Perceiving Small Contrasts: Xitsonga's 'whistled' fricative [s̪] vs. palatal fricative [ʃ]

Aaron Braver (*Texas Tech University*)
aaron@aaronbraver.com
Seunghun J. Lee (*Central Connecticut State University*)
juliolee@gmail.com

Phonology of contrast

- Contrast
  - Segments are in contrast when the distinction can change the meaning of a word
    - *sip* vs. *zip* *hit* vs. *hot*
  - Speakers know (*have the awareness of*) which segments contrast in their own language, and which do not
  - Contrast is not universal, but language-specific
- Production and perception of contrasts
  - Speakers may substitute sounds of a foreign language with one in their own language. This substitution sometimes results in the neutralization of the contrast.
    - *read* and *lead* as [rido] by Japanese speakers

Perception of contrast

- Speakers have difficulty peceiving contrasts that are not present in their own language.
  - Non-native speakers of Korean have difficulty hearing the laryngeal contrast in Korean
    - /pul/ 'grass'
    - /pʰul/ 'fire'
    - /pʻul/ 'horn'
  - Non-native speakers of Xitsonga have difficulty hearing the fricative contrast in Xitsonga
    - *xilo* 'a thing'
    - *swilo* 'things'

Small contrast?

- Acoustic features
  - When acoustically similar sounds are contrastive, the contrast can be said to be small.
  - The first sounds in *xilo* and *swilo* are both fricatives and they are acoustically similar (but not the same).
    - Acoustic measurements such as M1 (mean), L3 (skewness), L4 (kurtosis) are similar.
    - M2 (variance) suggests that the spectra is flatter in *xilo* than in *swilo*, and dynamic amplitude (A_d) is higher is *swilo*.
- How can we define (small) contrast in acoustic terms?
  - Number of acoustic measurements
  - Significant acoustic properties not commonly used
  - or else

Small contrast?

- Phonological features
  - Distinctive features
    - Place of articulation
    - Manner of articulation
    - Laryngeal settings
    - Height
    - Frontness
    - Roundedness
- How would we define (small) contrast in phonological terms?
  - existence of an IPA symbol
  - secondary articulation features
  - or else
- Yet, are there perceptual features that define (small) contrast?

Roadmap

- Findings of articulatory and acoustic study of whistled fricative and non-whistled fricative in Xitsonga
- Our study on perception of these fricatives
- Discussion

Xitsonga (S. 53)

- a Southern Bantu language
- spoken in South Africa, Mozambique, Zimbabwe and Lesotho
- one of the eleven official languages in South Africa.
- ca. 2 million speakers

The fricatives

- Palatal fricatives [ʃ] and whistled fricatives [ʂ] in Xitsonga distinguish singular (class 7) from plural (class 8)
  - [ʃi]-lo 'a thing' [ʂi]-lo 'things'
- Whistled fricatives are typologically very rare (Shosted, 2006). Impressionistically, whistled fricatives and palatal fricatives sound very similar.

Whistled fricatives

- Whistled fricatives are also called "bilabio-alveolar fricatives" (Janson, 2001), indicating that lip rounding is involved.
- However, in the acoustic study of Changana - a Xitsonga dialect spoken in Mozambique - Shosted (2011) shows that lip rounding is not a crucial component of whistled fricatives.
- A retroflex gesture has been hypothesized for the whistled fricatives in the literature (Carter & Kahari 1979, Laver 1994, Sitoe 1996)

Lingual and Labial data

(Lee-Kim, Kawahara & Lee 2014)

ARTICULATORY STUDY OF XITSONGA FRICATIVES

Methodology

- A female speaker of Xitsonga in her twenties
- Stimuli are selected from Cuenod 1967

	Target words In IPA	Xitsonga Orthography [Cuenod, 1967]	Gloss
	[sàngù]	*sàngù*	'sleeping mat'
/a/	[ʂàtà]	*swàtà*	'fall or dive into'
	[ʃàndzùkà]	*xàndzuka*	'to abandon one's family'
	[sìlà]	*sìlà*	'to grind on a stone'
/i/	[sìŋwè]	*swin'wè*	'together'
	[ʃìnànmù]	*xìnànmù*	'procrastination'
	[sùsà]	*sùsà*	'to take away'
/u/	[ʂúkútá]	*swúkuta*	'to chase away'
	[ʃùvùrù]	*xùvùrhù*	'uncircumcised male'

Ultrasound imaging - setting

- Speaker's position
  - a comfortable pose in a sound-attenuated booth at the Phonetics and Experimental Phonology Laboratory at New York University
- Head stabilization
  - a moldable head stabilizer (Comfort Company) on the wall, with a Velcro strap [Davidson and De Decker, 2005; Davidson, 2006]
- Transducer location
  - under the speaker's chin
  - adjusted until clear midsagittal images were captured.
- The participant was first asked to swallow water to extract the palate image [Epstein and Stone, 2005].

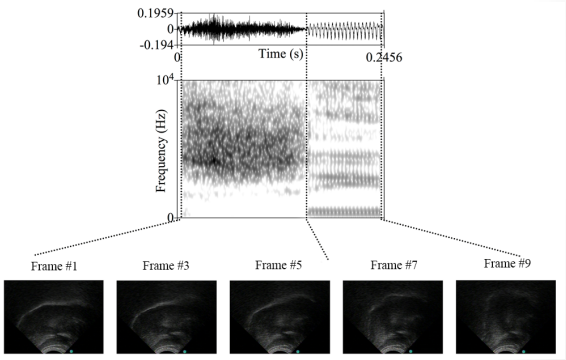
Ultrasound imaging

- a Sonosite Titan portable ultrasound
- a 5-8 MHz Sonosite C-11 transducer with a 90º field of view set at a depth of 8.2 cm
- Frame rate = 29.97 frame/s (one frame ≈ every 33.4 ms)
- Microphone: Audio Technica AT-813
- Synchronization: Canopus ADVC-1394 capture card and Adobe Premiere

Annotation of ultrasound images

- 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

Ultrasound images and acoustic signal

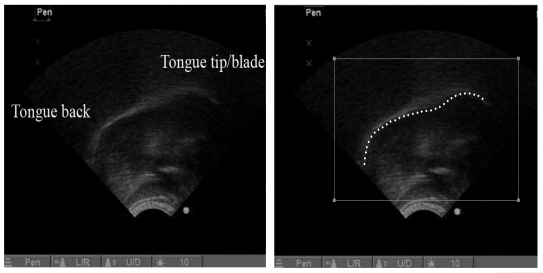


Frame selections for statistics

- The frame that shows maximal consonantal constriction
  - Whistled fricative /ʂ/
    one frame before the release of the tongue tip/blade
  - Palatoalveolar fricative /ʃ/
    one frame before slight tongue body lowering from the palate
  - Dental fricative /s/
    one frame before lowering of tongue just in back of the tip

Edgetrak (Li et al. 2005)

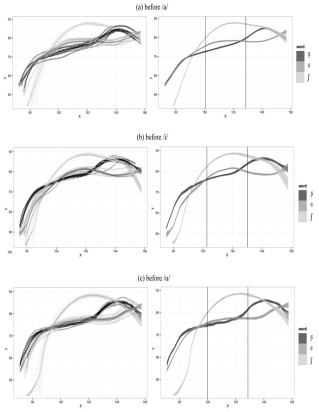
- automatically tracks tongue configuration by extracting x-y coordinates of the target region from the upper edge of the tongue.



Smoothing spline ANOVA (SS ANOVA)

- Returning of 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.
  - 95% Bayesian confidence intervals around the smoothing splines
  - SSANOVA was implemented using the gss package in R [Gu, 2012].

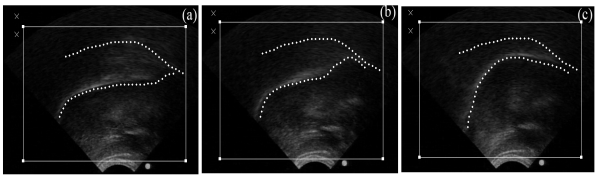
Results – ultrasound imaging



- Whistled fricative
  - retracted tongue back
  - lowest tongue middle
  - highest tongue tip
- Palatal fricative
  - fronted tongue back
  - highest tongue body
- Dental fricative
  - varying degree of tongue back retraction (coarticulation effects)
  - lowest tongue tip

Tongue and the palate

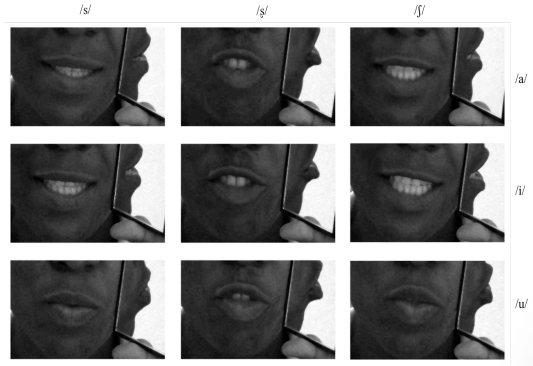
- The tongue shapes of /s/, /ʂ/ and /ʃ/ at maximal constriction in /a/ vowel context with the palate trace



Methods – video recording

- The speaker held a hand-mirror on the left side of her lips to examine lip protrusion as well as lip rounding [Shosted 2011]
  - A Sony PAL DCR-SX21 digital video camera recorder was mounted on a tripod two feet from the speaker.
  - Frame rate: 25 frame/s (one frame ≈ every 40 ms)
  - Video file: MPEG format.
  - Audio signal: 16 bit with a 44 kHz sampling rate
- The middle frame was selected out of the 4-6 frames of the acoustic realization of the fricatives

Sagittal images of labial gestures



Results – labial data

- Whistled fricatives
  - Weak lip protrusion (with horizontal narrowing)

	Whistled	Non-whistled (non-rounded V)
lower lip	raised	not-raised
lower teeth	covered	visible
upper lip	slightly raised	neutral state
upper teeth	exposed	partial exposure

Interim summary

- Lingual data – Ultrasound imaging
  - Whistled fricative
    - retracted tongue back
    - lowest tongue middle
    - highest tongue tip
- Labial data – Video recording
  - Lower lip is raised covering lower teeth
  - Upper lip is raised exposing upper teeth

(Lee-Kim, Kawahara & Lee 2014)

ACOUSTIC STUDY OF XITSONGA FRICATIVES

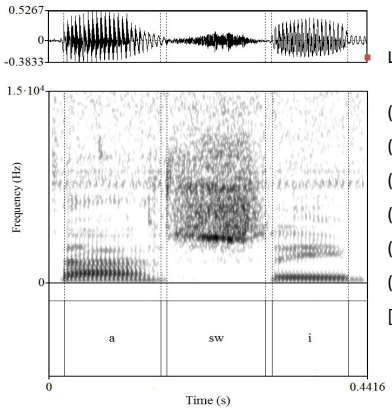
Experiment - Stimuli

- Stimuli
  - Preceded by
    - [ʃi] (class 7, singular)
    - *or*
    - [si] (class 8, plural)

-témpé	stamper	-lò	thing
-léti	slate (loan)	-hlòká	axe
-nkwámá	bag, pocket	-rhàmí	coldness
-fáki	mealie cob	-hèngé	pineapple
-pápá	cotton	-kwávává	lemon
-dyóhó	sin	-bámù	gun
-lótlélò	key	-mìlànà	plant
-hánánó	offering	-timèlà	train
-fánísó	picture	-àmbàlò	cloth
-tìnà	brick	-bèlèkèlò	belly, womb
-tébè	sleeping mat		
-hári	wild animal		

- Recording
  - two female (CB, SM) and two male (CM, HM) speakers of Xitsonga
  - carrier phrase: "ni tirisa __ kan'we" (I use __ again)
  - 3 times of repetition in random orders
  - sampling rate: 44,100 Hz

Experiment - Procedure



Labeled events

(1) onset of [a]
(2) offset of [a]
(3) frication onset
(4) frication offset
(5) onset of [i]
(6) offset of [i]

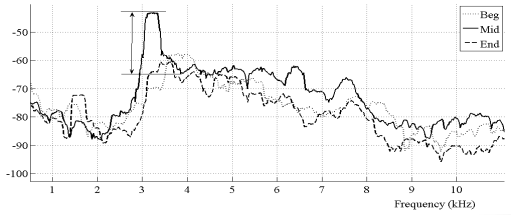
[Iskarous et al. 2011]

Quantification of noise spectra - Comparing /s/, /ʂ/, and /ʃ/

- Spectral peak *F* [cf. Jesus and Shadle 2002]
  - the frequency where the maximum amplitude occurs
  - association with the first resonance frequency of the front cavity
    - a longer front cavity -> a lower spectral peak
  - multitaper spectral analysis in Matlab
    - spectral normalization & computation of the spectral moments at Beg-Mid-End of frication
- Spectral moments [Forrest et al. 1988]
  - mean (M1)
  - variance (M2)
  - skewness (L3)
  - kurtosis (L4)

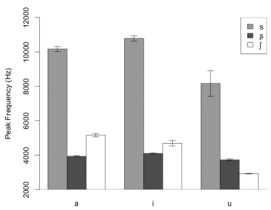
Testing the whistle using noise spectra

- Whistling mechanism
  - an oscillation in the source spectrum is stabilized through coupling into the resonance frequency of the cavity
  - a high-amplitude narrow-bandwidth peak [Shadle 1983, 2010]

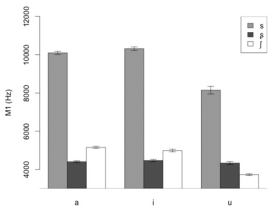


Spectral measurements - /s/ vs. /ʂ, ʃ/

Mean peak frequency *F* (Hz)



Spectral mean *M1* (Hz)

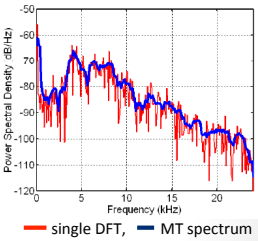


- Significant main effect of the sibilant type
- /s/: higher F and M1 than other fricatives
- Significant interaction between the sibilant type & the vowel type

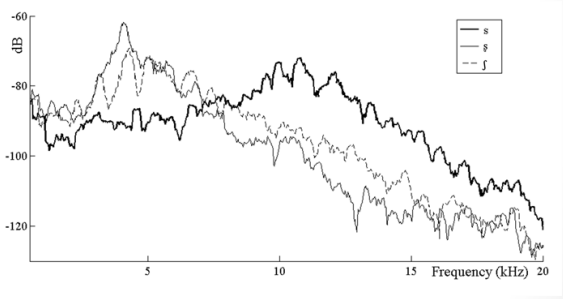
Multitaper analysis

- Computing a discrete Fourier transform (DFT) from a single windowed interval results in a spectral estimate with a large error; spectral averaging needed. (Shadle 2006: 449)

- **Multitaper analysis** (Blacklock 2004)
  - A single short segment is used, and multiplied by different windows, called tapers. Then, each DFT is computed and averaged.
  - A small error with good time and frequency resolution.
  - No assumption of an ensemble or of stationarity.



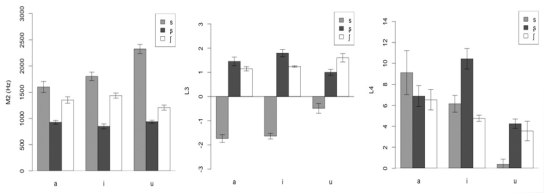
Results – Multitaper spectra - /s/ vs. /ʂ, ʃ/



- /a/ context taken at mid-phase of frication noise

Spectral moments - /s/ vs. /ʂ, ʃ/

- A more diffuse energy distribution in the dental spectra [s] (high M2 [variance])
- the spectral energy concentrated at higher frequencies (negative L3 [skewness])



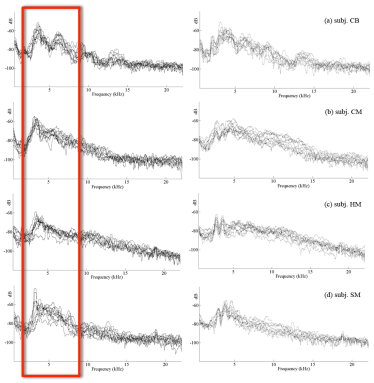
Mean spectral moment M2, L3, and L4 of the three sibilants /s, ʂ, ʃ/ in three vowel contexts /a, i, u/.

Comparing /ʂ/, and /ʃ/

- Dynamic amplitude (A_d)
  - the difference in amplitude between the spectral peak and the spectral trough that occurs between the cutoff frequency (500 Hz) and the spectral peak
  - An index of sibilancy
  - the more strident sound -> higher A_d values [Jesus & Shadle 2002]

- Formant values of the surrounding vowels
  - a 20 ms window centered at the midpoint of the vocalic intervals
  - maximum frequency was set to 5 kHz for male and 5.5 kHz for female
  - values extracted using the Berg algorithm in Praat

Multitaper spectra: /ʂ/ vs. /ʃ/



/ʂ/ /ʃ/

2nd peak at ca. 7 kHz

Spectral peak frequency -/ʂ/ vs. /ʃ/

- Location of peak frequency is not different.
- Higher estimated peak frequency of the whistled fricative at beginning and end phases.

		Estimate	SE	t value	*p*
BEG	Intercept (/ʃ/)	3851.3	163.8	23.5	
	sib: /ʂ/	371.7	108.8	3.4	< 0.001*
MID	Intercept (/ʃ/)	3792.0	290.3	13.1	
	sib: /ʂ/	111.2	190.2	0.6	0.5586
END	Intercept (/ʃ/)	3457.5	229.0	15.1	
	sib: /ʂ/	397.3	146.3	2.7	< 0.01*

Spectral moments
-/ʂ/ vs. /ʃ/

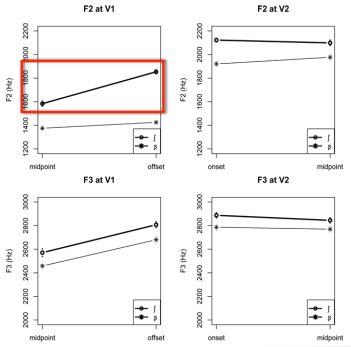
- M1 (mean), L3 (skewness), L4 (kurtosis) is NOT significantly different in all three phases
- M2 (variance) is significantly lower for the whistled fricative than for the palatoalveolar fricative (flatter spectra of the palatoalveolar fricative)

Dynamic amplitude
-/ʂ/ vs. /ʃ/

- The estimated dynamic amplitude of the whistled fricative is significantly higher by 6.4 dB
  - the peak is higher in amplitude for the whistled fricative

F2 and F3
-/ʂ/ vs. /ʃ/

- F2 values were significantly higher next to /ʃ/ than next to /ʂ/ at all four acoustic landmarks (all $p < .05$)
- Steep F2 transitions of the palatoalveolar
- F3 values patterned similarly



Whistling peak

- Varying results of the overall percentage of whistled peaks
  - One speaker — whistling 20% of the time
  - Other two speakers — 7% of the time
  - One speaker — no whistling

	Whistled/(tokens*phases)	% of whistled
CB	6/(30*3)	7%
CM	12/(21*3)	19%
HM	0/(18*3)	0%
SM	9/(43*3)	7%

- Infrequent whistled peaks in the whistled fricative
- Nevertheless, the location of the whistled peak coincides with the location of the main peak without exception

Discussion
- aerodynamic models

- Edge tone model (our model)
  - the teeth serves as an 'edge' and the lingual tongue constriction creates a turbulence jet
  - a whistle occurs when oscillation formed around the sharp edge couples into the resonance frequency of the cavity between the teeth and the lingual constriction
- Hole tone model
  - the rounded lips form an orifice to create an unstable jet
  - Changana whistled fricatives reported in Shosted (2011)

Summary
- articulation and acoustics

- Whistled fricatives in Xitsonga
  - an apical retroflex fricative with a retracted tongue back, a lowered tongue middle and a raised tongue tip/blade (ultrasound)
  - raising of the lower lip and horizontal narrowing toward the upper teeth, with little lip rounding or protrusion (video)
- Whistled fricatives and palatoalveolar fricatives
  - lower in M2
  - higher in dynamic amplitude
  - peak frequency F, spectral moment M1, L3 and L4 only show individual variation.

OUR PERCEPTION STUDY OF XITSONGA FRICATIVES

Hypothesis

- Xitsonga speakers will be able to perceive the distinction between [ʂ] and [ʃ] with a high degree of accuracy
- English speakers will be much less able to perceive this distinction

Stimuli

- 12 Xitsonga nouns
- Each noun was given the singular class 7 prefix <xi-> and the plural class 8 prefix <swi->, resulting in 24 total stimulus items.
- 6 native speakers of Xitsonga were asked to read the singular and plural nouns in a carrier sentence. From these 6 speakers, the productions from one male speaker and one female speaker were chosen to be used in the experiment.

List of stimuli

ximilana	swimilana
xihloka	swihloka
xirhami	swirhami
xinkwama	swinkwama
xitina	switina
xihenge	swihenge
xifaki	swifaki
xiharhi	swiharhi
xiambalo	swiambalo
xibamo	swibamo
xifaniso	swifaniso
xihanano	swihanano

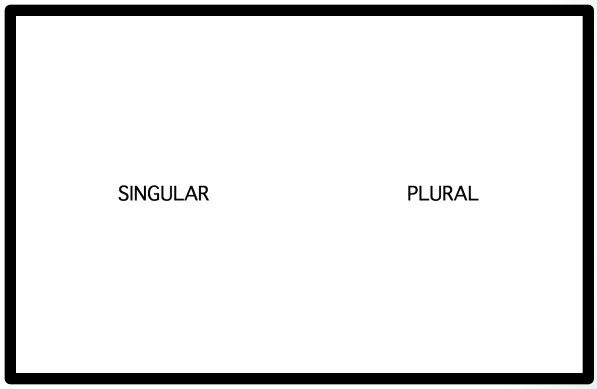
Participants (Xitsonga & English)

- Xitsonga
  - 21 native speakers
  - all from Mhinga or Boxahuku in Limpopo
  - high school degree or above
- English
  - 15 native speakers
  - all from Connecticut, USA
  - high school degree or above

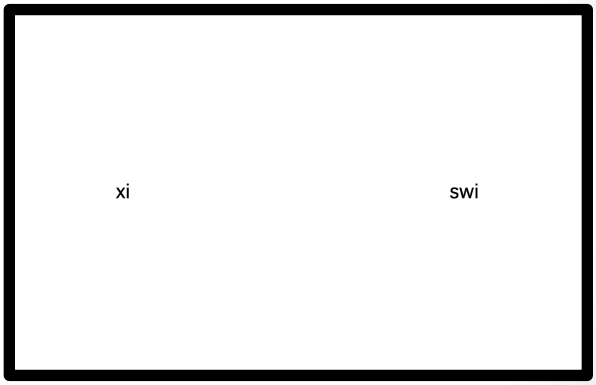
Experimental design (Xitsonga speakers)

- Identification task
  - Speakers heard one word at a time, either singular ([ʃ]) or plural ([ʂ])
  - Participants pressed a button indicating whether the word they heard was singular or plural
  - Variations within this task (we collapse all the results in this presentation):
    - Tokens from female speaker vs. male speaker
    - Stimuli with vowel in the singular/plural prefix either cut out or replaced with a tone in order to reduce acoustic cues to the contrast

Procedure – Xitsonga, full sentence



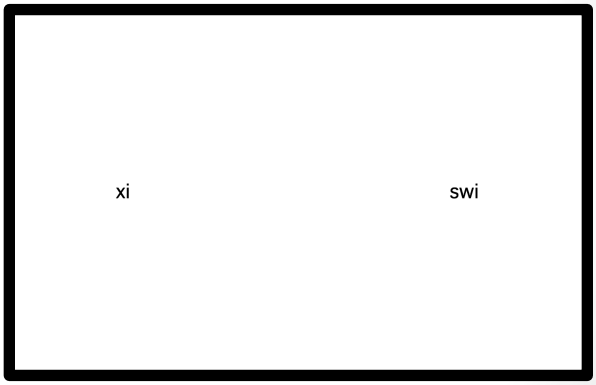
Procedure – Xitsonga, sound only



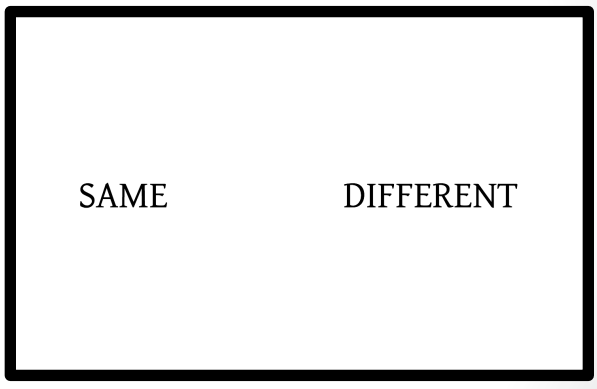
Experimental design (English speakers)

- Identification task
  - Speakers heard one word at a time, either singular ([ʃ]) or plural ([ʂ])
  - Participants pressed a button indicating whether the word they heard was [ʃ] or [ʂ]
- AX task
  - Speakers heard two words per trial
  - Pressed a button indicating whether the words had the same sound (both [ʃ] or both [ʂ]), or the words had different sounds.

Procedure – English identification



Procedure – English AX



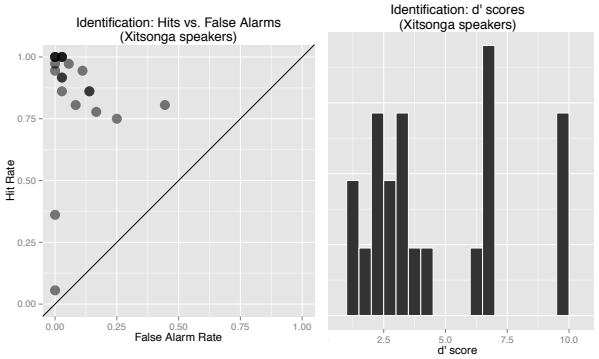
Hit rate, false alarms, d'

- Hit rate: how often does a participant say they heard "xi" when they heard "xi"
- False alarm rate: how often does a participant say they heard "xi" when they heard "xwi"
- d'-scores
  - z(Hit rate) – z(False alarm rate) = d'
  - z() is a function that fits this values to a normal distribution
  - Higher d' score means better discrimination (0 = no discrimination)
- Hit rate vs. false alarm rate
  - Higher d' when you have high hit rate and low false alarm rate (scores in top left of graph on following slide)

Results: Xitsonga Identification

- Speakers have high d' scores
  - Hit rates are high, false alarm rates are low
  - Most speakers are in the top-left corner of the graph
- d' scores are all above zero (and in fact much higher), suggesting a high degree of discriminability

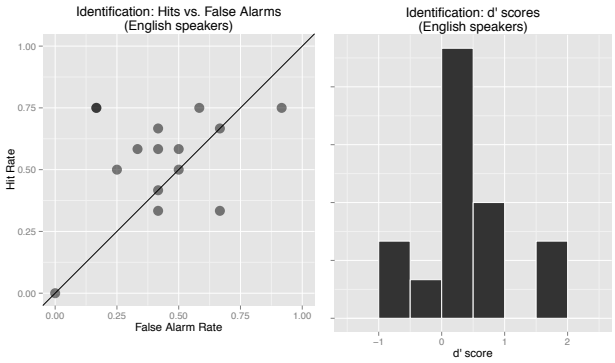
Results (Xitsonga speakers)



Results: English Identification

- Speakers have low d' scores
  - Hit rates are approximately equal to false alarm rates—speakers said "xi" regardless of whether they heard "xi" or "swi"
  - Speakers are centered around the hit rate = false alarm rate diagonal line
- d' scores center around zero, with a few speakers doing a little better or a little worse

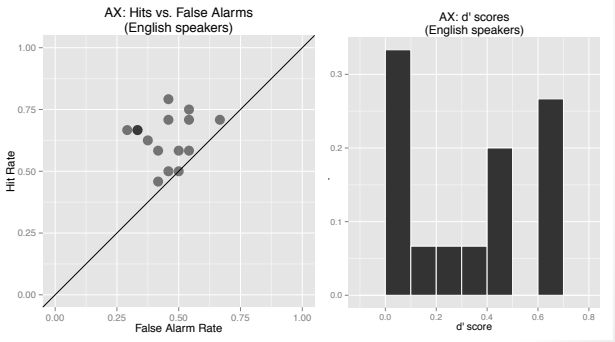
Results (English speakers: ID)



Results: English AX

- Speakers have low d' scores, but above zero
  - Hit rates are generally higher than false alarm rates, but not by a lot (indicated by speakers above the hit rate = false alarm rate line)
- d' scores center above zero, but nowhere near as high as Xitsonga speakers
- English speakers do better in the easier AX task than in Identification, since in an AX task speakers can compare the two sounds heard on each trial.

Results (English speakers: AX)



Summary – Perception study

- Xitsonga speakers can identify [ʂ] and [ʃ] at near ceiling levels
- English speakers ability to discriminate [ʂ] and [ʃ] is much worse:
  - Identification: speakers' d' scores average just above zero
  - AX: speakers' hit rates tend to be above their false alarm rates, but not nearly as high as Xitsonga speakers

Discussion

- In Xitsonga, there is a contrast between <x> and <sw>. In acoustic terms, the differences between these two sounds lie in M2 (variance of the frication noise) and dynamic amplitude
- This acoustic difference, which seems to be the source of the contrast, was not perceivable by English native speakers (even after a round of training of the sounds).
- What may count as a "small difference" in one language is perceptible at near-perfect levels in another.
  - Question: How do we define "small difference" if it is language-specific?

Acknowledgements

- We thank the consultants who participated in the experiment. We also thank Clementinah Burheni, Emson Mathebula, Walter Boloyi for recruiting Xitsonga participants and Manny Solario for running the English experiment.
- We also thank Will Bennett for comments on an earlier version of this study.

References

- Blacklock, O.S.B.: Characteristics of variation in production of normal and disordered fricatives using reducedvariance spectral methods; PhD thesis University of Southampton (2004).
- 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).
- 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.; De Decker, P.: Stabilization techniques for ultrasound imaging of speech articulations. J. acoust. Soc. Am. 117: 2544 (2005).
- Epstein, M.A.; Stone, M.: The tongue stops here: ultrasound imaging of the palate. J. acoust. Soc. Am. 118: 2128–2131 (2005). Li, M.; Kambhamettu, C.; Stone, M.: Automatic contour tracking in ultrasound images. Clin. Ling. Phonet. 19: 545–554 (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).

References

- Gu, C.: Gss: general smoothing R package version 2.0-9. http://cran.r-project.org/web/packages/gss/index.html (2012).
- Jesus, L.M.T.; Shadle, C.H.: A parametric study of the spectral characteristics of European Portuguese fricatives. J. Phonet. 30: 437–464 (2002).
- Laver, J.: Principles of phonetics (Cambridge University Press, Cambridge 1994).
- Lee-Kim, S.; Kawahara, S.; Lee, S. J.: The 'whistled' fricative in Xitsonga: its articulation and acoustics. Phonet. 71: 50-81 (2014).
- Macmillan, N.A., C. D. Creelman: Detection Theory: A User's Guide (Lawrence Erlbaum, Mahwah, NJ 2005).
- Shadle, C.H.: Experiments on the acoustics of whistling. Physics Teacher 21: 148–154 (1983).
- Shadle, C.H.: The aerodynamics of speech; in Hardcastle, Laver, The handbook of phonetic science, pp. 39–80 (Wiley-Blackwell, Chichester 2010).
- Shosted, R.K.: Just put your lips together and blow? Whistled fricatives in Southern Bantu. (2006)
- 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).
- Sitoe, B.: Dicionário Changana-Português (Instituto Nacional do Desenvolvimento da Educação, Maputo 1996).

Inkomu!!