average energy per frame), and then normalized by dividing by the average low frequency energy per frame over the utterance (ibid). For the YAAPT analysis of the Yoloxóchitl Mixtec data, a 35 ms frame window was used with a 10 ms time step. F0 range was adjusted by gender after manually inspecting each sound file. A range of 80–300 Hz was used for male speakers and a range of 110–450 Hz was used for female speakers. The other algorithm control parameters followed those used in Zahorian and Hu (2008).

This method of voicing detection differs from methods used in previous research on voicing lenition where researchers have either manually categorized segments as voiced, partially-voiced, or voiceless (Hualde et al., 2011; Warner & Tucker, 2011) or relied on the autocorrelation-based voicing detection algorithm (i.e. in Praat (Boersma & Weenink, 2016)) (Davidson, 2018). YAAPT outperforms Praat's (Boersma & Weenink, 2016) autocorrelation method for pitch-tracking across a variety of SNR conditions (Zahorian & Hu, 2008). Using YAAPT instead of Praat also allowed us the ability to quantify the proportion of voicing during the obstruent duration.

2.2.2. Statistical analyses

For the continuous, quantitative measure of duration on polysyllabic words, we fitted Bayesian hierarchical linear models with two fixed effects (syllable position and consonant) and their two-way interaction using the Stan modeling language (Carpenter et al., 2017) and the package brms (Buerkner, 2016). For the voicing data, a zero-one inflated beta regression was set as the response distribution (Liu & Kong, 2015). This distribution is bounded (0–1) and includes values at the margins to reflect the nature of the data – many tokens are fully voiced and many are completely devoiced. For the model encoding stress, syllable (stem) position consisted of two levels (non-final and final) and consonant consisted of six levels (the consonants/t̪, k, kʷ, ʧ, s, ʃ/) which were orthographically coded as "t, k, kw, ch, s, x", respectively. For a model encoding initial position in trisyllabic and longer words, syllable (stem) position consisted of three levels (initial, medial, final). All factors were refactored with sum contrast coding. We excluded the glottal fricative (/h/) from our analysis since it never surfaced in final syllables of disyllabic or polysyllabic words in the corpus. The models included the maximal random effects structures that were justified by the design. Random effects included by-item and by-speaker intercepts; and by speaker random slopes for syllable position, consonant, and their interaction term.

For the durational models, we used a normal prior for the mean of the reference cell (final position) (μ = 0, σ = 100) and a Student's t-distribution prior for the likelihood function (υ = 3, μ = 0, σ = 100). Unbiased priors were used for regression coefficients. We used the default priors from the brms package for standard deviations of random effects (a Student's t-distribution with υ = 3, μ = 0 and σ = 20), as well as for correlation coefficients in interaction models (LKJ η = 1). For the voicing models, we used a normal prior for the mean of the reference cell (final position) (μ = 0, σ = 1) and a Student's t-distribution prior for the likelihood function (υ = 3, μ = 0, σ = 100). For all models, four sampling chains ran for 2000 iterations with a warm-up period of 1000 iterations for each model, thereby yielding 4000 samples for each parameter tuple. For each of the relevant factors, we report the beta estimates under the posterior distribution and their 95% credible intervals (CIs). We also report the posterior probability that the difference δ is larger than zero. If a hypothesis states that $\delta > 0$, we judge there to be compelling evidence for this hypothesis if zero is (by a reasonably clear margin) not included in the 95% CI of δ and the posterior $P(\delta > 0)$ is close to one.

Separate Bayesian hierarchical linear models were fit for monosyllabic words. Here, we examined a different fixed effect – whether the word was monomoraic (and therefore a functional particle) or bimoraic (and therefore usually a lexical root); alongside consonant type. Only three consonants occurred in monomoraic words (/ʧ, k, t/ (ch, k, t)), so the model fit here included only these consonants. The motivation for including the monomoraic effect was the observation that function words are frequently reduced more often than content words (Gahl et al., 2012). Since the former is phonologically distinct in Yoloxóchitl Mixtec, we believed that examining this distinction might offer a way to tease apart variation related to consonant duration and the degree of voicing lenition observed in words where syllable position can not be examined. Though we are evaluating the effect of word size in this study, it should be noted that the distribution of words by size varied substantially within the corpus. Out of 7,159 obstruents, just 678 (9.5%) came from polysyllabic words. The majority of the obstruents came from disyllabic words (3,687, 51.5%) and monosyllabic words (2,794, 39.0%). R 3.6.3 was used for all statistical analyses (R Development Core Team, 2020).

2.3. Results I: Consonant duration

Fig. 1 shows the effect of stem position, word size, and consonant type on consonant duration in Yoloxóchitl Mixtec for each of the consonants. Grouping all consonants together, the mean duration for onset consonants in non-final syllable position was 74.1 ms whereas the mean duration for those in final position was 98.1 ms. There is compelling evidence for the durational difference between these consonants (β = 14.7, 95% CI: [9.2, 19.7], $P(\delta > 0) = 1$) and, in fact, there is quite compelling evidence that this difference in duration is at least 10 ms ($P(\delta < -10) = 0.96$). The 95% CI around the estimate of this effect did not include zero, providing compelling evidence for the direction and strength of the effect.

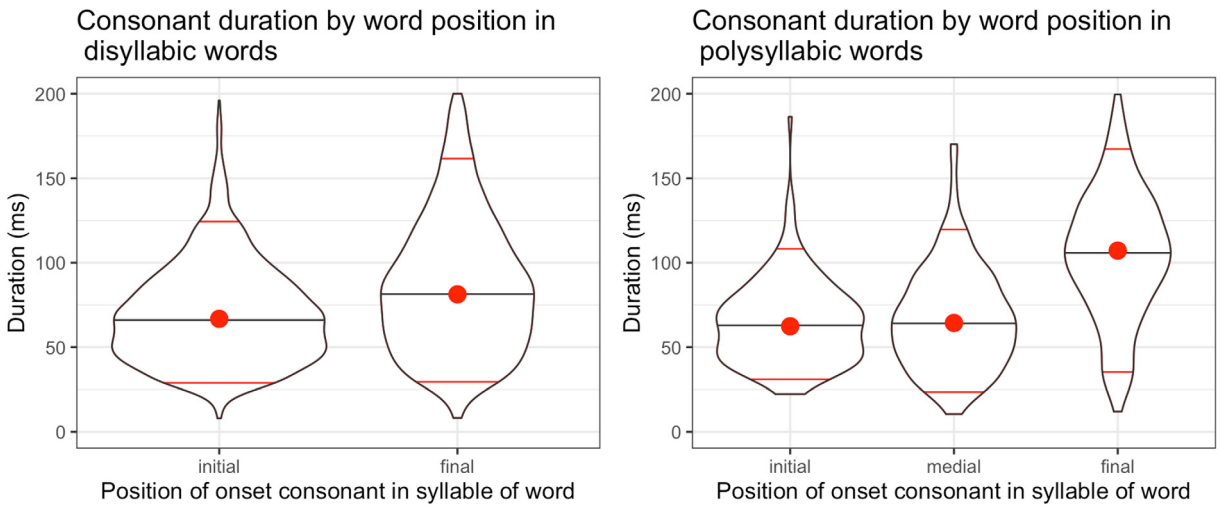

Fig. 1. Consonant duration by word size and stem position in Yoloxóchitl Mixtec. Red lines indicate boundaries at 5% and 95% of the distribution. The black line indicates the median value and the red dot indicates the mean value.

For the fixed effect of consonant type, the consonant/s#x032A;/ was coded as the intercept since its value was closest to the mean duration for all consonants. Two patterns were worth noting in relation to this fixed effect. First, there is compelling evidence that the average duration of /t̪/ and /tʃ/ were longer than the weighted mean duration of the consonants (for /t̪/: β = 10.0, 95% CI: [4.2, 15.6], P(δ > 0) = 0.993; for /tʃ/: β = 18.3, 95% CI: [11.7, 26.3], P(δ > 0) = 0.999). Second, there is compelling evidence that the average durations of /k/ and /kʷ/ were shorter than the weighted mean duration of the consonants (for /k/: β = −16.5, 95% CI: [-25.7, −7.5], P(δ < 0) = 0.993; for /kʷ/: β = −14.7, 95% CI: [-26.9, −1.9], P(δ < 0) = 0.986). The observations regarding velar shortening were specifically hypothesized by past work discussing patterns of velar lenition (see Section 1.2). For all consonants except /tʃ/, the expected relation held where the consonant in final syllable position was longer than the consonant in non-final position. Apart from this, the degree of this effect varied a little by consonant. There is compelling evidence that the difference in duration by word position was stronger for /t̪/ (β = 7.2, 95% CI: [1.1, 13.2]) than for other consonants.

An additional model was run consisting of only polysyllabic words with the same effect structure but where the main effect of stem position was split into three categories instead of two - word-initial, word-medial, and word-final. Within this model, there is compelling evidence for the durational difference between final and non-final consonants (β = −13.9, 95% CI: [-28.0, −0.1], P(δ > 0) = 0.98) but no evidence that the durations of word-initial and medial onset consonants differed (β = −7.2, 95% CI: [-25.9, 15.2], P(δ > 0) = 0.78). Though most words in the corpus (and in the language) are disyllabic, we observe no difference in duration between unstressed word-initial and word-medial consonants in trisyllabic or longer words.

Fig. 2 shows the effect of functional status and consonant type on consonant duration in Yoloxóchitl Mixtec monosyllabic words. All monomoraic monosyllabic words are functional mor-

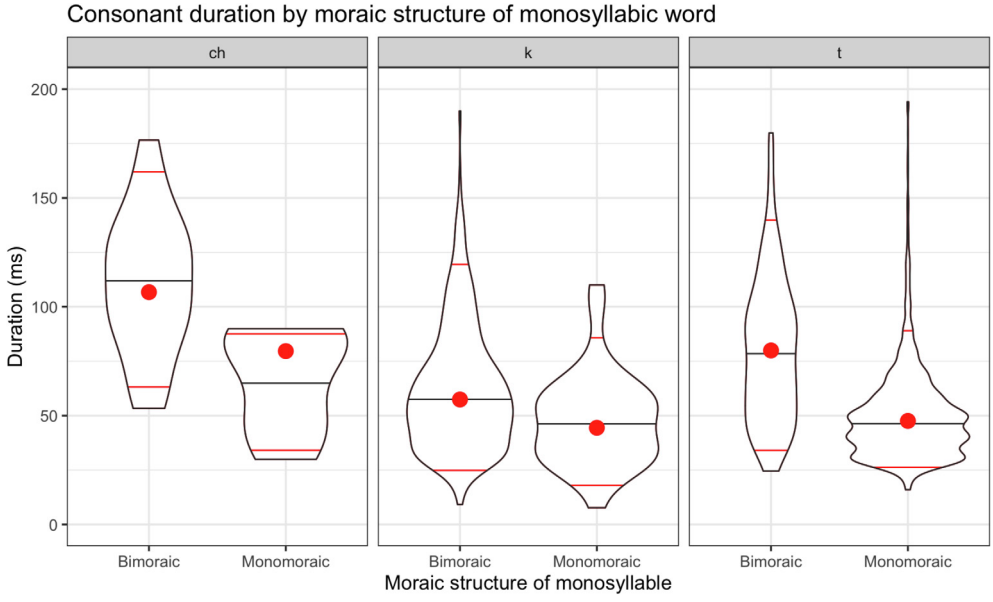

Fig. 2. Consonant duration by moraic structure in Yoloxóchitl Mixtec. Consonants /tʃ, k, t̪/ are represented orthographically as “ch, k, t” here. Monomoraic words are all functional morphemes. Red lines indicate boundaries at 5% and 95% of the distribution. The black line indicates the median value and the red dot indicates the mean value.

phemes in Yoloxóchitl Mixtec and onset consonants in these words are uniformly shorter (mean = 53 ms) than those found in bimoraic monosyllabic words (mean = 68 ms). There is compelling evidence for this durational difference (β = 12.2, 95% CI: [3.7, 20.7], P(δ > 0) = 0.998). The distribution of consonant type varied substantially: only 26 monosyllabic tokens containing the consonant /ʧ/ were observed in the data, whereas 2247 monosyllabic tokens containing the consonants /t̰/ or /k/ were observed. Of these latter two consonants, just 32/933 tokens of /k/ occurred in monomoraic words while 1099/1314 tokens of /t̰/ occurred in monomoraic words. Despite this imbalance, the general pattern of monomoraic consonant shortening/reduction appears to hold across each consonant type examined. No interactions between functional status and consonant type were found.

2.4. Results II: Voicing lenition

Fig. 3 shows the effect of stem/word position, stem/word size, and consonant type on consonant voicing in Yoloxóchitl Mixtec for each of the consonants. Grouping all consonants together, on average voicing for onset consonants in non-final syllable position occurred over 29.4% of the consonant duration whereas the voicing for those in final position occurred over 24.8% of the duration. Though this difference is small, there is compelling evidence for an effect by position (β = −0.097, 95% CI: [-0.172, −0.030], P(δ < 0) = 0.988). The 95% CI around the estimate of this effect did not include zero, providing compelling evidence for the direction and strength of the effect.

There was compelling evidence for one consonant-specific effect as well - the affricate/ʧ/ underwent less passive voicing during closure than the other consonants (β = −0.168, 95% CI: [-0.313, −0.040], P(δ < 0) = 0.992). Recall that this affricate was also longer in duration than the other consonants. The amount of voicing during closure may therefore be consistent across consonant types but comprise a smaller proportion of the entire constriction for the affricate. Perhaps because of this, we also found compelling evidence for an interaction between position and consonant for /ʧ/ (β = 0.146, 95% CI: [0.016, 0.289]). Though the mean proportion of voicing for this affricate varied by position in the same direction as the main effect (22% in non-final position, 16% in final position), these differences by position were not as robust.

Unlike the durational data for monosyllabic words above, there no compelling statistical evidence for an effect of functional status on the degree of passive voicing during constriction. Though, the trend in the data is identical to the pattern observed for polysyllabic words above. That is, more voicing is observed in monomoraic roots, all of which are function words and of shorter duration, than in the onset consonants of bimoraic roots. The pattern is shown in Fig. 4.

2.5. Results III: Does duration predict voicing lenition?

Shorter duration is one common thread connecting the pattern of voicing lenition found in function words and in non-final syllables in Yoloxóchitl Mixtec. In both cases, shorter obstruents are more likely to undergo voicing lenition. While grammatical and phonological constraints determine durational patterns in the language, what is the specific relation between duration and voicing lenition? A Bayesian regression model which examines this directly can reveal the specific relation-

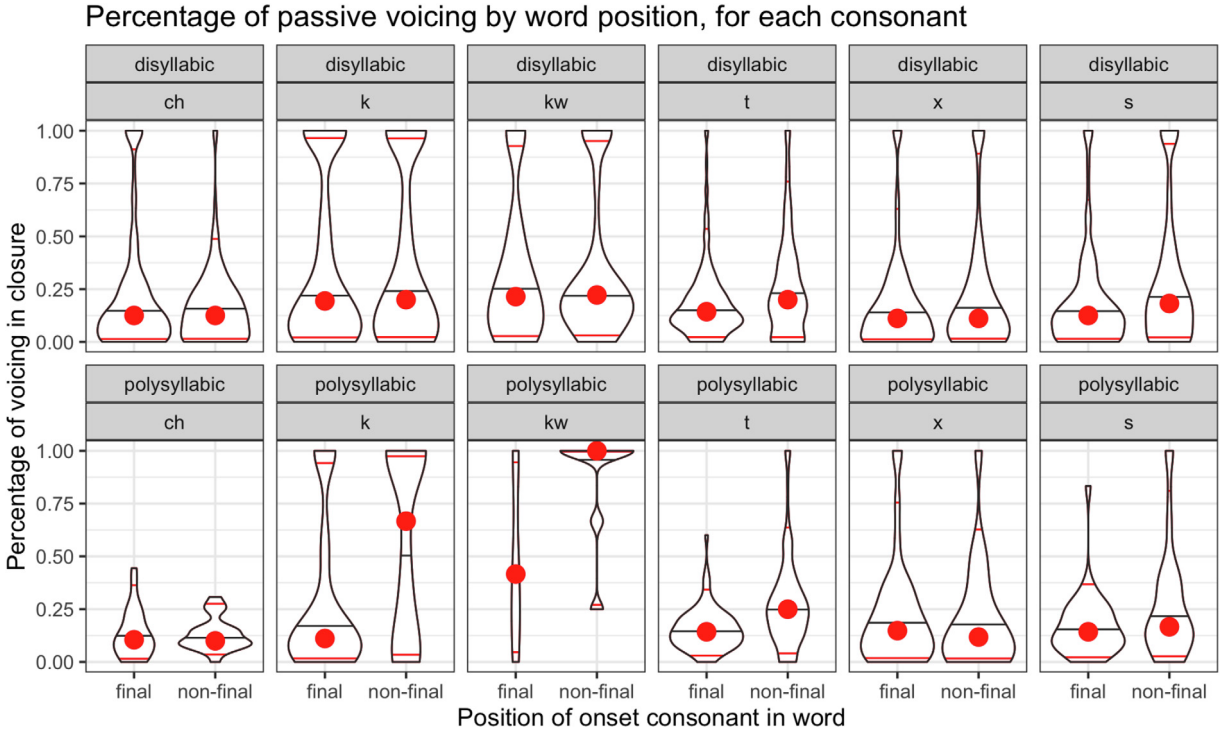

Fig. 3. Percentage of voicing during closure/constriction by word size and word position in Yoloxóchitl Mixtec. Red lines indicate boundaries at 5% and 95% of the distribution. The black line indicates the median value and the red dot indicates the mean value.

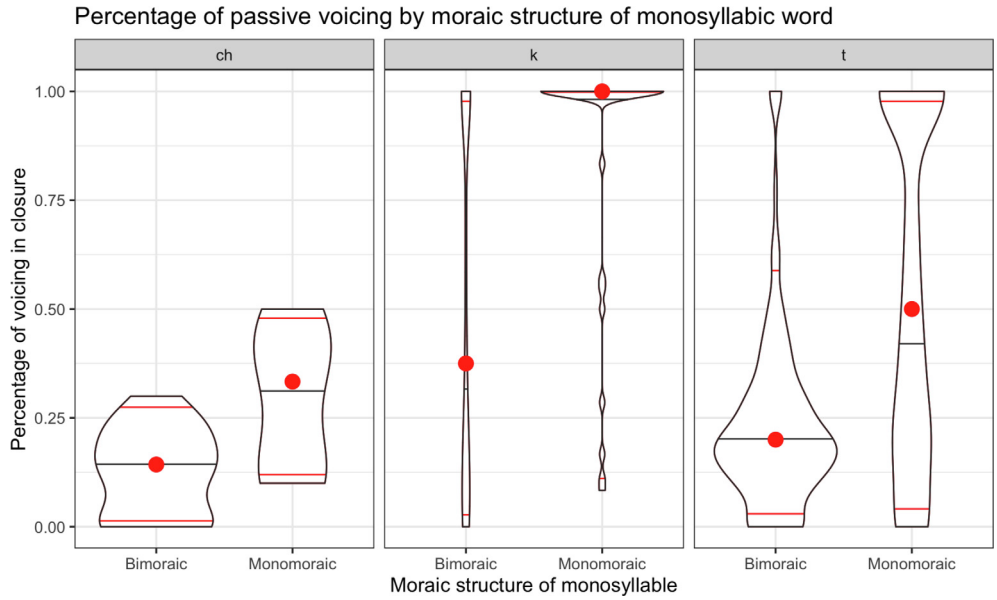

Fig. 4. Percentage of voicing during closure/constriction by moraic structure in Yoloxóchitl Mixtec. Red lines indicate boundaries at 5% and 95% of the distribution. The black line indicates the median value and the red dot indicates the mean value.

ship. Fig. 5 plots the relationship between voicing and constriction duration for each of the consonants examined.

One clear trend from Fig. 5 is that this relationship is non-linear. To model the relationship between voicing percentage and duration, we fitted a Bayesian hierarchical model with two fixed effects, log(duration) and consonant; and their two-way interaction using the Stan modeling language (Carpenter et al., 2017) and the package brms (Buerkner, 2016). As above, a zero-one inflated beta regression was set as the response distribution (Liu & Kong, 2015). The remaining model specifications were identical to those described above for the voicing analysis in Section 2.2.2. For each of the relevant factors, we report the beta estimates under the posterior distribution and their 95% credible intervals (CIs). We also report the posterior probability that the difference δ is larger than zero. If a hypothesis states that $\delta > 0$, we judge there to be compelling evidence for this hypothesis if zero is (by a reasonably clear margin) not included in the 95% CI of δ and the posterior $P(\delta > 0)$ is close to one.

There is compelling evidence for a main effect of log(duration) on voicing ($\beta = -0.93$, 95% CI: [-1.00, -0.86], $P(\delta < 0) = 1$). The 95% CI around the estimate of this effect did not include zero, providing compelling evidence for the direction and strength of the effect. There was compelling evidence for a consonant-specific effect in this model - less voicing overall was found for /k/ ($\beta = -0.364$, 95% CI: [-0.641, -0.073], $P(\delta > 0) = 0.959$) relative to the weighted mean for other consonants. We found compelling evidence for an interaction between duration and consonant for /k/ ($\beta = -0.128$, 95% CI: [-0.222, -0.027]), for /kʷ/ ($\beta = 0.333$, 95% CI: [0.147, 0.513]),

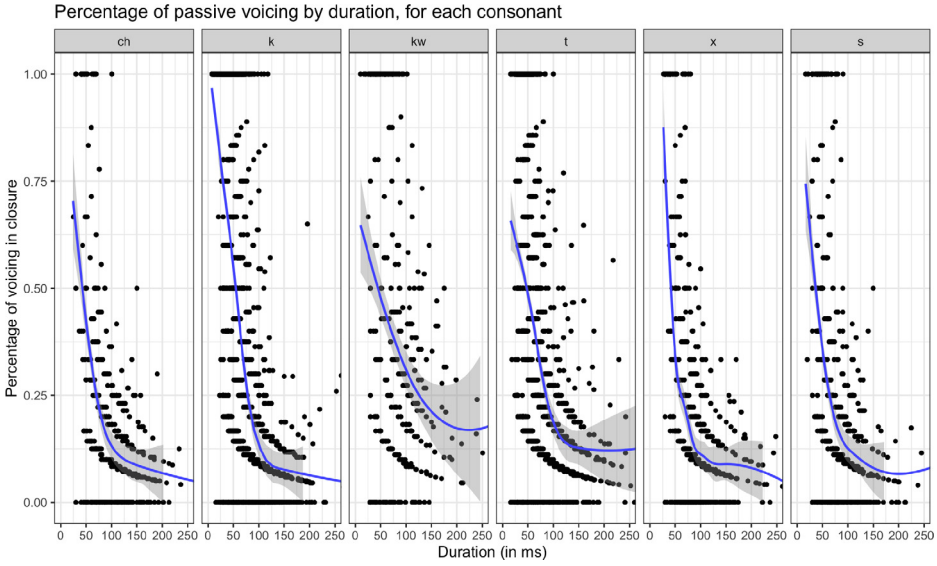

Fig. 5. Percentage of voicing during closure/constriction duration in Yoloxóchitl Mixtec. The blue line traces a smoothed conditional mean using locally estimated scatterplot smoothing (LOESS) to approximate the trajectory and the grey outline indicates the standard error around the local mean value.

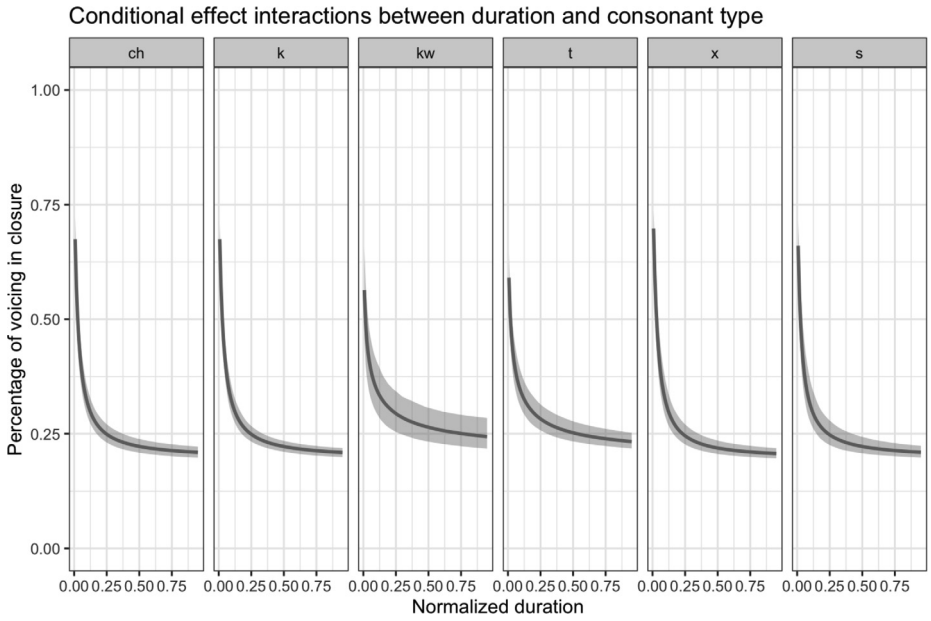

Fig. 6. Model conditional fits for the interaction between duration and consonant type on voicing percentage. The solid line traces a smoothed median value and the grey outline indicates the standard error around the local mean value.

and for /t̰/ (β = 0.236, 95% CI: [0.135, 0.345]). These interactions reflect the nature of the corpus data. There are more /k/ tokens in the corpus with a shorter duration and more tokens of /kʷ/ and/t/ in the corpus with a longer duration. As a result, the logarithmic fits for these consonants are distinct from each other, as the plot of the conditional fits demonstrates in Fig. 6.

2.6. Interim discussion

The duration results for polysyllabic words indicate that obstruents in stem-final syllables in Yoloxóchitl Mixtec are longer in duration than obstruents in non-final syllables (including both word-initial and word-medial onsets). This finding is in general agreement with observations made previously regarding stress placement (DiCanio et al., 2018; DiCanio et al., 2020; DiCanio, Benn, & Castillo García, 2021). However, unlike most of the previous work, we have found the effect in a corpus of spontaneous speech. While the effect of word position on duration held for all obstruents, no effect was found for the affricate /tʃ/, which also had longer average duration than the other obstruents. The voicing results mirrored the durational results - less passive voicing was observed in the obstruent onset of stem-final, stressed syllables than in those in non-final syllables. With respect to the hypotheses laid out above, we find that voicing lenition is less common in the onset of a stressed syllable than in the onset of an unstressed syllable, in-line with findings from Bouavichith and Davidson (2013) for English. Obstruents in non-pre-tonic positions (in any non-final syllable) underwent more passive voicing than those in word-medial pre-tonic position (in the final syllable onset).

The results for the monosyllabic words revealed a difference in duration for the onset obstruents of functional morphemes relative to content morphemes. However, unlike the results for the polysyllabic words, there was no compelling evidence that obstruents in functional morphemes were realized with more passive voicing than those in content morphemes. Durational shortening of function morphemes has been found in many languages (see Gahl et al. (2012)), but to our knowledge, this is the first instrumental evidence that it occurs in an Otomanguean language.

Some degree of passive voicing (cf. Westbury & Keating, 1986) occurs in most obstruents in Yoloxóchitl Mixtec running speech (since most are not utterance-initial), but the relationship between duration and voicing is rather different for obstruents shorter than (about) 100 ms in duration in comparison with those longer than 100 ms. Duration is a strong predictor of voicing probability. This finding is in line with similar findings regarding reduction in Kakadelis (2018) for Arapaho, Bardi, and Sierra Norte de Puebla Nahuatl, all languages like Yoloxóchitl Mixtec which lack a phonological voicing contrast in obstruents.

A non-linear relationship between voicing and duration is expected if the overall duration of passive voicing is held constant ($y = c/x$ is an inverse function). However, if voicing duration were constant, then we would anticipate similar average voicing duration values regardless of total obstruent duration. This is not the case. If we bin only those obstruents with total duration < 50 ms, the average voicing duration is 22.4 ms, while those with duration values between 50–100 ms have an average voicing duration of 19.9 ms and those with duration values above 100 ms have an average voicing duration of 15.9 ms. Voicing duration during constriction is not constant, but it increases slightly with shorter duration obstruents.

3. Experiment II: Categorical lenition patterns

3.1. Method

Voicing lenition is just of the types of lenition which occurred within the Yoloxóchitl Mixtec spontaneous speech corpus. Many stop consonants were also realized with incomplete closure. As mentioned above, García (2007) describes a variable process of velar lenition, where velar and labialized velar stops will be realized as frictionless continuants [ɣ̥, ɣ̥ʷ]. Yet, impres-

sionistically this lenition process does not appear to be limited to velar stops. The second analysis utilized a qualitative, discrete measure of lenition where the two most-common stop consonants (/t, k/) were individually coded and examined by the third author along a series of possible lenition categories. A total of 4,552 stops were categorized into eight groups by visual inspection of the third author: (1) voiceless stop, (2) partially voiced stop, (3) voiced stop, (4) voiced fricative, (5) voiced approximant, (6) nasal stop, (7) Tap, (8) Deletion. These categories were not chosen arbitrarily, but came out of the first and third author's observations that additional groupings were needed to capture the variability that was observed in the data. Stops were identified by the presence of a visible burst release and voicing was identified by the presence of a visible periodicity in the waveform. If a stop contained at least one period of voicing but voicing ceased prior to release, it was categorized as partially voiced. If neither noticeable formant transitions nor noticeable amplitude transitions were present, the stop was counted as "deleted." Categories (4) and (5) were mainly distinguished by the annotator's observation of steady state frication (indicating a voiced fricative) versus the absence of a steady state (indicating a voiced approximant).[6]

Categorization of stops was done via the use of a script written for Praat which tracked phonological variant type. This script scans a sound recording for the target phone and displays the speech waveform and spectrogram corresponding to the labelled interval. It then allows users to select the variant realization. These variant realizations (groups 1–8, above) are saved to a log file alongside the durational interval of the stop. The user also specified the position of the obstruent in the word (word-initial vs. word-medial). The encoding of word position here was different than the more precise labelling in Experiment 1 where word length and exact word position were recorded.

The categorization performance of the third author was checked against a subset of the data categorized by the first author, comprising 953/4552 tokens or 21% of the corpus. Since so many categories are included, the overall agreement between researchers was somewhat low (44.5%), however most disagreements occurred in deciding whether (a) a token was deleted or there was evidence of some formant transitions (b) disagreements between whether a stop was realized as a voiced approximant or voiced fricative, and (c) whether the dental stop was realized as a tap or voiced fricative. In other words, most disagreements related to the subtler distinctions in allophonic variation. With respect to the larger distinctions, such as whether the stop was realized with closure or not, the researchers in fact showed a high level of agreement (93.4%). With respect to whether voicing was present during closure, the researchers also showed a fairly high level of agreement (83.5%). Some examples of variant realizations and their categorizations are given in Fig. 7.

We report a listing of all the variant types and their frequencies here. For statistical analysis, variant types were categorized in terms of whether the target stop involved closure or not. Groupings (1), (2), (3), (6), and (7) above were placed in the *closure* group while groupings (4), (5), and (8) were placed in the *no closure* group. A generalized logistic mixed effects model (glmer with a binomial fit) was then used to examine whether word position (initial/medial) and segment type (/t̪, k/) were significant predictors of closure grouping. This model included word position, stop place, and their interaction as fixed effects and random intercept for speaker. Models including random slopes for word position and stop place did not converge.

3.2. Results

Fig. 8 displays the variant realizations of the two stops analyzed here. One notable observation is that stop realizations without voicing were relatively rare in the spontaneous speech corpus for both /t̪/ and /k/. For /t̪/, voiceless or partially voiced stops make up 52% of all stop realizations and for /k/, voiceless or partially voiced stops make up just 37% of all stop realizations. For /t̪/, a number of additional novel realizations were observed in the corpus data, including a nasal stop [n] and a tap [ɾ] though these were rarer relative to the variants varying in constriction degree and voicing ([t̪, d̪, ä, a̤]). Nasal realizations were most frequent in contexts adjacent to nasal vowels, e.g. /t̪ã^3/ 'and' > [nã^3] in running speech.

While the presence of full voicing in obstruents was observed in the previous experiment in quantitative detail, the qualitative data here provides information on full vs. incomplete closure. The relationship between stem/word position and closure type is given in Fig. 9. We observe that more stops in word-initial (non-final) position were realized with incomplete closure than in word-medial (pre-tonic) position. Stress position was a significant predictor of closure, β = −0.56, z = −6.4, p <.001. Place of articulation was also a predictor of closure, where incomplete closure was significantly more common in velar stops than in dental stops, β=-1.34, z = −15.8, p <.001. A significant interaction between stem position and place of articulation was also observed, β=-2.08, z = −5.6, p <.001. Word-initial/k/ was realized without closure in 47% of the tokens in word-initial position but in 34% of the tokens in word-medial position (a difference of 13%). Word-initial/t/ was realized without closure in 20% of the tokens in word-initial position but in just 1.6% of the tokens in word-medial position (a difference of 18%). The interaction between stem position and place of articulation reflects this differences in magnitude (13% vs. 18%).

3.3. Interim Discussion

The results here suggest that final stress patterning may lead to an effective pattern where lenition occurs more often in non-final syllables. Voiceless stops are more likely to be realized without closure in non-final, non-pre-tonic positions, but more likely to be realized with closure in pre-tonic final syllables.

4. Experiment III: Modelling allophonic detail using deep neural networks

While it is possible for trained human speech scientists and phoneticians to categorize surface phonetic variants as in Experiment 2, can these categories be detected by a computational model? A current bottleneck in corpus phonetics is the ability to provide phonetic detail below the level of the transcribed phone. Speech corpus segmentation typically reflects

[6] A reviewer notes that this distinction is probably unreliable. We are in general agreement with this assessment, but considered these to be important categories for our initial categorization. We address this specific point further in Section 5.2.

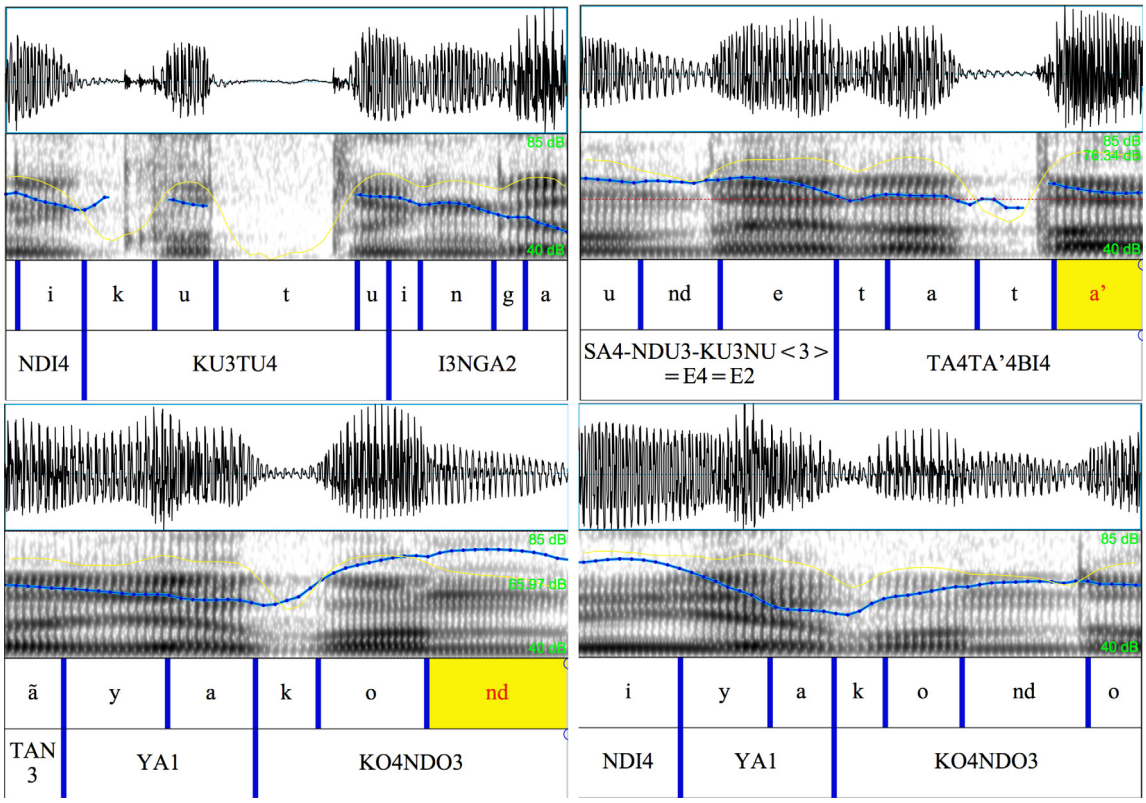

Fig. 7. Examples of variant realizations of stop consonants in Yoloxóchitl Mixtec. In the first, upper panel spectrogram, the initial /k/ in the word /ku^{3}ʯ^{4}/ 'to fill up' is realized with partial voicing while the medial /ʯ/ is realized with no voicing. In the second, upper panel spectrogram, the word-initial /ʯ/ in /ʯa^{4}ʯa^{4}ʔ^{4}βi^{4}/ 'of both sides' is realized as a voiced approximant, while the following stop is realized with complete closure but partial voicing. In the first, lower panel spectrogram, the initial /k/ in /ko^{4} ^{n}do^{3}/ '$_{HAB}$.be' is realized with a steady state period of constriction and categorized as a voiced fricative. The same word is realized as an approximant in the panel to the right.

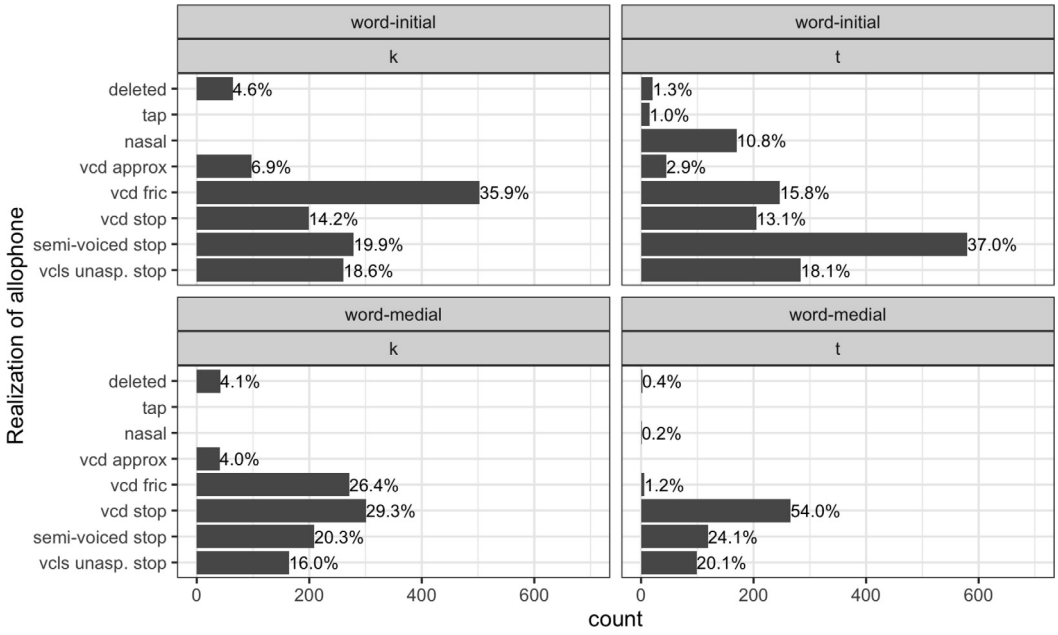

Fig. 8. Counts of variant realizations of stop consonants /k, ʯ/ in the Yoloxóchitl Mixtec corpus.

phonemic categories after forced alignment has been applied (Babinski et al., 2019; DiCanio et al., 2013; Tang & Bennett, 2019; Yuan & Liberman, 2008; Yuan & Liberman, 2009). The aligned phonemic transcriptions typically reflect careful speech productions and do not include the vast amount of reduction that occurs in spontaneous speech (Warner, 2019). It is therefore left to researchers to apply additional methods to determine, among other things, whether stops are realized with closure, syllables are omitted, and vowels undergo coalescence. A possible future state in phonetics is one where we

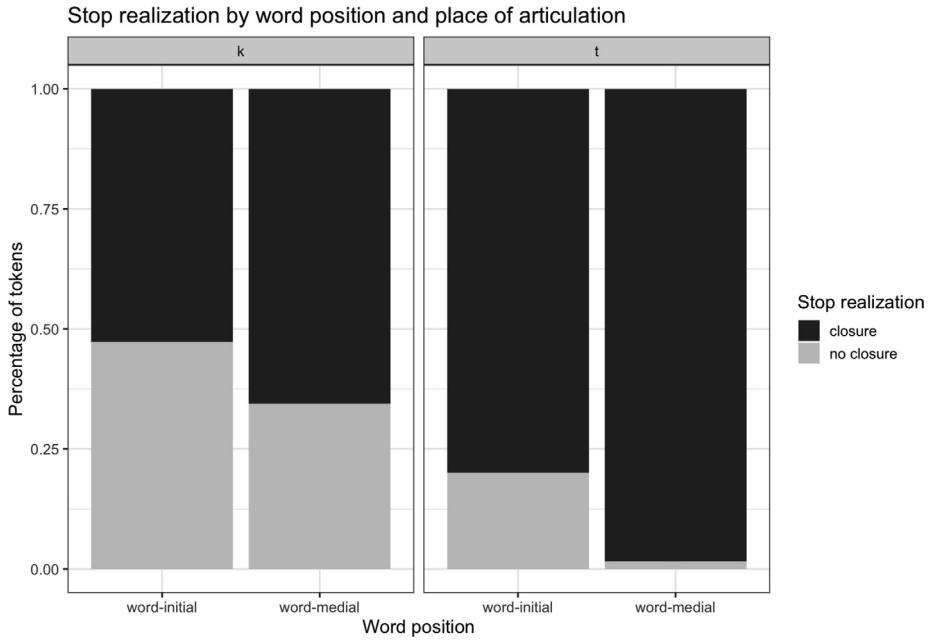

Fig. 9. Realization of stop consonants /t̲, k/ in the Yoloxóchitl Mixtec corpus.

are able to not only force align a corpus with high accuracy but also to predict surface phonetic variants and provide additional annotations which include this information.

4.1. Methods

In the third experiment, we trained three deep neural network models (DNN) to predict 2-way, 3-way and 4-way allophonic realizations from acoustic signals, separately for each of the Mixtec stops investigated in Experiment 2. We used the categorization provided in Experiment 2 in the modeling process. The 2-way models only predict the distinction between stops and non-stops. The 3-way models further distinguishes obstruent (i.e., 'fricative') and sonorant (i.e., 'nasal' for /t̲/ and 'approximant' for /k/) in the non-stop category. The 4-way models expands further, separating voiced and voiceless stops. Table 1 summarizes the allophonic categories in each DNN model.

The acoustic signals in each instance of the stop segments were extracted and tapered by a 10 ms Hanning-window with 2 ms stepping. For each window, 20 Mel-frequency cepstral coefficients (MFCCs) were calculated, standardized in z-scores and re-scaled between 0 and 1. These MFCCs across all windows in a segment were then time-normalized (one-dimensional cubic spline interpolation) into 20 frames, yielding a grid of 20 x 20 standardized MFCC features for each segment. Such a feature extraction routine ignores the temporal information, which is also an important cue for allophonic variation. So, we added the duration of the segment as an additional feature, resulting in 401 input features for each segment. The architecture of each DNN that we selected was a four-layer Deep Belief Network with two fully-connected hidden layers (250 neurons in each layer) and a Softmax layer, using a sigmoid activation function. Further deeper architectures were tested and did not improve the models. The dataset was randomly separated into 80% as the training set, 10% as the cross-validation set, and the remaining 10% as the test set. The number of samples in the three sets for each of the models are presented in Table 2. The training set further underwent a resampling algorithm (ADASYN –"adaptive synthetic sampling approach for imbalanced learning" (He, Bai, Garcia, & Li, 2008)) to balance the number of samples in each category. L2 regularizations were added to all weights to penalize over-fitting. Cross-entropy loss was chosen as the loss function. The hyper-parameters were fine-tuned with the cross-validation set. The optimized hyper-parameters were the sparsity parameter (set to 0.1), the weight decay parameter (set to $3x10^{-3}$), and the weight of sparsity penalty term (set to 3). The accuracies of the trained models were calculated as the number of correctly predicted categories over the total number of samples in the untrained test set.

Table 1
Summary of contrasts in 2-way, 3-way, and 4-way DNN models for/t/ and/k/.

	DNN models		
	2-way	3-way	4-way
/t̲/	stop vs. non-stop	stop vs. fricative vs. nasal	-voi stop vs. +voi stop vs. fricative vs. nasal
/k/	stop vs. non-stop	stop vs. fricative vs. approximant	-voi stop vs. +voi stop vs. fricative vs. approximant

4.2. Results

Fig. 10 shows the results of the DNN models on categorizing surface stop allophones. The two-way models distinguishing between variants with and without stop closure performed with high accuracy (98% for /t̲/ and 95% for/k/). For the 3-way models, we chose the most frequent allophones varying in manner of articulation for each stop. For /t̲/, this included stop, fricative, and nasal allophones. Stops and nasal realizations were predicted with high accuracy (96% and 91%, respectively), but the fricative allophone was frequently confused with the nasal allophone. One possibility for this is that short dura-

Table 2
Number of samples in the training, cross-validation and test sets for the DNN models.

Target	Model	Training	Cross-valid.	Test	Total
/t/	2-way	713	89	89	891
/t/	3-way	1106	139	138	1383
/t/	4-way	646	81	81	808
/k/	2-way	1029	129	128	1286
/k/	3-way	1023	128	128	1279
/k/	4-way	1411	177	176	1764

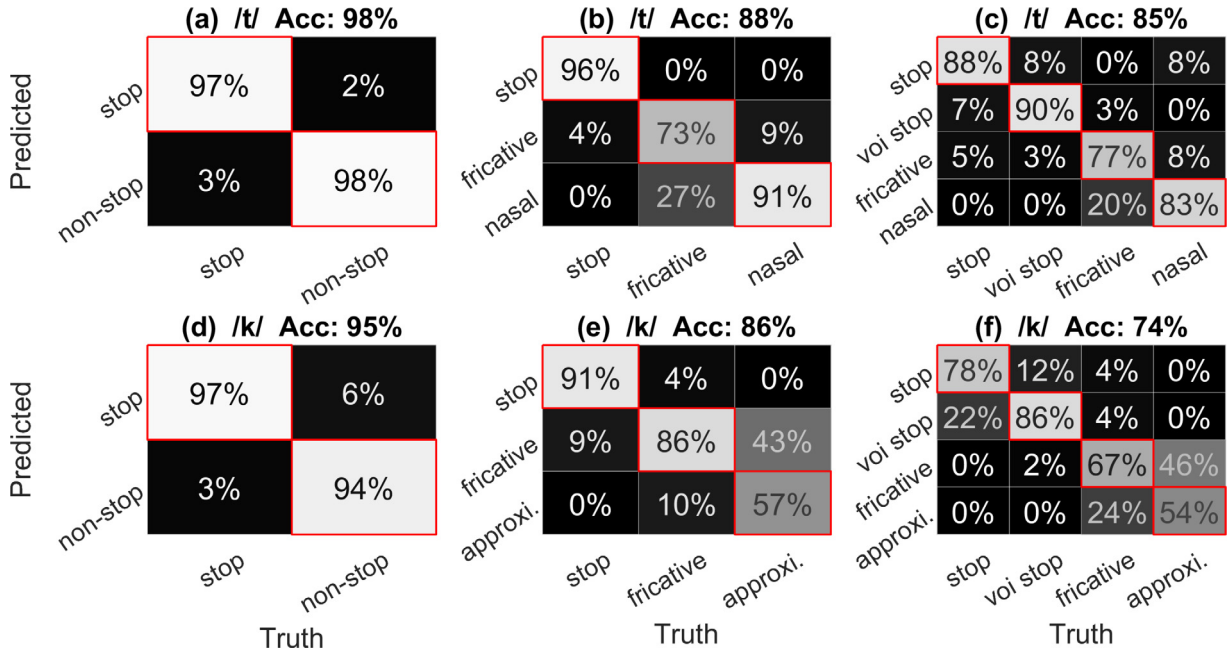

Fig. 10. Predicted allophones for /t̪/ (top row) and /k/ (bottom row) by DNN models with 2-way (left), 3-way (middle) and 4-way (right) contrasts in a confusion matrix. 'Truth' indicates the transcriber's original categorization, while 'predicted' indicates the categorization predicted by the model. Along the main diagonal, darker cells indicate lower accuracy in prediction and brighter cells indicate higher accuracy. Outside of the main diagonal, darker cells indicate greater rejection accuracy. Predictions shown here were based on the held-off test sets only.

tion [ð] and [n] lack strong distinguishing acoustic cues relative to the stop-fricative distinction (see below). For /k/, the three most common allophones differing in manner of articulation were stops, fricatives, and approximants. The stop and fricative categories were predicted with reasonably good accuracy (91% and 86%, respectively), but the approximant category was often identified by the model as the fricative category (43% of cases). The most probable cause for this was that the distinction between fricative and approximant realizations of [ɣ] and [ɣ̞] reflects a continuum that is very difficult to distinguish perceptually.

In both four-way models, we included the distinction between voiced and voiceless stops. For the /t̪/ model, the voiceless stop, voiced stop, and nasal allophones were predicted with reasonably good accuracy (88%, 90%, and 83%, respectively), and the (voiced) fricative allophone was predicted with 77% accuracy. Interestingly, the model showed improved accuracy in distinguishing between reduced nasal and fricative allophones (73 vs. 27% in the 3-way model, but 77 vs. 20% in the 4-way model). This suggests that including a voiced stop category forced the model to exclude those features most closely related to voicing here. This might have the consequence of forcing the model to rely more on spectral features related to nasal resonance. The voiced and voiceless stops were confused with each other about 7∼8% of time.

For the 4-way/k/ model, the voiceless stop and voiced stop allophones were predicted with reasonably good accuracy (78%, and 86%, respectively). Though, the distinction between the reduced fricative-approximant realizations remained an issue as both were still not well-predicted by this model (67%, and 54%, respectively). As the number of categories were increased, the overall accuracy of each DNN decreased.

5. Discussion

5.1. Patterns of lenition

The results from experiment 1 show that voicing lenition is more common in non-final syllables in Yoloxóchitl Mixtec than in consonants in the onsets of stem-final, stressed syllables. Stem-final syllables are longer than non-final syllables and much of the durational difference between non-final and final syllables occurs within the consonant onset, not the vowel. The durational findings here replicate results from three previous studies on Yoloxóchitl Mixtec speech (DiCanio et al., 2018; DiCanio et al., 2020; DiCanio et al., 2021). Overall, duration is

a strong predictor for voicing during closure and this factor also accounts for the observation that obstruents in function words in the language undergo passive voicing in spontaneous speech more often than content words. The former are uniformly shorter in duration than the latter, a finding in agreement with work on English (Gahl et al., 2012).

The relationship between duration and voicing observed here closely parallels findings in Kakadelis (2018) on Arapaho, Bardi, and Siera Norte del Puebla Nahuatl. Each of these languages lacks a phonological voicing contrasts among obstruents. In languages with a voicing contrast, it seems that the degree of passive voicing is both less pervasive and less sensitive to prosodic position. For instance, in English, Davidson (2018) finds that voiceless stops tend to undergo rather little passive voicing. One explanation for these differences here is that voicing is more likely to be used to mark prosodically weak environments in languages lacking a phonological voicing contrast than in languages possessing a contrast. However, more research on the latter type of language is clearly needed to test this hypothesis.[7]

In terms of lenition patterning, the findings in experiment 2 closely match those in experiment 1. Like voicing lenition, spirantization is also more frequent in non-final syllables than in final, stressed syllables. This patterning is not too surprising since there is evidence that pre-tonic onsets are strengthened in the literature and final syllables are also stressed in Yoloxóchitl Mixtec. Similar findings regarding lenition and stress are observed in Bouavichith and Davidson (2013) for English, Lavoie (2001) for Mexican Spanish, and Lewis (2001) for Central Columbian and Bilbao Spanish.

The results here argue that stress is the primary determinant of patterns of lenition in Yoloxóchitl Mixtec and also demonstrate no differences in consonant production among non-final syllable onsets (word-initial vs. word-medial). Word-initial syllable onsets are not actively weakened relative to word-medial (non-pre-tonic) syllable onsets. If it appears that word-initial position is frankly "unremarkable" in the Yoloxóchitl Mixtec data, it begs the question of why word-initial strengthening might appear to be so robust based in past literature on more well-studied languages. Consider that stem-final stress is fairly common in the world's languages. Depending on the typological survey one cites, among languages with phonologically-predictable stress, between 18% - 31% have stem-final stress (Gordon, 2016). Yet, to our knowledge, there is little to no work on the positional effects of lenition among languages of this type. Though we believe that past phonetic research on initial strengthening is convincing and robust, it may also be unintentionally typologically biased against languages with stem-final stress.

Morphological structure is another important structural consideration that might explain the unique findings in Yoloxóchitl Mixtec. Word-initial strengthening has been found in English, French, Italian, Hungarian, and Korean, all of which are primarily suffixing languages. The other languages where word-initial strengthening has been observed are either isolating (Taiwanese) or primarily suffixing with some prefixation on verbs (Bardi) (Bowern, 2012). Yoloxóchitl Mixtec has only prefixal inflectional morphology on verbs, historical derivational prefixes on nouns, and pronominal enclitics which can freely apply to most parts of speech (García, 2007). This patterning is common among Otomanguean languages – it is found in Triqui (DiCanio, 2020; DiCanio, 2016), in Zapotecan languages (Beam de Azcona, 2004; Campbell, 2014), and other Mixtecan languages (Macaulay, 1996).

If the goal of initial strengthening is to ensure reliable cues to word segmentation (Katz & Fricke, 2018; White et al., 2020), then it seems particularly important that speakers should ensure clear acoustic/articulatory cues in the initial portion of words which happens to be co-extensive with a lexical stem, i.e. in most of the languages where word-initial strengthening has been found. However, in languages where the initial portion of the word does not immediately inform listeners of the stem's identity, it is less important for speakers to strengthen it in the natural context of speech communication. In the process of speech production, Yoloxóchitl Mixtec speakers lengthen stem-final syllables because they carry phonological information that is more important to the stem's identity than non-final syllables do (see DiCanio et al. (2018) and Section 1.2 for a discussion of phonological distributional asymmetries with respect to stress). Though there are prononimal enclitics, many are vocalic and involve final vowel replacement, e.g. /ju^{3}-βa^{4}/ 'father', /ju^{3}βa^{42}/ 'my father', /ju^{3}βõ^{4}/ 'your father', etc (García, 2007; Palancar et al., 2016). The stem-final syllable is therefore often also the locus of additional morphological information. Grosso modo, it would appear to be effective to lengthen final syllables for word-level parsing in Yoloxócitl Mixtec.

Word-initial syllables in polysyllabic words may be less informative because this position is less likely to be co-extensive with either the lexical stem or the stressed syllable. The empirical data presented here demonstrate that word-initial strengthening is not a universal in speech production, contra White et al. (2020). If the hypothesis regarding morphological headedness is verified, it raises the possibility that word-initial strengthening is structure-dependent. This possibility has potential repercussions for models of speech recognition which rely primarily on word onset identification (Norris & McQueen, 2008).

5.2. Modeling lenition

The patterns of manner lenition observed in Experiment 2 were modeled using deep neural networks in Experiment 3. The results from this work demonstrate that it is possible to use limited training data to create a fairly accurate classifier for surface phonetic variants of stop phonemes. The model's accuracy in identifying variants with complete and incomplete closure was high. Though, its accuracy in distinguishing among variants differing in degree of constriction (fricative vs. approximant) was lower. However, we note that human rater accuracy in distinguishing between voiced fricative and voiced approximant allophones may also not be very high. Indeed, among the two human categorizers in this study, inter-rater agreement was lowest between these allophones. Poor categorization among humans is matched by less accurate prediction from the DNN model, suggesting that this dis-

[7] We exclude for the moment analyses of languages like Spanish which, from a phonetic perspective mainly contrast voiced frictionless continuants with voiceless unaspirated stops. Voiced stops in most Spanish varieties are limited to utterance-initial or post-nasal position.

tinction may be altogether unreliable in the spontaneous speech data we examined.

The DNN models performed reasonably well in categorizing voiced vs. voiceless stop allophones (78–90%). Interestingly, voiced stop allophones were slightly better categorized than voiceless stops (86–90% vs. 78–88%), possibly due to the fact that stops with some degree of passive voicing were included in the voiceless stop category for the 4-way DNNs. This has the effect of making the voiceless stop category more phonetically heterogeneous. One future possibility would be to include those partially-voiced stops in their own category separate from fully voiceless or fully voiced stops. Just how much partial/passive voicing is required for inclusion in such a category is currently unclear though.

We believe that this modeling research has important practical consequences for phonetic science. There is a persistent annotation bottleneck for phonetic research in spontaneous speech (in endangered or non-endangered languages). As researchers seek to examine larger quantities of speech data, there is an increasing need to include useful speech annotations. Forced alignment will never provide sufficient allophonic detail of the speech signal since pronunciation dictionaries usually only include single, carefully-produced word transcriptions. Speech reduction is largely probabilistic and non-deterministic, so it is usually excluded from pronunciation dictionaries. Though, using a DNN classifier on some limited phonetic variant training data, it should be possible for researchers to predict surface allophones with greater accuracy. These allophones could then be included as additional annotation.

This surface phonetic information makes it possible to examine the extent to which speech reduction is found in running speech and whether such reduction is typical. For instance, among other things, developmental apraxia is marked by an individual's inability to create complete closure in stops (Davis, Jakielski, & Marquardt, 1998). Though incomplete stop closure is a normal part of speech reduction in many languages, including English (Bouavichith & Davidson, 2013; Warner & Tucker, 2011; Warner, 2019), it may be present in higher or lower percentages in clinical populations. Combined with a forced aligner, the DNN classifier for surface allophonic detail could make it possible to assess the degree of speech reduction in a spontaneous speech corpus for a variety of speech communities.

5.3. Study limitations and future directions

One limitation of the current study is the fact that higher-level prosodic groupings were not considered in the analysis of domain-level effects on strengthening. In practice, it is exceedingly difficult to use cues like F0, duration, or phonation type to indicate more subtle prosodic boundaries in Yoloxóchitl Mixtec. The nine lexical tones in the language are assigned to moras and, as a result, up to 20 possible tonal melodies may occur on a single monosyllabic root (DiCanio et al., 2014). In addition to this, there is a surface-level contrast in vowel length, where monosyllabic roots have obligatory long vowels and disyllabic roots have obligatory short vowels. Thus, a prefixed monosyllabic root might have the shape CV-CVV and contrast with a disyllabic root with the shape CVCV, e.g. /ko^{3}-^{n}da^{2}a^{2}/ 'to take care of' vs. /ka^{3}ʔ^{n}da^{2}/ 'to scratch.'[8]. Listeners need to parse stems in addition to inflected words in the process of speech perception and would resultingly rely on duration as a cue. Glottalization is also contrastive and orthogonal to the tonal contrasts in the language, e.g. /^{n}do^{1}ʔo^{4}/ 'basket' vs. /^{n}do^{1}o^{4}/ 'sugarcane.'.

There is evidence for stronger final syllable lengthening effects in pre-verbal focused constituents (DiCanio et al., 2018) and in utterance-final position when compared with utterance-medial position (DiCanio et al., 2021). In the former case, it is unclear whether the focused constituent is lengthened due to focus or due to positional effects on word production (left dislocation). In the latter case though, the effect is clearly positional where, in utterance-final position, the onset consonant lengthened slightly (34 ms lengthening) and the vowel lengthened more substantially (> 100 ms). However, it remains unclear how any other type of prosodic constituents smaller than the utterance contribute to patterns of lengthening and lenition at the present moment and; because of the functional load of various suprasegmentals in the language, we predict any additional prosodic cues in the phonetic signal to be quite weak.

The current study utilized a smaller corpus of hand-corrected speech from a language documentation corpus which now comprises over 200 h of transcribed and phonologically-transduced recordings from over 50 speakers (Amith & García, 2019; Amith & García, 2021). Future work on the phonetics of the language should be able to take advantage of a much larger data set than the 85.5 min we analyzed in the current paper consisting of more speakers. Moreover, recently-published work has utilized mediation analysis to determine more definitively how causal different factors are in explaining patterns of lenition in spontaneous speech data, i.e. what is the relative weighting of duration, prosodic position, information content in explaining the observed variation (Priva et al., 2020). Though the current study suggests the duration plays a central role in determining patterns of lenition in Yoloxóchitl Mixtec, newer statistical methods will allow us to verify the relative weight of this and other factors.

6. Conclusions

An examination of speech reduction in a corpus of Yoloxóchitl Mixtec spontaneous speech revealed that both stress and functional status influenced obstruent duration, a finding in line with previous work on speech reduction (Gahl et al., 2012; Kakadelis, 2018; Lavoie, 2001; Parrell & Narayanan, 2018). However, the language is unique in relation to past research on prosodic strengthening because it demonstrates a pattern of stress-induced strengthening and a pattern of relative weakening of all non-pre-tonic onset consonants (including those in word-initial position). All non-final onset obstruents in polysyllabic words, including those in word-initial position, were shorter relative to final, stress-initial onset obstruents. This finding is in agreement with recent research on Yoloxóchitl Mixtec speech production (DiCanio et al., 2018; DiCanio et al., 2020; DiCanio et al., 2021), but extends this work by also demonstrating that this durational shortening correlates with

[8] This is in fact a historical prefix related to the formation of irrealis stems and is non-productive in the language (Palancar et al., 2016)

increased voicing and manner lenition. As such, Yoloxóchitl Mixtec represents an exception to research examining the influence of prosodic boundaries on speech production since initial strengthening is absent. While we hypothesize on the causal factors responsible for this pattern in Yoloxóchitl Mixtec, more research on speech reduction within a variety of prosodically and morphologically distinct languages will be a fruitful avenue of future research.

The research here also demonstrates the utility of applying computational tools, like deep neural networks, to the analysis of speech reduction data. With limited training data, it is possible to use automatic categorization to improve the annotation of spontaneous speech data. This has the potential to improve the amont of fine-phonetic detail often missing from larger-scale phonetic corpora. Future research in this area is bound to improve upon the findings presented here. The ultimate outcome of this work will add value to archival recordings and spontaneous speech corpora for research on both endangered and non-endangered languages.

CRediT authorship contribution statement

Christian DiCanio: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Wei-Rong Chen:** Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Joshua Benn:** Investigation, Data curation. **Jonathan Amith:** Resources, Data curation. **Rey Castillo García:** Resources, Data curation, Funding acquisition.

Acknowledgments

This work was supported by NSF Grant 1603323 (DiCanio, PI) at the University at Buffalo and NIH Grant DC-002717 (Whalen, PI) at Haskins Laboratories. The corpus data from Yoloxóchitl Mixtec was supported by NSF Awards 0966462, 2123578, and 1761421 (Amith, PI) as well as ELDP Projects PPG0048 and MDP0201 (Amith, PI) at Gettysburg College.

References

Amith, J., & Castillo García, R. (2021). Audio corpus of Yoloxóchitl Mixtec with accompanying time-coded transcriptions in ELAN. In *Open Speech and Language Resources (OpenSLR)*. .

Amith, J. D., & Castillo García, R. (2019). *Documentation of Yoloxóchitl Mixtec (Glottocode: Yolo1241; ISO 639–3: xty): Corpus, Lexicon, and Grammar*. Philadelphia: Linguistic Data Consortium.

Babinski, S., Dockum, R., Goldenberg, D., Craft, J. H., Fergus, A., & Bowern, C. (2019). A Robin Hood approach to Forced Alignment: English-trained algorithms and their use on Australian languages. In *Proceedings of the* 93*rd Annual Meeting of the Linguistic Society of America. Linguistic Society of America*. .

Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech, 52*(4), 391–413.

Beam de Azcona, R. (2004). *A Coatlán-Loxicha Zapotec Grammar* PhD thesis. Berkeley: University of California.

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics, 49*(2), 259–284.

Boersma, P. and Weenink, D. (2016). Praat: doing phonetics by computer [computer program]. www.praat.org. .

Bouavichith, D., & Davidson, L. (2013). Segmental and prosodic effects on intervocalic voiced stop reduction in connected speech. *Phonetica, 70*, 182–206.

Bowern, C.L. (2012). A Grammar of Bardi, volume 57 of Mouton Grammar Library. Walter de Gruyter GmbH & Co. KG, Berlin/Boston. .

Browman, C., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology 1: between the grammar and the physics of speech* (pp. 341–376). Cambridge University Press.

Buerkner, P.-C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*(2), 149–180.

Byrd, D., & Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics, 24*, 263–282.

Campbell, E. W. (2014). *Aspects of the Phonology and Morphology of Zenzontepec Chatino, a Zapotecan Language of Oaxaca, Mexico* PhD thesis. University of Texas at Austin.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32.

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. .

Cheng, C., & Xu, Y. (2015). Mechanism of Disyllabic Tonal Reduction in Taiwan Mandarin. *Language and Speech, 58*(3), 281–314.

Cho, T., & Keating, P. (2001). Articulatory and acoustic studies of domain-initial strengthening in Korean. *Journal of Phonetics, 29*(2), 155–190.

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically-driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics, 35*, 210–243.

Cohen Priva, U., & Gleason, E. (2020). The causal structure of lenition: a case for the causal precedence of durational shortening. *Language, 96*(2), 413–448.

Dalby, J. (1984). *Phonetic structure of fast speech in American English* PhD thesis. Indiana University.

Davidson, L. (2018). Phonation and laryngeal specification in American English voiceless obstruents. *Journal of the International Phonetic Association, 48*(3), 331–356.

Davis, B. L., Jakielski, K. J., & Marquardt, T. P. (1998). Developmental apraxia of speech: Determiners of differential diagnosis. *Clinical Linguistics & Phonetics, 12*(1), 25–45.

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

DiCanio, C., Benn, J., & Castillo García, R. (2018). The phonetics of information structure in Yoloxóchitl Mixtec. *Journal of Phonetics, 68*, 50–68.

DiCanio, C., Benn, J., & Castillo García, R. (2021). Disentangling the effects of position and utterance-level declination on the production of complex tones in Yoloxóchitl Mixtec. *Language and Speech, 64*(3), 515–557.

DiCanio, C., Nam, H., Amith, J. D., Castillo García, R., & Whalen, D. H. (2015). Vowel variability in elicited versus spontaneous speech: evidence from Mixtec. *Journal of Phonetics, 48*, 45–59.

DiCanio, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., & 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., & Whalen, D. H. (2015). The interaction of vowel length and speech style in an arapaho speech corpus. In *Proceedings of the* 18*th International Congress of the Phonetic Sciences Glasgow, Scotland* (pp. 513–517).

DiCanio, C., Zhang, C., Whalen, D. H., & Castillo García, R. (2020). Phonetic structure in Yoloxóchitl Mixtec consonants. *Journal of the International Phonetic Association, 50*(3), 333–365.

DiCanio, C. T. (2012). The Phonetics of Fortis and Lenis Consonants in Itunyoso Trique. *International Journal of American Linguistics, 78*(2), 239–272.

DiCanio, C. T. (2016). Abstract and concrete tonal classes in Itunyoso Trique person morphology. In E. Palancar & J.-L. Léonard (Eds.), *Tone and Inflection: New Facts and New Perspectives, volume 296 of Trends in Linguistics Studies and Monographs, chapter 10* (pp. 225–266). Mouton de Gruyter.

DiCanio, C.T. (2020). Aspecto verbal en triqui de Itunyoso. In Swanton, M., San Giacomo Trinidad, M., and Hernández Mendoza, F., editors, Estudios fonológicos de idiomas mixtecanas. Universidad Nacional Autónoma de México. .

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*(6), 3728–3740.

Gahl, S. (2008). "Time" and "thyme" are not homophones: Word durations in spontaneous speech. *Language, 84*, 474–496.

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

Gordon, M. K. (2016). *Phonological Typology*. Oxford University Press.

Gurevich, N. (2004). Lenition and contrast: the functional consequences of certain phonetically conditioned sound changes. New York: Garland, Outstanding Dissertations in Linguistics. .

Gurevich, N. (2011). Lenition. In M. van Oostendorp, C. J. Ewen, E. V. Hume, & K. Rice (Eds.), *Blackwell Companion in Phonology, volume 3, chapter 66* (pp. 1559–1575). Malden, MA: Wiley-Blackwell.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *In 2008 IEEE International Joint Conference on Neural Networks. IEEE World Congress on Computational Intelligence*. .

Hualde, J., Simonet, M., & Nadeu, M. (2011). Consonant lenition and phonological recategorization. *Laboratory Phonology, 2*(2), 301–329.

Hualde, J. I., & Nadeu, M. (2011). Lenition and phonemic overlap in Rome Italian. *Phonetica, 68*, 215–242.

Jansen, W. (2004). *Laryngeal Contrast and Phonetic Voicing: A Laboratory Phonology Approach to English, Hungarian, and Dutch* PhD thesis . Groningen University.

Josserand, J. K. (1983). *Mixtec Dialect History* PhD thesis . Tulane University.

Jun, S.-A. (1995). Asymmetrical prosodic effects on the laryngeal gesture in Korean. In B. Connell & A. Arvaniti (Eds.), *Phonology and Phonetic Evidence: Papers in laboratory phonology 4, chapter 17* (pp. 235–253). Cambridge University Press.
Kakadelis, S. M. (2018). *Phonetic Properties of Oral Stops in Three Languages with No Voicing Distinction* PhD thesis . Graduate Center, City University of New York.
Kasi, K., & Zahorian, S. A. (2002). In *Yet another algorithm for pitch tracking In Proceedings of ICASSP02* (pp. 361–364). Orlando.
Katz, J. (2016). Lenition, perception, and neutralisation. *Phonology, 33*, 43–85.
Katz, J., & Fricke, M. (2018). Auditory disruption improves word segmentation: A functional basis for lenition phenomena. *Glossa, 3*(1), 1–25.
Katz, J., & Pitzanti, G. (2019). The phonetics and phonology of lenition: A Campidanese Sardinian case study. *Journal of Laboratory Phonology, 10*(1), 1–40.
Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2003). Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in Laboratory Phonology VI* (pp. 145–163). Cambridge, UK: Cambridge University Press.
Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2004). Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology 6* (pp. 145–163). Cambridge University Press.
Keyser, S. J., & Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language, 82*(1), 33–63.
Kingston, J., Diehl, R. L., Kirk, C. J., & Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: the [voice] contrast. *Journal of Phonetics, 36*, 28–54.
Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication, 22*, 43–65.
Ladd, D., & Scobbie, J. (2003). External sandhi as gestural overlap? counterevidence from Sardinian. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in Laboratory Phonology VI* (pp. 164–182). Cambridge, UK: Cambridge University Press.
Lavoie, L. M. (2001). *Consonant Strength: Phonological Patterns and Phonetic Manifestations. Outstanding Dissertations in Linguistics*. Garland Publishing Inc..
Lewis, A. M. (2001). *Weakening of intervocalic/ptk/ in two Spanish dialects: Toward the quantification of lenition processes* PhD thesis . University of Illinois at Urbana-Champaign.
Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.). (2015). *Ethnologue: Languages of the World* (Eighteenth edition. Dallas, Texas: SIL International. Available online at http://www.ethnologue.com/.
Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In W. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Dordrecht: Kluwer.
Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language, 33*, 42–49.
Lisker, L. (1986). Voicing in English: a catalogue of acoustic features signaling/b/ versus/p/ in trochees. *Language and Speech, 29*(3), 3–11.
Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20*, 384–422.
Liu, F., & Kong, Y. (2015). zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression. *The R Journal, 7*(2), 34–51.
Macaulay, M. (1996). A Grammar of Chalcatongo Mixtec, volume 127 of University of California Publications in Linguistics. University of California Press. .
Moon, S.-J., & 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.
Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357–395.
Palancar, E. L., Amith, J. D., & Castillo García, R. (2016). Verbal inflection in Yoloxóchitl Mixtec. In E. L. Palancar & J.-L. Léonard (Eds.), *Tone and Inflection: New Facts and New Perspectives, chapter 12* (pp. 295–336). Mouton de Gruyter.
Parrell, B. (2014). *Dynamics of consonant reduction* PhD thesis . University of Southern California.
Parrell, B., & Narayanan, S. (2018). Explaining coronal reduction: Prosodic structure and articulatory posture. *Phonetica, 75*, 151–181.
Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language, 87*(3), 539–558.
Pitt, M. A., Johnson, K., Hume, E. V., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication, 45*(1), 89–95.
R Development Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 3.6.3 edition. Version 3.6.3, retrieved from http://www.R-project.org/.
Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.
Schwarz, M., Sonderegger, M., & Goad, H. (2019). Realization and representation of Nepali laryngeal contrasts: Voiced aspirates and laryngeal realism. *Journal of Phonetics, 73*, 113–127.
Shaw, J., Carignan, C., Agostini, T. G., Mailhammer, R., Harvey, M., & Derrick, D. (2020). Phonological contrast and phonetic variation: The case of velars in Iwaidja. *Language, 96*(3), 576–617.
Shih, C., Möbius, B., & Narasimhan, B. (1999). Contextual effects on consonant voicing profiles: A cross-linguistic study. In *Proceedings of the* 14th *International Congress of the Phonetic Sciences, San Francisco* (pp. 989–992).
Stevens, K. N. (2000). *Acoustic Phonetics* first edition. MIT Press.
Stevens, M. & Hajek, J. (2004). Comparing voiced and voiceless geminates in Sienese Italian: what role does preaspiration play? In Proceedings of the 10th Australian International Conference on Speech Science & Technology, pages 340–345. Macquarie University, Sydney, Australian Speech Science & Technology Association Inc.
Tang, K. & Bennett, R. (2019). Unite and conquer: bootstrapping forced alignment tools for closely-related minority languages Mayan. In Calhoun, S., Escudero, P., Tabain, M., and Warren, P., editors, Proceedings of the International Congress of Phonetic Sciences (ICPhS) 2019, pages 3584–3552. Canberra, Australia: ASSTA.
Torreira, F., & Ernestus, M. (2011). Realization of voiceless stops and vowels in conversational French and Spanish. *Journal of Laboratory Phonology, 2*, 331–353.
Torreira, F., & Ernestus, M. (2012). Weakening of intervocalic/s/ in the Nijmegen Corpus of Casual Spanish. *Phonetica, 69*, 124–148.
Ussishkin, A., Warner, N., Clayton, I., Brenner, D., Carnie, A., Hammond, M., & Fisher, M. (2017). Lexical representation and processing of word-initial morphological alternations: Scottish Gaelic mutation. *Journal of Laboratory Phonology, 8*(1), 1–34.
Verhoeven, J., De Pauw, G., & 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.
Warner, N. (2019). Reduced speech: All is variability. WIREs. *Cognitive Science, e1496*, 1–7.
Warner, N., & Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *Journal of the Acoustical Society of America, 130*(3), 1606–1617.
Westbury, J. R. (1983). Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *Journal of the Acoustical Society of America, 73*, 1322–1336.
Westbury, J. R., & Keating, P. A. (1986). On the naturalness of stop consonant voicing. *Journal of Linguistics, 22*, 145–166.
White, L., Benavides-Varela, S., & Mády, K. (2020). Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *Journal of Phonetics, 81*, 1–14.
White, L., Mattys, S. L., Stefansdottir, L., & Jones, V. (2015). Beating the bounds: Localized timing cues to word segmentation. *Journal of the Acoustical Society of America, 138*(2), 1214–1220.
Wittenburg, P., Brugman, H., Russel, A., Klassman, A., and Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, http://tla.mpi.nl/tools/tla-tools/elan/.
Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In Proceedings of Acoustics - 2008.
Yuan, J., & Liberman, M. (2009). Investigating/l/ variation in English through forced alignment. In *Interspeech - 2009* (pp. 2215–2218).
Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *Journal of the Acoustical Society of America, 123*(6), 4559–4571.