Abstract

Temporal Articulatory Stability, Phonological Variation, and Lexical Contrast Preservation in Diaspora Tibetan

Christopher Alden Geissler

2021

This dissertation examines how lexical tone can be represented with articulatory gestures, and the ways a gestural perspective can inform synchronic and diachronic analysis of the phonology and phonetics of a language. Tibetan is chosen as an example of a language with interacting laryngeal and tonal phonology, a history of tonogenesis and dialect diversification, and recent contact-induced realignment of the tonal and consonantal systems. Despite variation in voice onset time (VOT) and presence/absence of the lexical tone contrast, speakers retain a consistent relative timing of consonant and vowel gestures.

Recent research has attempted to integrate tone into the framework of Articulatory Phonology through the addition of tone gestures. Unlike other theories of phonetics-phonology, Articulatory Phonology uniquely incorporates relative timing as a key parameter. This allows the system to represent contrasts instantiated not just in the presence or absence of gestures, but also in how gestures are timed with each other. Building on the different predictions of various timing relations, along with the historical developments in the language, hypotheses are generated and tested with acoustic and articulatory experiments.

Following an overview of relevant theory, the second chapter surveys past literature on the history of sound change and present phonological diversity of Tibetic dialects. Whereas Old Tibetan lacked lexical tone, contrasted voiced and voiceless obstruents, and exhibited complex clusters, a series of overlapping sound changes have led to some modern varieties that have tone, lack clusters,

and vary in the expression of voicing and aspiration. Furthermore, speakers in the Tibetan diaspora use a variety that has grown out of the contact between diverse Tibetic dialects. The state of the language and the dynamics of diaspora have created a situation ripe for sound change, including the recombination of elements from different dialects and, potentially, the loss of tone contrasts.

The nature of the diaspora Tibetan is investigated through an acoustic corpus study. Recordings made in Kathmandu, Nepal, are being transcribed and forced-aligned into a useful audio corpus. Speakers in the corpus come from diverse backgrounds across and outside traditional Tibetan-speaking regions, but the analysis presented here focuses on speakers who grew up in diaspora, with a mixed input of Standard Tibetan (*spyi skad*) and other Tibetan varieties. Especially notable among these speakers is the high variability of voice onset time (VOT) and its interaction with tone. An analysis of this data in terms of the relative timing of oral, laryngeal, and tone gestures leads to the generation of hypotheses for testing using articulatory data.

The articulatory study is conducted using electromagnetic articulography (EMA), and six Tibetan-speaking participants. The key finding is that the relative timing of consonant and vowel gestures is consistent across phonological categories and across speakers who do and do not contrast tone. This result leads to the conclusion that the relative timing of speech gestures is conserved and acquired independently. Speakers acquire and generalize a limited inventory of timing patterns, and can use timing patterns even when the conditioning environment for the development of those patterns, namely tone, has been lost.

Temporal Articulatory Stability, Phonological Variation, and Lexical Contrast Preservation in Diaspora Tibetan

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Christopher Alden Geissler

Dissertation Director: Jason Anthony Shaw

June 2021

Table of Contents

3 Corpus study

3.1 Introduction

3.1.1 Toward phonetic predictions

The descriptions of Tibetan surveyed in Chapter 2 present a language with unique opportunities for the study of the interaction between tone, aspiration, and voicing. The approach adopted here uses phonetic data to make inferences about the representation of consonants and tones in Tibetan. In this chapter, a corpus of acoustic recordings are analyzed in terms of the primary phonetic correlates of aspiration and tone: voice onset time (VOT) and fundamental frequency (F0). The results motivate a constrained set of hypotheses regarding the articulatory timing of Tibetan consonant and vowel gestures, which are tested in Chapter 4.

3.1.1 Aspiration and VOT

Languages tend to exhibit either two, three, or four oral stop contrasts that make use of voicing and aspiration (Lisker & Abramson 1964). Four-contrast languages such as Hindi and Marathi make use of phonetically voiced, aspirated, unaspirated, and voiced-aspirated stops, while three-contrast languages like Thai and Eastern Armenian use phonetically voiced, unaspirated, and aspirated stops. However, two-category languages can either exhibit a contrast that is primarily voiced/voiceless as in Dutch and Spanish, or unaspirated/aspirated as in English and Cantonese (Lisker & Abramson 1964). These typologies are commonly explained through the co-occurence of two features, as schematized in Fig. 3.1, and this characterization will be adopted throughout this chapter. The use of separate [SG] and [voice] features follows the “laryngeal realist” approach (e.g. Halle & Stevens 1971, Lombardi 1991/1994, Iverson & Salmons 1995, Honeybone 2005). This approach uses distinctive features corresponding to stop categories, particularly [voice] for languages with a two-way voiced/voiceless contrast and [SG] for languages with a two-way aspirated/unaspirated contrast, rather than using different phonetic realizations of a single phonological feature. The laryngeal realist approach was adopted here because it can account for three- and four-category systems using only binary features (Schwartz et al 2019).

Two-way contrast (English)

		[+voice]	[-voice]
	[+SG]		aspirated
(a)	[-SG]		plain

Two-way contrast (Dutch)

		[+voice]	[-voice]
	[+SG]		
(b)	[-SG]	prevoiced	plain

Three-way contrast (Thai)

		[+voice]	[-voice]
	[+SG]		aspirated
(c)	[-SG]	prevoiced	plain

Four-way contrast (Hindi)

		[+voice]	[-voice]
	[+SG]	breathy	aspirated
(d)	[-SG]	prevoiced	plain

Figure 3.1. [Spread Glottis] and [Voice] features for four types of contrast systems, following a “laryngeal realist” perspective.

Of course, the four typologies presented in Fig. 3.1 do not include all the stops in the worlds’ languages. Other series, such as implosives and ejectives, require additional features to instantiate the contrasts. Proposals include the articulatory laryngeal features of Halle and Stevens (1971) and Gallagher's (2011) [long VOT] feature that groups aspirated and ejective stops. The latter is grounded in acoustics rather than articulation, and motivated by the phonological patterns of Quechua. The tension between articulatory, acoustic, and abstract aspects of phonological representation has remained a consistent theme throughout the history of research on laryngeal phonology (e.g. Chomsky & Halle 1968, Keating 1984, Lombardi 1991, Iverson & Salmons 1995).

Phonetically, VOT has been used as a primary acoustic correlate of these contrasts, but languages instantiate the same phonological contrasts using very different VOT values (Lisker & Abramson 1964, Cho & Ladefoged 1999, Abramson & Whalen 2017, Hussain 2018). However, (Central) Tibetan is unique in that it has been described with three degrees of positive VOT length—aspirated, unaspirated, and an intermediate value—without ejective or breathy-voiced series (e.g. Denwood 1999, Tournadre & Dorje 2003, Tsering 2011). The research presented in this chapter provides the first acoustic measurements to quantify this claim.

3.1.2 Tone, F0, and VOT

Tone, and its primary phonetic exponent of F0, is closely related to consonantal laryngeal contrasts and VOT. Both F0 and VOT rely on the larynx for articulation, both are frequently involved as secondary cues for the other, and both can be reanalyzed as the other diachronically.

From an articulatory perspective, VOT and F0 are naturally related through sharing articulation at the glottis, and extensive research has investigated the effects of phonological voicing/aspiration and phonetic VOT on

F0. Broadly speaking, two major hypotheses have been proposed: vocal fold tension and aerodynamics. According to the vocal fold tension hypothesis (e.g. Halle & Stevens 1971, Ohde 1984), the vocal folds are slackened to facilitate voicing and stiffened to inhibit voicing; stiffer vocal folds vibrate faster, causing a higher F0. According to the aerodynamic hypothesis, F0 is correlated with the rate of airflow across the glottis, so F0 would rise near high-airflow productions such as aspirated stops, and lower near low-airflow productions such as voiced stops.

From an acoustic perspective, automatically-arising phonetic correlates can serve as cues to phonological contrast, as in the case of lowered F0 with voiced stops. However, Kingston and Diehl (1994) argue that speakers, aware of these associations, can marshal secondary cues to support a contrast. Taken further, this could lead to a later generation reanalyzing the cues, leading to tonogenesis, as surveyed in Sections 1.3.1 and 2.5.

The kind of precise modulation of phonetic parameters as discussed by Kingston and Diehl (1994), however, relies on speakers adjusting the details of highly-practiced articulations. In contrast, the form of enhancement discussed in Quantal Theory (Keyser & Stevens 2006, Stevens & Keyser 2010) explicitly involves the addition of an articulatory gesture. The role of enhancement in the present study is discussed in Section 3.4.1, and Section 3.4.2 analyzes the patterns in terms of gestures.

3.2 Corpus Methods

3.2.1 Participants

Data was collected from nineteen native speakers of Tibetan living in Kathmandu, Nepal, as part of a larger study of Tibetan in diaspora. Recruitment took place through social networks of Tibetans known to the author. Sixteen were born in Nepal, and three were born in the Tibet Autonomous Region (U-Tsang dialect regions: Lhasa and Kyirong) but came to Nepal as children. Eight were women and eleven were men; age ranged from 21-33 years (median 22; mean 23.8). All spoke at least some Nepali and many were fully bilingual; all also reported knowing at least some English, and several also had some knowledge of Chinese, Hindi, or another language.

3.2.2 Procedures

Interviews took place in a range of locations according to the comfort and availability of the speaker. Locations included speaker's homes, monasteries, a school, and a spare room in a Tibetan-owned apartment building. All interviews were conducted with both the author and one of two native-speaker interviewers present. As the author is not a native speaker, the interviewer was primarily responsible for interacting with the speaker, in order to facilitate communication, maximize speaker comfort, and minimize foreigner-talk. Recordings were made on a Zoom H4N recorder at 48 kHz sampling rate with an Audio-Technica ATM73a headset microphone; an Audio-Technica AT2020 microphone on a small tripod was also present, but only recordings from the ATM73a headset microphone were analyzed.

Interviews were conducted according to a standard sociolinguistic interview format, with tasks proceeding from more- to less-structured. Following basic demographic questions, the items used in this study appeared in a wordlist presented in the Tibetan orthography, for which speakers were asked to repeat each item twice. Following the wordlist, the interview continued with a choice task involving light verbs, a short reading passage, storyboard elicitations, and free speech/narration. Twenty-two words from the 64-item wordlist were used in this study. The order of items was randomized once and the same order used for all speakers.

3.2.3 Stimuli

The twenty-two items used in this study included examples of all combinations of register tone values (high and low) and word-initial aspiration values (aspirated and unaspirated). All attested places of articulation for stops were represented (bilabial, dental, retroflex, palatal, and velar), though unbalanced and in combination with varying vowels, tones, and aspiration categories, and vowels.

Taken together, 15 items were aspirated and 7 unaspirated, while 9 were high-tone and 13 were low-tone. Table 3.1 shows the distribution of these items across tone and aspiration categories: 7 items were low-tone and aspirated, 6 items were low-tone and unaspirated, 8 items were high-tone and aspirated, but only one item was high-tone and unaspirated. This distribution is depicted in Table 3.1, below.

	Aspirated	Unaspirated	Total
High-tone	8	1	9
Low-tone	7	6	13
Total	15	7	

Table 3.1. Items of interest by aspiration and voicing.

Each place of articulation was represented by four or five items, though the full four-way aspiration/tone contrast was only represented in the dental series, which comprised two minimal pairs for aspiration: the high-tone pair /tá.mák/ *rta dmag* ‘cavalry’ and /t^{h}á.mák/ *tha mag* ‘cigarette,’ and the low-tone pair /tǒm/ *dom* ‘bear’ and /dǒm/ *sdom* ‘spider.’ Further detail about the distribution of items is depicted in Table 3.2, below.

	Bilabial	Dental	Retroflex	Palatal	Velar	Total
High-tone, Aspirated	1	1	0	2	4	8
High-tone, Unspirated	0	1	0	0	0	1
Low-tone, Aspirated	3	2	2	0	0	7
Low-tone, Unaspirated	0	1	2	3	0	6
Total	4	5	4	5	4	

Table 3.2. Items of interest by tone/aspiration and place of articulation of initial consonant.

3.2.4 Data analysis

Measurements were taken using Praat (Boersma and Weeninck 2011). VOT values were measured from hand-selected TextGrid intervals, calculated by subtracting the time index of the first appearance of a release burst from of the beginning of periodicity in the waveform. Most VOT values were positive, though some negative values indicating pre-voicing were observed (see Fig. 3.2, below).

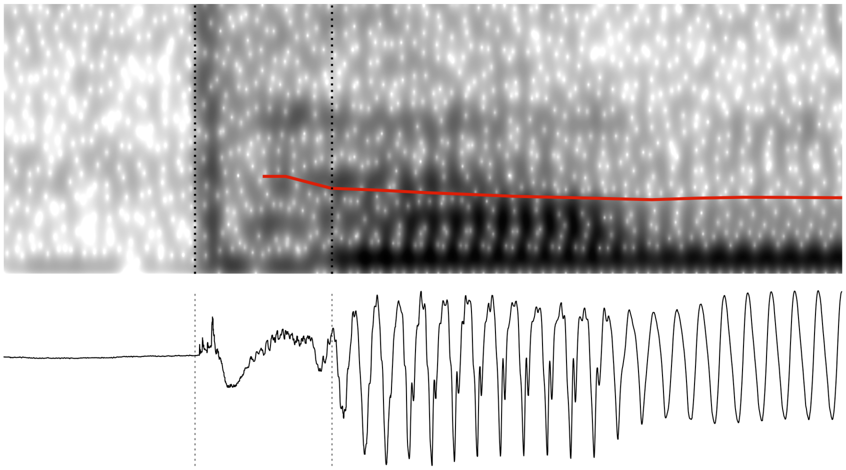

Figure 3.2. Sample spectrogram, waveform, and pitch track for the first syllable of [t^{h}á.mák] 'cavalry.' Vertical dotted lines indicate identified aspiration, and the red lines indicate the pitch track. F0 was measured at the end of aspiration, i.e. where the dotted line crosses the pitch track.

F0 was measured using Praat's built-in pitch tracker, with a time step of 0.01 seconds and a pitch range of 75-500 Hz for all speakers. Whenever possible, the pitch value recorded was that interpolated at the first period following the release burst; in most cases, this coincided with the endpoint of the VOT. When the pitch tracker had not interpolated a pitch value at this point, as well as for pre-voiced tokens, F0 was measured in the first period at which a pitch value was available.

Average values for VOT and F0 were calculated within and across the various phonological categories. Z-scores were also calculated by speaker in order to effectively compare VOT and F0 values relative to other tokens produced by the same speaker.

3.3 Results

3.3.1 F0 and Tonality

A first task was to establish the status of the tone contrast across speakers. Given the limited data analyzed so far, F0 at the onset of voicing was compared across the two tones for the various speakers. Since the high and low tones are phonetically high-level and low-rising, the contrast should be most robust at the beginning of a word; if there is a significant difference at this point, the speaker can be considered as using a tone contrast. F0 at the onset of voicing, by speaker and tone, is presented in Fig 3.3, below.

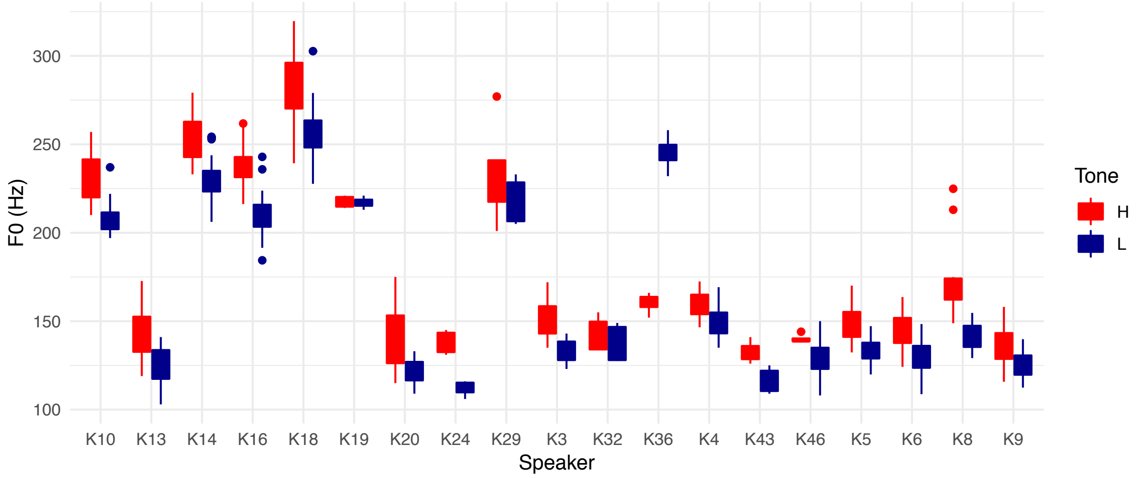

Figure 3.3. F0 at the onset of voicing by speaker and tone. Tokens presented are the same as analyzed for VOT later in this chapter. F0 values in Hertz.

Visual inspection of Fig. 3.3 shows that F0 at the onset of voicing is generally higher for high-tone words than for low-tone words, but with significant overlap and variability by speaker. In order to quantify this, a linear mixed-effects model was constructed using the _lme4_ package (Bates et al. 2014) in R (R core team 2013) to predict these F0 values (in Hz). A random effect was included for word (lexical item), fixed effects for speaker and putative tone category, and an interaction of speaker and putative tone category. Post-hoc analysis (Chi-square tests with Holm-adjusted _p_-values) was conducted using the _phia_ package (De Rosario-Martinez 2015) to determine which speaker*tone interactions were significant. The results are presented in Table 3.3.

Speaker	Coefficient	Chisq	_p_-value	Speaker	Coefficient	Chisq	_p_-value
*K10	23.144	28.5104	1.305E-06	K32	13.900	3.2099	0.1463923
*K13	19.086	19.6039	0.0001143	K36	-76.100	96.2155	< 2.2e-16
*K14	23.302	29.4728	8.506E-07	K4	10.931	6.3785	0.0412732
*K16	28.916	45.0427	3.47E-10	K43	24.872	9.0594	0.0179229
*K18	27.517	40.4372	3.249E-09	K46	19.900	6.5791	0.0412732
K19	9.400	1.4679	0.2256740	*K5	17.356	16.0929	0.0006031
*K20	28.900	13.8759	0.0017576	*K6	14.513	10.6203	0.0089482
*K24	34.900	20.2356	8.901E-05	*K8	34.085	42.7435	1.061E-09
K29	23.400	9.0969	0.0179229	K9	12.730	8.5759	0.0179229
*K3	18.768	19.3094	0.0001223				

Table 3.3. Post-hoc analysis of speaker*tone interactions. Chi-square values vary substantially, with larger values corresponding to larger differences between high and low tones. 11 of 19 speakers, indicated with asterisks (*) and in bold, have p > .01 and positive coefficients, indicating higher F0 for H tone than for L tone. Only one speaker, K36, has a negative coefficient, indicating unexpectedly lower F0 in H than in L.

As shown in Table 3.3, 17 of 19 speakers (all except K19 and K32) have a $p<.05$, though $p>.01$ for K29, K43, K46, and K9. Additionally, K36 shows a coefficient in the opposite direction, with H tones lower than L tones. The majority of speakers thus show evidence of the expected tone contrast, though for some speakers this may be attenuated or absent.

3.3.2 VOT

The data was hypothesized to exhibit clustering in VOT predictable by the phonological feature [SPREAD GLOTTIS] (henceforth, [SG]) and High/Low tone. However, analysis of the data indicated that some stops were pre-voiced as well, as shown in Figure 3.4, below:

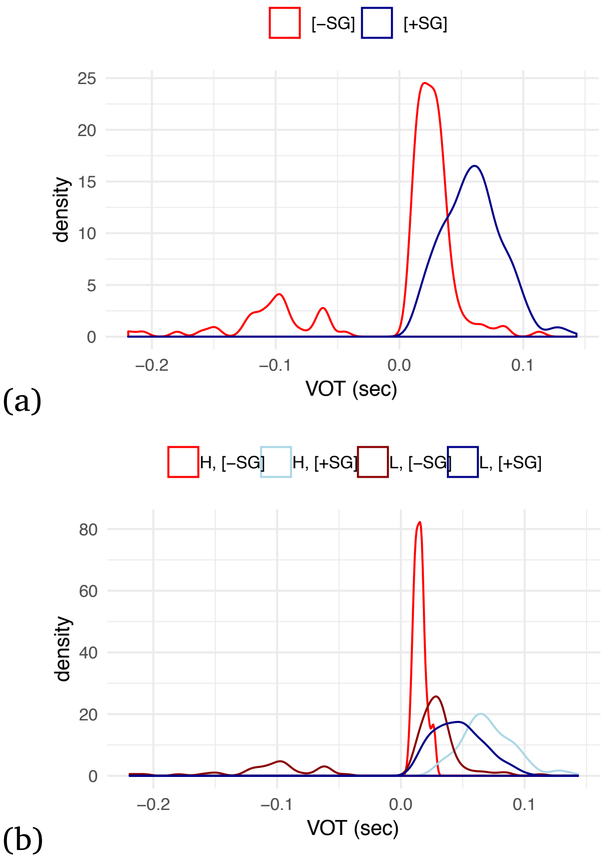

Figure 3.4. Density plot of voice onset time (sec) by onset category. (a) VOT by [SG]. (b) VOT by [SG] and tone.

As expected, most items exhibit positive VOT values, and [+SG] stops show longer VOT than [-SG] stops. Prevoicing (negative VOT) is observed in a subset of the [-SG] stops with low tone. Unsurprisingly, the [-SG] stops exhibit shorter VOT than the [+SG] stops. However, the [+SG] stops of Fig. 3.4(a) appear as two distinct clusters in Fig. 3.4(b): [+SG] stops with high tone show a longer VOT than their counterparts with low tone.

The prevoiced tokens were produced by different speakers, and no items were consistently voiced by all speakers. The frequency of voicing by item was as follows: 3/17 tokens of /cà.gé/ ‘Chinese language’; 3/18 tokens of /cà.mí/ ‘Chinese person’; 4/37 tokens of /cà.t͡ʃá/ ‘pheasant’; 4/23 tokens of /tŏm/ ‘spider’; 12/28 tokens of /ʈŭ/ ‘barley’; and 10/24 tokens of /ʈŭk/ ‘dragon’, all of which were non-[SG], as well as two [SG] tokens: 1 token of /p^{h}à.t͡ʃúk/ ‘cow’ and 1 token of /t^{h}ŏm/ ‘bear.’ In both of these latter cases, the voicing may be interpreted as a reading error on the part of the speaker, since the same orthographic characters are used for both aspirated and unaspirated low-tone series. These two tokens have been excluded from all subsequent analyses. Otherwise, the voiced tokens were spread across all low-tone, [-SG] items measured. The frequency of voicing preceding retroflex consonants was somewhat higher than other places of articulation, though it is not clear that there is any principled reason for this.

Since tokens with prevoicing exhibit relatively long intervals of prevoicing, this is interpreted as a distinct alternate realization of [-SG] stops with low tone, rather than as incidental, limited vibration of the vocal folds during the closure. This alternative is available to many speakers, as 12 speakers produced at least one pre-voiced token. Treating pre-voicing as rooted in a different phonological form, three categories present themselves with regards to voicing and aspiration: “voiced” for low-tone, [-SG], pre-voiced tokens; “unaspirated” for [-SG] tokens with either tone but without pre-voicing; and “aspirated” for [+SG] tokens with either tone. The effect of this categorization on VOT is shown in Fig. 3.5:

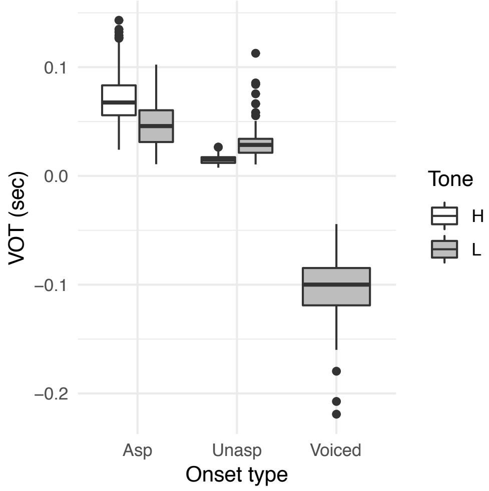

Figure 3.5. Effect of Phonological aspiration, voicing, and tone on VOT (sec). The x-axis depicts the three phonological categories of (left-to-right) aspirated, unaspirated, and voiced tokens, broken down into low-tone and high-tone tokens. Note that voicing only co-occurs with low tone.

Figure 3.5, above, shows the distribution of VOT across aspirated, unaspirated, and voiced categories. When tone category (high or low) is considered, as in Fig. 3.4(b), it becomes apparent that VOT is longer among high-tone aspirated tokens than low-tone aspirated tokens, which is in line with predictions of the gestural theory. However, there is an unexpected result of a shorter VOT for high-tone unaspirated tokens than low-tone unaspirated tokens, where no difference was predicted. It is worth noting that only one item, /tá.mák/ ‘cavalry,’ was high-tone and unaspirated.

In order to assess the robustness of these results, a series of linear mixed-effects models were fit to the data drawn from all places of articulation. The first model is a baseline model, with random effects for Speaker, Word, and Place (bilabial, dental, retroflex, palatal, and velar). The second model adds a factor for [SG], and a third model adds a factor for tone; these interact in the fourth model. The summary of the model comparison appears in Table 3.4.

(1) VOT ~ (1|Place) + (1|Speaker) + (1|Word)

(2) VOT ~ SG + (1|Place) + (1|Speaker) + (1|Word)

(3) VOT ~ Tone + SG + (1|Place) + (1|Speaker) + (1|Word)

(4) VOT ~ SG*Tone + (1|Place) + (1|Speaker) + (1|Word)

Model	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	*p*-value
Baseline	5	-2,844.69	-2,823.37	1,427.35	-2,854.69	NA	NA	NA
SG	6	-2,859.83	-2,834.24	1,435.92	-2,871.83	17.14	1	3.472E-05
SG + Tone	7	-2,863.84	-2,833.99	1,438.92	-2,877.84	6.01	1	0.0142
SG*Tone	8	-2,865.87	-2,831.75	1,440.94	-2,881.87	4.03	1	0.0447

Table 3.4. Summary of Model Comparison. All models include random effects of Place, Speaker, and Word.

Crucially, comparing these models reveals a significant effect of the interaction between SG and Tone. This interaction indicates that, for example, for a single value of SG, changing the value of Tone (from ‘0’ to ‘1’, i.e. from Low to High) leads to an improvement in the model—precisely what is predicted if High-tone conditioned a longer VOT in aspirated segments but not unaspirated segments..

3.3.3 VOT and F0

To what degree is the patterning of VOT and F0 in Tibetan under phonological control? Section 3.3.1 found that lexical tone predicts F0 at the onset of voicing for most speakers, with variation, while Section 3.3.2 found that VOT is affected by both the [SPREAD GLOTTIS] feature and by tone. This section investigates the covariation of the two phonetic parameters, VOT and F0. A scatterplot of the two, grouped by phonological categories of [SG] and tone, is presented in Fig. 3.6.

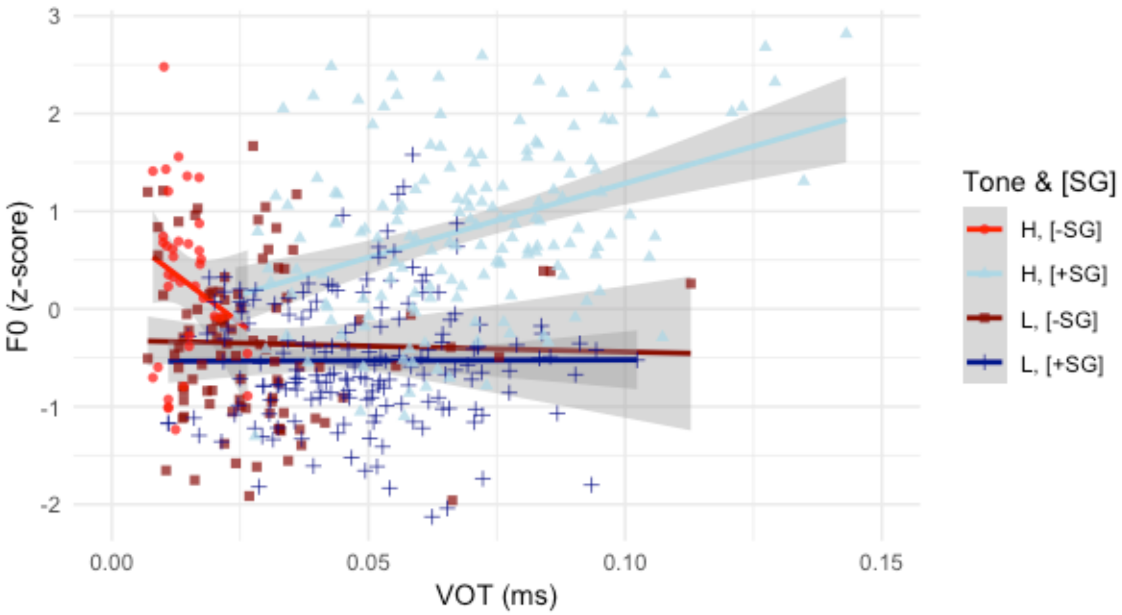

Figure 3.6. VOT and F0 (z-score by speaker) at onset of voicing. Negative-VOT items are excluded, and data from all nineteen speakers is aggregated. Linear regression lines and 95% confidence intervals are included..

The previously-established relationships are visible in Fig. 3.6. High-tone items have generally higher F0 than low-tone items, [+SG] stops have generally longer VOT than [-SG] stops, and low-tone, [+SG] stops have an intermediate VOT value. However, the covariation of VOT and F0 differs by the interaction of tone and [SG]. Only among high-tone aspirated stops is longer VOT associated with higher F0. No relation is found among the low-tone aspirated or low-tone unaspirated stops, and the trend is reversed among high-tone unaspirated stops.

3.4 Discussion

The present study investigates the relationship between VOT, tone, and F0 in Tibetan. Based on recordings of nineteen speakers, it was found that word-initial VOT varied significantly across phonological categories of both tone and aspiration. Prevoicing was only observed for unaspirated stops in low-tone words, and only variably. Positive VOT values were conditioned not only by whether the stop in question was aspirated or unaspirated, but also by the tone of the word. Among aspirated stops, VOT was shown to be longer in high-tone words than in low-tone words. However, among unaspirated stops, a small effect was observed in the opposite direction: VOT was slightly shorter in high-tone words than in low-tone words. Finally, the covariation of VOT with F0 differed according to the interaction of tone and [SG], providing evidence of different physiological mechanisms instantiating these phonological categories.

It is important to consider how the presentation of stimuli using the Tibetan orthography might affect results. Tibetan speakers are well aware that a word's spelling differs from its pronunciation, and that dialects vary in their pronunciation. The orthography is still related to pronunciation, however, and so bias is possible. Research on other languages indicate that orthographic factors can induce effects resembling phonetic/phonological processes such as incomplete neutralization (Warner et al. 2006). The context where orthographic effects seem most plausible is low-tone stops, which are all written with the same letter (the historically voiced series *b d g*) irrespective of aspiration.

Indeed, the aggregated VOT data presented in Fig. 3.4(b) and Fig. 3.5 show that low-tone aspirated stops have a shorter VOT than high-tone aspirated stops—that is, that low-tone aspirated stops fall between the other aspirated stops and the other low-tone stops. An orthographic explanation is unlikely for two reasons. First, the prevoicing observed on some stops only ever occurs with the low-tone unaspirated stops; speakers do not produce prevoicing on the low-tone aspirated stops even though they are written with the same graphemes. Second, interspeaker variation is likely, and this topic is explored in Chapter 4 using the larger number of tokens per speaker in the EMA experiment (see FIg. 4.5). Therefore, the apparent intermediate value is the result of data aggregation necessitated by the small number of tokens per speaker. Perhaps orthography has contributed to the fact that some speakers maintain short VOT for the low-tone "aspirated" stops, but this should be understood as a historical rather than synchronic factor. The long VOT with low tone is a relatively recent sound change (see section 2.9), so it is the speakers with long-VOT low-tone stops who have undergone a change. It is conceivable that the orthography contributed to other speakers not adopting this change, though the fact remains that the orthography does not appear to influence the synchronic phonology of the speakers in this study.

3.4.1 Enhancement account

Could the observed effects of VOT on tone be explained with reference to phonetic enhancement? According to the Quantal Theory account of enhancement (Keyser & Stevens 2006, Stevens & Keyser 2010), enhancement consists either of adding a gesture either to increase the perceptual saliency of a contrast, or to introduce a new parameter to support the existing contrast.

The latter provides an account of the prevoicing for unaspirated stops with low tone: the addition of prevoicing furnishes an additional phonetic parameter to aid the perception of the unaspirated + low tone items. Since no other word-initial stops are prevoiced, prevoicing helps distinguishes these items both from high-tone unaspirated stops and low-tone aspirated stops. The optionality of prevoicing is also consistent with its role in enhancement. The use of voicing as this parameter follows from the history of the contrast: the subphonemic lowering of F0 after voiced stops was reanalyzed as a tone contrast, but the voicing was able to remain and be recruited for the purpose of enhancement.

Why, then, would tone cause a difference in aspirated stops? Before tonogenesis, the words now produced with low tone and aspirated stops were voiced, but this voicing has been entirely lost among these speakers and those of other Central Tibetan dialects. Low-tone aspirated stops risk confusion with their low-tone unaspirated and high-tone aspirated counterparts; the high-tone unaspirated stops are already distinguished along two parameters. Languages may develop so this contrast is staggered along the VOT axis. Low-tone unaspirates are already at near-zero (or negative) VOT, so introducing a longer VOT would support the tone contrast; this appears to be the diachronic origin of aspiration with low tone. However, to avoid confusion with the high-tone aspirated stops, the low-tone aspirated stops could be produced with an intermediate value, longer than the unaspirated stops but not as long as the high-tone aspirated stops.

While perceptual distinctiveness motivates the intermediate VOT of low-tone aspirated stops, a complete account requires an additional mechanism to explain how the contrast is articulated. The examples used to illustrate enhancement by Keyser and Stevens (2006) involve the addition of a gesture, but it is not clear in this case what such a gesture would be. Instead, the difference could be one of intergestural timing.

VOT is not an articulatory measure; it is an acoustic consequence of articulatory timing. If tonal gestures are timed with onset consonant and vowel gestures in a similar way to the second member of a consonant cluster (as in Gao 2008, Karlin 2014, Hu 2016; see sections 1.4 and 4.1.1), this should cause the consonant gesture to begin earlier in time relative to the following vowel. No change in VOT would result if the glottal spreading gesture moved along with the oral consonantal gesture. However, in other languages it has been observed that a single glottal spreading gesture might be shared across a consonant cluster, overlapping with the oral gestures of multiple consonants. It is thus possible to hypothesize that the effect of a tone gesture on consonant-vowel timing might cause a VOT difference in aspirated stops.

The type of acoustic evidence presented in this chapter is not sufficient to test this hypothesis. If the effect of tone on VOT is mediated by consonant-vowel timing, articulatory evidence of consonant-vowel timing would be required. Chapter 4 presents the EMA study conducted to gather this evidence. Further implications for theories of temporal coordination will be discussed in Chapter 5.

3.4.2 Gestural scores

The results presented in Section 3.3 describe how VOT is conditioned by tone in Tibetan. In this section, those results are interpreted in terms of gestural scores, a mechanism for relating phonetic production with phonological representation. (More detail on intergestural coordination can be found in sections 1.4 and 4.1.1.)

Broadly speaking, three main aspects of a gestural score can be invoked to explain the differences in VOT. First, the gestures themselves might be different: different glottal gestures correspond to different laryngeal postures, such as a spread glottis (for aspiration), a critical opening (for voicing), and a partly open glottis (for voiceless unaspirated stops) (Esling & Harris 2003, Edmondson & Esling 2006). Second, the gestures may differ in in their duration. Third, the gestures may differ in temporal coordination with each other. In the remainder of this section, two possible accounts of the Tibetan data are discussed: one based on differing gestural activation durations, the other two based on different gestural coordination. The four Tibetan VOT categories according to the gesture duration account are presented in Fig 3.7(a-b) and (c-d), while those according to the gestural coordination account are presented in Fig. 3.7(a-b) and (e-f).

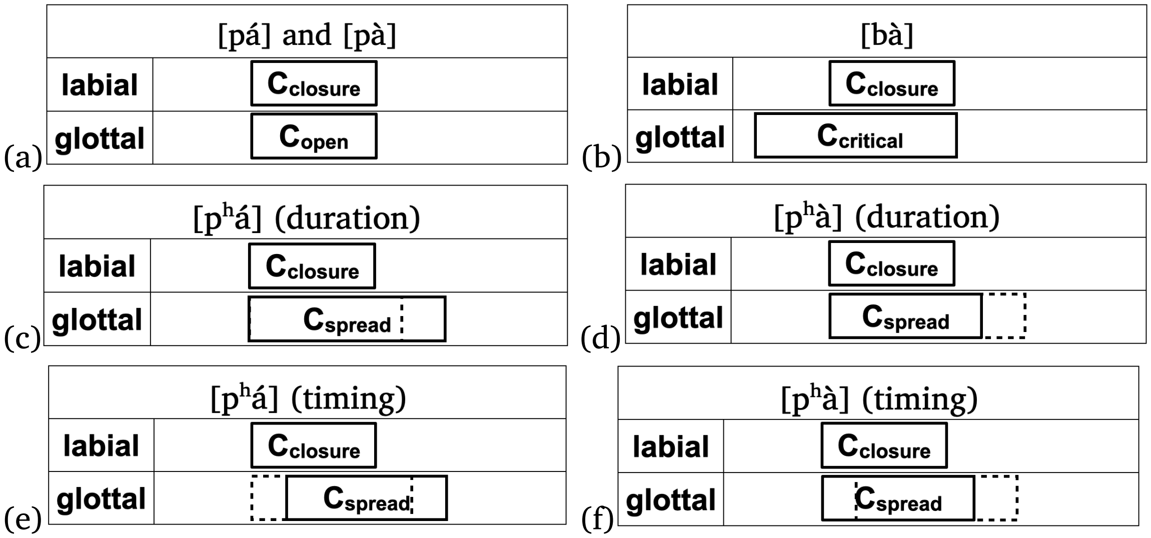

Figure 3.7. Partial gestural scores for Tibetan onset stops. For each pair of (c)-(d) and (e)-(f), dotted lines represent the temporal arrangement of the glottal gesture in the other member of the pair. (a) Short-VOT stop features glottal opening synchronous with oral closure. (b) Negative VOT stop features critical glottal gesture for prevoicing (this is the prevoiced alternant of the short-VOT stop with low tone). (c-f) The long-VOT and intermediate-VOT stops feature a glottal spreading gesture that could be (c) longer in the high-tone long-VOT stop and (d) shorter in the low-tone intermediate-VOT stop. Alternatively, the glottal spreading gesture could be the same duration in both cases, but (e) begin after the start of the oral closure for the long-VOT stop and (f) begin synchronous with the oral closure for the intermediate-VOT stop.

The gestural scores presented in Fig. 3.7 correspond to the four onset-tone categories of Tibetan. As in a Thai-style 3-way VOT contrast, three glottal gestures are used: a phonation-inhibiting glottal opening gesture for the short-VOT stops in Fig. 3.7(a), a critical glottis gesture for the voiced stop variant in Fig. 3.7(b), and a glottal spreading gesture for the long- and intermediate-VOT stops in Fig. 3.7(c-f). Fig.3.7(c-d) account for the difference between long and intermediate VOT as the result of different activation durations of the glottal spreading gesture. As the gesture is active for longer in 3.7(c) than in 3.7(d), the resulting VOT is longer. Alternatively, Fig. 3.7(e-f) accounts for the difference as a result of gestural timing. The glottal spreading gesture has the same activation duration for both stops; however, it begins and ends later in (e) than in (f), resulting in a longer duration of this gesture after the conclusion of the oral closure, which produces a longer VOT.

How might these accounts be evaluated? The most straightforward test would directly image the glottis to observe the nature and timing of its movements. However, this was found not to be feasible due to to the physical and technical difficulty of imaging the larynx. Instead, indirect tests are required. The remainder of this section presents and evaluates several predictions of the glottal gesture and glottal duration accounts.

In terms of phonology, the descriptions of Central Tibetan reviewed in Chapter 2 (including Denwood 1999, Tournadre & Dorje 2003, and Tsering 2011) group the stops and affricates into "aspirated" and "unaspirated" categories, both of which occur with high and low tone. In these descriptions, the different VOT by tone is a matter of surface-level phonetics, not the underlying contrast. This is most similar to the gestural timing account, which uses the same gesture—a unit of phonological contrast—for the high-tone long-VOT and low-tone intermediate-VOT stops. The different temporal coordination, then, instantiates the surface-level difference. The phonological descriptions are not consistent with the gesture duration account, as its four different gestures effectively posit four consonants rather than three. This abandons the attempt to maximize parsimony and even relegates tone to a redundant status. Previous phonological literature is thus more consistent with the temporal coordination account.

Another line of evidence comes from typological comparison. The world's languages vary tremendously in the duration of their VOT contrasts (e.g. Lisker

& Abramson 1964, Cho & Ladefoged 1999), but a language with three positive VOT contrasts remains unattested. It is thus more consistent with the typological literature to derive the intermediate-VOT stops of Tibetan from another category of stops, as in the gestural timing account. Again, the gesture duration account would effectively create another class of consonants, which would be unique among known languages.

Diachronically, the intermediate-VOT stops are derived from the historically-voiced simplex onsets (see Section 2.5). As such, the voicing was reanalyzed as low tone, but the speakers at the time of this tonogenesis would not have heard this series produced with long VOT. This means that the low-tone intermediate-VOT stops would have passed through a stage where they had lost prevoicing, but not yet developed longer VOT—a stage retained in many Eastern (Kham) dialects such as Dege (Tsering 2011), Bathang (Tsering 2011), and Thebo (Lin 2014). (In these dialects, contrast with the low-tone unaspirated series is maintained by consistent voicing in the latter, rather than the variable voicing in Central Tibetan). How would this series have gained a longer VOT? For reasons of enhancement (see Section 3.4.1), a variant featuring a longer VOT and glottal spreading gesture could have increased distinctiveness and spread, particularly if prevoicing was lost or became a less reliable cue. The difference between this series and the long-VOT stops would be maintained by tone, and further enhanced by a different glottal gesture or temporal coordinaiton. In light of these historical developments, both duration and timing accounts are diachronically plausible.

Finally, the two accounts generate articulatory predictions that can be tested as indirect evidence for the glottal gestures. In particular, the gestural timing account involves different timing of the glottal gesture in the intermediate- and long-VOT stops, which could be associated with timing differences in other gestures as well. Given the implausibility of directly observing glottal gestures, a mechanism is needed to generate predictions for oral gestures instead. The competitive coupling model of tone surveyed in Section 1.4.3 (Gao 2008) furnishes these predictions.

The coupling graphs in Fig 3.8. depict possible sets of coupling relations among the following types of gestures in a CV syllable: oral consonantal (C), glottal consonantal (G), vowel (V), and tone (T). Fig. 3.8(a) depicts the coupling relations in a toneless CV syllable as per Goldstein et al. (2009): in-phase C-V and C-G coupling[2] leading to simultaneous start times of the three gestures. The kind of tonal syllable presented in Gao (2008) is shown in Fig. 3.8(b): it retains the coupling relations of Fig.3.8 (a), adding in-phase V-T and anti-phase C-T coupling to model partial overlap. Finally, Fig. 3.8(c) represents an alternative structure for the tonal syllable: the glottal gesture here has a second in-phase coupling relation with the tonal gesture, reflecting the cluster-like relationship of consonant and tone and sharing a glottal gesture across such a cluster.

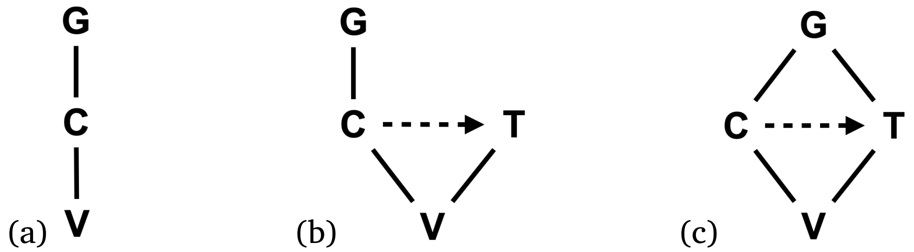

Figure 3.8. Predicted coupling graphs with competitive coupling of tone. C refers to oral consonant gesture; V refers to oral vowel gesture; T refers to tonal gesture; G refers to glottal gesture. Solid lines indicate in-phase coupling and dotted arrows indicate anti-phase coupling. (a) Mandarin-like tonal syllable. (b) Tonal syllable with

[2] In-phase C-G timing is used for voiceless and aspirated stops (e.g. Goldstein et al. 2009); voiced stops and other consonants require different treatment, but the present discussion concerns only the voiceless stops of Tibetan.

glottal gesture coordinated in-phase to both consonant and tone gestures. (c) Syllable without tone gestures.

The three coupling graphs of Fig. 3.8 offer three scenarios for the relationship between tonality, VOT, and C-V lag. Fig. 3.8(a) is a CV syllable with no tone gesture, and predicts in-phase C-V timing. Fig. 3.8(b) and Fig. 3.8(c) are CV syllables with tone, and predict the C gesture will begin before the V gesture, a difference known as C-V lag. This C-V lag should also covary with the duration of the C gesture as a result of anti-phase timing (Shaw et al. 2019). Where these two differ is the timing of the glottal gesture: in Fig. 3.8(b) the C and G gestures begin simultaneously, while in Fig 3.8(c) the C gesture begins before the G gesture. As a result, the difference in time between the end of the C and G gestures is longer for Fig. 3.8(c) than for Fig. 3.8(b)—a difference corresponding to a longer VOT for Fig. 3.8(c).

Finally, these coupling graphs and their predictions are applied to the partial gestural scores in Fig. 3.7, resulting in the more complete gestural scores in Fig. 3.9, below. The two accounts presented above can now be evaluated on the basis of phonetic predictions.

The gestural duration account, whose gestural scores are presented in Fig. 3.9(a-d), relies on the original gestural model of tone coupling graph, Fig. 3.8(b). As a result, all syllables are predicted to exhibit C-V lag that varies dynamically with C duration.

The gestural timing account, presented in Fig. 3.9(e-h), is instead based on the coupling graphs of Fig. 3.8(a) and (c). Here, C-V lag that covaries with C duration is predicted for three of the four conditions, but not the low-tone intermediate-VOT condition. The gesture timing difference between long-VOT and intermediate-VOT conditions in Fig. 3.9(g-h) is caused by the presence of a high tone gesture, but with no specified low tone gesture. Therefore, the high-tone syllables are predicted to exhibit a C-V lag (covarying with the C gesture duration), but the low-tone syllables are predicted to exhibit C-V simultaneity.

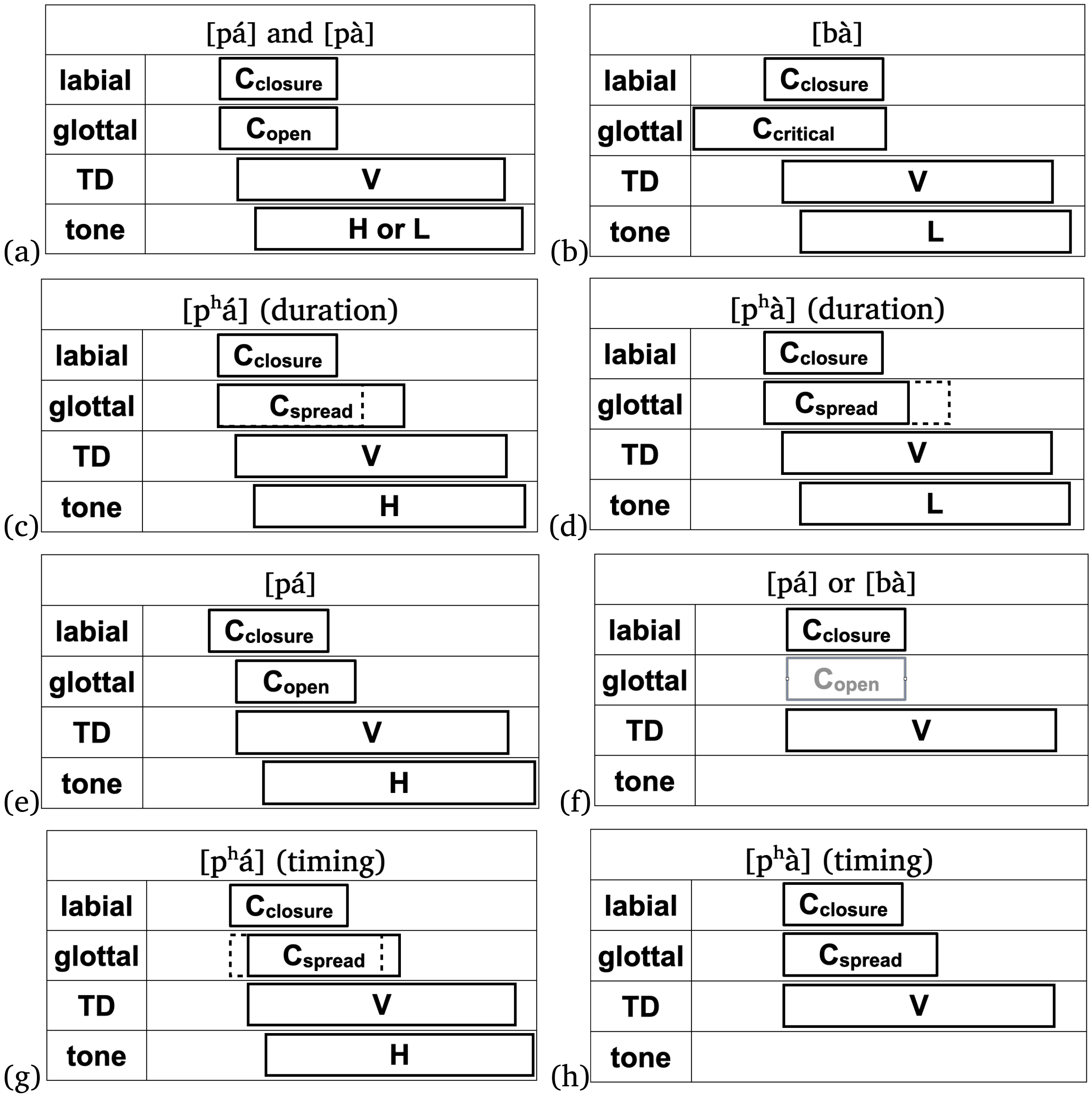

Figure 3.9. Gestural scores for Tibetan onset stops. (a-d) Gestural scores according to gesture duration account: (a) Short-VOT stop with C-G synchrony and C-V lag; (b) Negative-VOT stop with C-V lag; (c) High-tone long-VOT stop with long glottal spreading gesture; (d) Low-tone intermediate-VOT stop with shorter glottal spreading gesture. (e-h) Gestural scores according to gestural timing account. (e) High-tone short-VOT stop with C-V lag; (f) Low-tone short-VOT stop (as shown) or negative-VOT stop with C-V synchrony; (g) High-tone long-VOT stop with C-V lag (dotted lines

indicate timing of glottal gesture without competitive coupling); (h) Low-tone intermediate-VOT stop with C-V simultaneity.

The key difference in the gestural scores in Fig. 3.9 lies in C-V timing. Under the gesture duration account of Fig. 3.9(a-d), all words are predicted to exhibit similar, synchronous start times of C and V gestures. Under the gestural timing account of Fig 3.9(e-h), the timing would differ across tones: high-tone words would show a longer C-V lag than low-tone words due to the presence of a specified high-tone gesture. These predictions form the basis of the EMA experiment developed in Chapter 4.

The basic predicted difference is that C-V lag will be similar across tones for the gesture duration account, but will differ by tone under the temporal coordination account. This difference also applies to some variations of these accounts. For example, the version of the gesture duration account presented relies on the coupling graph in Fig. 3.8(b); if instead the coupling graph in Fig. 3.8(c) is used, the C-V timing would not be synchronous, but would still be consistent across tones. Likewise, the gesture duration account also does not predict differences in C-V lag by tone. However, such possibilities are not consistent with the established observation that VOT differs by tone in the aspirated stops.

The above discussion relies on an “H-L” or “H-LH” analysis of tone (see Section 2.10). If instead an “H-∅” analysis is used, C-V lag would be predicted to differ across tones because the ∅-tone condition, lacking a gesture, would remove the gesture coordinated anti-phase, and the remaining coupling relations would all be in-phase. Finally, the “L-∅” analysis is less plausible on phonological grounds, but it would result in longer VOT with *low-tone* words, the opposite of the results presented in Section 3.3. The relationship between the accounts and predictions are summarized in Table 3.5, below.

Onset series	**Tone**	**Glottal gesture (duration)**	**Glottal gesture (timing)**
short~negative VOT	L	open~critical	open~critical
short VOT	H	open	open
intermediate VOT	L	spread (shorter)	spread (earlier)
long VOT	H	spread (longer)	spread (later)
Predictions consistent with:			
Phonological description		No	Yes
Typology		No	Yes
Diachronic plausibility		Yes	Yes
Articulatory prediction		**C-V timing not different (simultaneous) by tone**	**C-V lag longer with H tone than with L tone**

Table 3.5. Summary of glottal gesture and predictions of proposed accounts.

The predictions presented here have dealt with the basics of gestural timing, but have abstracted away from a substantial amount of detail. For example, the timing-based accounts include gestural scores where the C closure begins before the glottal opening/spreading gesture. This would predict a short period of voicing leakage at the beginning of these consonants, which has not yet been observed. As for the tone gestures, this discussion only investigates them inasmuch as they interact with the other gestures, and does not make particular claims about the targets (e.g. F0 trajectories) of these gestures. Given the rising F0 trajectory of the low tone, the H-∅ analysis may still require a high tonal target in the "∅" or low-tone condition, perhaps coupled anti-phase to the vowel. All gestures were treated as unitary entities, though the coupling graphs and gestural scores could be constructed in a number of different ways, such as with the split-gesture hypothesis (Nam 2007). Nevertheless, the current framing is sufficient to establish viable hypotheses for the EMA experiment conducted in Chapter 4.

3.5 Ongoing corpus development

The analysis of this chapter has touched on only a small portion of the corpus: just the wordlist data from the nineteen diaspora-raised speakers. The remaining portions of these speakers' interviews, which include spontaneous-speech data, has not been analyzed, nor has the data from the other speakers. Thanks to a Doctoral Dissertation Research Improvement Grant from the National Science Foundation, I have been able to hire a native Tibetan speaker to transcribe the interviews, and another research assistant to help with forced-alignment of the corpus. This will allow the analysis of this chapter to be extended to a larger and more naturalistic set of data, as well as allow comparison across speakers of other dialects also living in Kathmandu.

3.6 Chapter summary

This chapter investigates the relationship between Tibetan speakers' phonetic parameters of F0 and word-initial VOT, and phonological contrasts of tone and aspiration. With a two-way contrast in tone and a two-way contrast between aspirated and unaspirated stops, it was found that tone affected VOT. Aspirated stops in high-tone words had a longer VOT than aspirated stops in low-tone words. Prevoicing was present, in a variable manner, only for unaspirated stops with low tone. The pattern of prevoicing was explained as a phonetically-natural and diachronically-plausible enhancement gesture. However, the tonal interaction that causes three positive VOT lengths is more difficult to explain. Two accounts are presented: one based on different gesture activation durations, the other based on different gestural timing caused by competitive coupling between consonant and tone gestures. The two accounts differ in their articulatory predictions: the first predicts no effect of tone on the timing of consonant and vowel gestures, while the second predicts different tones could affect C-V timing. An EMA study testing these predictions is described in chapter 4.

3.7 Chapter bibliography

Abramson, Arthur S. & Douglas H. Whalen. 2017. Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of phonetics*. Elsevier 63. 75–86.

Boersma, Paul & David Weenink. 2018. Praat: Doing phonetics by computer [Computer software]. Version 6.0. 43.

Cho, Taehong & Peter Ladefoged. 1999. Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics* 27(2). 207–229. https://doi.org/10.1006/jpho.1999.0094.

De Rosario-Martinez, Helios, John Fox, R. Core Team & Maintainer Helios De Rosario-Martinez. 2015. Package 'phia.' *CRAN repository*. *Retrieved* 1. 2015.

Gallagher, Gillian. 2011. Acoustic and articulatory features in phonology – the case for [long VOT]. *The Linguistic Review* 28(3). https://doi.org/10.1515/tlir.2011.008. https://www.degruyter.com/doi/10.1515/tlir.2011.008 (1 September, 2020).

Gao, Man. 2008. Tonal alignment in Mandarin Chinese: An articulatory phonology account. *Unpublished Doctoral Dissertation (Linguistics), Yale University, CT*.

Hu, Fang. 2016. Tones are not abstract autosegmentals. In *Speech Prosody*, 302–306.

Hussain, Qandeel. 2018. A typological study of Voice Onset Time (VOT) in Indo-Iranian languages. *Journal of Phonetics*. Elsevier 71. 284–305.

Iverson, Gregory K. & Joseph C. Salmons. 1995. Aspiration and laryngeal representation in Germanic. *Phonology* 12(3). 369–396. https://doi.org/10.1017/S0952675700002566.

Karlin, Robin. 2014. The articulatory TBU: gestural coordination of tone in Thai. In *Cornell Working Papers in Linguistics*.

Keyser, Samuel Jay & Kenneth N. Stevens. 2006. Enhancement and Overlap in the Speech Chain. *Language* 82(1). 33–63. https://doi.org/10.1353/lan.2006.0051.

Kingston, John & Randy L. Diehl. 1994. Phonetic Knowledge. *Language* 70(3). 419–454. https://doi.org/10.1353/lan.1994.0023.

Kingston, John, Randy L. Diehl, Cecilia J. Kirk & Wendy A. Castleman. 2008. On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics* 36(1). 28–54. https://doi.org/10.1016/j.wocn.2007.02.001.

Lisker, Leigh & Arthur S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*. Taylor & Francis 20(3). 384–422.

Ohde, Ralph N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America* 75(1). 224–230. https://doi.org/10.1121/1.390399.

Stevens, Kenneth Noble & Samuel Jay Keyser. 2010. Quantal theory, enhancement and overlap. *Journal of Phonetics* 38(1). 10–19. https://doi.org/10.1016/j.wocn.2008.10.004.

Warner, Natasha, Erin Good, Allard Jongman & Joan Sereno. 2006. Orthographic vs. morphological incomplete neutralization effects. *Journal of Phonetics* 34(2). 285–293. https://doi.org/10.1016/j.wocn.2004.11.003.