Auditory and acoustic-phonetic mechanisms of adaptation in the perception of sibilant fricatives

JOHNS HOPKINS UNIVERSITY

Eleanor Chodroff and Colin Wilson

Department of Cognitive Science, Johns Hopkins University

4pSC66

Introduction

Previous research has demonstrated that speech perception is highly dependent on preceding acoustic context (e.g., Ladefoged & Broadbent, 1957; Mann, 1980), and suggested that this reflects spectral contrast effects (e.g., Kingston & Diehl, 1995; Lotto & Kluender, 1998) or adaptation to the long-term average spectrum (LTAS; e.g., Holt, 2006).

Contrast effects have been related to general auditory mechanisms that could facilitate perceptual adaptation to a novel talker (e.g., Holt, 2006; Laing et al., 2012). The spectral contrast account of talker adaptation can be summarized as follows:

Spectral contrast account

- *High* frequency energy in a preceding sound should enhance *low* frequency energy present in a subsequent sound (and vice versa), shifting perception *contrastively*
- Adaptation should occur only when context sounds have energy in the frequency ranges that are relevant for perception (discrimination or categorization) of targets
- Non-speech contexts should elicit the same effects as matched speech context

(e.g., Lotto & Kluender, 1998; Holt, 2005, 2006; Laing et al., 2012)

Contribution of this study: we compare spectral contrast and two alternative accounts of **extrinsic talker adaptation** with respect to the perception of **fricatives**.

Cue-based normalization account

- Members of a natural class of sounds can be characterized by a common set of acoustic/auditory *cues* (e.g., formants for vowels, burst spectra & transitions for stops)
- Cue values for each sound in a class are represented relative to a cue-specific *mean*
- Talker adaptation involves determining the talker's mean for each cue and appropriately *shifting* the observed tokens of *all* class members (i.e., mean subtraction)

(e.g., Lobanov, 1971; Nearey, 1978; McMurray & Jongman, 2011)

Covariation account

- Members of a natural class have cue values that *covary* across talkers (to varying degrees). Ex. Talker mean COGs for [s] and [z] are highly correlated (cf. [s] and [v])
- Listeners infer talker-specific parameters for each sound in a way that takes into account such covariation relations. Ex. If observe high COG [z], infer high COG [s]

(e.g., Chodroff et al., 2015; Chodroff & Wilson, under review)

Experimental manipulation:

Test [s]-[ʃ] categorization after manipulating the spectral center of gravity for several types of context sound: [z], [v], speech-shaped noise, speech + noise

Acoustic-phonetic covariation

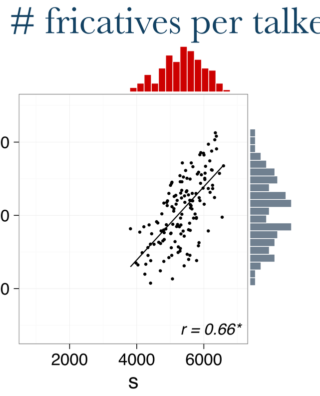
Center of gravity (COG): energy-weighted mean frequency

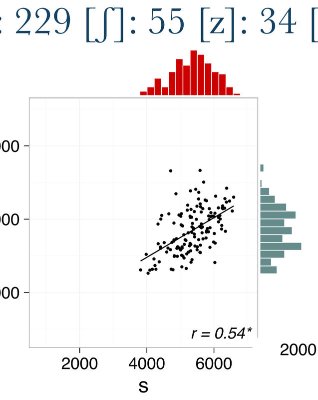
Measured in Hertz using a multitaper spectrum after high-pass filtering at 550 Hz

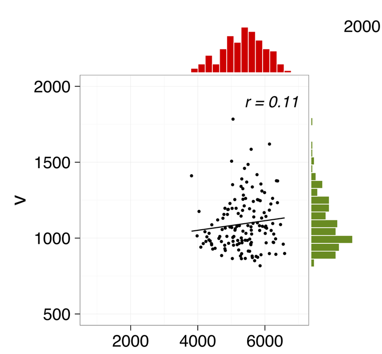
Mixer-6 corpus

141 talkers | read sentences| 16 kHz

median # fricatives per talker [s]: 229 [ʃ]: 55 [z]: 34 [v]: 98



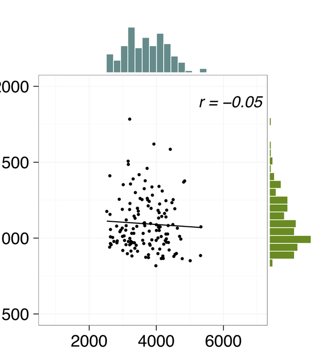




[s-z] $r = 0.66$* | [s-v] $r = 0.20$

[s-ʃ] $r = 0.54$* | [ʃ-v] $r = 0.17$

[ʃ-z] $r = 0.26$ | [z-v] $r = 0.51$*



Laboratory speech corpus

13 female talkers | fricative-initial CVC syllables | 44.1 kHz

median # fricatives per talker [s z v]: 24 [ʃ]: 21

[s-z] $r = 0.88$* | [s-ʃ] $r = 0.56^{\dagger}$ | [ʃ-z] $r = 0.52^{\shortparallel}$

[s-v] $r = 0.20$ | [ʃ-v] $r = 0.17$ | [z-v] $r = 0.40$

*$p < 0.001$, $^{\dagger}p < 0.05$, $^{\shortparallel}p < 0.1$

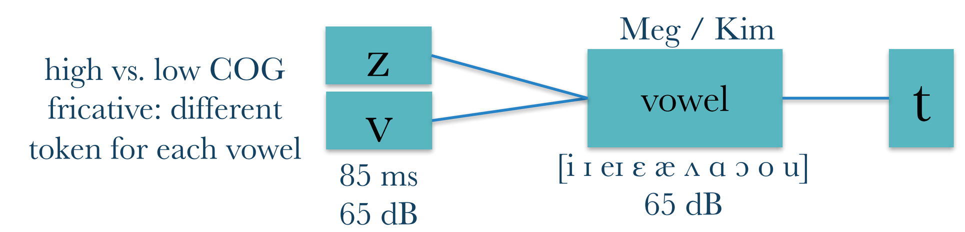
Methods

Speech contexts: fricative-initial CVC syllables created by concatenating natural recordings from 4 female speakers selected from laboratory speech corpus

Two female speakers with relatively neutral fricative COGs: "Meg" & "Kim"

One female speaker with high COG [z] (Exp. 1) or high COG [v] (Exp. 2)

One female speaker with low COG [z] (Exp. 1) or low COG [v] (Exp. 2)



Noise contexts: white noise matched in LTAS, duration, and amplitude to CV portion of [z]-initial syllable (Exp. 3)

Procedure:

Exposure
HIGH OR LOW COG*

trial: 1 exposure stimulus (2 reps) + 1 test

block: 20 trials

Test
[s] or [ʃ]

Exp. 1 (6 blocks): $[zæt]_i$ $[zæt]_i$

1000 ms delay

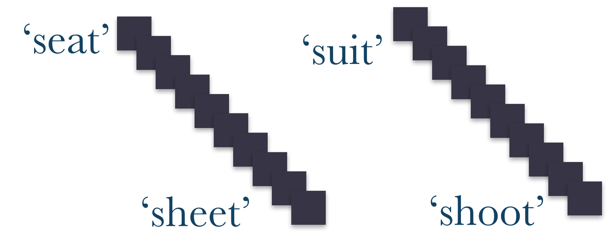
s/ʃ

Exp. 2 (6 blocks): $[væt]_i$ $[væt]_i$

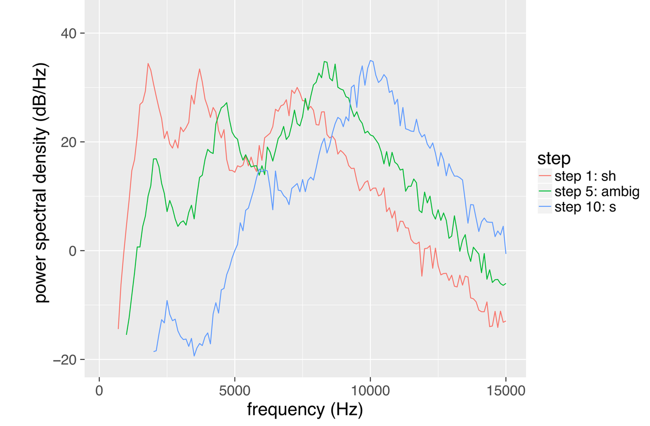
Exp. 3 (4 blocks): $noise_i$ $noise_i$

Exp. 4 (4 blocks): $[zæt]_i$ $noise_j$ $[zæt]_i$ $noise_j$

High-Low/Low-High alternation



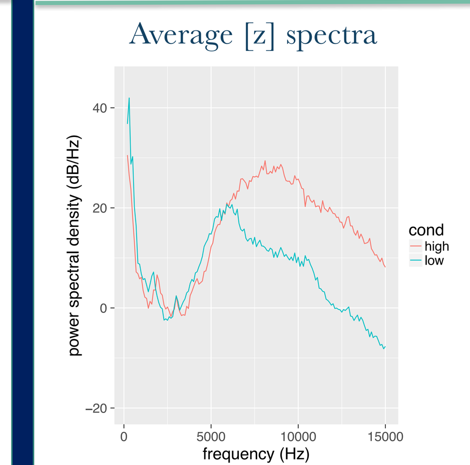
Synthesized [s] - [ʃ] continuum: 10-step Bark interpolation between standard spectral peaks and slopes for [s] and [ʃ]



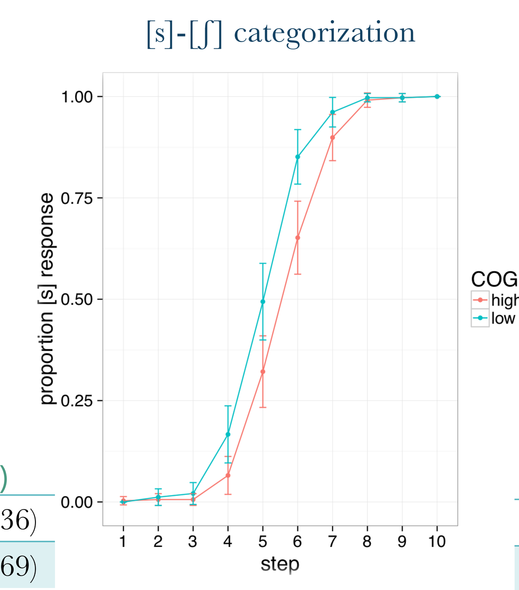
Each participant received opposite COG manipulations for the two speakers, Meg and Kim, with condition-speaker combination and condition order counterbalanced

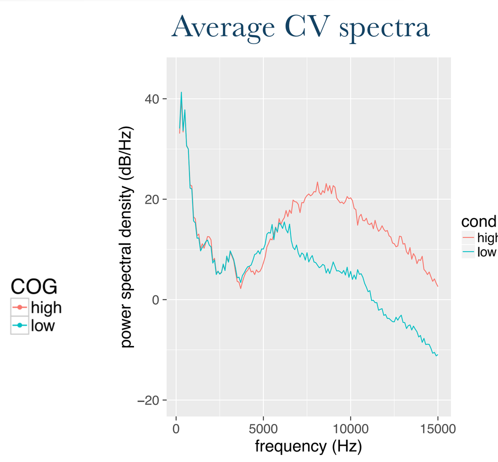
Exp. 1–3: 28 participants in each | Exp. 4: 32 participants

Exp. 1: Exposure to [z]



Mean COG values in Hz (sd)	
High COG	7026 (1136)
Low COG	1195 (1069)
High COG > 550 Hz	8576 (587)
Low COG > 550 Hz	6217 (785)





Mean COG values in Hz (sd)	
High COG	1711 (393)
Low COG	594 (224)
High COG > 550 Hz	5429 (2430)
Low COG > 550 Hz	2122 (1241)

Mixed-effects logistic regression

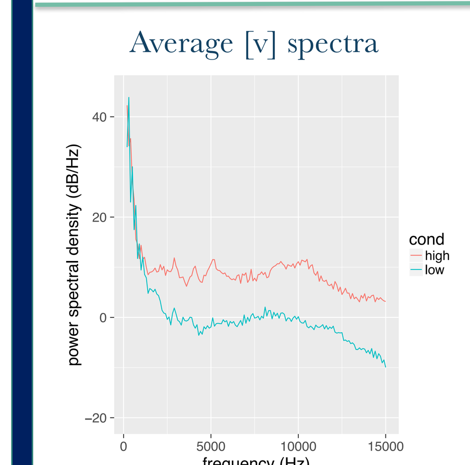
choose.s ~ 1 + continuum.step + vowel + spkr + cog.cond + (1 + cog.cond |subj)

Intercept: 0.55 | Step: 7.12* | Vowel: 2.15* | Speaker: -0.17 | **Condition: -1.39***

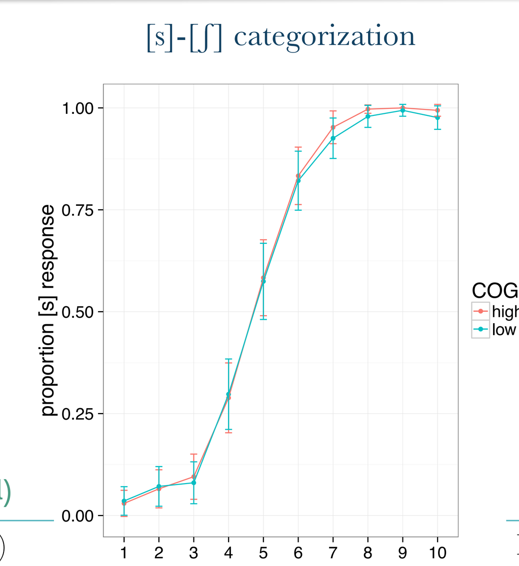
Listeners less likely to choose [s] after exposure to high COG [z] than low COG [z]

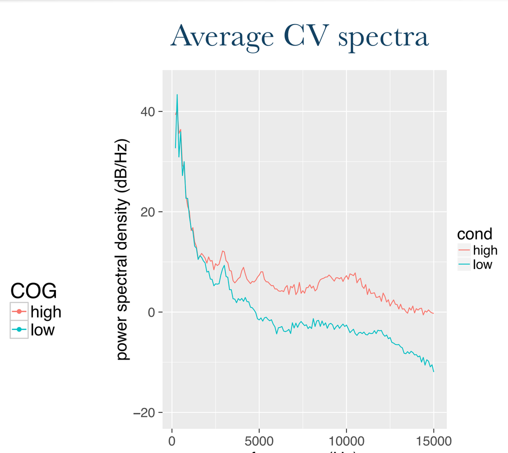
*$p < 0.001$, $^{\dagger}p < 0.01$, $^{+}p < 0.05$

Exp. 2: Exposure to [v]



Mean COG values in Hz (sd)	
High COG	583 (189)
Low COG	263 (23)
High COG > 550 Hz	6075 (1408)
Low COG > 550 Hz	2906 (1229)





Mean COG values in Hz (sd)	
High COG	485 (90)
Low COG	388 (84)
High COG > 550 Hz	2684 (2004)
Low COG > 550 Hz	1708 (1288)

Intercept: 1.22* | Step: 4.38* | Vowel: 1.07* | Speaker: 0.02 | **Condition: 0.13**

Listeners *not* less likely to choose [s] after exposure to high COG [v] than low COG [v]

Combined analysis of Exp. 1 and Exp. 2

choose.s ~ 1 + step + vowel + spkr + cond*experiment + (1| subj)

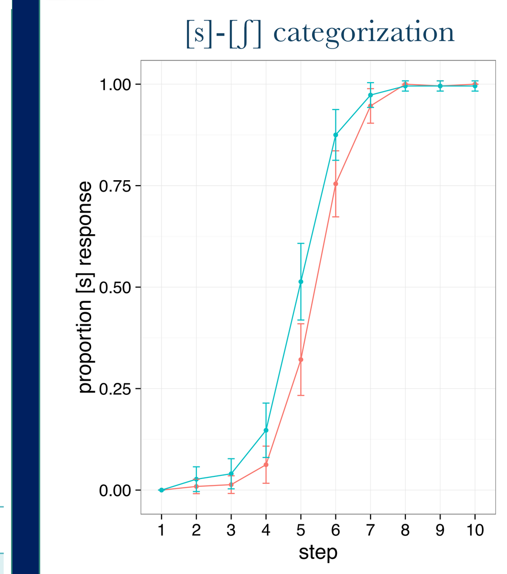
Intercept: 0.91* | Step: 5.11* | Vowel: 1.40* | Speaker: -0.05 | Condition: -0.42* |

Exposure: -1.03† | Condition x Exposure: -1.15*

Significantly greater number of [s] responses overall after exposure to [v] (vs. [z])

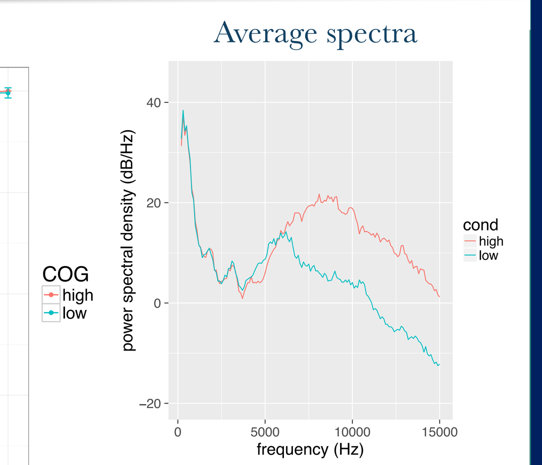
Significant interaction between condition (high-low) and fricative context ([z]-[v])

Exp. 3: CV-matched noise



Intercept: 0.77*| Step: 6.66* | Vowel: 2.10* | Speaker: 0.61^{+}|

Condition: -1.07*



Mean COG values in Hz (sd)	
High COG	2274 (685)
Low COG	698 (320)
High COG > 550 Hz	5920 (2271)
Low COG > 550 Hz	2552 (1292)

Listeners less likely to choose [s] after exposure to high COG noise than low COG noise

Combined analysis of Exp. 1 and Exp. 3

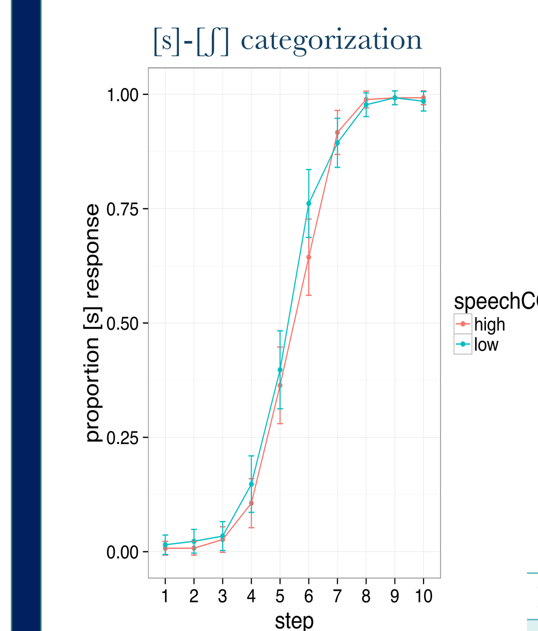
Intercept: 0.62* | Step: 6.49* | Vowel: 2.00* | Speaker: 0.15 | Condition: -1.15* |

Exposure: -0.25 | Condition x Exposure: -0.22

No significant difference in effect of condition (high-low) on categorization for speech (Exp. 1) and noise (Exp. 3)

Exp. 4: Speech + Noise

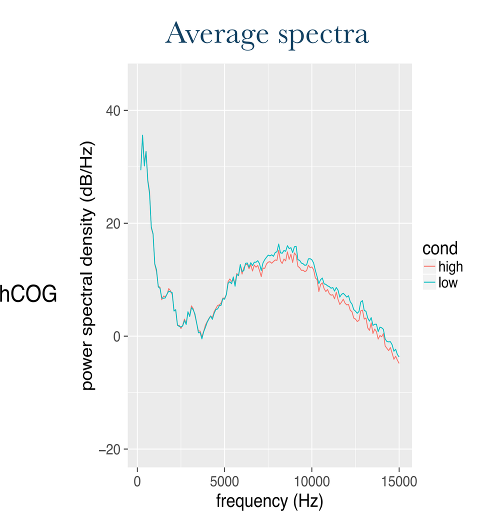
When spectra of speech and noise conflict, does speech have a stronger influence on [s]-[ʃ] categorization than noise?



Intercept: 0.24| Step: 5.08* |

Vowel: 1.64* | Speaker: 0.08 |

Condition: -0.32



Mean COG values in Hz (sd)	
High COG	1076 (238)
Low COG	1419 (353)
High COG > 550 Hz	4455 (2360)
Low COG > 550 Hz	5024 (2411)

Opposing spectra from speech and noise 'cancel out' (consistent with equal averaging of two contexts)

Discussion & Future Directions

Cue-based normalization account:
Exposure to any fricative should affect the overall COG mean, resulting in an [s]-[ʃ] boundary shift
✓Exp. 1 ✗ Exp. 2 — Exp. 3 — Exp. 4

Covariation account:
Only exposure to a fricative that is correlated with [s] or [ʃ] in the population should result in an [s]-[ʃ] boundary shift
✓Exp. 1 ✓Exp. 2 — Exp. 3 — Exp. 4

Spectral contrast account:
Exposure to any sound (speech or non-speech) with energy in a frequency range relevant for [s]-[ʃ] categorization will affect perception of [s]-[ʃ] continuum
✓Exp. 1 ✓Exp. 2 ✓ Exp. 3 ✓ Exp. 4

- Exp. 1: higher (lower) frequency concentration of energy in a preceding syllable contrastively enhances lower (higher) frequencies in a continuum member
- Exp. 2: spectra of 'high' [v] contexts does not have sufficiently high frequency energy to affect [s]-[ʃ] categorization (relative to 'low' [v] contexts)
- Exp. 1 vs. Exp. 2: high frequency energy in [z] contexts overall enhances low frequency components of continuum stimuli (or: low frequency energy in [v] contexts overall enhances high frequency components of continuum stimuli)
- Exp. 3 & 4: effect on categorization from noise equal to that of corresponding speech

Spectral contrast accounts for the findings of adaptation after immediate exposure to both speech and non-speech stimuli, provided the *relevant range of frequencies* is correctly specified.

Covariation account makes accurate predictions regarding the speech context experiments, but does not make a prediction about experiments with non-speech contexts.

Cue-based normalization account incorrectly assigns equal relevance in adaptation to all segments with a shared cue (and also does not predict shifts with non-speech contexts). Experiments demonstrate that [z] has a greater effect on [s]-[ʃ] categorization than [v].

Further questions:

Do listeners interpret turbulent noise as being sufficiently similar to a fricative? Would high-frequency tone sequences have as strong an effect on fricative continuum perception?

How does long-term *learning* of talker characteristics affect perception? How does knowledge of the talker interact with local spectral context effects?

What are the *relevant frequency ranges* for each speech sound, and how does dampening energy in a particular frequency range affect perception? (see ambiguity in interpretation for Exp. 1 vs. Exp. 2 interpretation)

Can the present results be accounted for with a formal model of spectral contrast?

Acknowledgments: Thanks to Lisa Davidson for recording the laboratory speech corpus and Matt Winn for providing a Praat script to create the fricative continuum. We also thank Alessandra Golden and Gigi Edwards for their help in data collection. This work was supported by a Distinguished Science of Learning Pre-doctoral Fellowship from the JHU Science of Learning Institute and a fellowship from the Dolores Zohrab Liebmann Fund.

Selected References: Chodroff, E., Godfrey, J., Khudanpur, S., Wilson, C. (2015). Structured variability in acoustic realization: A corpus study of voice onset time in American English stops. *Proc. 18th ICPhS*, 1–5. / Chodroff, E., Wilson, C. (under review). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. / Holt, L. (2006). The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorization. *J. Acoustic. Soc. Am.*, *120*(5), 2801–2817. / Ladefoged, P., Broadbent, D. (1957). Information conveyed by vowels. *J. Acoustic. Soc. Am.*, *29*(1), 98–104. / Laing, E., Liu, R., Lotto, A., Holt, L. (2012). Tuned with a Tune: Talker Normalization via General Auditory Processes. *Front. Psych.*, *3*, 1–9. / McMurray, B., Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psych. Rev.*, *118*(2), 219–246.