Original Paper

Phonetica 2014;71:50–81
DOI: 10.1159/000362672

Received: October 8, 2013
Accepted after revision: March 25, 2014
Published online: June 5, 2014

The 'Whistled' Fricative in Xitsonga: Its Articulation and Acoustics

Sang-Im Lee-Kim[a] Shigeto Kawahara[c] Seunghun J. Lee[b]

[a]Department of Linguistics, New York University, New York, N.Y., and [b]Central Connecticut State University, New Britain, Conn., USA; [c]Keio University, Tokyo, Japan

Abstract

The present study examines the articulation and acoustics of the typologically rare and understudied 'whistled' fricative sound in Xitsonga, a Southern Bantu language. Using ultrasound imaging and video recording, we examine the lingual and labial articulation of the whistled fricative. For the acoustic analysis, we employ the multitaper spectral analysis, which ensures reliable spectral estimates. The results revealed an interplay between multiple articulators involved in the production of the sound: the retroflex lingual gesture and the narrowing of the lower lip toward the upper teeth. Acoustically, the spectra of the whistled fricative are more peaked and compact than the acoustically similar palatoalveolar fricative, and the differences manifest themselves most clearly in two acoustic parameters, dynamic amplitude (A_d) and M2 (variance). The acoustic differences are also manifested in F2 and F3 in the surrounding vowels. Additionally, the 'whistled' fricative in Xitsonga is not quite whistled, contrary to the label given to the sound in previous studies. Building on the current articulatory and acoustic results, we discuss two different aerodynamic models for the whistled fricatives in Southern Bantu languages and conclude that the whistled fricative in Xitsonga is best characterized as a retroflex segment accompanied by weak whistling.

1 Introduction

A number of Southern Bantu languages are reported to have a typologically rare 'whistled' fricative sound, which contrasts with the dental /s/ and palatoalveolar sibilants /ʃ/ [see Shosted, 2006, for a full list of the languages]. The complex nature of the whistled fricatives – both in terms of their articulation and acoustics – is hinted by the fact that the previous literature provided various kinds of transcription of this sound; for example, focusing on different aspects of the sound, /s̫, z̫/ were used to indicate labialization [Bladon et al., 1987; Ladefoged and Maddieson, 1996], /s̢, z̢/ to indicate retroflexion [Sitoe, 1996], and /s͎, z͎/ to indicate pathological whistling [ICPLA, 1994]. While previous phonetic studies addressed some aspects of the whistled fricatives, the

KARGER

E-Mail karger@karger.com
www.karger.com/pho

© 2014 S. Karger AG, Basel
0031–8388/14/0711–0050
$39.50/0

Sang-Im Lee-Kim
Department of Linguistics
New York University
10 Washington Place, New York, NY 10003 (USA)
E-Mail sangim@nyu.edu

goal of the current study is to provide a comprehensive picture of both articulation and acoustics of the whistled fricative. In doing so, we aim to expand our understanding of the aerodynamic mechanism for whistling used in human speech.

1.1 The Whistled Fricative in Xitsonga

Of the various languages and dialects that are reported to have whistled fricatives, we study the voiceless whistled fricative in Xitsonga (S53) [Guthrie, 1967/1971]. Xitsonga is a language of the Tsonga people, spoken in South Africa, Zimbabwe, Mozambique and Swaziland. The Xitsonga language is one of the eleven official languages of South Africa and spoken by about 2 million people (4% of the South African population) [Leholha, 2003]. Most South African Xitsonga speakers live in the Limpopo province and use Xitsonga as a means of communication in their everyday life. Some people receive education in Xitsonga up to postgraduate degrees.

The three sibilants (dental /s/, whistled /ʂ/, and palatoalveolar /ʃ/) are all phonemic in Xitsonga as shown by the minimal triplet: [sìr̤á] 'tomb (class 5)', [ʂìr̤á] 'disasters (class 8)', and [ʃìr̤á] 'disaster (class 7)'.[1] The contrast between /ʂ/ and /ʃ/ is most prominently observed, when they are used in noun class prefixes: the singular noun class 7 prefix /ʃi-/ and the plural noun class 8 prefix /ʂi-/. The sibilants in the prefixes do not show any evident phonotactic restrictions with the initial segment of the following stem; any possible consonants of Xitsonga can occur in this position without any phonological alternations, unlike Zulu, which shows consonant harmony of sibilants [Bennett, 2013]. Unlike Shona, the voiceless whistled fricative /ʂ/ in Xitsonga does not have a voiced counterpart /ʐ/, but the other sibilants /s, ʃ/ have voiced counterparts /z, ʒ/ [Baumbach 1987, pp. 10–11].

1.2 Controversies over the Articulation of the Whistled Fricatives

There is general agreement that the whistled fricatives in Xitsonga and other languages involve some sort of labial gesture, but descriptions on the specifics of the labial gesture vary. While Ladefoged and Maddieson [1996, p. 171, pp. 358–360] describe Shona's whistled fricatives as involving extreme lip rounding, Maddieson [2003, p. 27] stresses a vertical narrowing of the lips without particular lip protrusion for Shona and Kalanga. Bladon et al. [1987] report that the whistled fricative in Shona lacks lip rounding or protrusion, but rather the lower lip raises toward the upper teeth, completely covering the lower teeth. They therefore labeled this labialization as 'labiodental', as opposed to 'labiolabial' as in ordinary lip rounding. On the other hand, Shosted [2011] observes a substantial lip rounding and protrusion for Changana whistled fricative.

While the properties of labial gestures of the whistled fricatives have been discussed extensively in the literature, there has not been much discussion on their lingual

[1] Xitsonga nouns are divided into 18 noun classes. As in other Bantu languages, the odd-numbered classes represent singulars and the even-numbered classes represent plurals. For more information about the class prefixes in Xitsonga, see Baumbach [1987, ch. 4]. Throughout this article, we use the retroflex fricative symbol /ʂ/ to represent the whistled fricative for reasons that become clear once we present the results of the ultrasound study.

articulation. Some previous studies have hypothesized a retroflex gesture for the whistled fricative [Carter and Kahari, 1979; Laver, 1994; Sitoe, 1996; Shosted, 2011], but it has not been examined instrumentally, presumably due to the unavailability of an appropriate instrumental technique. Shosted [2011], however, stresses the necessity of instrumental investigation into the lingual gesture of the whistled fricative. He conjectures that labialization in the whistled fricative is not the main source for creation of whistling. By measuring the relative lip rounding, he found that the nonwhistled fricatives are more rounded in the rounded vowel environment; nevertheless, they do not show whistling even in this environment. Thus, labialization might be merely secondary to whistling, and the defining feature of the whistled fricative may lie in the lingual articulation.

Building on these previous studies, the current study is designed to examine the lingual and labial gesture of the whistled fricative in Xitsonga. Although our articulatory study tested only 1 female Xitsonga speaker due to inherent difficulties of conducting instrumental studies of indigenous languages, we provide a complete picture of the interplay between multiple articulators involved in the whistled fricative, using ultrasound imaging and video recording. We complement these articulatory studies with acoustic examination of data obtained in fieldwork with 4 native speakers.

More specifically, the current study uses ultrasound imaging to examine the retroflex gesture of the whistled fricatives, hypothesized in some of the literature reviewed above [Carter and Kahari, 1979; Laver, 1994; Sitoe, 1996; Shosted, 2011]. Given the considerable variability in tongue tip/blade gesture involved in the segments that are described under the name of ‘retroflexes’ across languages, however, we start our discussion by clarifying what is meant by ‘retroflexes’. The retroflex stops are further classified as subapical or apical with respect to the fine-grained gestural differences in the tongue tip/blade gesture [Ladefoged and Bhaskararao, 1983; Ladefoged and Maddieson, 1996]. Specifically, Ladefoged and Maddieson [1996, p. 27] identify the subapical retroflex stops for the sounds in Dravidian languages where the underside of the tongue tip makes contact with the postalveolar region. In contrast, Hindi retroflexes are identified as apical retroflex stops in which the apical edge of the tongue tip forms a constriction along the alveolar ridge. While they propose two separate IPA symbols ([ɖ] vs. [ḍ], respectively), they also admit that there seem to be no languages that use these two types of sounds contrastively.

When it comes to the retroflex fricatives, however, the subapical articulation is not found even in the Dravidian languages [Hamann, 2003]. Rather, finer distinctions are made within the region of the tongue front: apical or laminal. To represent this difference, Ladefoged and Maddieson [1996, p. 160] introduce two different symbols: [ʂ] for the apical retroflex fricatives found in Toda and [s̠] for the laminal retroflex fricatives found in Tamil. However, instrumental studies cast doubt on whether this distinction is necessary. Crucially, the specific tongue front gesture varies depending on individual speakers, coarticulatory environment, and other factors [Hamann, 2003]. For example, a real-time magnetic resonance imaging study with 4 Mandarin speakers [Proctor et al., 2012] found that 1 speaker used the apical gesture, while the other 3 speakers showed a more laminal gesture for the retroflex fricative [ʂ]. Similarly, in X-ray studies of Polish retroflex fricatives, cross-study comparison suggests free variation between apical [Biedrzycki and Gontarczyk, 1974, p. 21] and laminal gestures [Wierzchowska, 1980, p. 64].

These findings therefore speak against the necessity of introducing two distinctive symbolic representations of the retroflex fricatives. Rather, it seems more appropriate to characterize retroflexes as having a particular global configuration of the tongue, as proposed by Hamann [2003]. That is, the retroflexes are characterized by a retracted tongue back and lowered tongue middle and postalveolar constriction.[2] Regardless of the specific tongue tip/blade gestures, these characteristics seem to remain stable for retroflexes. As Hamann [2003] argues, the tongue middle is stretched and pulled backward in order to displace the tongue front toward the postalveolar region. The tongue back, as a result, also retracts, because it is physically connected to the tongue middle. Therefore, we use 'retroflex' as a general term and leave a possibility that there might also be variation between apical and laminal articulation for the Xitsonga whistled fricative. In this regard, ultrasound imaging is particularly adequate, because it can capture the entire tongue configuration involved in the retroflex gesture. It further enables us to examine the relative tongue shape of the whistled fricative compared to other sibilants in Xitsonga.

Regarding the labial gesture, there are two video-recording studies that document the specific labial gestures involved in the whistled fricatives. Changana, a dialect of Xitsonga spoken in Mozambique, was studied by Shosted [2011], and Shona, spoken in an area geographically close to Zimbabwe, was studied by Bladon et al. [1987]. As mentioned earlier, the labial gestures of the two languages appear to be different: substantial lip rounding and protrusion are found for Changana whistled fricatives, whereas the labial gesture is mainly manifested as a closure between the upper teeth and the lower lip in Shona. This difference may reflect dialectal differences in the production of the whistled fricatives in the region, and the current study aims to add more data by documenting labial articulation of the Xitsonga whistled fricative using video-recording.

In summary, this article attempts to study the labial and lingual gestures of Xitsonga whistled fricative. In addition to its descriptive values, the current study also offers a better understanding of the aerodynamic mechanism of a whistle in human speech, especially in terms of how the labial gesture relates to a whistle mechanism. Shadle [2010] proposes an 'edge tone' model specifically for the whistled fricatives in which the teeth serve as an edge and the lingual tongue constriction creates a turbulence jet. A whistle occurs when oscillation formed around the sharp edge (i.e. the teeth) couples into the resonance frequency of the cavity between the teeth and the lingual constriction. Another possible mechanism for a whistle is the 'hole tone' model [Shadle, 2010]. Under this model, the rounded lips and the lingual constriction form two orifices and the whistle frequency is determined by the cavity between the two. With a longer cavity, this whistle can resonate at a very low resonance frequency. Given the considerably different descriptions on the labial gestures involved in Changana and Shona whistled fricatives, the two distinct acoustic mechanisms proposed by Shadle [2010] may be at work for whistles in the two languages. We compare our results on Xitsonga with the previous studies and discuss the possibility that there are indeed two distinct mechanisms for creating a whistle in Southern Bantu languages.

[2] Reetz and Jongman [2009] divide the tongue body into three subparts: front of the tongue body, center of the tongue body, and back of the tongue body. This three-way division of the tongue body is unnecessary for the purpose of our study, and we consistently refer to the first two parts of the tongue body as the 'tongue middle' and the last part as the 'tongue back'.

1.3 Predictions regarding the Acoustics of the Whistled Fricatives

In addition to the articulatory study, we examine the acoustic properties of the Xitsonga whistled fricative with reference to other sibilants in the language. Given the variable properties of the frication noise, the first step for the acoustic analysis of the noise spectra is to ensure reliable estimates of spectral shapes. We employed a multitaper spectral analysis for the acoustic analysis of the whistled fricatives. This method is particularly suitable for running speech, since it does not rely on the assumptions of stationarity or ergodicity of frication noise and thus ensures a small error with good time and frequency resolution. Ever since this method was first introduced for the acoustic analysis of speech signals [Blacklock and Shadle, 2003; Blacklock, 2004], a growing number of studies have used this method to analyze obstruents in many languages [Lee-Kim, 2011; Shosted, 2011; Lousada et al., 2012; Koenig et al., 2013; Żygis et al., 2012]. In particular, Shosted [2011] used this method to analyze the whistled fricatives in Changana. Applying the same method to the whistled fricative in Xitsonga is thus expected to yield results which can be compared to those of previous studies based on the multitaper analysis.

Based on the spectra, we examine the defining acoustic characteristics of the whistled fricative that are distinct from other sibilants [s, ʃ] in Xitsonga. Using acoustic recordings from the same speaker of ultrasound imaging and video recording, we first examine peak frequency and spectral moments for a three-way acoustic comparison. Spectral peak F is defined as the frequency where the maximum amplitude occurs [e.g. Jesus and Shadle, 2002]. Following Forrest et al. [1988], the spectral moments are computed over normalized spectra, treating them as random probability distribution.[3] See Forrest et al. [1988, p. 117] for example calculation formula. The first moment M1 represents the mean of the spectral energy distribution. Many studies used this parameter to distinguish English /s/ (high M1) from /ʃ/ (low M1) [e.g. Shadle and Mair, 1996; Jongman et al., 2000]. As a measure of the degree of dispersion of spectra around the spectral mean, M2 becomes smaller when the energy distribution is more compact. M2 is known to distinguish sibilants (low M2) from nonsibilants (high M2) [e.g. Shadle and Mair, 1996]. The third moment L3 is an index of skewness, which encodes whether the spectral energy is concentrated at higher (negative L3) or lower frequencies (positive L3). The fourth moment L4 is a measure of peakedness of the spectral peak relative to the spectral tail: a more peaked spectral peak gives rise to higher L4.

The parameter peak frequency F is correlated with the properties of the filter (i.e. the first resonance frequency of the front cavity); the longer the front cavity, the lower the spectral peak. The spectral mean M1 (also called center of gravity) works similarly, since the locus of the energy concentration depends on the length of the front cavity. Given its very short front cavity, these two parameter values of the dental /s/ are predicted to be much higher than other sibilants. In the acoustic study of Shona by Bladon et al. [1987], for example, /s/ has a much higher spectral peak (about 19 Bark) than /s̹/ and /ʃ/ (about 16 Bark for both). The same pattern is attested in languages where there is a similar three-way place contrast among sibilants. In an ensemble-average analysis of Polish sibilants [Nowak, 2006], for example, /s/ has an average spectral peak of

[3] Forrest et al. [1988] originally used L1–L4 to refer to each moment, but different studies have used different… abbreviations. Following the most recent study by Koenig et al. [2013], we use M1, M2, L3, and L4 to refer to each spectral moment. For more details of the history of the abbreviations, see Koenig et al. [2013, fn. 3].

8,000 Hz, while the spectral peak of other sibilants (i.e. /ʂ/ and /ɕ/) ranges between 3,500 and 5,500 Hz. The spectral mean of /s/ is also higher (about 6,500 Hz) than that of others (about 5,500 Hz). Likewise, in Mandarin Chinese, /s/ has an overall energy distribution at considerably high frequencies, having a spectral mean at about 9,000 Hz [Lee-Kim, 2011]. Although the spectral mean of alveolopalatal /ɕ/ in Mandarin Chinese is significantly higher (about 7,000 Hz) than that of the retroflex /ʂ/ (about 4,500 Hz), /ʂ/ is still closer to /ɕ/ than it is to /s/.

Just as these acoustic studies point to a general grouping of /ʂ/ and /ʃ/ (or /ɕ/) distinct from /s/, the previous perception studies have also shown that /ʂ/ and /ʃ/ are actually perceptually similar. While the salient noise spectral property was sufficient for the identification of the dental /s/, vocalic transitions play a crucial role in the identification of /ʂ/ and /ʃ/. In Nowak's [2006] perception study, Polish speakers identified the sibilant where the vocalic transition period was cross-spliced between /s, ʂ/ and /ɕ/. While the dental /s/ was correctly identified regardless of the mismatching transitions due to its strong cues in frication noise, the alveolopalatal /ɕ/ was often incorrectly identified as a retroflex /ʂ/, when it was not followed by proper transitions. Likewise, the discrimination study by Bladon et al. [1987] with Shona speakers showed that the choice of the whistled and palatoalveolar sibilant was heavily affected by the presence of the natural formant transitions: when the cues in the frication noise were ambiguous, listeners tended to make a decision relying on the transitional cues.

In agreement with the previous work which shows the perceptual similarities between /ʂ/ and /ʃ/, Xitsonga whistled fricative is impressionistically very similar to /ʃ/ as well. Against this background, this study undertook a thorough examination of the defining acoustic characteristics of the whistled fricative /ʂ/ from /ʃ/. In order to capture presumably subtle but reliable differences across speakers, data from multiple speakers are necessary. We therefore supplemented our ultrasound data with acoustic recordings from 4 Xitsonga speakers collected in the field. In addition to the acoustic parameters mentioned earlier, we further examined two additional parameters: dynamic amplitude A_d and formant transitions.

A prediction about spectral peak and mean is not clear from a comparison of the previous studies on different languages. For example, the retroflex and alveolopalatal sibilant in Polish are reported to be similar on these values [e.g. Nowak, 2006], whereas the retroflex in Mandarin has much lower values than the alveolopalatal sibilant [e.g. Lee-Kim, 2011]. It might be that the Mandarin retroflex has a larger sublingual cavity than the Polish retroflex sibilant [Ladefoged and Maddieson, 2006, p. 155], which would move the general energy distribution of the Mandarin retroflex to lower frequencies. Therefore, we remain neutral about our prediction of these two parameters for the comparison between Xitsonga whistled fricative and palatoalveolar fricative; these will be discussed in the results as post hoc analyses, as will the general properties of other spectral moments, without making at this point a priori predictions between the two Xitsonga sibilants.

The parameter dynamic amplitude (A_d) is defined as the difference in amplitude between the spectral peak and the spectral trough that occurs between the cutoff frequency (i.e. 500 Hz) and the spectral peak. This parameter reflects the strength of the noise source, i.e. an index of sibilancy [Jesus and Shadle, 2002]; the more strident the sound is, the higher the A_d. Studies showed that this parameter successfully distinguishes sibilants from nonsibilants in multiple languages [Shadle and Mair, 1996; Jesus and Shadle, 2002]. In addition to across-segment comparison, it has also been shown that a parameter similarly defined can be used to show the change in the strength of sibilancy over time in the production of English /s/ [Koenig et al., 2013]. Therefore, we conjecture that this parameter would reveal some systematic differences between the two sibilants. Assuming a retroflex articulation for the whistled fricative, it is likely that it has a stronger and more localized source than the palatoalveolar fricatives. That is, the former has an apical constriction that has a smaller constriction area, whereas the latter has a long palate channel with a larger constriction area. Assuming no systematic differences in volume velocity between the two sibilants, a smaller constriction area in the production of the whistled fricative should give rise to stronger source energy. Therefore, we predict that A_d is higher for the whistled fricative than for the palatoalveolar fricative.

In addition to the spectral properties of frication noise, we predict systematic differences between two sibilants with respect to formants and their transitions in surrounding vowels, because of coarticulation between the consonants and vowels. That is, due to the fronted tongue body of the palatoalveolar fricative, the adjacent low back vowels are likely to be produced with less retracted tongue and thus to show higher F2 and F3. In contrast, the retracted tongue back of the retroflex consonants [Hamann, 2003] would have a coarticulatory effect mostly to the adjacent high front vowels and thus lower their F2 and F3 values.

In addition to investigating the distinctive acoustic characteristics of the whistled fricatives, we examined the magnitude of the whistle in the whistled fricative by visually inspecting noise spectra [Shadle, 1983; Shadle and Scully, 1995; for whistled fricatives Shosted, 2011]. According to Shadle [1983, p. 2010], a whistle occurs when oscillation in the source spectrum is stabilized through coupling with the resonance frequency of the cavity. Since the coupling between the source and resonator amplifies the resonance frequency, a whistle is typically characterized by a narrow high-amplitude peak. Previous studies have observed a regular whistle for the whistled fricatives in other languages [Bladon et al., 1987 for Shona; Shosted, 2011 for Changana]: here we empirically examine the presence of the whistled peak in the Xitsonga whistled fricative.

1.4 Summary

In summary, the goals of the current study are twofold. First, in order to have a fuller picture of the complex interplay between tongue tip, middle and back, we use ultrasound imaging of the hypothesized retroflex gesture to examine the articulation of the whistled fricative in Xitsonga. If the whistled fricative indeed has a retroflex lingual articulation, we expect to see retracted tongue back, lowered tongue middle, and raised tongue tip/blade. We also use video recordings to verify the role of the labial gesture for the creation of whistle. Second, in order to single out characteristic acoustic features of the whistled fricative consistently present across multiple speakers, we use multitaper spectral analysis to ensure reliable spectral estimates in order to investigate multiple acoustic parameters, including peak frequency, spectral moments, and dynamic amplitude. Assuming a retroflex lingual gesture for the whistled fricative, we predict that dynamic amplitude would be greater for the whistled than for the palatoalveolar fricative due to the more localized apical constriction of the retroflex articulation. In addition to the differences in spectral properties, it is also predicted that the whistled fricative, compared with the palatoalveolar fricative, would lower the formant values of the surrounding vowels due to its retracted tongue back position. Based on the articulatory and acoustic results, we discuss aerodynamic models for potential subtypes of whistled sounds and offer an accordingly appropriate IPA transcription of the whistled fricative in Xitsonga.

Table 1. Target stimuli used in experiment 1

	Target words in IPA	Xitsonga orthography [Cuenod, 1967]	Gloss
/a/	[sàngù]	*sàngù*	‘sleeping mat’
	[ʂàtà]	*swàtà*	‘fall or dive into’
	[ʃàndzùkà]	*xàndzuka*	‘to abandon one’s family’
/i/	[sìlà]	*sìlà*	‘to grind on a stone’
	[ʂìŋwè]	*swìn’wè*	‘together’
	[ʃìnàmù]	*xìnàmù*	‘procrastination’
/u/	[sùsà]	*sùsà*	‘to take away’
	[ʂúkútá]	*swúkuta*	‘to chase away’
	[ʃùvùr̤ù]	*xùvùrhù*	‘uncircumcised male’

2 Articulation

In the articulatory study, we first examined the lingual gesture of the whistled fricatives using ultrasound imaging. Subsequently, a video recording was made to investigate labial gestures during the articulation of the whistled fricatives.

2.1 Methods

2.1.1 Participant

A female speaker of Xitsonga in her twenties participated in the ultrasound and video-recording study. Both of her parents are Xitsonga speakers from Mhinga, Limpopo in South Africa, and the speaker has lived all her life in Mhinga. She completed a college education in Xitsonga studies. She speaks English and has limited knowledge of Venda, another major language in the northeastern Limpopo province. English was used for communication during the experiment. The speaker reported no history of speech or hearing disorders.

2.1.2 Stimuli

The stimuli for the articulatory study included nine words that begin with three sibilants, /s, ʂ, ʃ/, before three different vowels, /a, i, u/. The target words had two or three syllables carrying low tone, except for one word with high tones (ʂúkútá). All the stimuli were chosen from the dictionary of Xitsonga [Cuenod, 1967], and the speaker confirmed that these were existing words in Xitsonga. Table 1 presents the target stimuli in both Xitsonga orthography and IPA.

2.1.3 Procedure

2.1.3.1 Ultrasound Imaging

The speaker was seated in a comfortable pose in a sound-attenuating booth at the Phonetics and Experimental Phonology Laboratory at New York University. The speaker’s head was then fit to a moldable head stabilizer (Comfort Company) on the wall, and her head was further stabilized with a Velcro strap [Davidson and De Decker, 2005; Davidson, 2006]. When used together with an ultrasound transducer that is fixed with a Bogen-Manfrotto Magic Arm, this method ensures that the same sagittal plane of tongue is imaged [Davidson, 2012]. The transducer was placed under the speaker’s chin, and adjusted until clear midsagittal images were captured. The participant was first asked to swallow water to extract the palate image [Epstein and Stone, 2005]. A list of nine target words plus nine fillers was presented seven times in random order. The speaker was asked to read the words in the following carrier sentence: [nì tìrìsà X kàŋwè] ‘I use X again’.

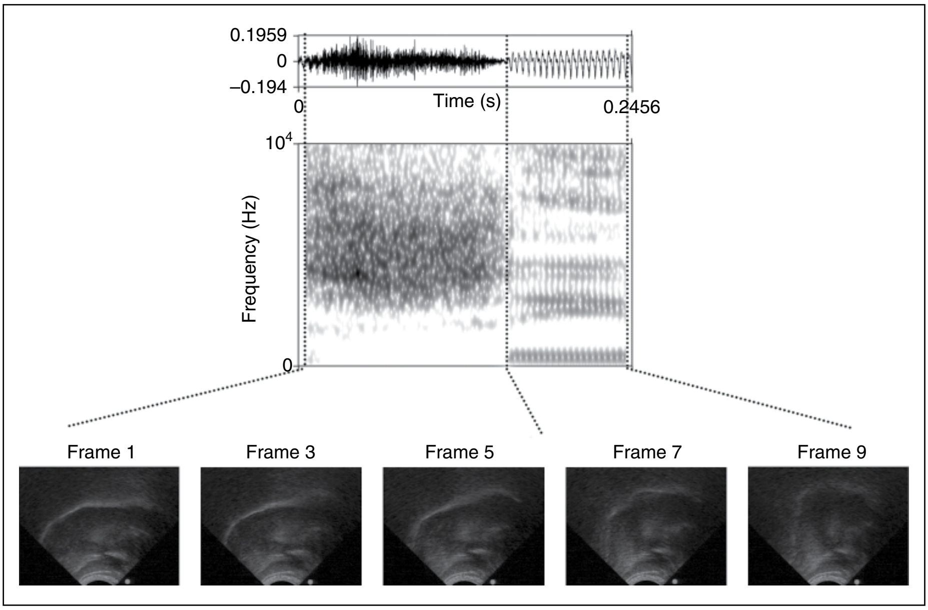

Fig. 1. A sample waveform and spectrogram of [ʂi] (top) and ultrasound frames obtained during the acoustic noise period (frames 1–5) and vocalic period (frames 6–9, bottom). Only odd-numbered frames are presented in this figure due to space limitation. In ultrasound images, the tongue back appears on the left, and the tongue tip/blade on the right. A method for visualizing and quantifying tongue shape is explained in the text.

Tongue images were recorded using a Sonosite Titan portable ultrasound, using a 5–8 MHz Sonosite C-11 transducer with a 90° field of view set at a depth of 8.2 cm. The frame rate of the ultrasound was 29.97 frames/s: one frame was captured approximately every 33.4 ms. The audio signal was collected using an Audio Technica AT-813 microphone and was synchronized with the video signal from the ultrasound machine using a Canopus ADVC-1394 capture card and Adobe Premiere. The audio files were then extracted from the movie file.

The boundaries of two sibilants were identified as the beginning and end of aperiodic noise in the waveform. The presence of F1 and F2 was used to locate the onset and offset of the vocalic segments. Praat [Boersma and Weenink, 2012] was used to identify acoustic landmarks for segmental boundaries. The tongue images captured during the acoustic realization of the target sibilant and the following vowel were extracted using Matlab. Figure 1 shows an example of the ultrasound frames extracted from the corresponding acoustic signal.

Among the frames extracted during the production of the entire syllable, the frame that shows maximal consonantal constriction was chosen for subsequent statistical analyses. For the whistled fricative /ʂ/, a clear tongue tip/blade movement toward the alveolar ridge and its release was consistently observed. Thus one frame before the release of the tongue tip/blade was chosen as the frame containing maximal constriction of /ʂ/. For the palatoalveolar fricative /ʃ/, its release was observed as slight tongue body lowering from the palate, and one frame before this movement was chosen as the maximal constriction frame. For the dental fricative /s/, it was not trivial to select the frame in which the dental articulation reaches its maximal constriction, because in ultrasound imaging the shadow of the jawbone obscures part of the tongue tip especially when the tongue is elevated [Stone, 2005]. However, lowering of the tongue just in back of the tip was consistently observed as its release, and one frame before this lowering movement was chosen as the frame containing maximal dental constriction. Although the very front of the tip gesture was

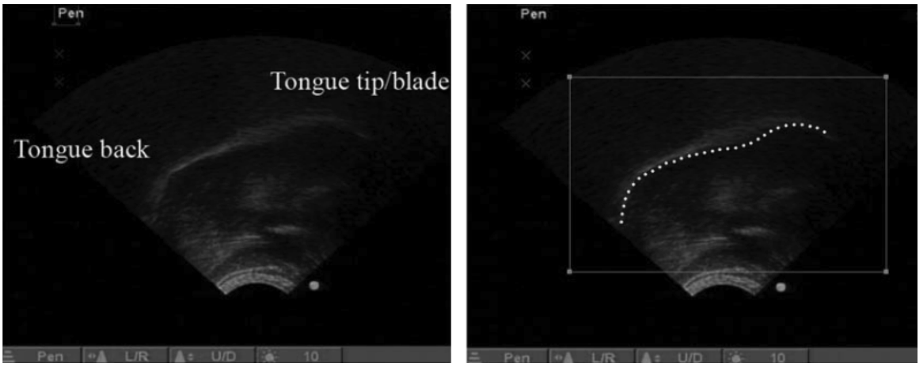

Fig. 2. A screenshot of the ultrasound image of the frame with maximal tongue tip/blade constriction (frame 5 in fig. 1) during the articulation of /ʂ/ followed by /i/ (left), and a screenshot of EdgeTrak (right). The white dots are the edge of the tongue found by the automated edge-tracking algorithm. The tongue tip is on the right and the tongue back is on the left.

not still captured, the exact tongue tip gesture of the dental fricative is not the main interest in our study. Because the release of the consonantal gestures can occur after the vowel gestures begin, some frames were taken at the end of the frication period or at the beginning of the following vowel. Given the inherent mismatch between articulatory and acoustic events, using the articulatory landmark was considered to provide more consistent tongue configurations across multiple tokens.

The selected frames were then fed into EdgeTrak [Li et al., 2005], which automatically tracks tongue configuration by extracting x-y coordinates of the target region from the upper edge of the tongue. One hundred equidistant points of the frame with maximal consonantal constriction were then extracted for the statistical analysis. In addition, the tongue curves of a series of frames during the production of the sibilants were extracted to visualize how the tongue movement proceeds. A screenshot of the tongue curve tracking in EdgeTrak is shown in figure 2 (right-hand image).

To assess the tongue configuration of the three fricatives, a smoothing spline ANOVA (SS ANOVA) was employed. SS ANOVA is a statistical procedure for investigating similarities and differences between different curves [Gu, 2002; Davidson, 2006; Wang, 2011]. It returns parameter values for the smoothing splines that show a best fit for all of the data at once and for the spline of the interaction, which represents the difference between the main effect splines and the spline that best fits all of the data. Since the statistical tests on the parameter values only indicate that there is a significant difference between curves in some area (but not exactly where), a more informative technique for testing statistical significance is to construct 95% Bayesian confidence intervals around the smoothing splines. The curves are significantly different where the confidence intervals do not overlap. As this use of confidence intervals has become the standard for the SS ANOVA analysis in the field [e.g., Davidson, 2006; Simonet et al., 2008; Chen and Lin, 2011; Mielke et al., 2011; De Decker and Nycz, 2012; Zharkova et al., 2012; Kochetov et al., 2013; Lee-Kim et al., 2013], we used the confidence intervals to determine significant differences between curves. SS ANOVA was implemented using the gss package in R [Gu, 2012].[4]

In addition to the statistical testing, we present the tongue position of each fricative relative to the palate. This was accomplished by superimposing the palate image on top of each ultrasound image. The palate image was captured while the participant swallowed water and traced using EdgeTrak [Epstein and Stone, 2005].

[4] The R source code can be found at https://files.nyu.edu/ld43/public/PEPLab/resources.html.

2.1.3.2 Video Recording

For the recording of the labial gesture, the speaker was seated in a chair in a quiet room in Limpopo, South Africa.[5] Following Shosted [2011], the speaker was asked to hold a hand-mirror on the left side of her lips to examine lip protrusion as well as lip rounding. A Sony PAL DCR-SX21 digital video camera recorder was mounted on a tripod and located about 2 feet from the speaker. The frame rate of the video camera was 25 frames/s, with one frame captured every 40 ms. The video file was saved in MPEG format. The audio signal was simultaneously captured by the microphone built into the video camera and recorded as 16-bit audio with a 44 kHz sampling rate. The wave files were extracted from the video files using Adobe Premiere. The audio files were then analyzed in Praat. The acoustic boundaries for the fricatives were segmented following the same criteria used for the ultrasound imaging. The corresponding video frames of the fricatives were then extracted using Adobe Premiere. Among the 4–6 frames extracted during the acoustic realization of the fricatives, the middle frame was chosen for presentation of the results. For even-numbered frames, the earlier frame of the two middle frames was chosen. The influence of coarticulation with the following vowel was evident after this middle frame; for instance, the lips are opening for the low vowel /a/ or rounding for the rounded vowel /u/.[6]

2.2 Results

2.2.1 Lingual Data

Figure 3 shows the smoothing splines based on the seven repetitions of each sibilant /s, ʂ, ʃ/ at their maximal constrictions. The 95% Bayesian confidence intervals are marked by the shadings that are superimposed on each smoothing spline. Following Davidson [2006], we partitioned the tongue contour into thirds to assess the relative tongue shape with reference to individual articulators. The resulting regions roughly correspond to the tongue back, tongue middle, and the tongue tip/blade. This procedure is only a rough approximation for defining articulator boundaries. Reliability to some extent depends on the field of view and angle of the transducer. That is, if the tongue root is more completely imaged, the division of the tongue would be changed accordingly.[7] Nevertheless, we focus on the majority pattern within a single region. For example, the tongue curves cross within the first third in which the back of the tongue is displayed in figure 3 (right-hand column), and in this case, we interpret the results

[5] The ultrasound and video data were collected separately because of practical limitations.

[6] An anonymous reviewer raised a concern about using different temporal points for taking the lingual and labial measures; the selected ultrasound frame of maximal constriction was usually taken near the end of the acoustic period of the fricative or at the beginning of the following vowel, whereas the labial video frame in the middle of the acoustic period of the fricative was selected. We have two reasons for this choice. The first reason is empirical: as noted earlier, the consonantal articulation could persist through the end of the acoustic period of the consonants, whereas the anticipatory vowel gesture was found much earlier than the end of the acoustic period of the consonant. Second, although the time point was different for the two measures, our choice is principled in a way that we chose the frame that shows the maximal constriction for each lingual and labial articulation for the fricatives. It might be that this inherent mismatch in timing between lingual and labial gestures is due to the different level of coarticulatory inhibition. That is, the vocalic (tongue body) gesture might not be able to start early because of the consonantal gesture, whereas the labial gesture is free from this pressure and starts as early as it could. We believe that this entire issue merits its own research topic, and leave it for future study.

[7] Indeed, one might think that the vertical lines in figure 3 (b and c especially) should be moved backward slightly, so that the raised tongue front does not form a part of the tongue middle. While acknowledging this possibility, we do not implement this method, due to the lack of an objective principle to divide the different articulators.

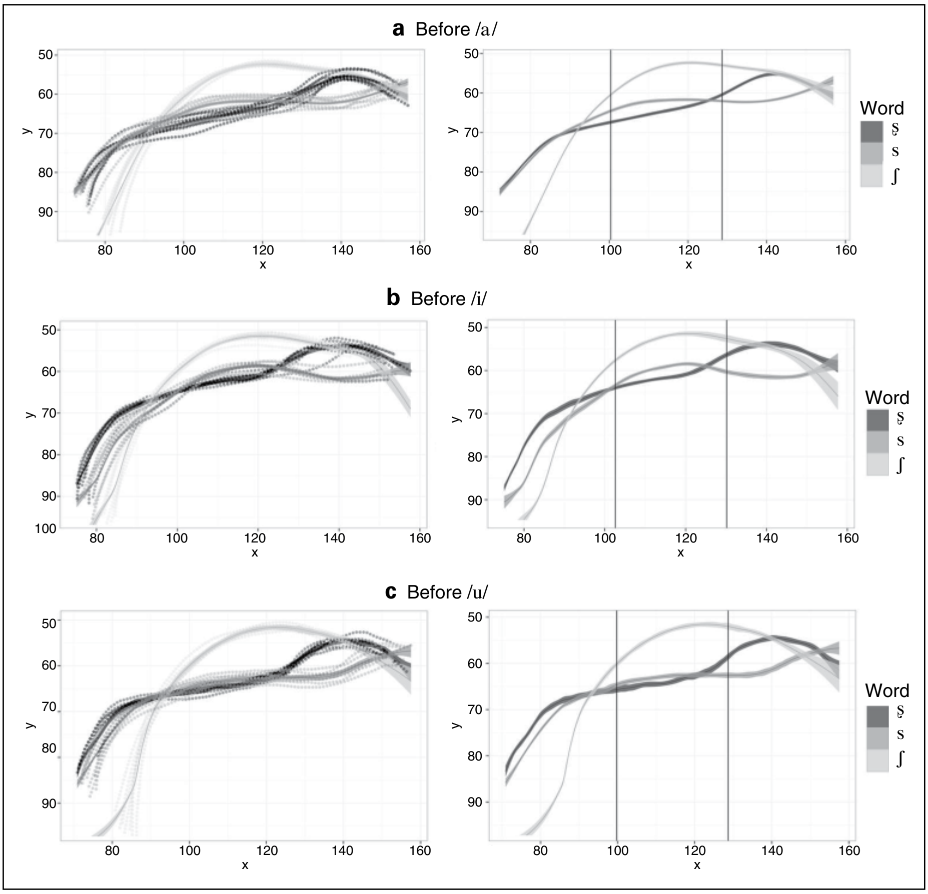

Fig. 3. Raw data points from seven repetitions and the smoothing spline estimate (solid lines) for comparison of the tongue shape of /s/, /ʂ/ and /ʃ/ in all vowel contexts (left-hand column). Smoothing spline estimate and 95% Bayesian confidence interval (right-hand column). The leftmost region in each image corresponds to the tongue back, the region in the middle corresponds to the tongue middle, and the rightmost region corresponds to the tongue tip/blade. The axes are in millimeters corresponding to the boxed-in region shown in figure 2 (right-hand column).

based on the majority pattern. While a more objective methodology may need to be developed, this method seems to be adequate for the current comparison where the data were not ambiguous.

The raw data points in figure 3 (left-hand column) show an articulatory stability among individual plots of the target sibilant. The partitions in figure 3 (right-hand column) show that the three sibilants have significantly different tongue shapes. Across almost all the contexts, the whistled fricative overall has the most retracted tongue back, the lowest tongue middle, and the highest tongue tip/blade. Exceptions include /s/ with similar degree of tongue back retraction, and /ʃ/ with similar degree of tongue tip/blade raising in the /a/ vowel context (fig. 3a, right-hand image). The results are consistent with our prediction that the whistled fricative has a retroflex lingual

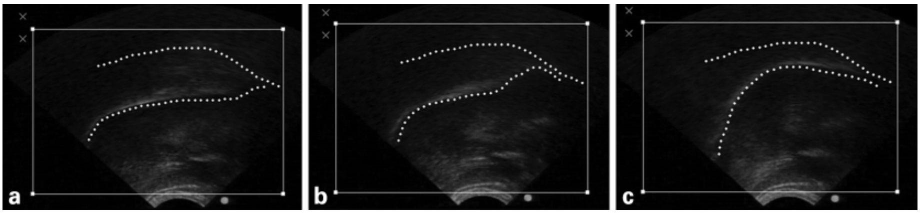

Fig. 4. The tongue shapes of /s/, /ʂ/ and /ʃ/ at maximal constriction in /a/ vowel context are reproduced with the palate trace shown on top of each figure.

articulation. Although also an alveolar consonant, the dental /s/ has a quite different tongue shape from /ʂ/ especially in the tongue tip/blade region: it is significantly lower than /ʂ/. Although ultrasound imaging cannot capture the very front part of the tongue tip due to the shadow of the jawbone, given the overall trajectory of the tongue front of the dental sound, it is reasonable to assume that the tongue tip would form a constriction at a more front area, not captured with current ultrasound imaging.[8] Lastly, the palatoalveolar fricative /ʃ/ has the most distinct tongue configuration from the other two sibilants, having the most fronted tongue back and the highest tongue body as its maximal constriction. It appears that the entire tongue front, including tongue tip and blade, makes a laminal constriction at around the alveolar ridge.

In order to examine the tongue positions relative to the palate, ultrasound images of each fricative at their maximal constriction are presented with the palate trace superimposed in figure 4. As in the SS ANOVA analysis, the relative difference in constriction site of /s/ and /ʂ/ stands out clearly: /s/ makes a constriction in front of the alveolar ridge, while /ʂ/ makes its constriction slightly behind the alveolar ridge. Again, /ʂ/ and /ʃ/ are similar with respect to the constriction site with the latter forming slightly more front constriction. However, the shape of the tongue front is quite distinct: an apical constriction by the tongue tip is prominent for /ʂ/, whereas a laminal constriction involving the entire tongue front is seen for the /ʃ/.

2.2.2 Labial Data

Figure 5 presents the labial gestures taken at the middle of the acoustic realization of each fricative in three vowel contexts. All the images are taken from the first repetition of the reading list. Figure 5 presents differences in the labial gesture between whistled and nonwhistled sibilants. Visual inspection suggests that vertical opening may be the most prominent gesture that distinguishes the whistled fricative from the other two fricatives. In the whistled fricative, the lower lip is raised, completely covering the lower teeth, while the lower teeth are still visible in nonwhistled fricatives for nonrounded vowel contexts. The upper lip of the whistled fricative is slightly raised, exposing the upper teeth more extensively, while in nonwhistled fricatives it stays in a neutral state, exposing the upper teeth only partially in nonrounded vowel contexts. The upper lip raising in the whistled fricative is to some degree retained even in the anticipation of the upcoming rounded vowel /u/ (fig. 5 on the bottom middle). The sagittal images (those on the mirrors) further confirm upper lip raising; the upper lip in the

[8] The degree of the tongue back retraction of /s/ varies the most among three fricatives depending on the vowel contexts, suggesting that /s/ is more susceptible to coarticulation with the surrounding vowels.

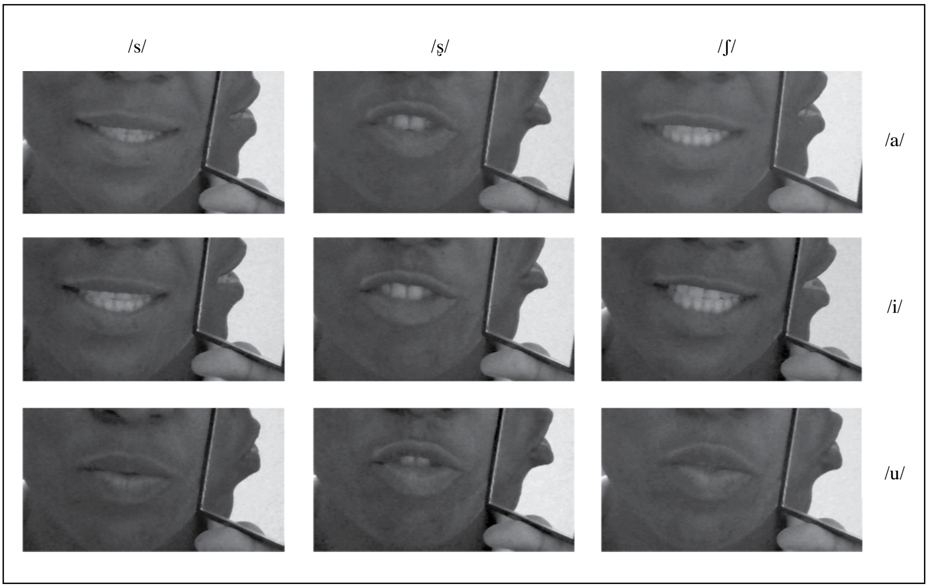

Fig. 5. Labial gestures during the articulation of the fricatives /s/, /ʂ/, and /ʃ/ in three vowel contexts /a/, /i/, and /u/. The frames presented here were captured from the center of the acoustic realization of each fricative. The sagittal images appear in the mirror, to the right of the face.

whistled fricative is slightly raised so that the edge of the upper lip is not completely captured within the current angle of the mirror. Overall, weak lip protrusion, if any, was found for the articulation of the whistled fricative, though horizontal narrowing was observed.

3 Acoustics

For the acoustic analysis of the whistled fricative, we first examined a three-way contrast among sibilants /s/, /ʂ/ and /ʃ/ using the acoustic recordings of the speaker who participated in the articulatory study. Subsequently, we examined the acoustic differences between /ʂ/ and /ʃ/ in further detail, using data of multiple speakers collected in the field.

3.1 Methods

3.1.1 Participants

For the three-way comparison among /s/, /ʂ/ and /ʃ/, the acoustic recordings of the speaker who participated in the ultrasound and video imaging were examined (see section 2). For the two-way comparison between /ʂ/ and /ʃ/, multiple speakers participated in elicitation: 2 female (C.B. and S.M.) and 2 male speakers (C.M. and H.M.). They were all in their twenties and were completing a college degree in Xitsonga studies at the time of recording. All of their parents are Xitsonga speakers and they lived all their lives in Xitsonga-speaking communities in Limpopo, South Africa. Their ordinary means of communication is Xitsonga, but they were also able to speak English and some Venda. Elicitation for the experiment was conducted in both English and Xitsonga, with the help of 2 Xitsonga research assistants. All data were collected in Limpopo, South Africa.

3.1.2 Stimuli

For the three-way comparison, the same stimuli used in the articulatory study were used for the acoustic recordings. As described earlier, the three sibilants occurred in three different vowel contexts, /a, i, u/. For more details about the speech material, see table 1. For the two-way comparison, however, we fixed the vowel environment with /i/, and contrasted a plural prefix [ʂi] with a singular prefix [ʃi]; the choice of this vowel was necessitated by the fact that we used a paradigm inflection method for elicitation, which worked best for Xitsonga speakers (see below). The data were extracted from a larger set of data collected from fieldwork by the third author, and the choice of the vowel context was practical rather than theoretical. The target prefixes were followed by the same stem noun, e.g. [ʃi]-témpé 'a stamper' (class 7) versus [ʂi]-témpé 'stampers' (class 8). There are no known phonological interactions between the prefix sibilant and stem-initial consonants. All target words were produced in the same frame sentence: [nì tìrìsà X kàŋwè] (where X = prefix + stem) 'I use X again'. All syllables preceding the target prefix as well as the prefix itself bore low tone and the following stem nouns varied in their tonal specification. The list of stem nouns is provided in the 'Appendix'.

3.1.3 Procedure: Elicitation in the Field

The acoustic recordings for the three-way comparison were made immediately after the video recording was made. The fieldwork with multiple speakers mainly consisted of translating English words to Xitsonga in a frame sentence. The speakers were first asked to provide a singular form of an English word (e.g. [ʃi]-témpé 'a stamper'), and then asked to provide a plural form of the corresponding singular form (e.g. [ʂi]-témpé 'stampers'). The order of the stem nouns was randomized during the elicitation. The speakers repeated the stimuli three times, although not all the repetitions were available because of occasional unexpected recording errors in the field. The acoustic analysis was based on the following numbers of tokens: speaker C.B.: /ʂ/ = 30, /ʃ/ = 30; C.M.: /ʂ/ = 21, /ʃ/ = 26; H.M.: /ʂ/ = 18, /ʃ/ = 17; S.M.: /ʂ/ = 43, /ʃ/ = 42. The elicitation took place in a quiet room. The speech signal was recorded at a 44k sampling rate and 16-bit quantization level using a Zoom H4 digital recorder. The microphone was a head-mounted unidirectional Shure WH-30. To avoid breath noise, the microphone was placed about 10 cm from the side of the lips.

3.1.4 Measurements

Following Iskarous et al. [2011], six acoustic landmarks were identified using Praat: onset and offset of each of the preceding vowel /a/, fricative, and the following vowel /i/.[9] The boundaries of fricatives were identified at the beginning and end of aperiodic noise in the waveform. The boundaries of both preceding and following vowels were also marked at the beginning and end of the periodic waveform.

For the multitaper spectral analysis, a 25-ms window of orthogonal taper functions was placed at the beginning, middle and end phases of each frication noise interval. Spectral estimates averaged over multiple tapers were then normalized to compute spectral moments. Following recent work [Jesus and Shadle, 2002; Lousada et al., 2012; Koenig et al., 2013], a low-frequency cutoff was applied at 500 Hz. Because the target sibilants were positioned intervocalically, partial voicing assimilation occasionally occurred even during the middle phase of the voiceless fricatives. The low-frequency cutoff was thus used to eliminate the effect of voicing. All the processes including the following measures were implemented in Matlab 2011 unless otherwise noted.

[9] The onset or offset of frication noise is not always perfectly aligned with the onset or offset of the surrounding vowels. There are slight temporal gaps between the fricatives and the vowels, and in order for complete acoustic analysis of the beginning and end phases of frication noise, we added two acoustic landmarks at the fricative-vowel juncture (e.g. one for frication offset, and the other for vowel onset). This segmentation procedure follows that of Iskarous et al. [2011]. No acoustic measurement was made during the interval between the frication offset and the vowel onset.

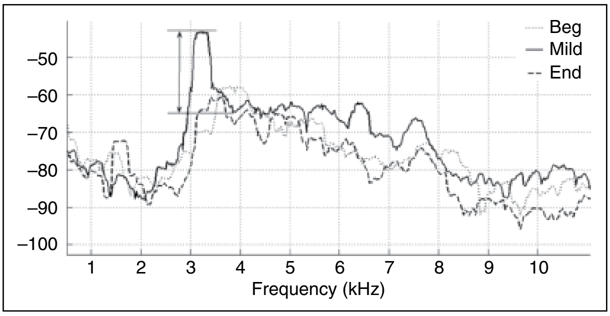

Fig. 6. An illustration of a whistled peak produced by speaker S.M. An unambiguous whistled peak is observed for the mid phase of this token: the narrow part of the high-amplitude peak has an amplitude extent greater than 15 dB (indicated with an arrow).

For the three-way comparison among /s/, /ʂ/, and /ʃ/, noise spectra were quantified according to spectral peak F and spectral moments. Recall that first, spectral peak F is defined as the frequency where the maximum amplitude occurs [e.g. Jesus and Shadle, 2002]. This parameter is associated with the first resonance frequency of the front cavity; a longer front cavity results in a lower spectral peak. Second, following Forrest et al. [1988], four spectral moments were computed over the normalized spectra: (i) mean (M1), (ii) variance (M2), (iii) skewness (L3), and (iv) kurtosis (L4).

For the two-way comparison between /ʂ/ and /ʃ/, we included two additional measures: dynamic amplitude A_d and formant values of the surrounding vowels. The parameter dynamic amplitude (A_d) is defined as the difference in amplitude between the spectral peak and the spectral trough that occurs between the cutoff frequency (i.e. 500 Hz) and the spectral peak. This parameter is known to reflect the strength of the noise source: i.e. as an index of sibilancy, the more strident sound shows higher A_d values [Jesus and Shadle, 2002]. Formant frequencies were measured by creating a 20-ms window centered at the midpoint of the vocalic intervals.[10] Onset and offset formant values were likewise calculated by creating a 20-ms window, located 10 ms from the boundaries between the fricatives and vowels. The maximum frequency for formant calculation was set to 5 kHz for male speakers and 5.5 kHz for female speakers. Formant values were extracted using the Berg algorithm in Praat.

Apart from investigating the quantitative differences in spectra among contrastive sibilants in Xitsonga, we tested the presence of a whistle in the whistled fricative by inspecting noise spectra [Shadle, 1983; Shadle and Scully, 1995; for whistled fricatives, Shosted, 2011]. A whistle occurs when oscillation in the source spectrum is stabilized through coupling into the resonance frequency of the cavity. Since the coupling between the source and resonator amplifies the resonance frequency, a whistle is typically characterized by a high-amplitude narrow-bandwidth peak [Shadle, 1983, 2010].

However, there has been no established method for unambiguously identifying a whistled peak, and thus in practice, identifying a whistled peak from spectral shape is not always straightforward. After an iterative examination and cross-comparison of spectra, we decided to use an unambiguous whistled peak as a reference point. Figure 6 shows the representative spectra of 1 speaker (S.M.), in which the mid phase of frication noise shows a whistled peak. Figure 6 represents a clear case of whistling, because (i) the peak is high-amplitude and narrow-bandwidth, and (ii) it occurs at a predicted frequency: a whistle couples into the first resonance frequency of the cavity [Shadle, 1983], and this high-amplitude narrow peak occurs precisely at the frequency that other nonwhistled peaks resonate (compare with the beginning and end phases of this token in fig. 6). As a rough approximation,

[10] For the vowels preceding the target fricatives, the preceding consonant was fixed as [s] as a part of a frame sentence. For the vowels following the target fricatives, the following consonants do vary across different items, but the two target fricatives were followed by the same set of consonants, thereby controlling for the potential effect of place of articulation differences. Since we used an inflection paradigm to elicit the current data, it was necessary to use real words, which made it impossible to control for the place of the consonants after the following vowel. We would like to leave it for a future study to address this issue, as it requires a new experiment using nonce words as stimuli.

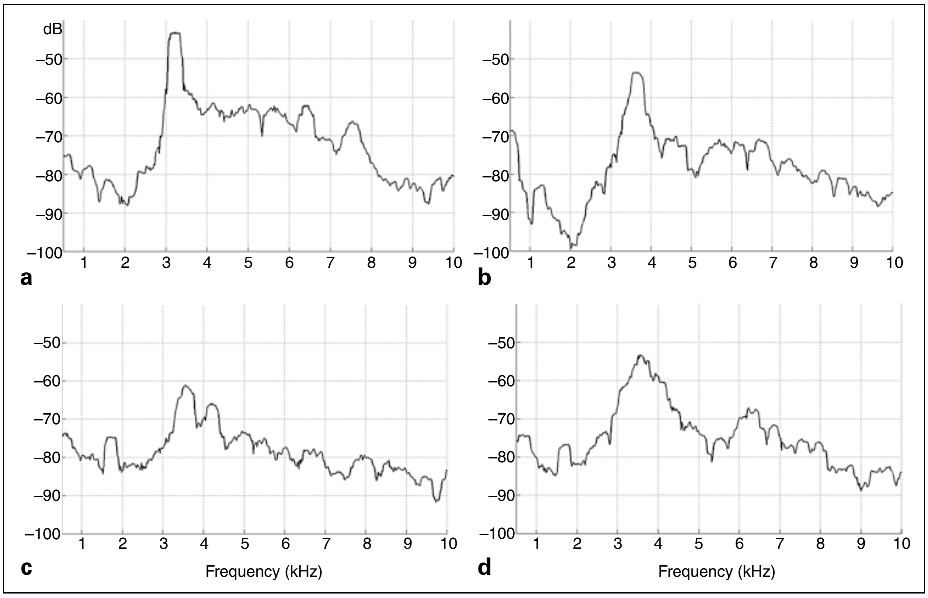

Fig. 7. Examples of the coding of whistling: S.M.'s whistling peak (**a**), C.M.'s whistled peak (**b**), H.M.'s nonwhistled peak (**c**), and C.B.'s nonwhistled peak (**d**).

therefore, the criterion by which a whistle is defined is 'a peak with a bandwidth smaller than 1 kHz and this narrow peak has an amplitude extent greater than 15 dB'. Peaks that fall outside of this criterion are considered to be nonwhistled. All the coding was done manually.[11]

Figure 7 presents some examples of this manual coding process. In figures 7a, b, the highest peaks fall within this criterion and thus are coded as whistled. In contrast, in figures 7c, d, the peaks are coded as nonwhistled, since the amplitude of the narrow part of the main peak was either not greater than 15 dB (fig. 7c), or the bandwidth of the peak was greater than1 kHz (fig. 7d).

3.2 Results

3.2.1 A Three-Way Comparison between /s/, /ʂ̡/, /ʃ/

Figure 8 shows representative multitaper spectra of the three sibilants in the /a/ vowel context taken at the mid phase of frication noise. The spectrum of the dental /s/ stands out most distinctively from that of the other two sibilants. While most energy distribution is concentrated at lower frequencies (around 4–5 kHz) for both /ʂ̡/ and /ʃ/, that of the /s/ spectrum is spread over a wide frequency range and concentrated at a much higher frequency (above 10 kHz). In comparison, the spectra of /ʂ̡/ and /ʃ/ appear to be quite similar to each other. Although the spectral peak may be more

[11] This method is based on our post hoc inspection of the data and does not admittedly have independent or principled motivations. Our method is thus inductive and exploratory. It has a virtue of at least being objective. However, it is also tentative. It is hoped that our coding method based on examination of empirical data will contribute to the establishment of a more objective, automatic method, but doing so with the current set of data is beyond the scope of this paper.

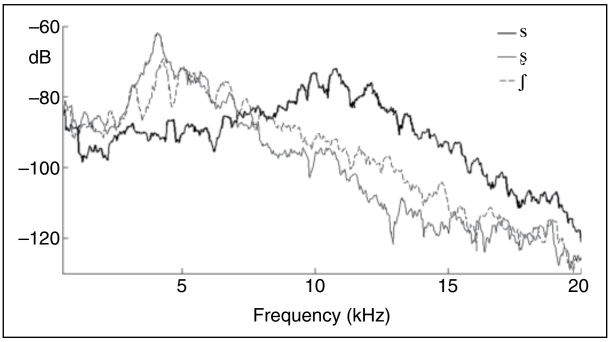

Fig. 8. Representative multitaper spectra of three fricatives /s, ʂ, ʃ/ in the /a/ vowel context taken at mid phase of frication noise.

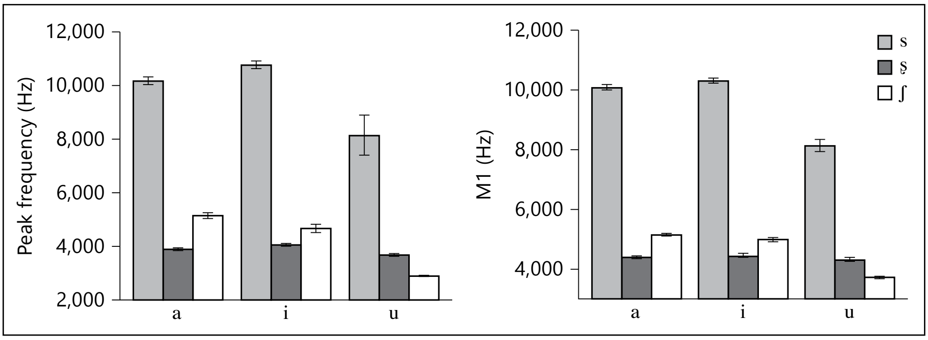

Fig. 9. Mean peak frequency F (Hz) and spectral mean M1 (Hz) of the three sibilants /s, ʂ, ʃ/ in three vowel contexts /a, i, u/. Error bars indicate standard error.

peaked for the /ʂ/ than for the /ʃ/, they are very similar in terms of the overall energy distribution. The observation is reflected in two parameters: spectral peak F and spectral moment M1. Figure 9 presents the mean values of F and M1 for each fricative and vowel context.

To examine the effect of the sibilant type on peak frequency, a linear regression model was implemented using the *lm* function in R. The independent variables were the sibilant type (three levels: /s, ʂ, ʃ/ with /ʂ/ as the baseline) and the vowel type (three levels: /a, i, u/ with /i/ as the baseline). The dependent variable was the peak frequency and spectral moment M1.

Results of both peak frequency F and spectral mean M1 showed significant main effect of the sibilant type. For the sibilant type, /s/ had significantly higher F and M1 than /ʂ/ (F: t = 10.3, $p < 0.001$; M1: t = 15.4, $p < 0.001$), but /ʃ/ did not differ from /ʂ/ (F: t = 0.9, n.s.; M1: t = 1.1, n.s.). While almost all interaction terms were nonsignificant, there was a significant interaction between the sibilant type (/s/ vs. /ʂ/) and the vowel type (/u/ vs. /i/) (F: t = –2.1, $p < 0.05$; M1: t = –3.1, $p < 0.05$), indicating that both F and M1 of /s/ were lowered in /u/ significantly more than that of the baseline /ʂ/. Additionally, the lowering of spectral peak F for /s/ in /a/ was significantly different than that of /ʂ/ (F: t = 2.0, $p < 0.05$). The statistical results suggest that /s/ has a significantly higher spectral peak and mean than the other sibilants, and that it is more vulnerable to coarticulation with the surrounding vowels.

Figure 10 summarizes the results of the other parameters. Statistical results regarding these parameters are not reported due to space limitation, but /s/ spectra showed unique differences, especially with regard to M2 (variance) and L3 (skewness). The pattern reflects a more diffuse energy distribution in the dental spectra [s] (i.e. high M2), and the spectral energy concentrated at higher frequencies (negative L3). L4, on the other hand, did not present any particular pattern for /s/ distinct from other sibilants.

3.2.2 A Two-Way Comparison between /ʂ̡/ and /ʃ/

Having established that /s/ stands out among the three sibilants, we now zoom in on the subtle differences between /ʂ̡/ and /ʃ/ based on the data obtained in the field. Figure 11 presents multitaper spectra of ten tokens of mid-phase /ʂ̡/ and /ʃ/ for each of the 4 speakers. Overall, the highest spectral peak occurs between 3.5 and 4 kHz for both sibilants, and the overall energy distribution is similar. A prominent second peak is found at around 7 kHz in speaker C.B.'s production of both sibilants, but the second peak is not consistently found across all speakers.

3.2.2.1 Spectral Peak Frequency F

Figure 12 presents the distribution of the spectral peak frequency of mid-phase noise for each speaker, and table 2 summarizes the statistical results of all three phases of frication noise. To examine the effect of the sibilant type on peak frequency, a linear mixed effect (LME) model was implemented using the *lmer* function in the *lme4* package [Bates and Maechler, 2009] in R. The fixed factor was the type of the sibilant (two levels: /ʂ̡/ or /ʃ/ with /ʃ/ as the baseline). The dependent variable was the peak frequency (Hz). Subjects and items were treated as random factors. Since there are no uncontroversial methods for obtaining p values for the mixed effect models due to the difficulty in computing degrees of freedom [Baayen, 2008], we report the p values directly estimated from the t scores.[12]

The statistical results showed that the two mid-phase sibilants are not different in location of peak frequency (n.s.). However, the estimated peak frequency of the whistled fricative was significantly higher than that of the palatoalveolar both at beginning ($p < 0.0001$) and end ($p < 0.01$) phases.[13]

3.2.2.2 Spectral Moments

Figure 13 presents the means of four spectral moments at the mid phase of frication noise for each speaker, and table 3 summarizes the statistical results of the spectral

[12] The R command is as follows: round(2* (1-pt(abs(coef(summary(x))[,3]), Inf)), where x is the model.

[13] This is partly because of the unusual development of noise in the whistled fricative; the estimated peak frequency was the highest at the beginning with a gradually decreasing trend over time, i.e. estimated peak frequency of /ʂ̡/: 4,222.6 (beginning) > 3,903.2 (middle) > 3,854.8 Hz (end). This is somewhat unexpected given the documentation in the literature in which the highest peak frequency is often found at the mid phase of noise spectra because of the lack of coarticulation with surrounding segments [e.g. Koenig et al., 2013]. While a comprehensive investigation is beyond the scope of this article, we conjecture that this unusual behavior of the whistled fricative might be due to the dynamic properties of the retroflex gesture. That is, at the beginning of the frication, the tongue tip/blade may be making an alveolar constriction while apical retraction for the full retroflex gesture is still under way, giving rise to a small front cavity and high resonance frequency.

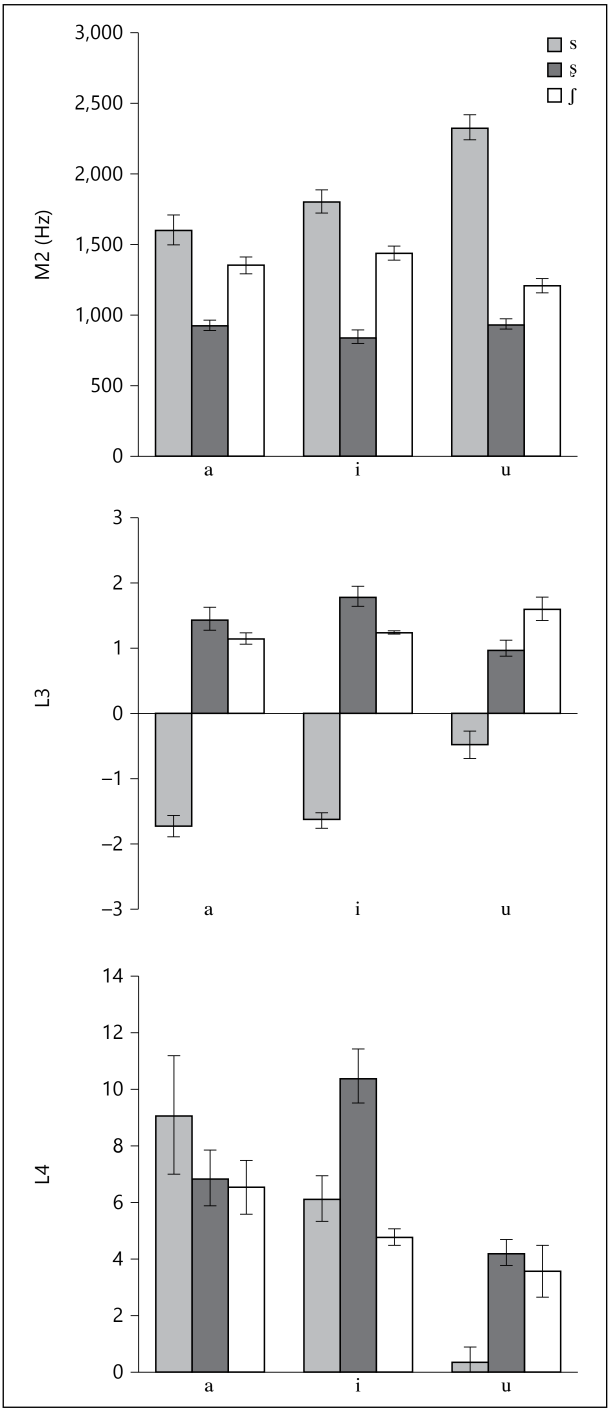

Fig. 10. Mean spectral moment M2, L3, and L4 of the three sibilants /s, ʂ, ʃ/ in three vowel contexts /a, i, u/. Error bars indicate standard error.

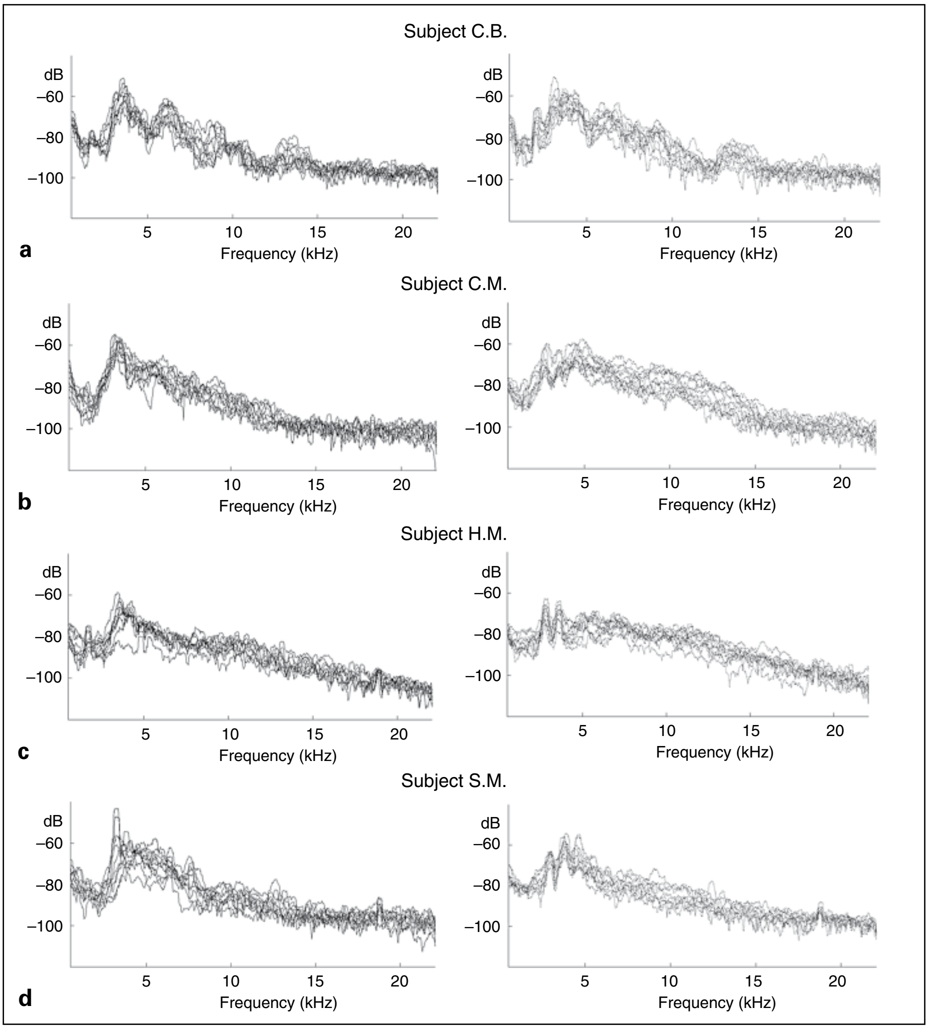

Fig. 11. Multitaper spectra of ten tokens of /ʂ/ (left-hand column) and /ʃ/ (right-hand column) of 4 speakers taken at mid phase of frication noise.

moments in all three phases. The results show that none of the M1 (mean), L3 (skewness), and L4 (kurtosis) is significantly different between the two sibilants. However, M2 (variance) is significantly lower for the whistled fricative than for the palatoalveolar fricative, indicating that the spectra of the palatoalveolar are less compact than that of the whistled fricative. The results were consistent across all phases of frication noise.

3.2.2.3 Dynamic Amplitude

Figure 14 summarizes the results of dynamic amplitude of two sibilants at the mid phase of frication noise. The estimated dynamic amplitude of the whistled fricative is significantly higher by 6.4 dB than that of the palatoalveolar fricative ($p < 0.001$),

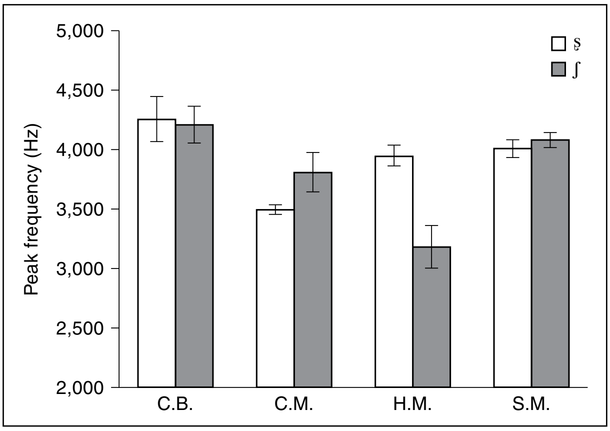

Fig. 12. Mean of the spectral peak frequency (Hz) at mid phase of two sibilants for each speaker. Error bars indicate standard error.

Table 2. Results of the mixed-effect linear regression for the peak frequency of the three phases of the target sibilant

		Estimate	SE	t value	p
BEG	intercept (/ʃ/)	3,851.3	163.8	23.5	
	sibilant: /s̨/	371.7	108.8	3.4	<0.001*
MID	intercept (/ʃ/)	3,792.0	290.3	13.1	
	sibilant: /s̨/	111.2	190.2	0.6	0.5586
END	intercept (/ʃ/)	3,457.5	229.0	15.1	
	sibilant: /s̨/	397.3	146.3	2.7	<0.01*

An asterisk indicates a significant result.

indicating that the peak is higher in amplitude for the whistled fricative than for the palatoalveolar fricative. Table 4 summarizes the statistical analyses of dynamic amplitude of the two fricatives. The results were consistent across all phases of frication noise.

3.2.2.4 F2 and F3

Figure 15 presents mean F2 and F3 values and their transitions averaged over 4 speakers in the preceding and following vowels /a/ and /i/ (recall that the target sibilants are followed by /i/, because the two prefixes examined here were /s̨i/ versus /ʃi/, and the preceding word in a frame sentence ends with /a/). Table 5 shows a summary of the statistical analysis. F2 values were significantly higher next to /ʃ/ than next to /s̨/ (all $p < 0.05$) at all four acoustic landmarks (i.e. vowel midpoints and onset/offset). In addition, steep F2 transitions of the palatoalveolar stand out. F3 values patterned similarly: they were significantly higher next to /ʃ/ than next to /s̨/ at all points measured (all $p < 0.05$).

3.2.3 Whistling Peak

Table 6 summarizes the overall percentage of whistled peaks for each speaker. One speaker showed whistling 20% of the time, the other 2 only 7%, and 1 last speaker showed no whistling. As these low numbers indicate, whistled peaks occurred

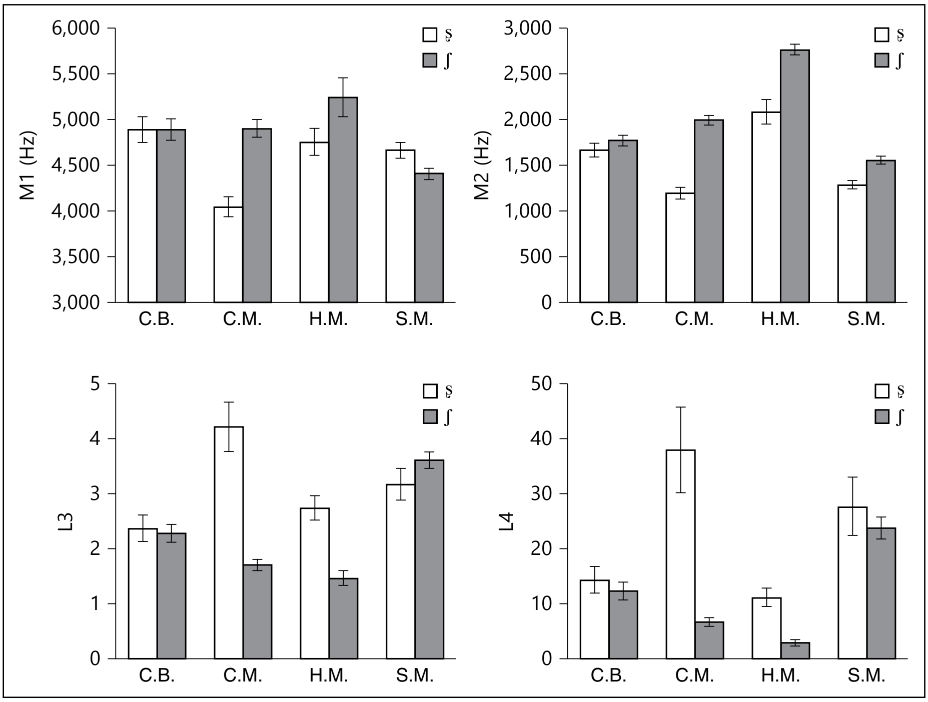

Fig. 13. Mean of the four spectral moments, M1, M2, L3, and L4, at mid phase of two sibilants for each speaker. Error bars indicate standard error.

somewhat infrequently for the whistled fricative. The results are surprising given the label assigned to the sound in the previous literature on Xitsonga [Carter and Kahari, 1979; Laver, 1994].

Nevertheless, the location of the whistled peak, if any, coincides with the location of the main peak without exception (e.g. fig. 7a, b). This indicates that the source spectrum for a whistle regularly couples into the first resonance frequency without jumping upward to higher resonance frequencies.

4 General Discussion

4.1 Summary and Remaining Issues

The articulatory study has identified the characteristic lingual and labial gestures of the whistled fricatives in Xitsonga. The ultrasound study confirmed that the Xitsonga whistled fricative is an apical retroflex fricative characterized by a retracted tongue back, a lowered tongue middle, and a raised tongue tip/blade. The video recording provided evidence for a unique type of labialization in the production of the whistled fricative. Unlike ordinary lip rounding, labialization in Xitsonga whistled fricative is manifested primarily in raising of the lower lip and horizontal narrowing toward the upper teeth, with little lip rounding or protrusion.

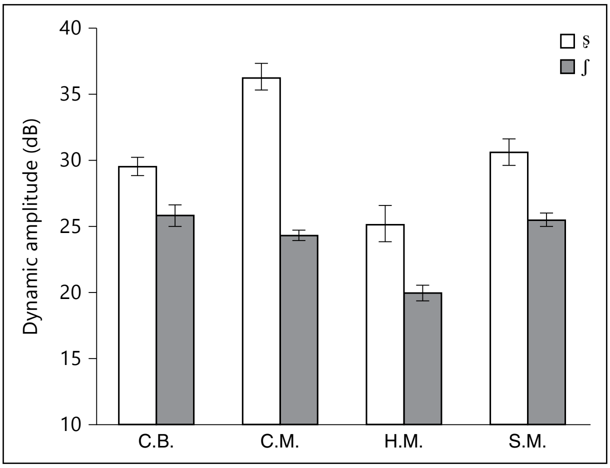

Fig. 14. Mean of the dynamic amplitude (A_d) at mid phase of two fricatives for each speaker. Error bars indicate standard error.

Table 3. Results of the mixed-effect linear regression for the four spectral moments of the three phases of the target sibilant

			Estimate	SE	t value	p
M1	BEG	intercept (/ʃ/)	4,773.94	189.55	25.2	
		sibilant: /ʂ/	17.37	257.07	0.068	0.9461
	MID	intercept (/ʃ/)	4,826.5	172.1	28.0	
		sibilant: /ʂ/	–283.0	248.3	–1.1	0.2544
	END	intercept (/ʃ/)	4,446.1	164.7	27.0	
		sibilant: /ʂ/	–108.0	302.3	–0.4	0.7209
M2	BEG	intercept (/ʃ/)	2,145.0	209.0	10.3	
		sibilant: /ʂ/	–397.3	195.3	–2.0	<0.05*
	MID	intercept (/ʃ/)	2,016.5	258.2	7.8	
		sibilant: /ʂ/	–462.1	168.8	–2.7	<0.01*
	END	intercept (/ʃ/)	2,062.3	162.0	12.7	
		sibilant: /ʂ/	–427.9	172.3	–2.5	<0.05*
L3	BEG	intercept (/ʃ/)	2.3133	0.3699	6.3	
		sibilant: /ʂ/	0.5890	0.5549	1.1	0.2885
	MID	intercept (/ʃ/)	2.2993	0.4880	4.7	
		sibilant: /ʂ/	0.8936	0.6609	1.4	0.1763
	END	intercept (/ʃ/)	2.8643	0.4637	6.2	
		sibilant: /ʂ/	1.2087	1.0900	1.1	0.2675
L4	BEG	intercept (/ʃ/)	11.600	3.181	3.6	
		sibilant: /ʂ/	6.980	3.866	1.8	0.0710
	MID	intercept (/ʃ/)	11.879	4.652	2.6	
		sibilant: /ʂ/	11.557	6.625	1.7	0.0811
	END	intercept (/ʃ/)	16.385	4.616	3.6	
		sibilant: /ʂ/	17.877	15.862	1.1	0.2597

An asterisk indicates a significant result.

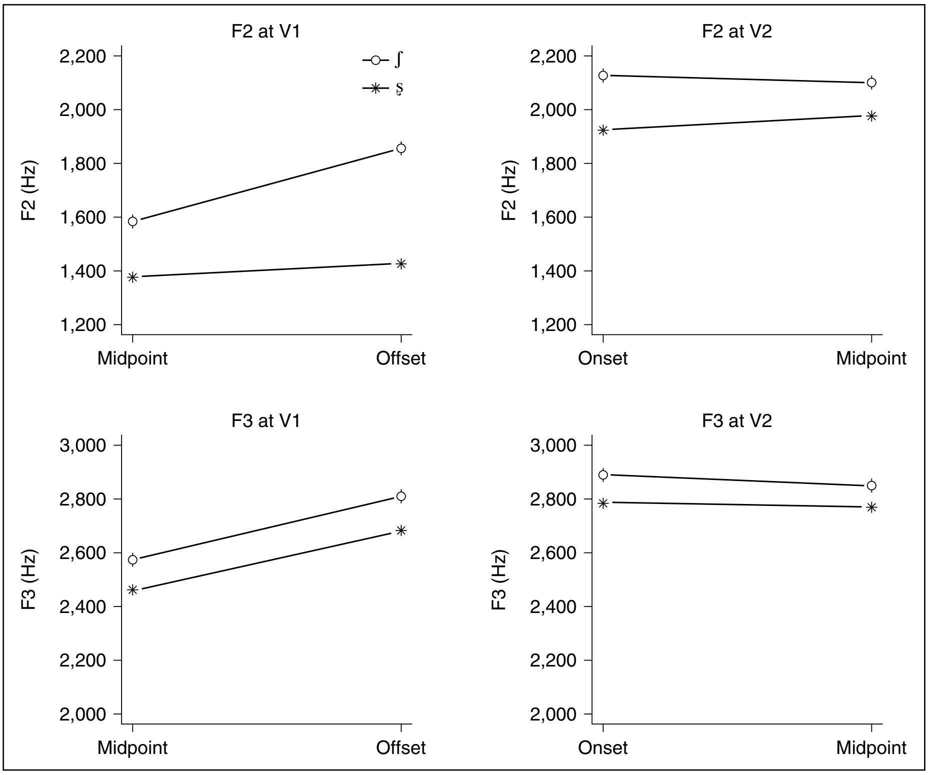

Fig. 15. F2 and F3 formant transitions of the surrounding vowels in the sequence /a-ʂ/ʃ-i/ averaged among all speakers. The figures on the left panel represent formant transitions of the preceding vowel /a/, and the figures on the right panel of the following vowel /i/. The top two figures represent F2 values, and the bottom figures F3 values, respectively. Error bars indicate standard error.

Table 4. Results of the mixed-effect linear regression for dynamic amplitude (A_d) of the three phases of the target sibilant

		Estimate	SE	t value	p
BEG	intercept (/ʃ/)	20.463	1.228	16.7	
	sibilant: /ʂ/	5.097	1.671	3.0	<0.01*
MID	intercept (/ʃ/)	24.027	1.283	18.7	
	sibilant: /ʂ/	6.515	1.781	3.7	<0.001*
END	intercept (/ʃ/)	22.593	1.036	21.8	
	sibilant: /ʂ/	4.964	1.149	4.3	<0.001*

An asterisk indicates a significant result.

Table 5. Results of the mixed-effect linear regression for F2 and F3 at mid and onset or offset of the surrounding vowels

		Estimate	SE	t value	p
F2_V1$_{mid}$	intercept (/ʃ/)	1,604.67	48.48	33.10	
	sibilant: /ʂ/	–212.92	66.12	–3.22	<0.01*
F2_V1$_{offset}$	intercept (/ʃ/)	1,851.84	43.92	42.16	
	sibilant: /ʂ/	–425.14	80.55	–5.28	<0.001*
F2_V2$_{onset}$	intercept (/ʃ/)	2,123.58	49.21	43.15	
	sibilant: /ʂ/	–200.52	24.10	–8.32	<0.001*
F2_V2$_{mid}$	intercept (/ʃ/)	2,090.70	67.93	30.78	
	sibilant: /ʂ/	–119.74	32.84	–3.647	<0.001*
F3_V1$_{mid}$	intercept (/ʃ/)	2,602.93	82.97	31.37	
	sibilant: /ʂ/	–102.28	43.37	–2.36	<0.05*
F3_V1$_{offset}$	intercept (/ʃ/)	2,785.10	123.41	22.57	
	sibilant: /ʂ/	–132.69	29.52	–4.50	<0.001*
F3_V2$_{onset}$	intercept (/ʃ/)	2,871.25	105.47	27.22	
	sibilant: /ʂ/	–121.23	55.01	–2.20	<0.05*
F3_V2$_{mid}$	intercept (/ʃ/)	2,820.05	96.29	29.29	
	sibilant: /ʂ/	–88.33	44.26	–2.00	<0.05*

An asterisk indicates a significant result.

Table 6. Percent of a whistled peak

	Whistled/(tokens×phases)	% of whistled
C.B.	6/(30×3)	7
C.M.	12/(21×3)	19
H.M.	0/(18×3)	0
S.M.	9/(43×3)	7

In the acoustic study, we first carried out a three-way analysis for the sibilants /s, ʂ, ʃ/ using the recordings of the speaker who participated in our articulatory study. While the overall energy distribution of /s/ stands out from /ʂ/ and /ʃ/, the latter two were close to each other for the parameters tested. The subsequent acoustic analyses using multiple speakers found that spectral moment M2 and the additional two parameters, dynamic amplitude and formant values in surrounding vowels, differentiate the two sibilants; /ʂ/ was shown to be lower in M2 and higher in dynamic amplitude than /ʃ/. In addition, we found significant differences in F2 and F3 values in surrounding vowels: formant values at both midpoint and onset/offset of the vowels were higher when next to /ʃ/ than next to /ʂ/. However, other acoustic parameters (peak frequency F, spectral moment M1, L3, and L4) showed individual variation rather than systematic differences.

Integrating articulatory and acoustic results, we attribute the differences in acoustics to the articulatory differences between /ʂ/ and /ʃ/. Articulatorily, the apical retroflex constriction is likely to have a smaller constriction area than the palate constriction by the palatoalveolar fricative. With a smaller constriction area, the whistled fricative would show a stronger source property, assuming a similar volume velocity for the two fricatives. The acoustic consequence of this is greater sibilancy for the whistled fricative than for the palatoalveolar fricative. As we have shown, this is manifested with higher peak amplitude (i.e. higher A_d) and more compact energy distribution (i.e. lower M2) in the spectra of the whistled fricative. The significantly lower F2 and F3 values of the vowels next to /ʂ̫/ compared to next to /ʃ/ receive a natural explanation from the articulatory perspectives as well: due to coarticulation with the retracted tongue back of the whistled fricative the following front vowel /i/ is more retracted, whereas the fronted tongue back of the palatoalveolar fricative has the opposite effect on the preceding nonfront vowel /a/.

Although the current study reveals aspects of the Xitsonga whistled fricative, there are some remaining issues for future study. First, the articulatory results should be verified with multiple speakers. Our study as a fieldwork was limited to the production of a single speaker. As noted in the 'Introduction', we expect some speaker variation, mainly in the exact tongue tip/blade gesture for the whistled fricative. Second, the vowel environment should be expanded to vowels /a/ and /u/ as well as /i/ in a future acoustic analysis. The acoustic differences observed in our study are predicted to be seen in other vowel environments as well. In theory, the acoustic properties of noise spectra can be attributed to source and/or filter. Lip rounding of the /u/ vowel is likely to change the filter property of the noise spectra by lengthening the resonance cavity with no considerable changes in the source property. As indexes of the source strength, the dynamic amplitude and M2 are thus likely to be stable despite the different vowel contexts. Based on the current results, we further elaborate on two issues below.

4.2 Aerodynamic Models for the Whistled Fricatives and Potential Dialectal Differences

Unlike ordinary lip rounding, labialization in Xitsonga whistled fricatives is manifested primarily in raising of the lower lip and horizontal narrowing toward the upper teeth with little lip rounding or protrusion. The current finding of the labial gestures involved in the Xitsonga whistled fricative is consistent with the descriptions of the labial gesture in the Zezuru dialect of Shona whistled fricative by Bladon et al. [1987]. In their video recording, the raising of the lower lip was also predominant, leading them to characterize the particular labialization as 'labiodentalized' (as opposed to 'labiolabial'), while there was little evidence for lip rounding or protrusion. On the other hand, the labial gestures of the whistled fricative in the Hlengwe dialect of Changana (S53/S511) video-recorded by Shosted [2011] appear to be substantially different: lip protrusion and rounding were both evident, much like in ordinary lip rounding. Since both our study and Shosted's [2011] involve only 1 speaker for video recording, the difference may be due to interspeaker variation rather than systematic dialectal variation. Although further study is needed, an anecdotal note on the labial gesture of the two languages, however, seems to point to the latter explanation. Our ultrasound speaker, when asked, produced the Changana whistled fricative with extreme lip rounding and protrusion. The speaker was well aware of the dialectal differences especially in labial gestures and could freely alternate the two whistled fricatives.

A close look at the acoustic property of the whistle reveals qualitative differences especially in the location of the whistled peak. That is, Changana is reported to have a whistled peak at 1.5 kHz [Shosted, 2011], whereas Shona is reported to have a whistled peak at 4 kHz [Bladon et al., 1987]. According to our findings, Xitsonga joins the latter group (peak frequency at about 3.5 kHz), as it did for the labial gesture. Assuming that the lingual gestures for the whistled fricatives are not considerably different, the discussion thus far leads to a conjecture that the particular labial gesture is a key to understanding of how different types of acoustic whistles are produced and used in human speech.

Theoretically, two acoustic mechanisms are conceivable for the whistled fricatives. One model is an 'edge tone' model in which the teeth serve as an 'edge' and the lingual tongue constriction creates a turbulence jet [Shadle, 2010]. A whistle occurs when oscillation formed around the sharp edge (i.e. the teeth) couples into the resonance frequency of the cavity between the teeth and the lingual constriction. Given the specific labial gesture found in our study, this aerodynamic model seems to be suitable for the Xitsonga whistled fricative. That is, the unstable jet around the upper teeth turns into a periodic oscillation as the flow speed goes up, presumably driven by the narrowing of the lower lip toward the upper teeth. When the source spectrum couples into one of the resonance frequencies of the cavity formed between the upper teeth and the tongue tip/blade constriction against the alveolar ridge, a whistle occurs. In principle, a whistle coupled into the first resonance frequency of the cavity may jump upward, coupling into the next higher resonance frequency of the cavity as the flow speed increases [Shadle, 1983]. In our data, however, the source coupled regularly into the first resonance frequency of the retroflex; a regular peak was found at a similar location (i.e. 3.5 kHz) when a whistled peak was not present.

Changana whistled fricatives provide an instance of another aerodynamic model for the whistled fricatives. According to Shadle [2010], a 'hole tone' model is also possible in which the rounded lips, instead of the teeth, could form an orifice to create an unstable jet. Together with the lingual constriction, they form two orifices of which resonance frequencies are determined by the cavity between the two. Since the lip constriction necessarily forms a longer front cavity, the resonance frequency of a whistle of this kind can be as low as F2. Changana whistled fricatives exemplify this model. Articulatorily, they involve ordinary lip rounding in which the lips are drawn to the center and substantially protruded along the sagittal plane. Acoustically, the whistled peak occurs at a much lower frequency around 1.5 kHz.

Although more instrumental studies are warranted, the current study in comparison with previous work hints at the possibility that the 'whistled' fricatives in Southern Bantu languages are nonmonolithic and may vary greatly in terms of articulation, aerodynamics, and acoustics. We conjecture that the observed differences in articulation and acoustics of the whistled fricatives may reflect the influences of the geographic limitations or sociopolitical interactions between neighboring areas. Geographically, Changana is spoken in Mozambique, while Xitsonga is spoken in South Africa. Although the languages form a dialect continuum, the two groups of people live in areas that have not been in direct contact since the Kruger National Park founded in 1898 became off-limits to foot traffic for wildlife reservation. The tumultuous political situations in South Africa and Mozambique during the latter part of 20th century further distanced the two speech communities. While the political circumstances have changed since the 1990s, the Kruger National Park still forms a barrier to direct contacts between Changana speakers and Xitsonga speakers. On the other hand, Shona speakers live in Zimbabwe across the Limpopo River, which is connected to South Africa with one of the busiest border crossings in southern Africa. Though considered as two different languages, the relatively frequent contacts between Shona and Xitsonga speakers may explain certain common features of the whistled fricatives reported in this study.

4.3 The Whistled Peak of the Whistled Fricative in Xitsonga and Its IPA Notation

Our data show that a whistle, if any, occurs at a location where the main peak of the nonwhistled fricative occurs, indicating that the turbulence oscillation coupled into the first resonance frequency of the front cavity. However, the results also show that whistling, contrary to the traditional label given to this sound, was extremely weak and occurred relatively infrequently across speakers. While the posited ‘edge tone’ model seems appropriate for the Xitsonga whistled fricative considering its articulation and acoustics, the considerably rare occurrence of a whistled peak in the whistled fricative suggests that a narrowing between the lower lips and the upper teeth does not regularly create the turbulence oscillation necessary for a whistle. Without a turbulence jet at the teeth, the only noise source left is the lingual constriction, namely, retroflex apical constriction. When the turbulence jet formed by the lingual constriction is excited at the resonance frequency of the cavity formed between lingual and labiodental constriction, it would give rise to the regular acoustic property of a simple retroflex. Our results suggest that this is most common in the Xitsonga whistled fricative. Therefore, the retroflex IPA transcription seems to be most appropriate, especially for the whistled fricative in Xitsonga. Nevertheless, it is possible that the jet formed at the lips or teeth is strong enough to regularly excite a whistled peak for the whistled fricatives in related languages. It is hoped that our preliminary observation on the whistled peak will contribute to more objective and automatic methods for future investigation of crosslinguistic and dialectal comparisons of whistled fricatives.

5 Conclusion

The present study examined both articulatory and acoustic characteristics of the ‘whistled’ fricative in Xitsonga in comparison with other sibilants. The lingual gesture of the whistled fricative is retroflex, confirming the impressionistic hypothesis in the previous literature by way of modern experimental technique. Labialization in the Xitsonga whistled fricative is primarily manifested in the raising of the lower lip toward the upper teeth, forming a potential source for an ‘edge tone’ whistle. The current acoustic study found the spectra of the whistled fricative to be higher in dynamic amplitude and lower in M2 than that of the palatoalveolar, indicating that the former has a stronger and more localized source than the latter. As the whistle was not prevalent in the whistled fricatives in Xitsonga, we proposed a retroflex IPA notation (i.e. [ʂ]) for the Xitsonga whistled fricative. We hope to examine crosslinguistic and dialectal differences on the articulation and the acoustics of the whistled fricatives in future studies.

Acknowledgments

We would like to thank Clementinah Burheni, Sean Martin, Christine Shadle, Donna Erickson, and audiences at Phonology Festa 7, NYU PEP Lab, and ACAL 43 for their helpful discussion. We also thank Melanie Pangilinan and other research assistants at the Rutgers Phonetics Lab for their help with the initial segmentation of the recordings. Finally, we thank the editor of *Phonetica* and the anonymous reviewers for the careful and thorough reviews of the manuscript. This work was supported by Faculty Development Grant, Dean's Research Initiative and University Research Grant from the Central Connecticut State University.

Appendix

Stem List for the Acoustic Study

The stem list used in experiment 2. The transcription is given in IPA and in Xitsonga orthography. The target consonants were placed in the number prefixes attached before these stems (e.g. [ʃi]-témpé 'a stamper' vs. [s̤i]-témpé 'stampers').

Tone	Stem in IPA	Stem in orthography	Gloss
HH(H)	-témpé	-témpé	stamper (loan)
	-létí	-létí	slate (loan)
	-^{ŋ}k^{w}ámá	-nkwámá	bag, pocket
	-fákí	-fákí	mealie cob
	-pápá	-pápá	snuff pouch
	-ʤóɦó	-dyóhó	sin
	-lótléló	-lótléló	key
	-ɦánánó	-hánánó	offering
	-t^{h}éβé	-thévé	narrow sleeping mat
HL	-tínà	-tínà	brick
	-ɦár̤ì	-hárhì	wild animal
	-fánísò	-fánísò	picture
L(HH)	-lò	-lò	thing
	-ɬòká	-hlòká	axe
	-r̤àmí	-rhàmí	coldness
	-ɦèŋgé	-hèngé	pineapple
	-k^{w}àβáβá	-kwàvává	lemon
LL(LL)	-bàmò	-bàmò	gun
	-mìlànà	-mìlànà	plant
	-tìmèlà	-tìmèlà	train (loan)
	-àmbàlò	-àmbàlò	cloth
	-bèlèkèlò	-bèlèkèlò	belly, womb

References

Baayen, R.H.: Analyzing linguistic data: a practical introduction to statistics (Cambridge University Press, New York 2008).

Bates, D.; Maechler, M.: Lme4: linear mixed-effects models using S4 classes (2009).

Baumbach, E.J.M.: Analytical Tsonga grammar (University of South Africa, Pretoria 1987).

Bennett, W.: Dissimilation, consonant harmony and surface correspondence; PhD thesis Rutgers, The State University of New Jersey (2013).

Biedrzycki, L.; Gontarczyk, S.: Abriß der polnischen Phonetik (Wiedza Powszechna, Warszawa 1974).

Blacklock, O.S.; Shadle, C.H.: Spectral moments and alternative methods of characterizing fricatives. J. acoust. Soc. Am. *113:* 2199 (2003).

Blacklock, O.S.B.: Characteristics of variation in production of normal and disordered fricatives using reduced-variance spectral methods; PhD thesis University of Southampton (2004).

Bladon, A.; Clark, C.; Mickey, K.: Production and perception of sibilant fricatives: Shona data. J. int. phonet. Ass. *17:* 39–65 (1987).

Boersma, P.; David W.: Praat 5.3.23: Doing phonetics by computer. http://www.praat.org (2012).

Carter, H.; Kahari, G.P.: Kuverenga Chishona, an introductory Shona reader with grammatical sketch (School of Oriental and African Studies, London 1979).

Chen, Y.; Lin, H.: Analysing tongue shape and movement in vowel production using ss ANOVA in ultrasound imaging. Proc. XVIIth ICPhS, Hong Kong, pp. 124–127 (2011).

Cuenod, R.: Tsonga-English dictionary (Sasavona Publishers & Booksellers, Johannesburg 1967).

Davidson, L.: Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. J. acoust. Soc. Am. *120:* 407–415 (2006).

Davidson, L.: Ultrasound as a tool for speech research; in Cohn, Fougeron, Huffman, The Oxford handbook of laboratory phonology, pp. 484–495 (Oxford University Press, Oxford 2012).

Davidson, L.; De Decker, P.: Stabilization techniques for ultrasound imaging of speech articulations. J. acoust. Soc. Am. *117:* 2544 (2005).

De Decker, P.M.; Nycz, J.R.: Are tense [æ]s really tense? The mapping between articulation and acoustics. Lingua *122:* 810–821 (2012).

Epstein, M.A.; Stone, M.: The tongue stops here: ultrasound imaging of the palate. J. acoust. Soc. Am. *118:* 2128–2131 (2005).

Forrest, K.; Weismer, G.; Milenkovic, P.; Dougall, R.N.: Statistical analysis of word-initial voiceless obstruents: preliminary data. J. acoust. Soc. Am. *84:* 115–123 (1988).

Gu, C.: Smoothing spline ANOVA models (Springer, New York 2002).

Gu, C.: Gss: general smoothing splines. R package version 2.0-9. http://cran.r-project.org/web/packages/gss/index.html (2012).

Guthrie, M.: Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages. 4 Vols. (Gregg International, Letchworth UK & Brookfield, VT, 1967/1971).

Hamann, S.R.: The phonetics and phonology of retroflexes; PhD thesis Utrecht University (2003).

ICPLA: extIPA symbols for disordered speech (1994).

Iskarous, K.; Shadle, C.H.; Proctor, M.I.: Articulatory-acoustic kinematics: the production of American English /s/. J. acoust. Soc. Am. *129:* 944–954 (2011).

Jesus, L.M.T.; Shadle, C.H.: A parametric study of the spectral characteristics of European Portuguese fricatives. J. Phonet. *30:* 437–464 (2002).

Jongman, A.; Wayland, R.; Wong, S.: Acoustic characteristics of English fricatives. J. acoust. Soc. Am. *108:* 1252–1263 (2000).

Kochetov, A.; Pouplier, M.; Truong, S.: A preliminary ultrasound study of Nepali lingual articulations. Proc. Meetings. Acoust., Montreal, vol. 19, pp. 1–9 (2013).

Koenig, L.L.; Shadle, C.H.; Preston, J.L.; Mooshammer, C.R.: Toward improved spectral measures of /s/: results from adolescents. J. Speech Lang. Hear. Res. *56:* 1175–1189 (2013).

Ladefoged, P.; Bhaskararao, P.: Non-quantal aspects of consonant production: a study of retroflex consonants. J. Phonet. *11:* 291–302 (1983).

Ladefoged, P.; Maddieson, I.: The sounds of the world's languages (Blackwell, Oxford 1996).

Laver, J.: Principles of phonetics (Cambridge University Press, Cambridge 1994).

Lee-Kim, S.-I.: Spectral Analysis of Mandarin Chinese Sibilant Fricatives. Proc. XVIIth ICPhS, Hong Kong, pp. 1178–1181 (2011).

Lee-Kim, S.-I.; Davidson, L.; Hwang, S.: Morphological effects on the darkness of English intervocalic /l/. Lab. Phonol. *4:* 475–511 (2013).

Leholha, P.: Census 2001: census in brief, report No. 03-02-03 2001 (Statistics South Africa, Pretoria 2003).

Li, M.; Kambhamettu, C.; Stone, M.: Automatic contour tracking in ultrasound images. Clin. Ling. Phonet. *19:* 545–554 (2005).

Lousada, M.L.; Jesus, L.M.T.; Pape, D.: Estimation of stops' spectral place cues using multitaper techniques. Documentação de Estudos em Lingüística Teórica e Aplicada *28:* 1–26 (2012).

Maddieson, I.: The sounds of the Bantu languages; in Nurse, Philippson, The Bantu languages, pp. 15–41 (Routledge, London 2003).

Mielke, J.; Olson, K.S.; Baker, A.; Archangeli, D.: Articulation of the Kagayanen interdental approximant: an ultrasound study. J. Phonet. *39:* 403–412 (2011).

Nowak, P.M.: The role of vowel transitions and frication noise in the perception of Polish sibilants. J. Phonet. *34:* 139–152 (2006).

Proctor, M.; Lu, L.H.; Zhu, Y.; Goldstein, L.; Narayanan, S.: Articulation of Mandarin sibilants: a multi-plane real time MRI study. 14th Australasian Int. Conf. on Speech Sci. Technol., Macquarie University, Sydney 2012.

Reetz, H.; Jongman, A.: Phonetics: transcription, production, acoustics, and perception (Wiley-Blackwell, Chichester 2009).

Shadle, C.H.: Experiments on the acoustics of whistling. Physics Teacher *21:* 148–154 (1983).

DOI: 10.1159/000362672

Shadle, C.H.: The aerodynamics of speech; in Hardcastle, Laver, The handbook of phonetic science, pp. 39–80 (Wiley-Blackwell, Chichester 2010).

Shadle, C.H.; Mair, S.J.: Quantifying spectral characteristics of fricatives. Proc. 4th ICSLP, Philadelphia, pp. 1521–1524 (1996).

Shadle, C.H.; Scully, C.: An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. J. Phonet. *23:* 53–66 (1995).

Shosted, R.K.: Just put your lips together and blow? Whistled fricatives in Southern Bantu.

Shosted, R.K.: Articulatory and acoustic characteristic of whistled fricatives in Changana. Proc. 40th Annu. Conf. Afr. Ling., Cascadilla Press, Somerville, pp. 119–129 (2011).

Simonet, M.; Rohena-Madrazo, M.; Paz, M.: Preliminary evidence for incomplete neutralization of coda liquids in Puerto Rican Spanish; in Colantoni, Steele. Proc. 3rd Conf. Lab. approaches to Span. phonol., 2008, pp. 72–86.

Sitoe, B.: Dicionário Changana-Português (Instituto Nacional do Desenvolvimento da Educação, Maputo 1996).

Stone, M.: A guide to analysing tongue motion from ultrasound images. Clin. Ling. Phonet. *19:* 455–501 (2005).

Wang, Y.: Smoothing splines: methods and applications (Chapman & Hall/CRC Press, Boca Raton 2011).

Wierzchowska, B.: Fonetyka i fonologia języka polskiego (Zakład Narodowy imienia Ossolińskich, Wrocław 1980).

Zharkova, N.; Hewlett, N.; Hardcastle, W.J.: An ultrasound study of lingual coarticulation in /sv/ syllables produced by adults and typically developing children. J. int. phonet. Ass. *42:* 193–208 (2012).

Żygis, M.; Pape, D.; Jesus, L.M.T.: (Non-)retroflex Slavic affricates and their motivation: evidence from Czech and Polish. J. int. phonet. Ass. *42:* 281–329 (2012).