# Structured Variation in the Phonetics of English and Czech Sibilant Fricatives

Eleanor Chodroff and Colin Wilson
Johns Hopkins University
Department of Cognitive Science

January 6, 2017 | Linguistic Society of America

Talkers vary significantly in the phonetic realization of speech sounds

Vowel formants
Stop consonant voice onset time (VOT)
Fricative spectral shape
Glottalization
Articulation of /l/ and /r/
etc.

e.g., Peterson and Barney, 1952; Redi and Shattuck-Hufnagel, 2001; Newman et al., 2001; Allen et al., 2003; Gick et al., 2006; Theodore et al., 2009

---

However, across talkers there is substantial similarity in the *patterns* of phonetic realization for different speech sounds

- Examples of structured variation
- Sources of structured variation
- Case study: sibilant fricatives in two languages

# Structured variation in phonetic realization

Covariation of vowel formants across talkers

Ex. Individuals form relatively congruent, but shifted vowel spaces

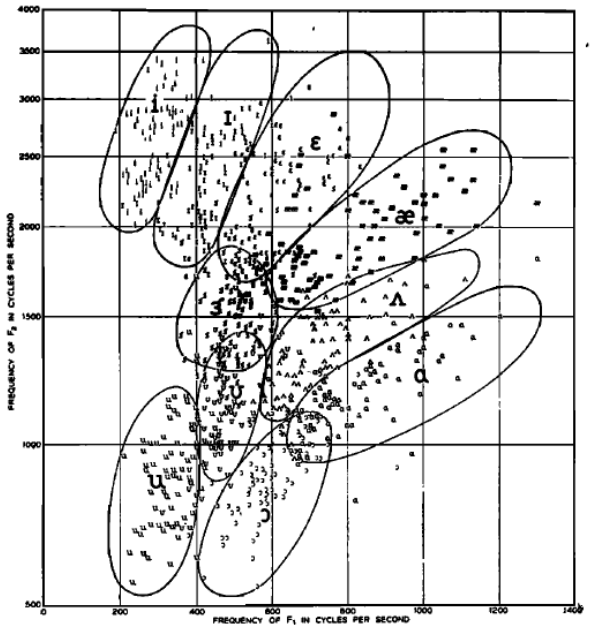
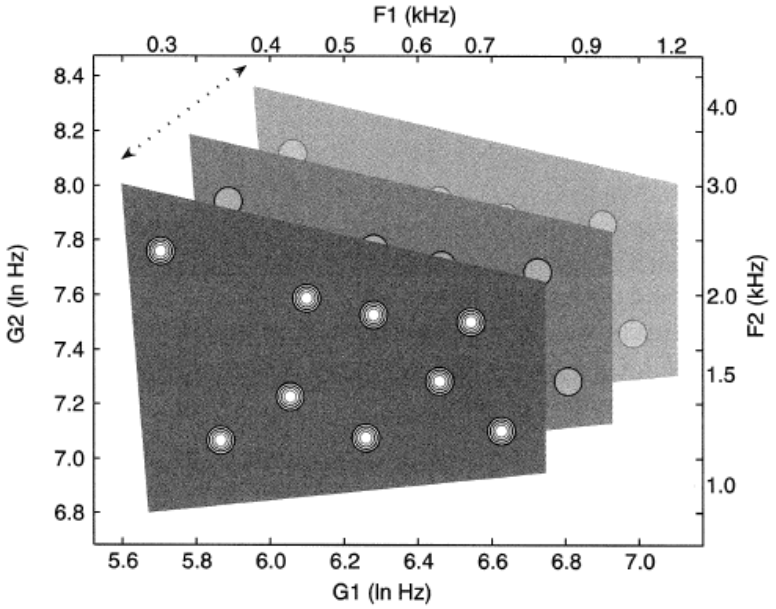



FIG. 8. Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.

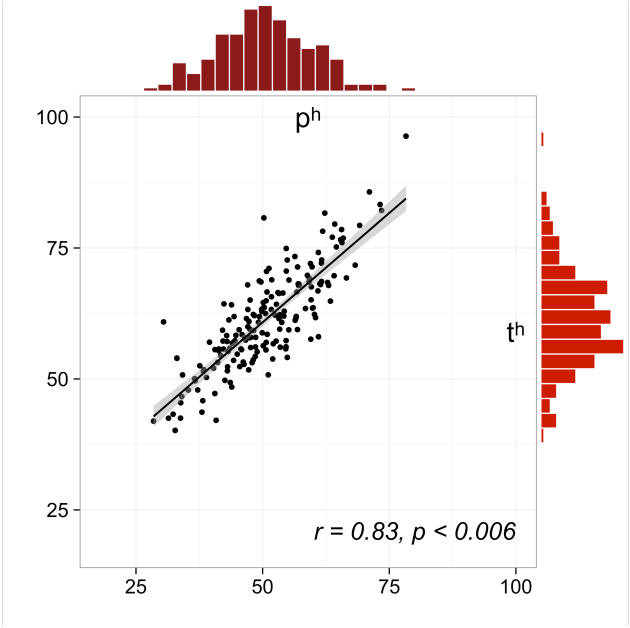


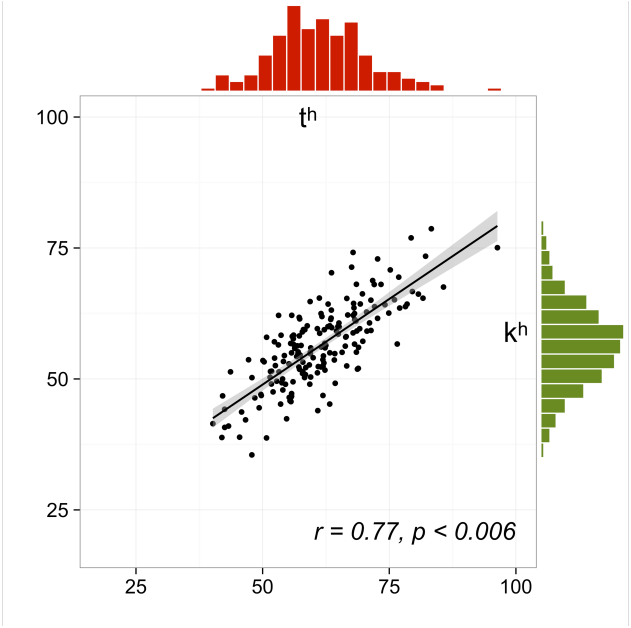

Joos, 1948; Peterson & Barney, 1952;
Nearey, 1978; Nearey & Assmann, 2007

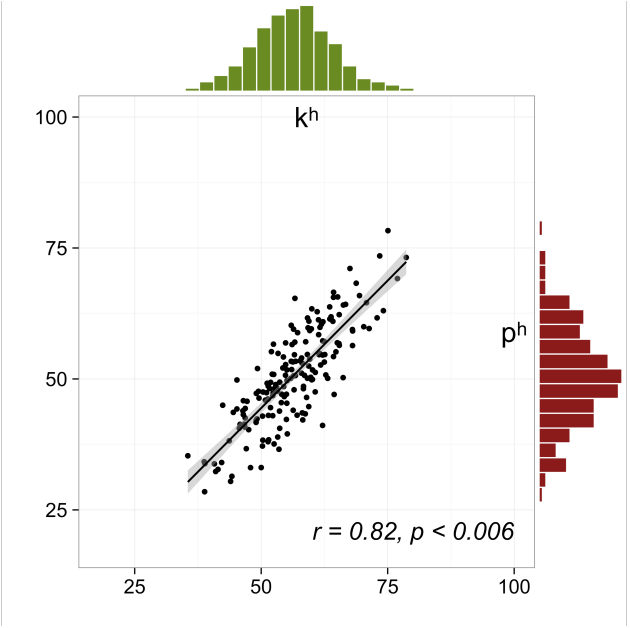
# Structured variation in phonetic realization

Covariation of mean VOT for stop consonants across talkers

Ex. Talkers with relatively long VOT for [$k^h$] also have a long VOT for [$t^h$]



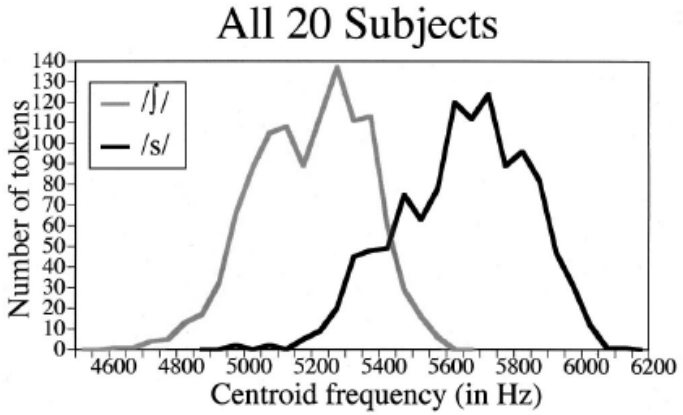



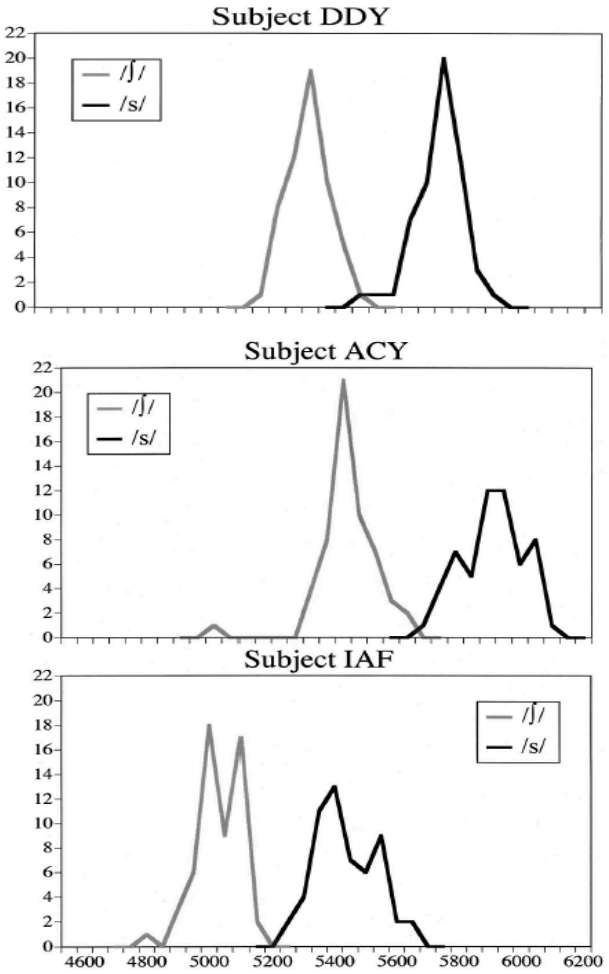

Chodroff & Wilson, to appear; Chodroff et al., 2015;
see also Koenig, 2000; Newman, 2003; Theodore et al., 2009

# Structured variation in phonetic realization

Systematic relations of talker means across sibilants: spectral center of gravity

Newman, Clouse, & Burnham (2001)



1. Introduction: Structured phonetic variation across talkers

2. Sources of structured variation
   - Talker anatomy
   - Uniformity constraints

3. Structured variation in American English sibilants

4. Structured variation in Czech sibilants

5. Discussion

# What gives rise to structured variation?

Talker-specific anatomy

- Anatomical properties of a talker (e.g., length of the vocal tract) will affect the articulation and resulting acoustics of many speech sounds (e.g., all vowel formants)

  e.g., Ladefoged & Broadbent, 1957; Nearey, 1978

- Factors that influence sibilant articulation / acoustics specifically:
  - Size of vocal tract, palate length and width, teeth

  e.g., Schwartz, 1968; Stevens, 1998; Fox & Nissen, 2005; Fuchs & Toda, 2010; Koenig et al., 2013

# What gives rise to structured variation?

Talker-specific anatomy

- Anatomical properties of a talker (e.g., length of the vocal tract) will affect the articulation and resulting acoustics of many speech sounds (e.g., all vowel formants)

  e.g., Ladefoged & Broadbent, 1957; Nearey, 1978

- Factors that influence sibilant articulation / acoustics specifically:
  - Size of vocal tract, palate length and width, teeth

  e.g., Schwartz, 1968; Stevens, 1998; Fox & Nissen, 2005; Fuchs & Toda, 2010; Koenig et al., 2013

But anatomy does not completely determine phonetic realization

- Language- / dialect- specific phonetics
- Conditioning by sociolinguistic variables

More generally, systematicity in the *physical outputs* of speech strongly indicates systematicity in the underlying *phonetic targets* (as determined by the language-/talker- specific grammar)

(Keating, 2003)

# What gives rise to structured variation?

Evidence for talker-specific control in phonetic realization of sibilant fricatives

- **Cross-linguistic variation**
  - Articulation of [s] (e.g., constriction width in English vs. German)

    Fuchs & Toda, 2010
  - Spectrum of sibilants

    Gordon et al., 2002; Heffernan, 2004; Li et al., 2007; Fuchs & Toda, 2010

- **Sociolinguistic variation**
  - Gender differences (beyond physiology)

    e.g., Flipsen et al., 1999; Strand, 1999; Stuart-Smith et al., 2003; Heffernan, 2004; Fuchs & Toda, 2010
  - Perceived and self-identified sexual orientation

    e.g., Linville, 1998; Munson et al., 2006; Campbell-Kibler, 2011; Brown, 2015
  - Socioeconomic status (SES)

    e.g., Stuart-Smith et al., 2003, 2007; Levon & Holmes-Elliot, 2013

# What gives rise to structured variation?

Two *uniformity constraints* could account for patterns in phonetic realization across talkers that are not completely determined by vocal tract morphology

1. **Uniformity of target**
   Within the phonetic grammar of an individual talker, the phonetic targets corresponding to a phonological feature value [αF] should be uniform (ideally, identical) for segments that are specified [αF]

   Ex. For a given a talker, [+spread glottis] should correspond to a glottal spreading gesture of uniform magnitude / relative timing for all [+s.g.] stops.

2. **Uniformity of contrast**
   Across talkers, the differences or ratios of phonetic targets corresponding to different values of the feature [F] should be uniform (ideally, identical).

See also Nearey, 1978

# Evaluating uniformity in sibilant fricatives

Examining uniformity in *place of articulation* targets for sibilant fricatives (i.e., targets corresponding to values of the feature [anterior])

Strong indicator of constriction location in sibilant acoustics

- Mid-frequency peak in spectrum ($Freq_M$)

Koenig et al., 2013

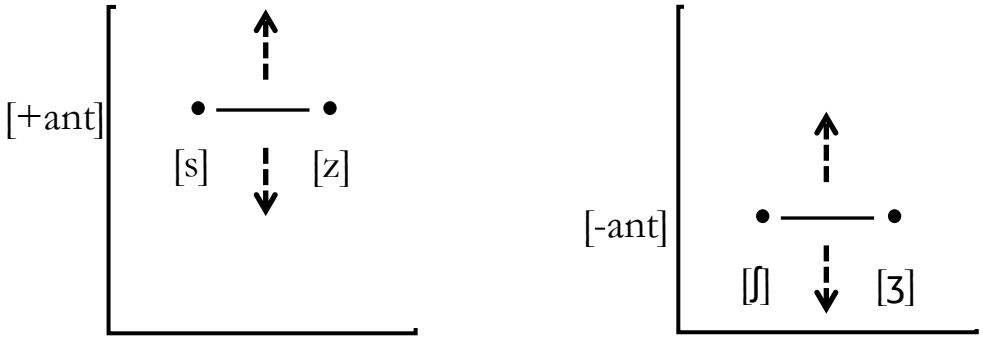
Two primary statistical methods

1. Covariation of talker mean $Freq_M$
2. Analysis of fixed effects and random effects for talker in mixed-effects models

# Evaluating uniformity in sibilant fricatives

**1. Uniformity of target**

place target
(constriction location)




Predictions:

Strong relations between [s] and [z], [ʃ] and [ʒ] across talkers

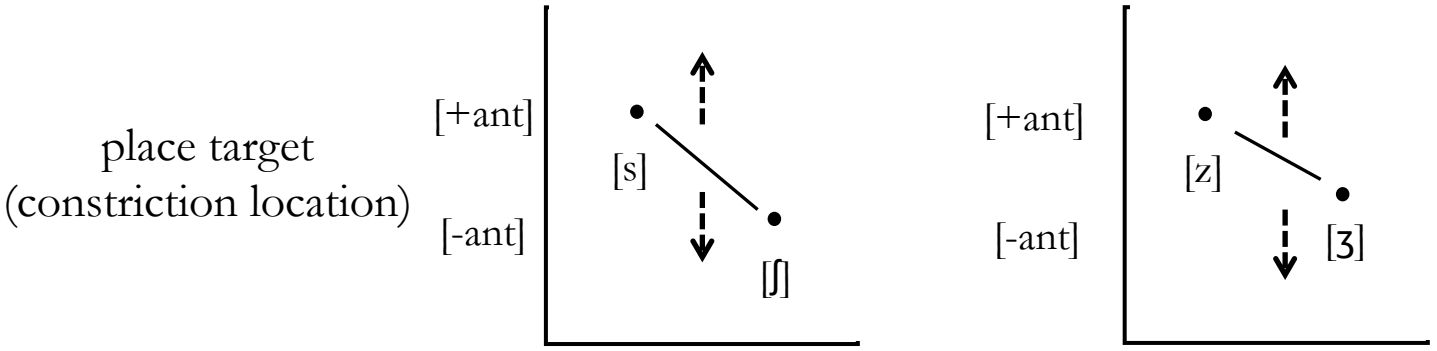
$$Freq_M \sim \beta_0 + \beta_1 * anterior + \beta_2 * voice + \beta_3 * anterior * voice +$$
$$z_0 + z_1 * anterior + z_2 * voice + z_3 * anterior * voice$$

Mixed-effects model:

- Place feature should account for the greatest variability in $Freq_M$ within and across talkers (minimal influence of [voice] and [anterior] × [voice] interaction)
- Variation across talkers should be found primarily in the grand mean (random talker intercept), with minimal variance of non-place feature [voice]

# Evaluating uniformity in sibilant fricatives

**2. Uniformity of contrast**




Predictions:

Covariation of [s] and [ʃ], [z] and [ʒ] across talkers

$$Freq_M \sim \beta_0 + \beta_1 * anterior + \beta_2 * voice + \beta_3 * anterior * voice +$$
$$z_0 + z_1 * anterior + z_2 * voice + z_3 * anterior * voice$$

Mixed-effects model:

- Variation across talkers should primarily be in the grand mean (random talker intercept), with minimal talker-specific effect of [anterior]

1. Introduction: Structured (patterned) variation

2. Account of structured variation
   - Talker physiology
   - Principle of uniformity

3. Uniformity in American English sibilants

4. Uniformity in Czech sibilants

5. Discussion

# American English: Mixer 6 Corpus

**Recordings**

Read speech – sentences selected from Switchboard

Sentence length: 1-17 words (median: 7)

3 separate sessions, ~15 minutes each

Sentences read in same order within each session

Sampled at 16,000 Hz

Available from the LDC

Sentences with reading/recording errors removed from analysis

**Talkers**

180 native talkers of American English

102 female, 78 male

Age: 19 – 86 years old (median: 27)

Corpus: Brandschain et al., 2010, 2013
Corpus audit: Chodroff et al., 2016
cf. corpus studies from: Byrd, 1993; Cole et al., 2004; Yao, 2007; Yuan & Liberman, 2008; Davidson, 2011; Gahl et al., 2012; Labov et al., 2013; Elvin & Escudero, 2015; Stuart-Smith et al., 2016

# Mixer 6 Fricatives

/s, z, ʃ/: word-initial, word-medial, a few word-final sibilants before vowels

Multitaper spectral analysis on middle 50% of fricative
Measured $Freq_M$: peak amplitude between 2000-7000 Hz

Adapted from Koenig et al. (2013)

Excluded tokens ±2.5 standard deviations from talker-specific category mean

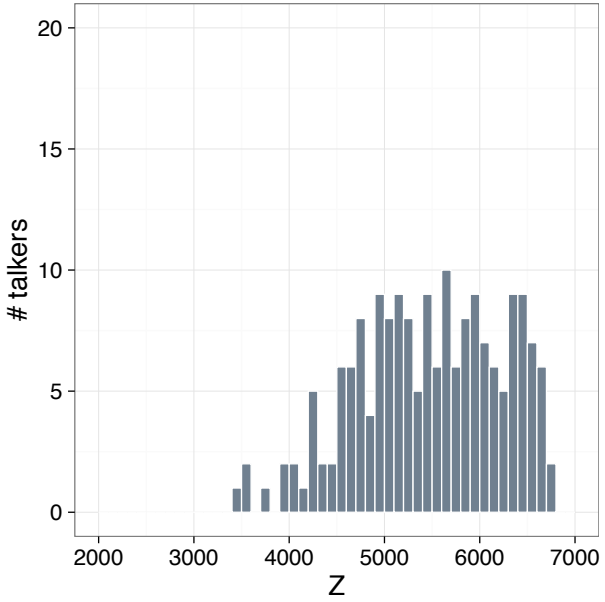
54,972 sibilants in $Freq_M$ analysis

| Fricative | Range per talker | Median # Tokens | Total |
|---|---|---|---|
| [s] | 110 - 314 | 222 | 39,192 |
| [z] | 21 - 43 | 33 | 5,972 |
| [ʃ] | 29 - 84 | 54 | 9,808 |

# Talker variation in mean $Freq_M$: American English

[z]

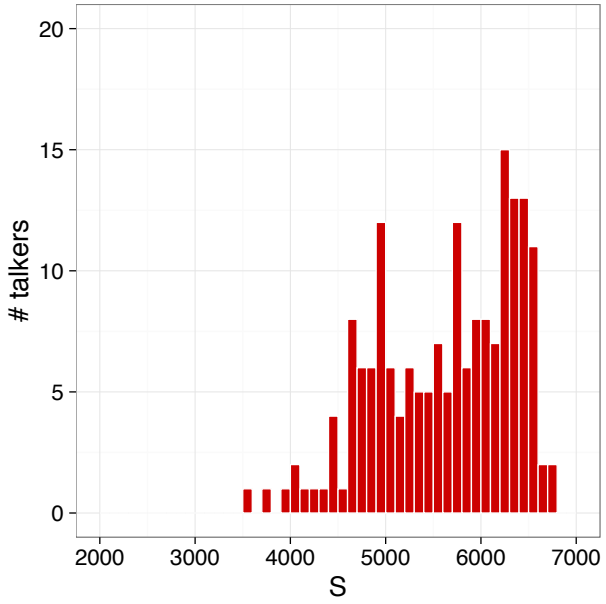
$\mu$ = 5474 Hz




Range of talker means
3487 – 6796 Hz

[s]

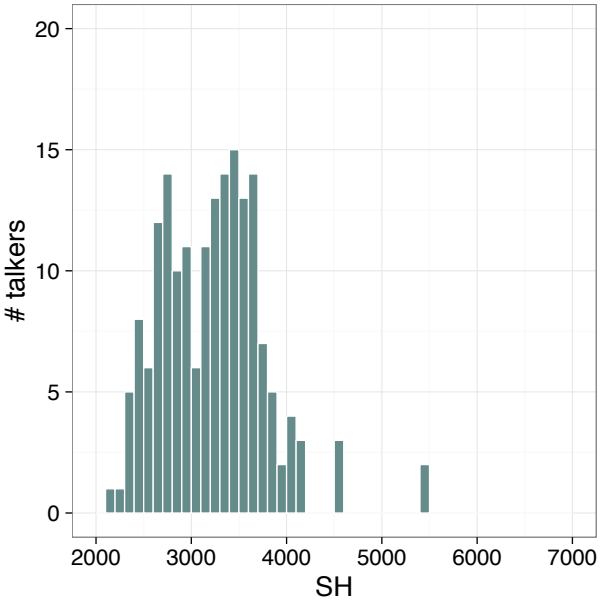
$\mu$ = 5633 Hz




Range of talker means
3506 – 6735 Hz

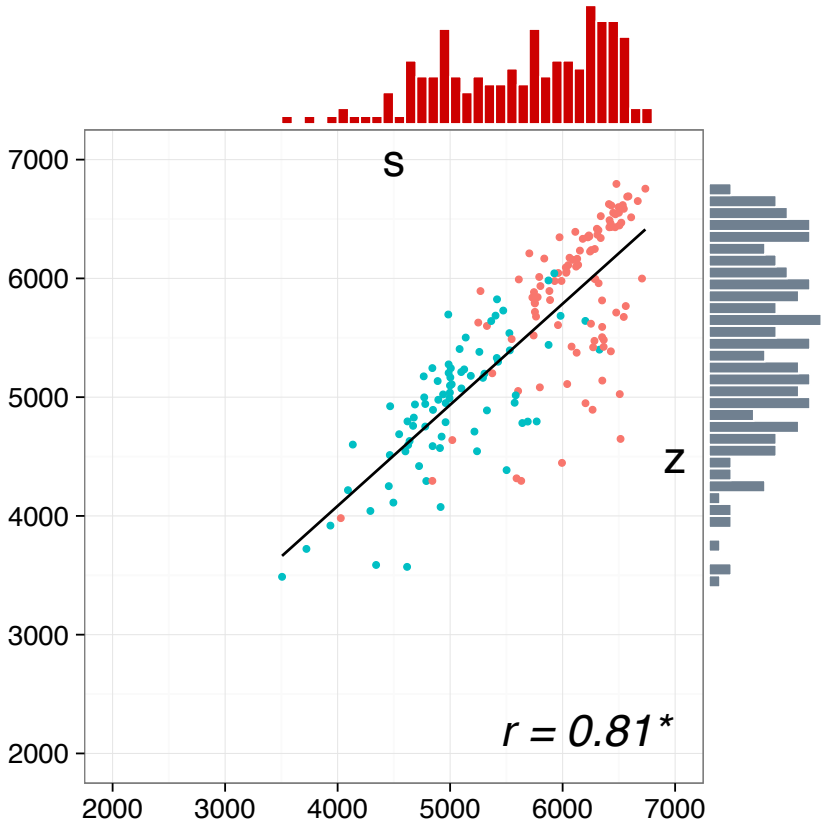
[ʃ]

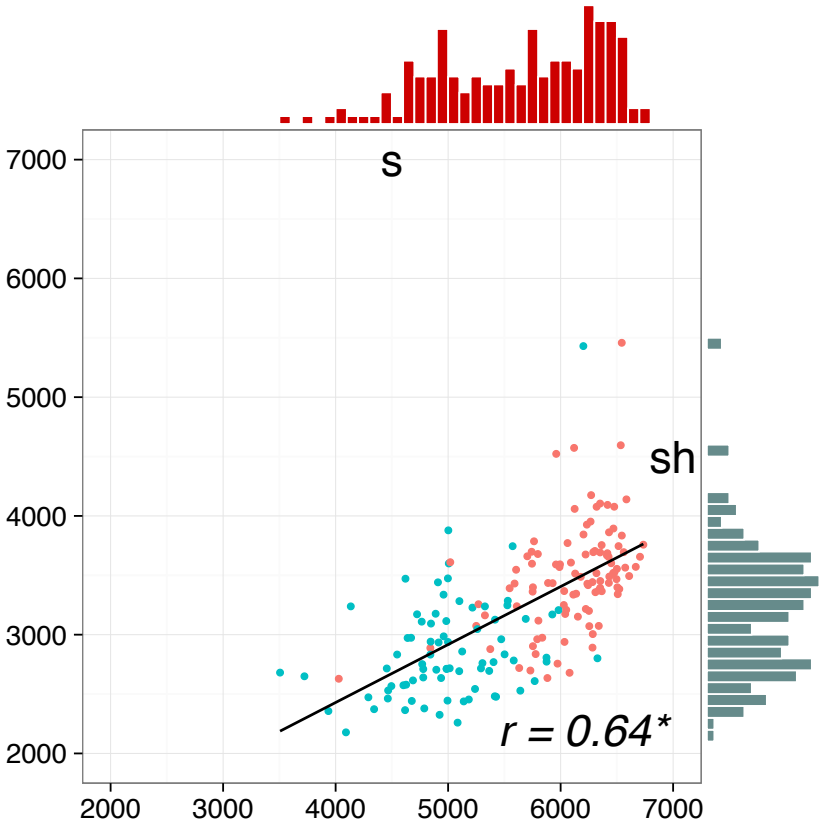
$\mu$ = 3226 Hz




Range of talker means
2178 – 5458 Hz

# Covariation of $Freq_M$ means: American English




[s] – [z]
[0.74, 0.86]
Females: $r = 0.60$*
Males: $r = 0.75$*

Female
Male

[s] – [ʃ]
[0.56, 0.71]
Females: $r = 0.44$*
Males: $r = 0.32$*

$ps < 0.001$

# Mixed-effects analysis: American English

**Fixed effects:**

place (+anterior, -anterior)

e.g., Hughes & Halle, 1956; Strevens, 1960; Forrest et al., 1988

voice (+voice, -voice)

e.g., Hughes & Halle, 1956; Jongman et al., 2000; Silbert & de Jong, 2008

vowel roundness (+round, -round)

e.g., Mann & Repp, 1980; Soli, 1981; Whalen, 1981; Johnson, 1991

gender (female, male)

e.g., Schwartz, 1968; Whiteside, 1996; Heffernan, 2004; Fox & Nissen, 2005; Fuchs & Toda, 2010

place x gender

Fox & Nissen, 2005

voice x gender

Different rate and extent of voicing by gender?

$Freq_M$ centered (subtracted the mean, collapsing across all talkers/sibilants)
All factors weighted-effect coded to correct for unequal sample sizes (Darlington, 1990)

$$Freq_M \sim 1 + place + voice + rounding + gender + place*gender + voice*gender + (1 \mid word) + (1 + place + voice \mid talker)$$

# Mixed-effects analysis: American English

$Freq_M$ ~ 1 + place + voice + rounding + gender + place*gender + voice*gender + (1 | word) + (1 + place + voice| talker)

| fixed effects | Beta | t-value |
|---|---|---|
| intercept | 16 | 0.4 |
| **place** | **459** | **52.3** |
| voice | 105 | 18.0 |
| vowel rounding | -13 | -0.3 |
| gender | 431 | 15.8 |
| place x gender | 37 | 6.2 |
| voice x gender | 6 | 1.8 |

$|\beta_{place}| > |\beta_{other}|$

| Talker random effect | sd |
|---|---|
| **intercept** | **418** |
| place | 88 |
| voice | 48 |

$\sigma_{intercept} > \sigma_{other}$

# Discussion: American English

Substantial variation in realization of $Freq_M$ across talkers

**Uniformity of target:**
Strong covariation of [s] and [z], with comparable constriction targets
$|\beta_{place}| > |\beta_{other}|$ indicates that the non-place feature [voice] has less influence on the constriction target (i.e., target is not very sensitive to intrasegmental context)

**Uniformity of contrast:**
Moderately strong correlation between [s] and [ʃ] across talkers

**Predictions of both:**
$\sigma_{intercept} > \sigma_{other}$ Greatest variation across talkers is in the grand mean (random talker intercept) as opposed to talker-specific effects of place or voice

Evidence for both types of constraint, but stronger indication of target ('within-feature') uniformity than of contrast ('between-feature') uniformity

1. Introduction: Structured (patterned) variation

2. Account of structured variation
   - Talker physiology
   - Principle of uniformity

3. Uniformity in American English sibilants

4. Uniformity in Czech sibilants

5. Discussion

# Why Czech?

Full place × voice contrast in sibilants: [s z ʃ ʒ]
Native words beginning with all four sibilants
Multi-talker corpus available

West Slavic language

10.6 million speakers

8 fricatives (4 sibilants)

5 monophthongs ([i e a ɔ u]), long and short
3 diphthongs ([au eu ou])

| **Labiodental** | **Alveolar** | **Post-alveolar** | **Velar** | **Glottal** |
|---|---|---|---|---|
| f | **s** | **ʃ** | x | |
| v | **z** | **ʒ** | | ɦ |

# Nijmegen Corpus of Casual Czech

**Recordings**
30 hours of speech

Spontaneous informal speech

Three same-sex friends per recording group

Sampled at 44,100 Hz

**Talkers**
60 native Czech talkers from Prague and Central Bohemia region

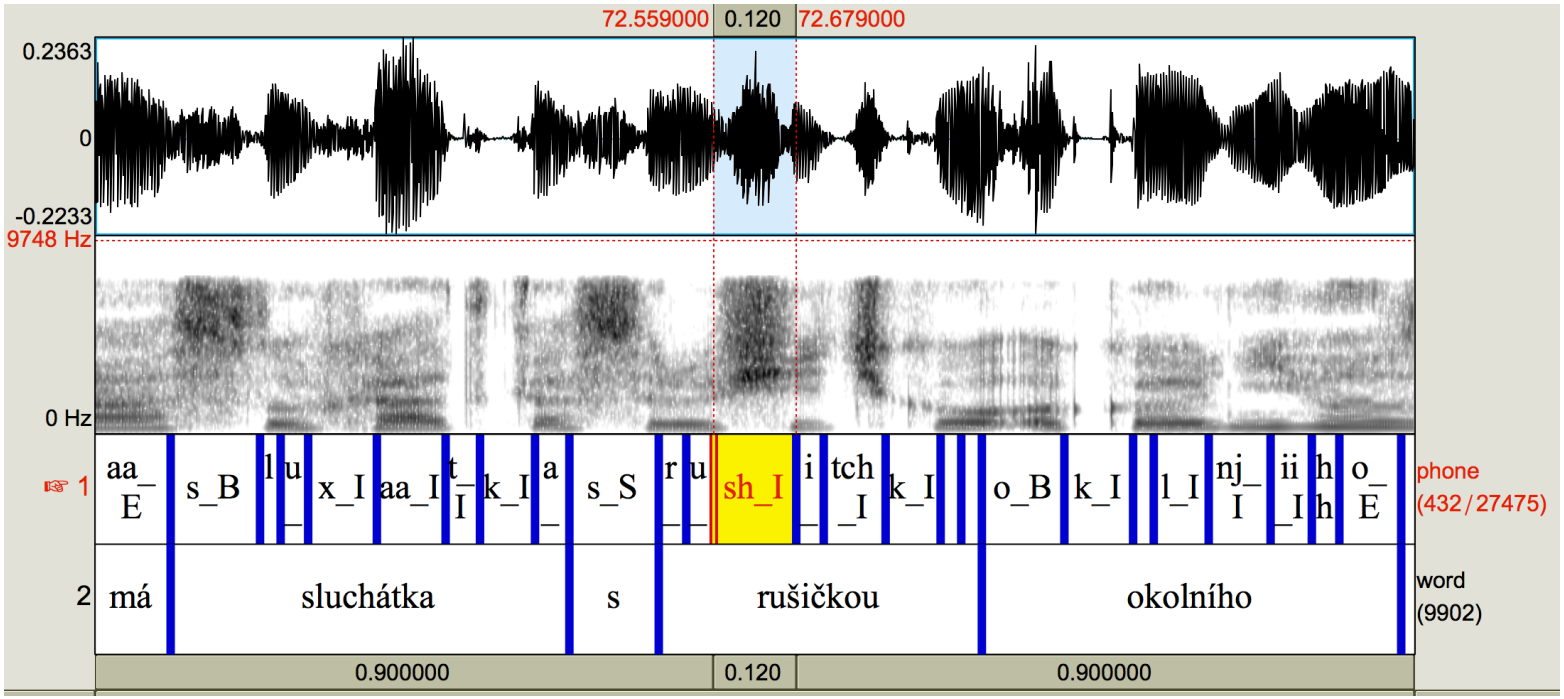
30 female, 30 male

Age: 19 – 26 years old

Corpus: Ernestus et al., 2014
Kaldi: Povey et al., 2011

# Nijmegen Corpus of Casual Czech

- Czech pronunciation lexicon developed with custom grapheme to phon(eme) conversion script enriched with phonological rules operating on orthographic representation
- Acoustic models trained using Kaldi ASR toolkit on cleaned version of the transcripts
- Fricatives extracted from output of Kaldi forced alignment




Kaldi tutorial: Google my name + Kaldi tutorial
http://pages.jh.edu/~echodro1/tutorial/kaldi/kaldi-intro.html

# Nijmegen Corpus of Casual Czech

/s, z, ʃ, ʒ/: word-initial, word-medial, a few word-final sibilants before vowels

Multitaper spectral analysis on middle 50% of fricative
Measured $Freq_M$ as for American English

Excluded tokens ±2.5 standard deviations from talker-specific category mean

51,793 sibilants for analysis

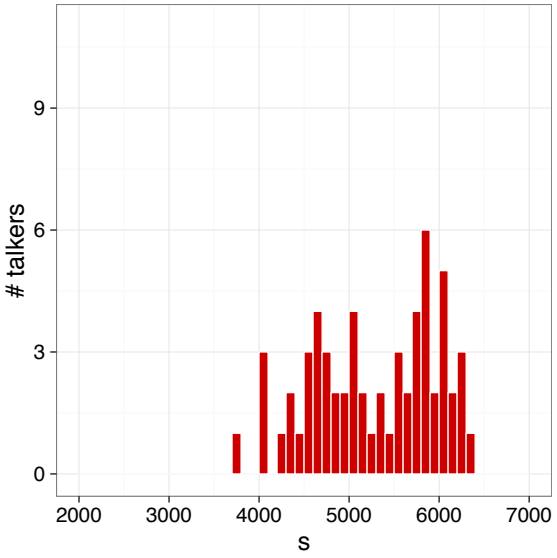
| Fricative | Range per talker | Median # Tokens | Total |
|---|---|---|---|
| [s] | 128 - 895 | 353 | 23,997 |
| [z] | 43 - 351 | 122 | 8,235 |
| [ʃ] | 24 - 179 | 69.5 | 4,813 |
| [ʒ] | 45 - 572 | 211 | 14,748 |

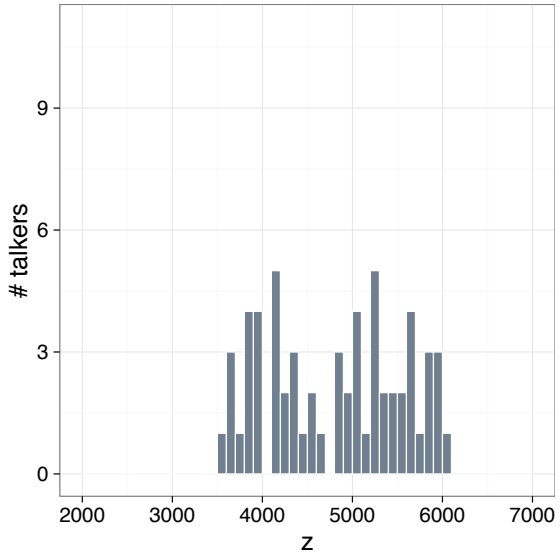
# Talker variation in mean $Freq_M$: Czech

[s]

$\mu$ = 5283 Hz

Range of talker means
3790 – 6382 Hz




[z]

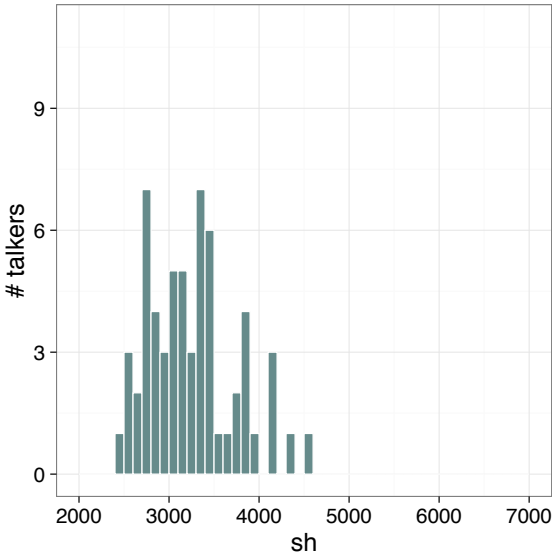
$\mu$ = 4800 Hz

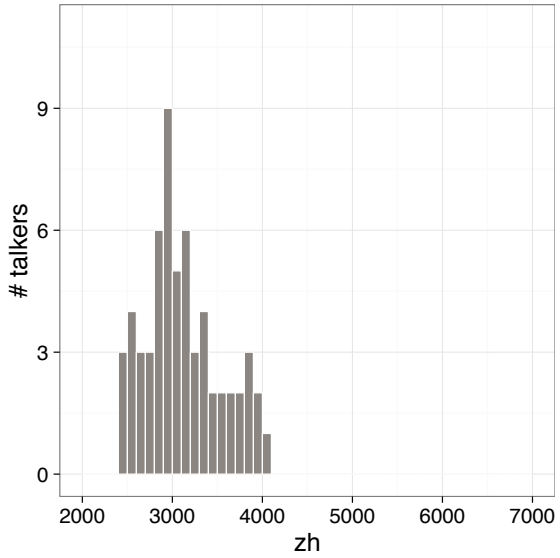
Range of talker means
3578 – 6017 Hz

[ʃ]

$\mu$ = 3252 Hz

Range of talker means
2400 – 4509 Hz



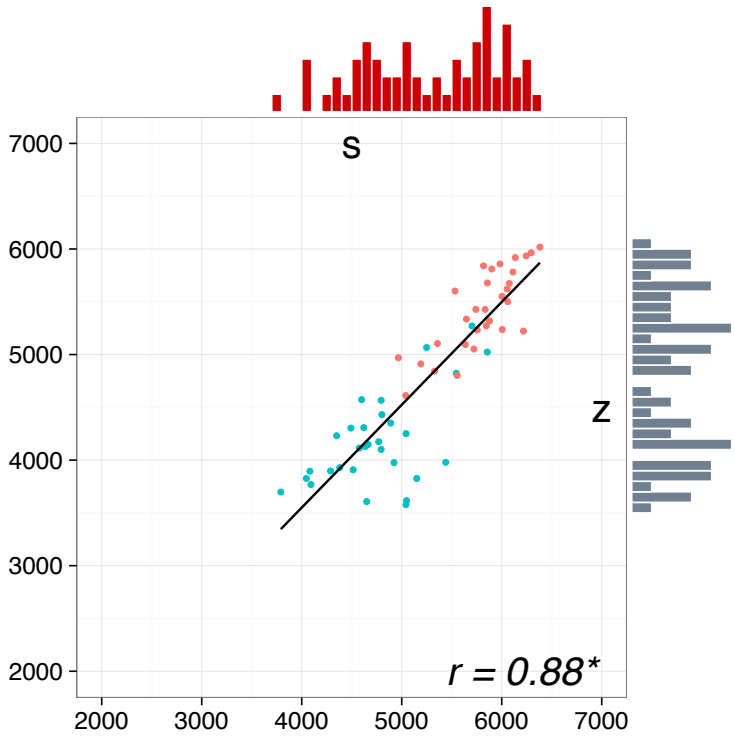

[ʒ]

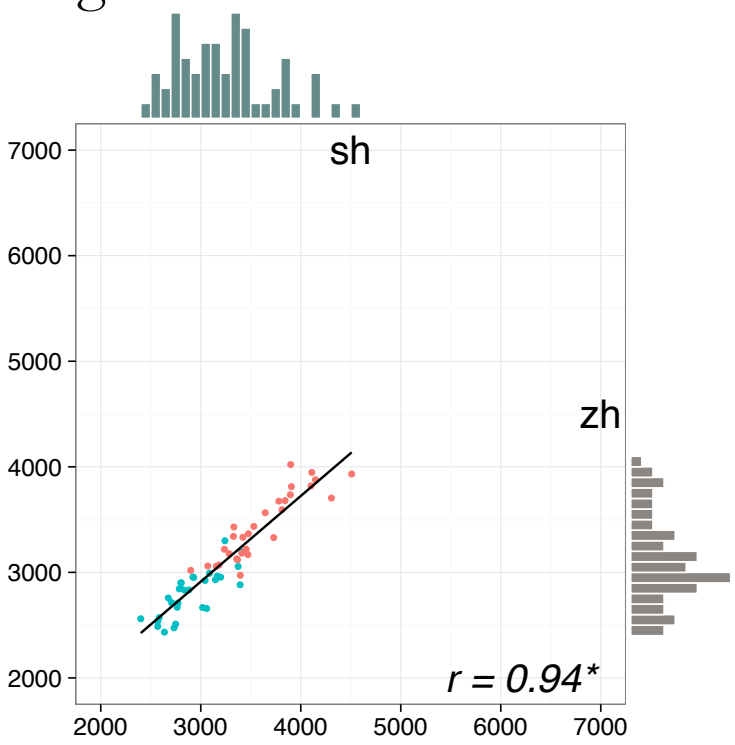
$\mu$ = 3117 Hz

Range of talker means
2434 – 4022 Hz

# Covariation of $Freq_M$ means: Czech

## Uniformity of target




[s] – [z]
[0.81, 0.92]

Females: $r = 0.79$*
Males: $r = 0.60$*

Female
Male

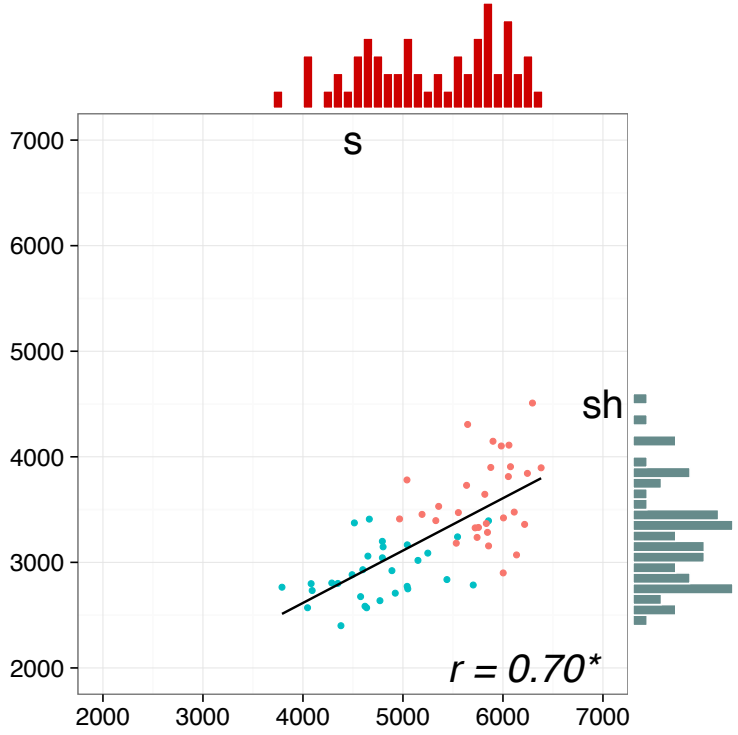
[ʃ] – [ʒ]
[0.90, 0.96]

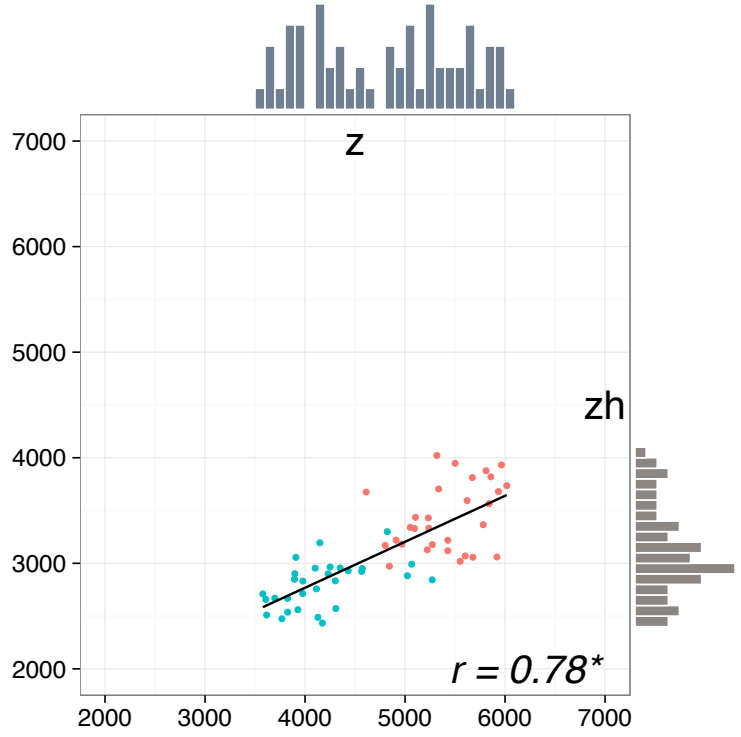
Females: $r = 0.89$*
Males: $r = 0.78$*

* = $p < 0.001$

# Covariation of $Freq_M$ means: Czech

## Uniformity of contrast




[s] – [ʃ]
[0.56, 0.79]

Females: $r = 0.23$, *n.s.*
Males: $r = 0.41$, *n.s.*

Female
Male

[z] – [ʒ]
[0.68, 0.86]

Females: $r = 0.35$, *n.s.*
Males: $r = 0.50$, $p < 0.01$

* = $p < 0.001$

# Mixed-effects analysis: Czech

Fixed effects:

place (+anterior, -anterior)

e.g., Hughes & Halle, 1956; Strevens, 1960; Forrest et al., 1988

voice (+voice, -voice)

e.g., Hughes & Halle, 1956; Jongman et al., 2000; Silbert & de Jong, 2008

vowel roundness (+round, -round)

e.g., Mann & Repp, 1980; Soli, 1981; Whalen, 1981; Johnson, 1991

gender (female, male)

e.g., Schwartz, 1968; Whiteside, 1996; Heffernan, 2004;
Fox & Nissen, 2005; Fuchs & Toda, 2010

place*voice*roundness*gender

$Freq_M$ centered (subtracted the mean, collapsing across all talkers/sibilants)
All factors weighted-effect coded to correct for unequal sample sizes (Darlington, 1990)

$$Freq_M \sim 1 + place*voice*rounding*gender + (1 \mid word) + (1 + place*voice \mid talker)$$

# Mixed-effects analysis: Czech

| fixed effects | Beta | t-value |
|---|---|---|
| intercept | 16 | 0.4 |
| **place** | **778** | **37.2** |
| voice | 187 | 11.4 |
| vowel rounding | -418 | -19.9 |
| gender | 489 | 12.0 |
| place x voice | 76 | 7.9 |
| place x rounding | -43 | -2.4 |
| place x gender | 104 | 5.4 |
| voice x rounding | -106 | -5.8 |
| voice x gender | -23 | -1.6 |
| place x voice x gender | -25 | -3.1 |
| place x voice x rounding | -42 | -2.7 |
| voice x rounding x gender | 28 | 2.3 |
| place x voice x rounding x gender | 28 | 2.7 |

$Freq_M \sim 1 + place*voice*rounding*gender + (1 \mid word) + (1 + place*voice \mid talker)$

| Talker random effect | sd |
|---|---|
| **intercept** | **325** |
| place | 153 |
| voice | 113 |
| place x voice | 59 |

$\sigma_{intercept} > \sigma_{other}$

$|\beta_{anterior}| > |\beta_{other}|$

# Discussion: Czech

Substantial variation in realization of $Freq_M$ across talkers

**Uniformity of target:**
Strong correlation between [s] and [z], [ʃ] and [ʒ]
$|\beta_{place}| > |\beta_{other}|$ indicates that the non-place feature [voice] has less influence on the constriction target (i.e., target is not very sensitive to intrasegmental context)

**Uniformity of contrast:**
Moderately strong correlation between [s] and [ʃ], [ʃ] and [ʒ]
Correlations break down within gender

**Predictions of both:**
$\sigma_{intercept} > \sigma_{other}$ Greatest variation across talkers is in the grand mean (random talker intercept) as opposed to talker-specific effects of place or voice

Evidence for both types of constraint, but stronger indication of target ('within-feature') uniformity than of contrast ('between-feature') uniformity

# Discussion: Constraints on realization

Largely comparable findings across American English and Czech in correlations and mixed-effects analyses of sibilant fricative peak frequencies

Evidence for **uniformity of target** stronger than for **uniformity of contrast**

**Uniformity of target** is a constraint on talker-specific phonetic grammars, not plausibly reducible to other hypothesized constraints such as:

- Articulatory ease
- Perceptual distinctiveness

e.g., Liljencrants & Lindblom, 1972; Lindblom, 1983, 1986, 1990; Flemming, 2004

**Uniformity of contrast** could exist as an independent "conformity" constraint, or its weaker effects may be reducible to a landscape of ease / distinctiveness trade-offs

Ex. Some talkers may favor ease relative to distinctiveness (or vice versa), variation allowed by OT/HG formalizations of phonology and phonetics.

e.g., Kirchner 1998, 2000; Padgett and Żygis, 2007; Boersma & Hamann 20006, 2008 ; Flemming, 2011

# Discussion: Talker adaptation in perception

Strong covariation across talkers also implies mutual predictability among speech sounds

Listeners could exploit structured variation to extrapolate from limited talker-specific evidence and refine a talker-specific model with further exposure.

- Joint (rather than independent) estimation of talker-specific phonetic properties
- Implications for cognitive models of talker adaptation

See also Lobanov, 1971; Nearey, 1978, Furui, 1980; Cox, 1995

Preliminary evidence that listeners generalize talker-specific spectral properties across fricative categories after minimal exposure.

- After exposure to a talker with a high COG [z], listeners shift boundary between [s] and [ʃ] to a higher COG

Chodroff et al., 2016 (ASA)

# Future Directions

Determine how broadly the principle of uniformity applies

- Does uniformity apply across the duration of fricatives (i.e., to dynamic targets)?
- Does uniformity of [-anterior] target extend to affricates such as [tʃ] and [dʒ]?
- Is place target observed for [s] and [z] uniform with place targets for homorganic stops ([t] and [d])?

Investigate talker covariation of $Freq_M$ among sibilants in other languages

Apply analysis to articulatory correlate of the place target (e.g., constriction location)

- Articulation may reveal targets more closely (less affected by laryngeal source)
- Some evidence, however, indicating that tongue body tends to be slightly lowered for voiced than voiceless consonants.
    - Is this an articulatory consequence of the targets for the [voice] feature or true context-sensitivity (intrasegmental) in the place target?

Suzy Ahn, p.c., Ahn & Davidson, 2016

# Conclusion

Variation in phonetic realization is extensive across talkers within a language.

Implications for theory of phonetic realization and models of perceptual adaptation:

- Uniformity constraint restricts variation in the phonetic implementation of speech sounds

- Prior knowledge of relations of mutual predictability among speech sounds may allow for rapid adaptation to novel talkers

More generally, structured variation and uniformity contribute to our understanding of the phonetic grammar:

- Should be evaluated along further acoustic and articulatory dimensions and for additional segments and languages
- Can be examined in any corpus with multiple talkers – laboratory or spontaneous
- Uniformity can be evaluated relative to other known constraints on the grammar (e.g., perceptual distinctiveness, articulatory ease)

# Thank you!

Thanks to:
Jack Godfrey
Sanjeev Khudanpur
Matthew Maciejewski
Christine Shadle
Paul Smolensky
Jan Trmal
Doug Whalen

This work was partially supported by:
JHU Distinguished Science of Learning Fellowship
Dolores Zohrab Liebmann Fund

# Model comparison: American English

$Freq_M$ ~ 1 + place + voice + rounding + gender + place*gender + voice*gender + (1 | word)

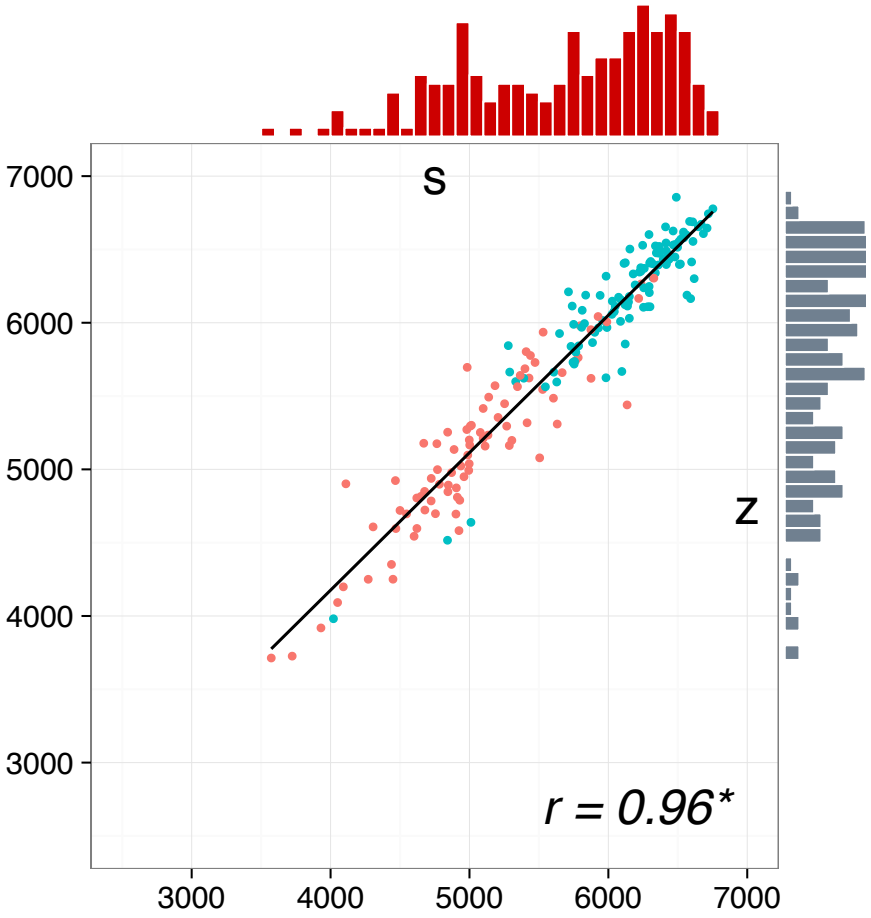
| Talker Random Effect Structure | BIC | ΔBIC |
|---|---|---|
| + 0 | 899,824 | ⌀ |
| + (1 \| talker) | 886,382 | 13,442 |
| + (1 + place\| talker) | 883,411 | 2,971 |
| + (1 + place + voice \| talker) | 881,949 | 1,462 |

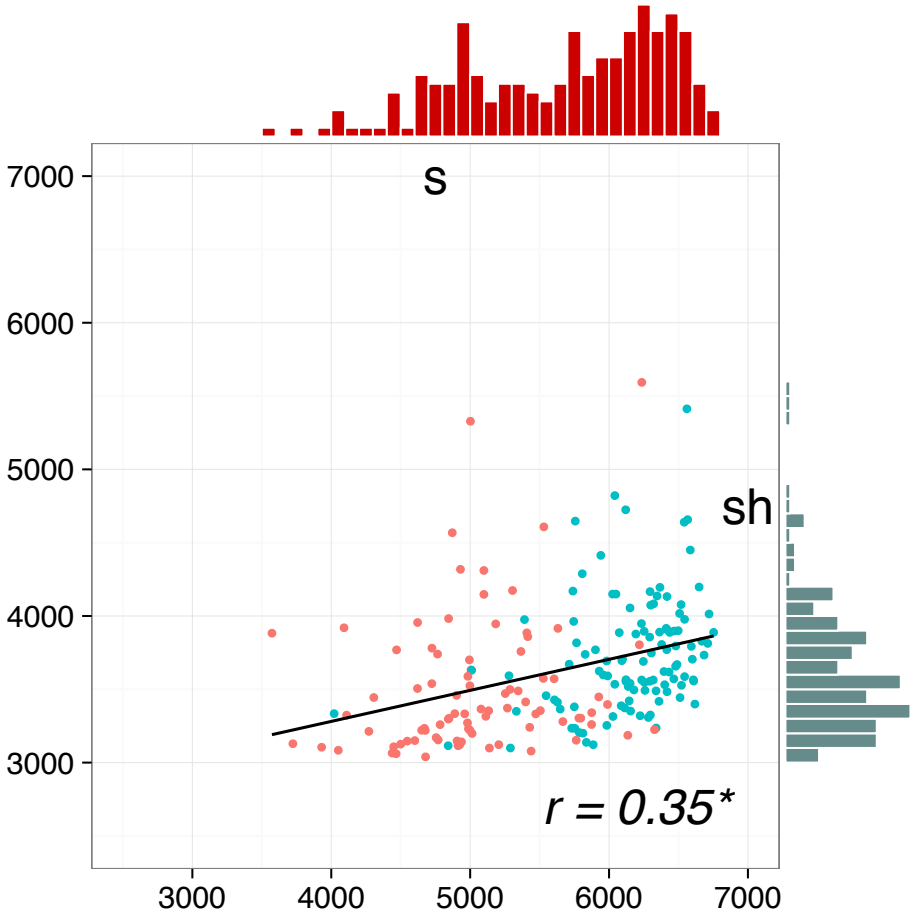
# Model comparison: Czech

$Freq_M$ ~ 1 + place*voice*rounding*gender +
(1 | word)

| Talker Random Effect Structure | BIC | ΔBIC |
|---|---|---|
| + 0 | 851,952 | ⌀ |
| + (1 \| talker) | 845,515 | 6,437 |
| + (1 + place\| talker) | 843,174 | 2,341 |
| + (1 + place+ voice \| talker) | 842,410 | 764 |
| + (1 + place*voice\| talker) | 842,123 | 287 |

# Mixer 6 Freq3k7k



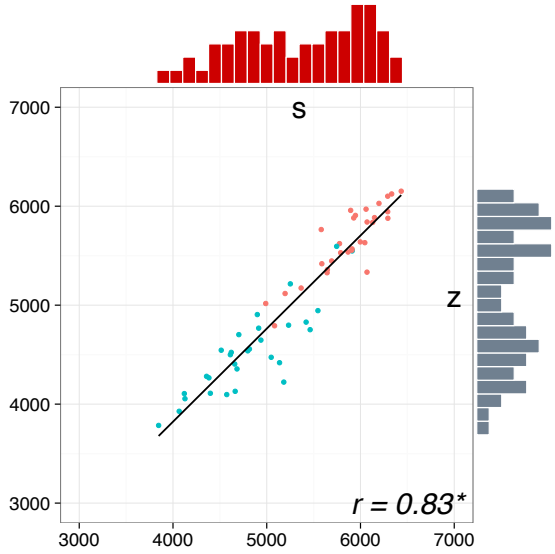




[s] vs [z]
[0.95, 0.97]

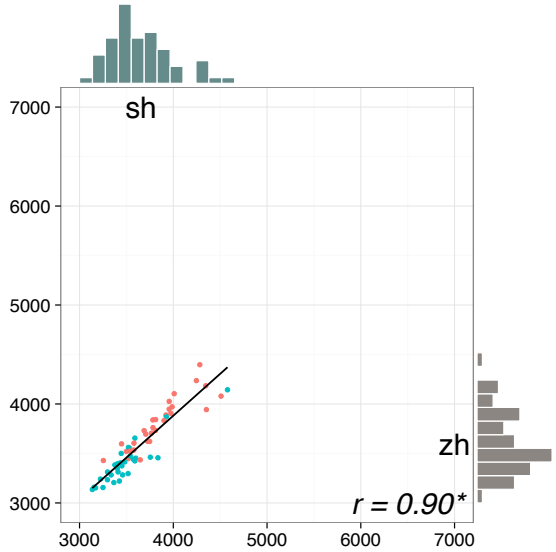
[s] vs [ʃ]
[0.21, 0.46]

*p*s < 0.001

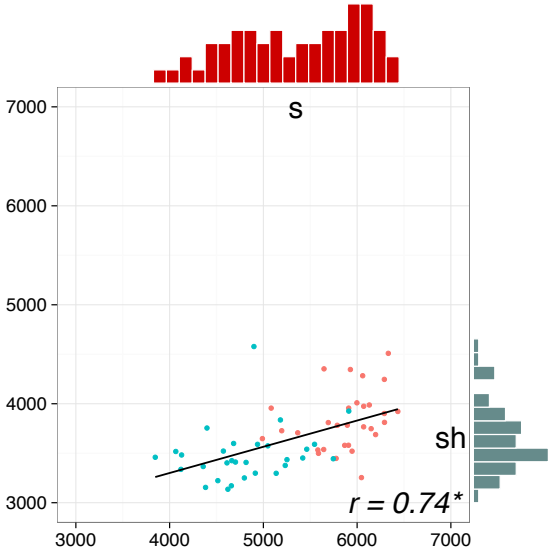
# Czech Freq 3k-7k (Hz)



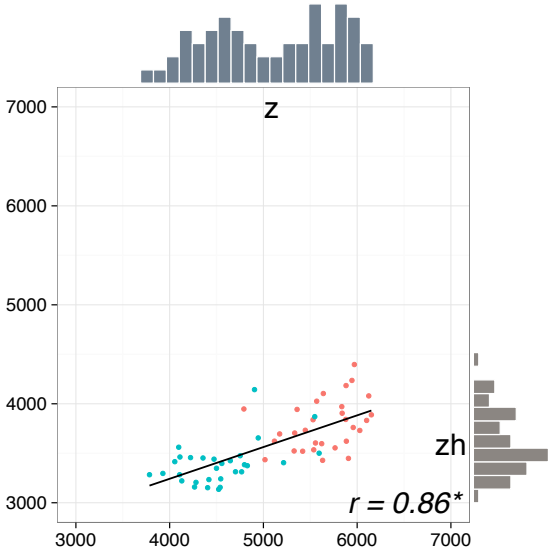

/s/ - /z/




/ʃ/ - /ʒ/




/s/ - /ʃ/




/z/ - /ʒ/

*p*s < 0.001