Bayesian interaction of phonetics and phonotactics in cluster production

Colin Wilson (colin@cogsci.jhu.edu)
Johns Hopkins University
Lisa Davidson (lisa.davidson@nyu.edu)
Sean Martin (sean.martin@nyu.edu)
New York University

LSA Annual Meeting
January 5, 2012

Cross-language speech production

Speakers make a variety of modifications when they attempt to produce non-native consonant clusters and syllable codas

(e.g., Broselow 1983, Broselow et al. 1998; Abrahamsson 1999; Davidson 2003, 2006, 2010)

- Epenthesis (of a transitional/excrescent vocoid)
  Ex. [kpaga] → [kⁱpaga]
- Prothesis
  Ex. [gnazi] → [ᵊgnazi]
- C1 feature change
  Ex. [ptasi] → [ktasi]
- C1 deletion
  Ex. [dbazo] → [bazo]
- . . . and others, including combination modifications
  Ex. [zgade] → [ᵊzⁱgade]

Cross-language speech production

These modification patterns reveal aspects of language-specific, and perhaps universal, phonetic and phonological knowledge

(e.g., Broselow 1983, Broselow et al. 1998; Hancin-Bhatt & Bhatt 1997; Eckman 1991; Davidson 2003, 2006)

. . . and results from production studies largely converge with other data on processing of non-native sound structures

- Discrimination, identification, and transcription

  (e.g., Werker and Tees 1984; Kuhl et al. 1992; Best 1995; Pitt 1998; Dupoux et al. 1999; Moreton 2002; Bent 2005; Berent et al. 2007; Yarmolinskaya 2010; Shaw & Davidson 2011)

- Loanword and L2 adaptations (e.g., Eckman 1977; Broselow and Finer 1991; Hancin-Bhatt 2000; Kang 2004, Kenstowicz & Uffmann 2006, Zuraw 2007; Peperkamp et al. 2008)

- Acceptability judgments

  (e.g., Greenberg and Jenkins 1964; Scholes 1966; Albright 2009, Daland et al. 2011)

Cross-language speech production

Modifications, misperceptions, and low ratings of non-native clusters have typically been attributed to relative abstract **phonotactics**

- sonority sequencing

  (e.g., Berent et al. 2007, 2008, 2009; Hayes 2007)

- syllable parsing

  (e.g., Kabak and Idsardi 2007)

- segmental phonotactics

  (e.g., Dupoux et al. 1999; Moreton 2002; Hayes & Wilson 2008)

- restrictions on gestural overlap

  (eg., Davidson 2003; Davidson et al. 2004; Davidson 2006)

We demonstrate that phonotactic knowledge is not sufficient to account for the detailed pattern of production modifications: fine-grained details of **phonetic realization** have a central role in predicting error types and rates

Outline

1. Production experiment in which the phonetic properties of non-native stimuli are parametrically manipulated
   - duration, amplitude of the transition between consonants
   - presence of modal voicing before obstruent constriction
2. Accounting for the experimental results with a Bayesian model that integrates knowledge of phonotactics with knowledge of phonetic realization
   - Prior: bias against phonotactically illegal structures
   - Likelihood: measure of acoustic-phonetic similarity between non-native stimuli and expected realizations of native structures
3. Summary and future directions

Cross-language cluster production experiment

English speaking participants ($N = 24$) heard and repeated critical items of the form [C1C2áCV] produced by a native Russian speaker, and corresponding fillers with initial schwa and medial schwa

Ex. [ptáke], [əptáke], [pətáke]; [zgámo], [əzgámo], [zəgámo]

	cluster type	C1 [-voice]	C1 [+voice]
SS	(stop-stop)	pt, tp, kp, kt	bd, db, gb, gd
SN	(stop-nasal)	pn, tm, km, kn	bn, dm, gm, gn
FS	(fricative-stop)		vd, vg, zb, zg
FN	(fricative-nasal)		vm, vn, zm, zn

- avoided [sC2] (many legal) and [fC2] clusters, [ŋ] (illegal in onset), and perfectly homorganic clusters ([pm, bm, tn, dn])
- each cluster appeared in 4 distinct [__áCV] items

Acoustic manipulations

Wilson & Davidson (2010) observed speaker- and cluster- internal phonetic variation that correlated with modification rate and type

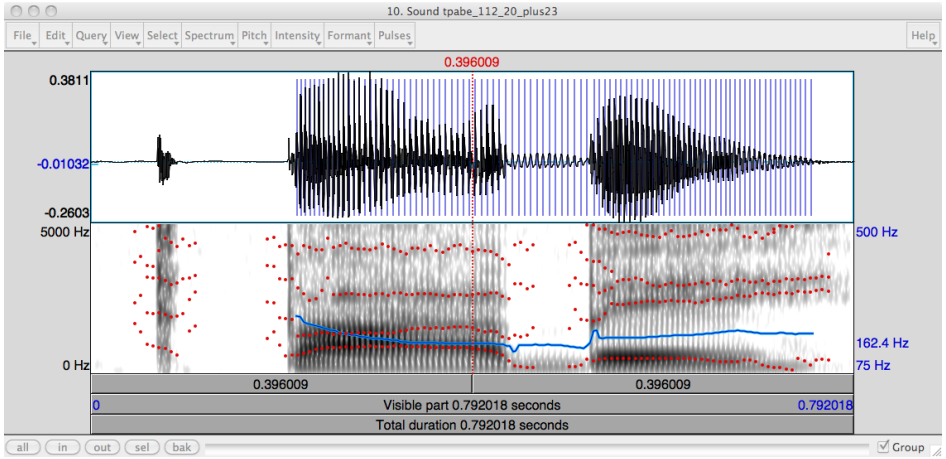
- DUR: duration of the acoustic transition (burst + aspiration) between an oral stop and closure of the following stop
  DUR longer $\rightarrow$ *more epenthesis*
- AMP: amplitude of the acoustic transition (burst+release) from an oral stop relativized to the following consonant onset
  AMP lower $\rightarrow$ *more deletion* (also more C1 change)
- POV (pre-obstruent voicing): modally-voiced interval that precedes the formation of a voiced obstruent constriction
  POV present $\rightarrow$ *more prothesis*

Different versions of the present stimulus items were created by systematically manipulating these acoustic-phonetic properties

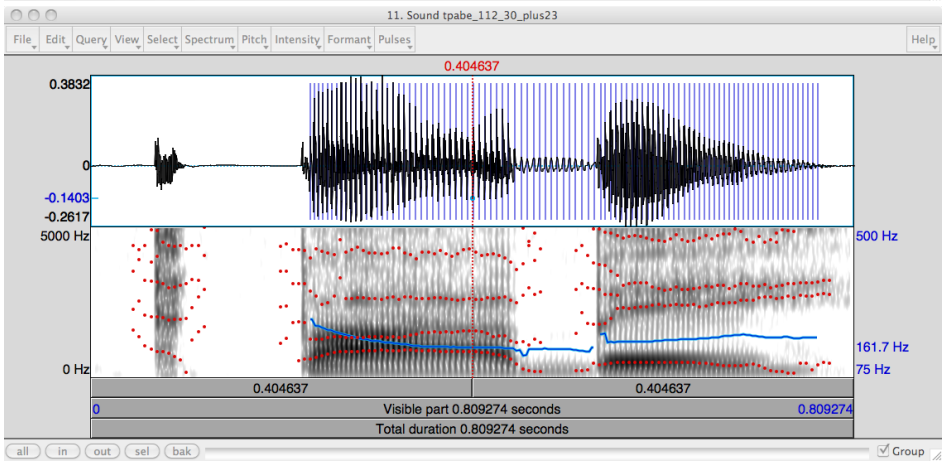
Acoustic manipulations

DUR (transition duration): S[-v]X clusters [pt tp kp kt; pn tm km kn] and S[+v]X clusters [bd db gb gd; bn dm gn]

[tpabe] play
DUR = 20ms



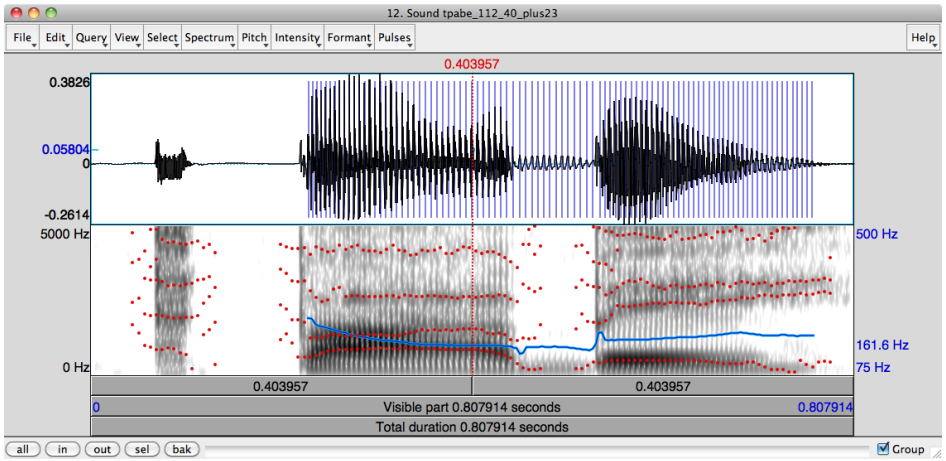
[tpabe] play
DUR = 30ms



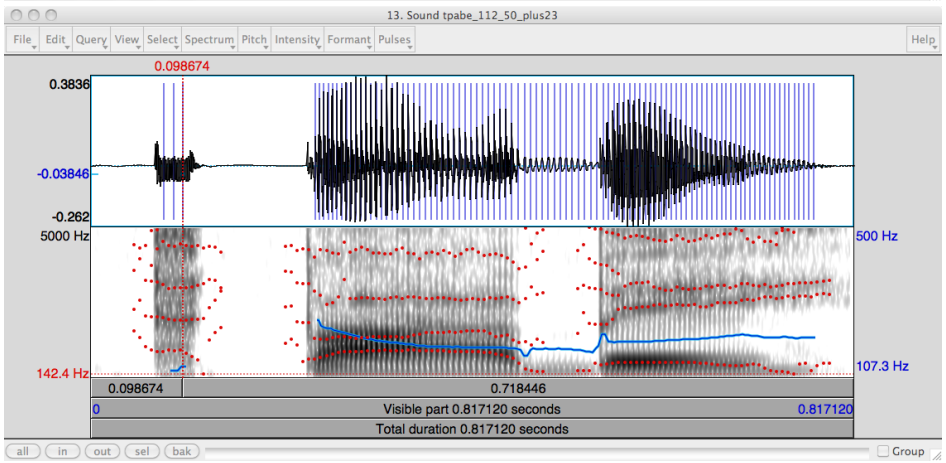
Acoustic manipulations

DUR (transition duration): S[-v]X clusters [pt tp kp kt; pn tm km kn] and S[+v]X clusters [bd db gb gd; bn dm gn]

[tpabe] play
DUR = 40ms



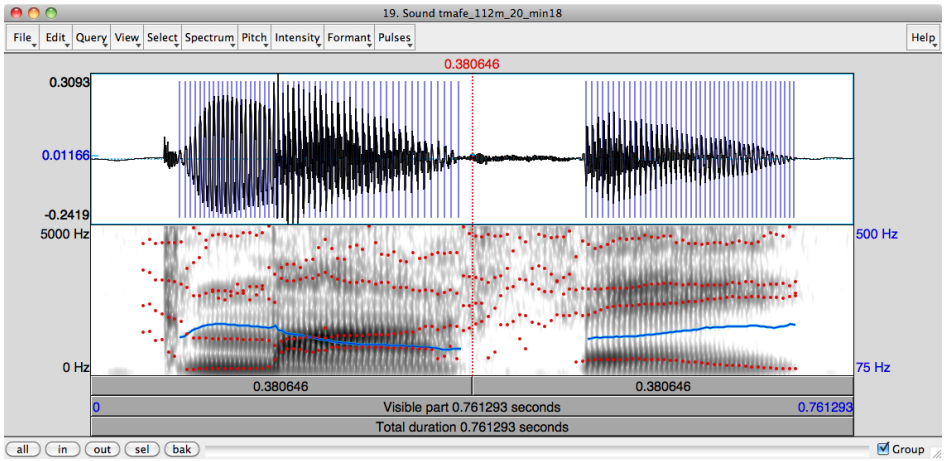
[tpabe] play
DUR = 50ms



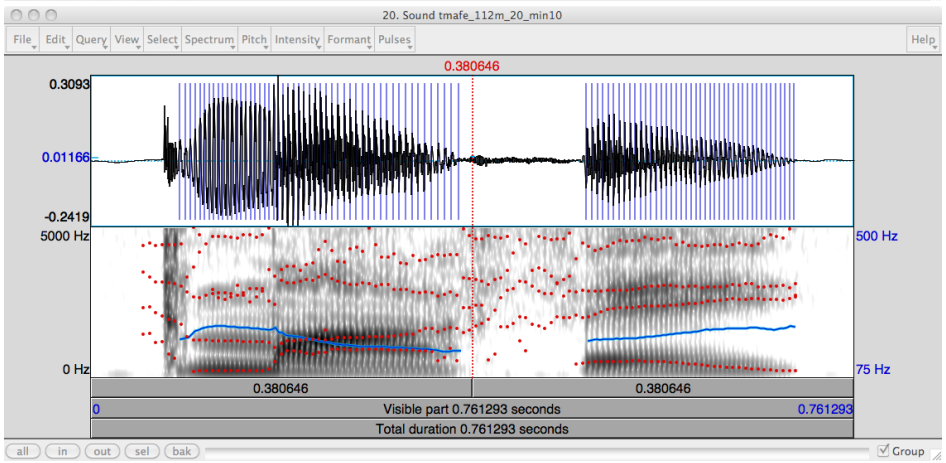
Acoustic manipulations

AMP (relative transition amplitude): S[-v]X and S[+v]X clusters

[tmafe] play
DUR = 20ms
AMP = -18dB
(baseline)



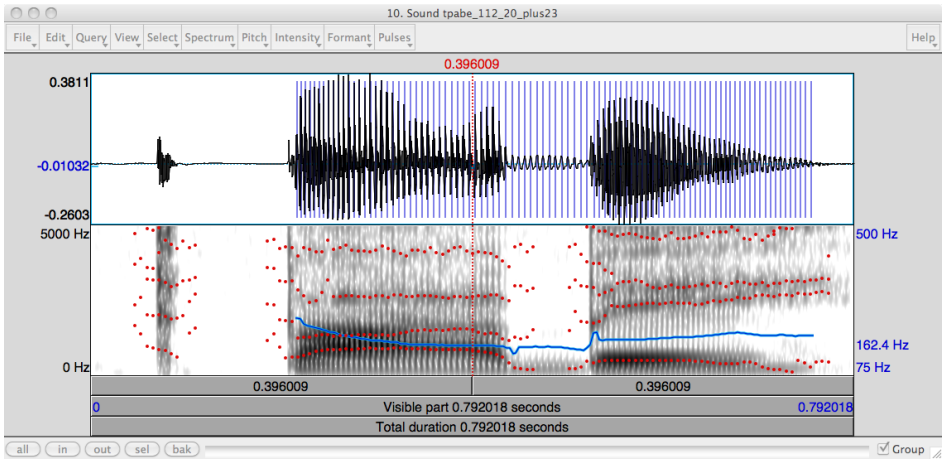
[tmafe] play
DUR = 20ms
AMP = -10dB
(raised)



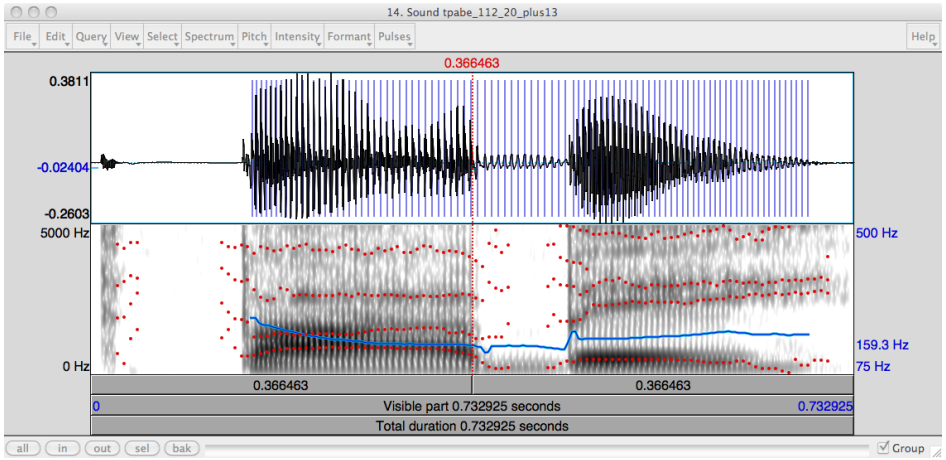
Acoustic manipulations

AMP (relative transition amplitude): S[-v]X and S[+v]X clusters

[tpabe] play
DUR = 20ms
AMP = +23dB
(baseline)



[tpabe] play
DUR = 20ms
AMP = +13dB
(lowered)



Acoustic manipulations

AMP manipulations for S[-v]X and S[+v]X clusters

S[-v]N	base: -18dB	amp: -10dB	(raised)
S[-v]S	base: +23dB	amp: +13dB	(lowered)

S[+v]N	base: -7dB	amp: 0dB	(raised)
S[+v]S	base: 0dB	amp: -7dB	(lowered)

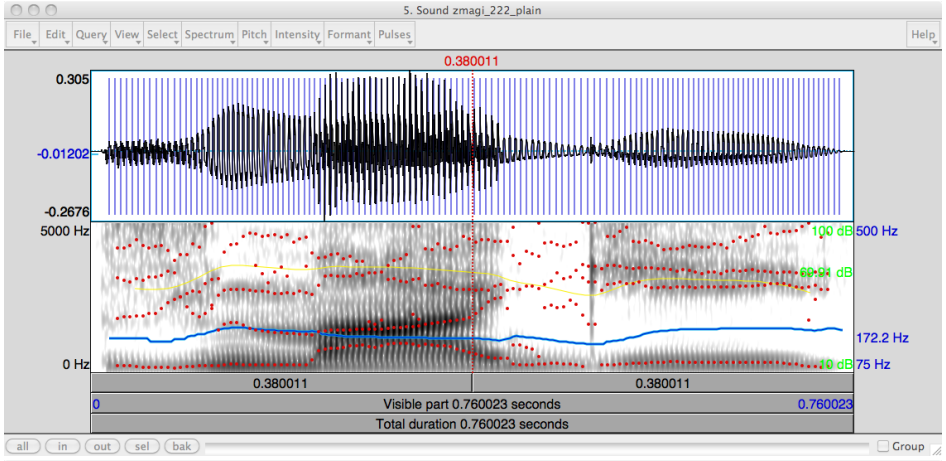
Praat script by Sean Martin scaled sound pressure level of C1 transition so that average intensity was a certain dB value above or below the average intensity of the following C2 closure

(on importance of relative burst intensity see also Stoel-Gammon et al. 1994; Sundara 2005; Vicenik 2010)

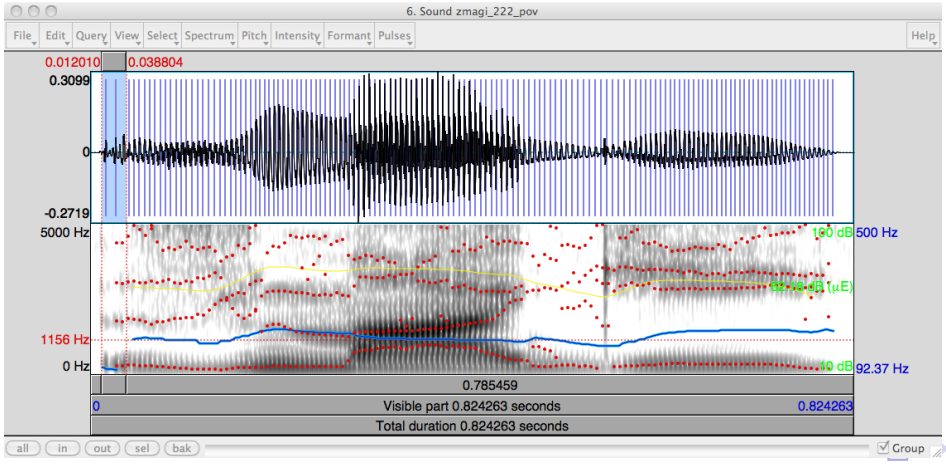
Acoustic manipulations

POV (pre-obstruent voicing): FX clusters [vd vg zb zg; vm vn zm zn] and S[+v]X clusters [bd db gb gd; bn dm gn]

[zmagi] play
POV absent



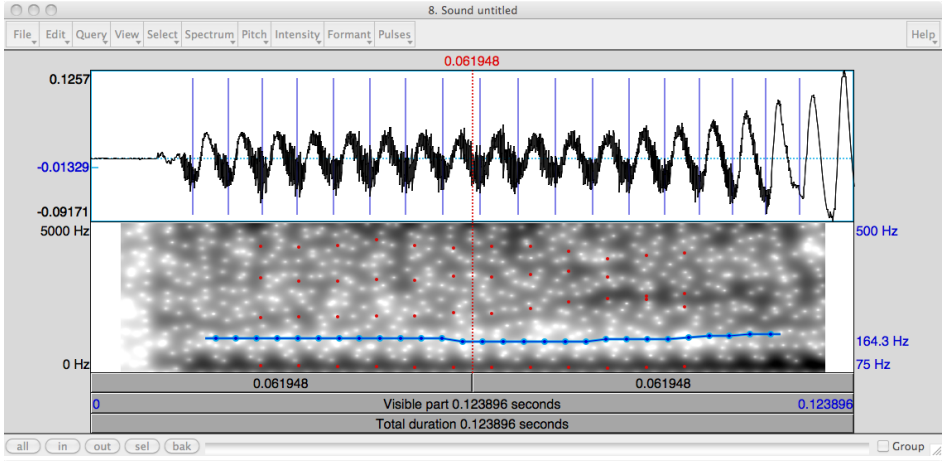
[zmagi] play
POV present



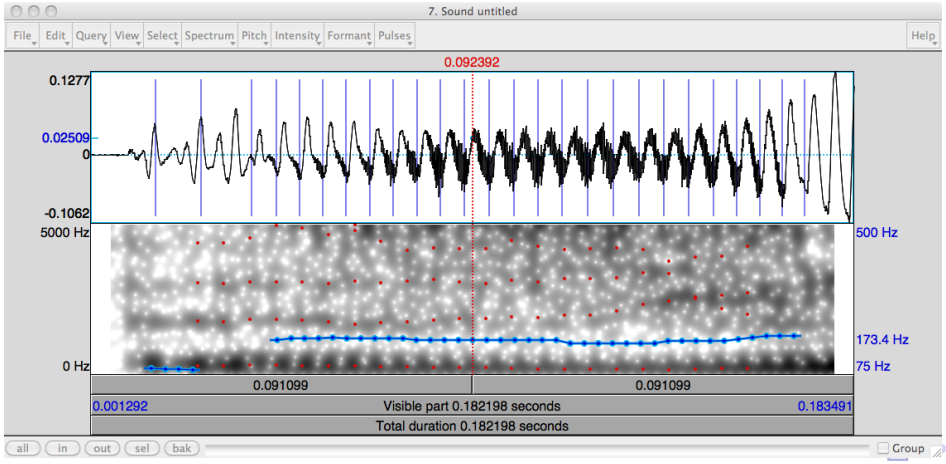
Acoustic manipulations

POV (pre-obstruent voicing): FX clusters [vd vg zb zg; vm vn zm zn] and S[+v]X clusters [bd db gb gd; bn dm gn]

[zmagi] play
POV absent



[zmagi] play
POV present



Summary of acoustic manipulations

S[-v]X clusters

DUR (20ms, 30ms, 40ms, 50ms) × AMP (base vs. lowered/raised)

(recall AMP denotes raised before N and lowered before S)

S[+v]X clusters

DUR (20ms, 30ms, 40ms, 50ms) × POV (absent vs. present)

(recall AMP denotes raised before N and lowered before S)

DUR (20ms, 30ms, 40ms, 50ms) × AMP (base vs. lowered)

FX clusters

POV (absent vs. present)

+ fillers = 800 total sound files, distributed across 12 experimental lists so that each critical stimulus occurs equally often across the lists

Cross-language cluster production experiment

Order of events in a trial

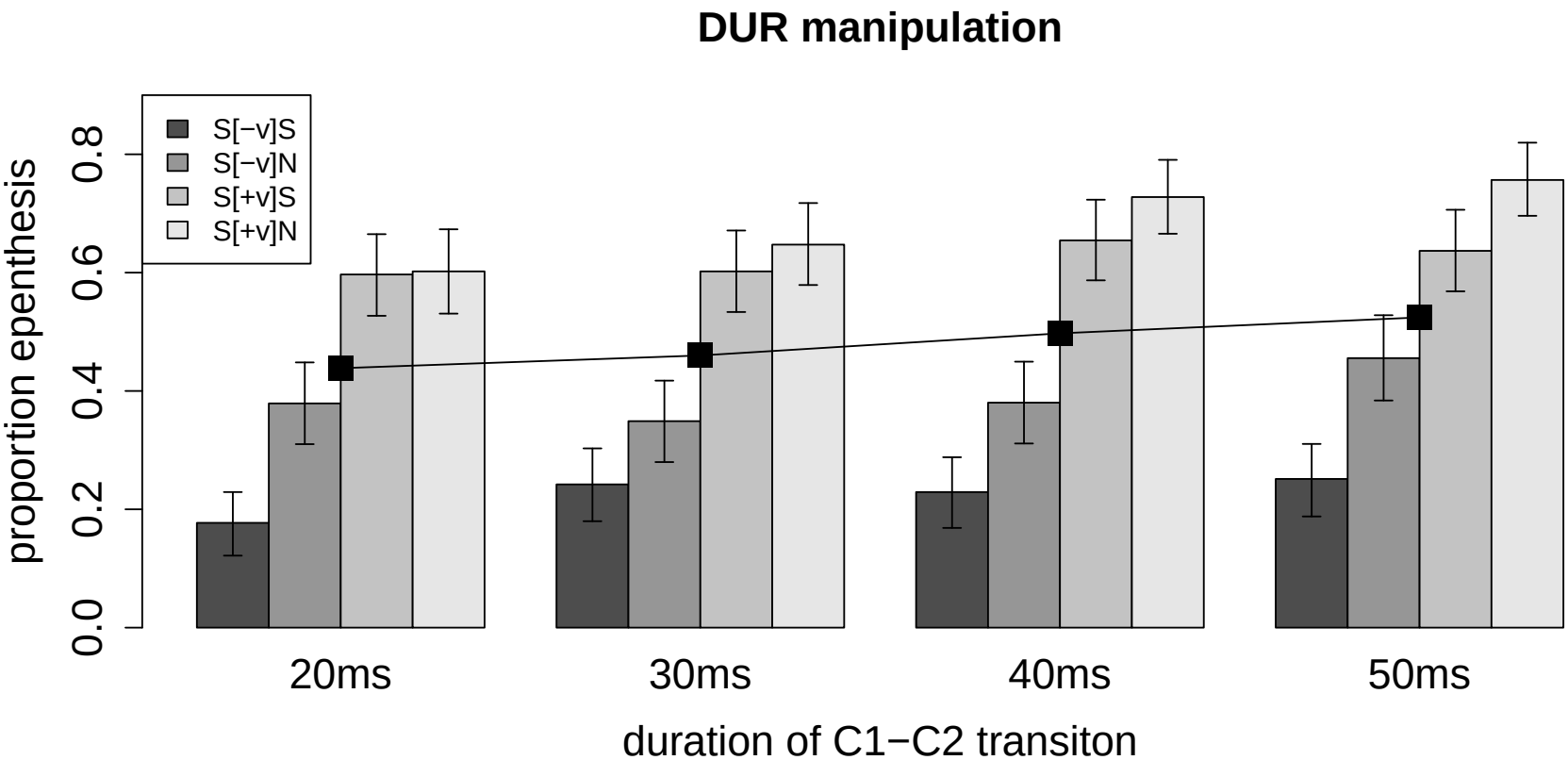
- single recording of a stimulus item was played twice
- participant repeated the stimulus once

Each participant heard and produced 288 total items

- 64 SN, 64 SS (each approx. half S[-v]X, half S[+v]X)
- 32 FN, 32 FS (all F[+v]X)
- 48 CəCX, 48 əCCX fillers

Two versions of each item were heard and produced by a given participant; versions were counterbalanced across participants

Results and statistical analysis



- **Less epenthesis at shortest duration** ($\beta = -.23, p < .01$)
- More epenthesis for voiced clusters ($\beta = .92, p < .001$)
- More epenthesis before nasals ($\beta = .33, p < .001$)

Bootstrap confidence intervals (`boot`). All statistics based on mixed-effects logistic regression (`lme4`); no sig. interactions.

Results and statistical analysis

Pairwise comparisons of DUR levels

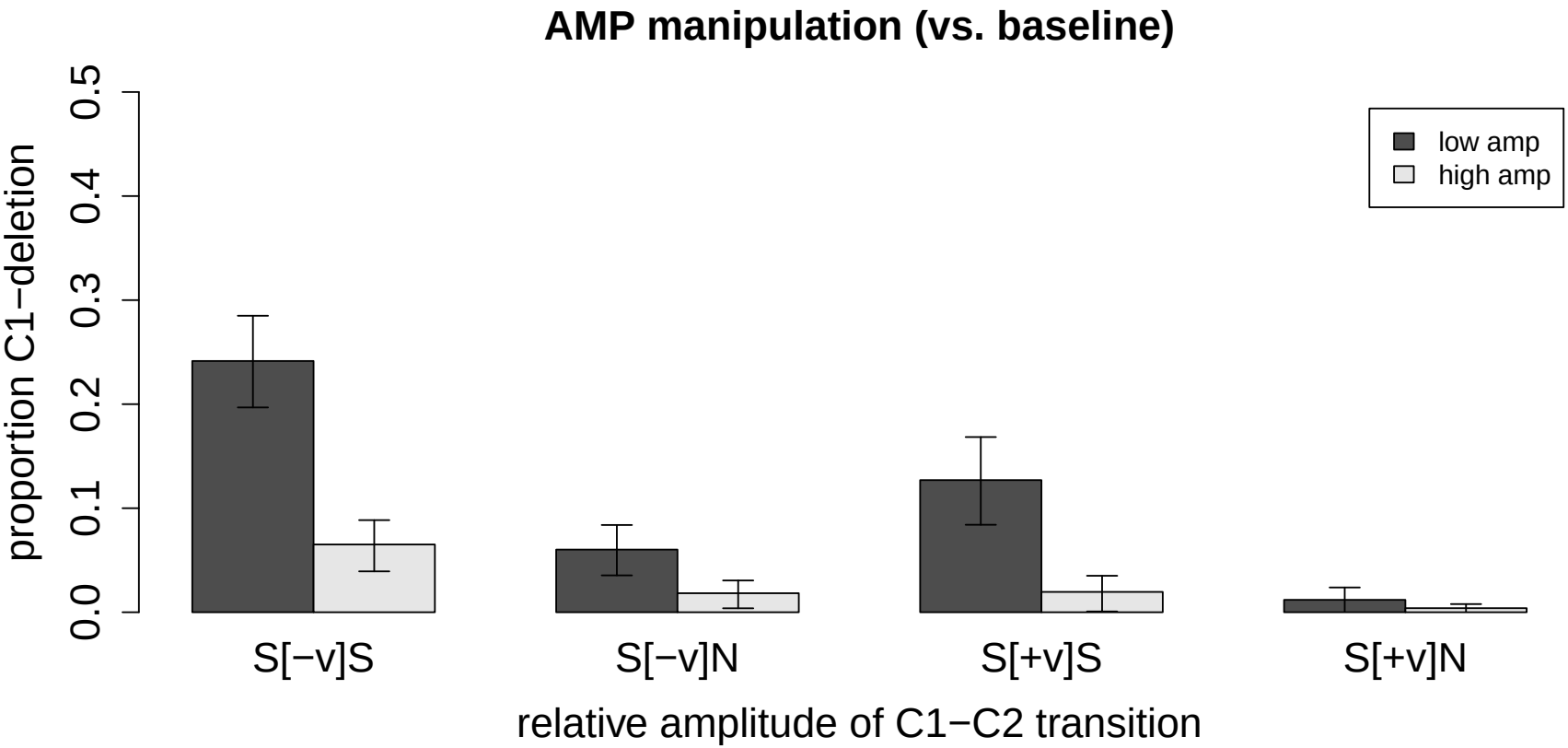
	20ms	30ms	40ms	50ms
20ms	—	n.s.	$p < .05$	$p < .001$
30ms		—	n.s.	$p < .05$
40ms			—	n.s.

Significance levels reflect single-step *p*-value adjustment (`multcomp`)

Epenthesis rates at lowest vs. highest DUR values

	20ms	50ms	ratio (50ms/20ms)
S[-v]S	.177	.251	1.42
S[-v]N	.379	.455	1.20
S[+v]S	.597	.637	1.07
S[+v]N	.602	.757	1.26

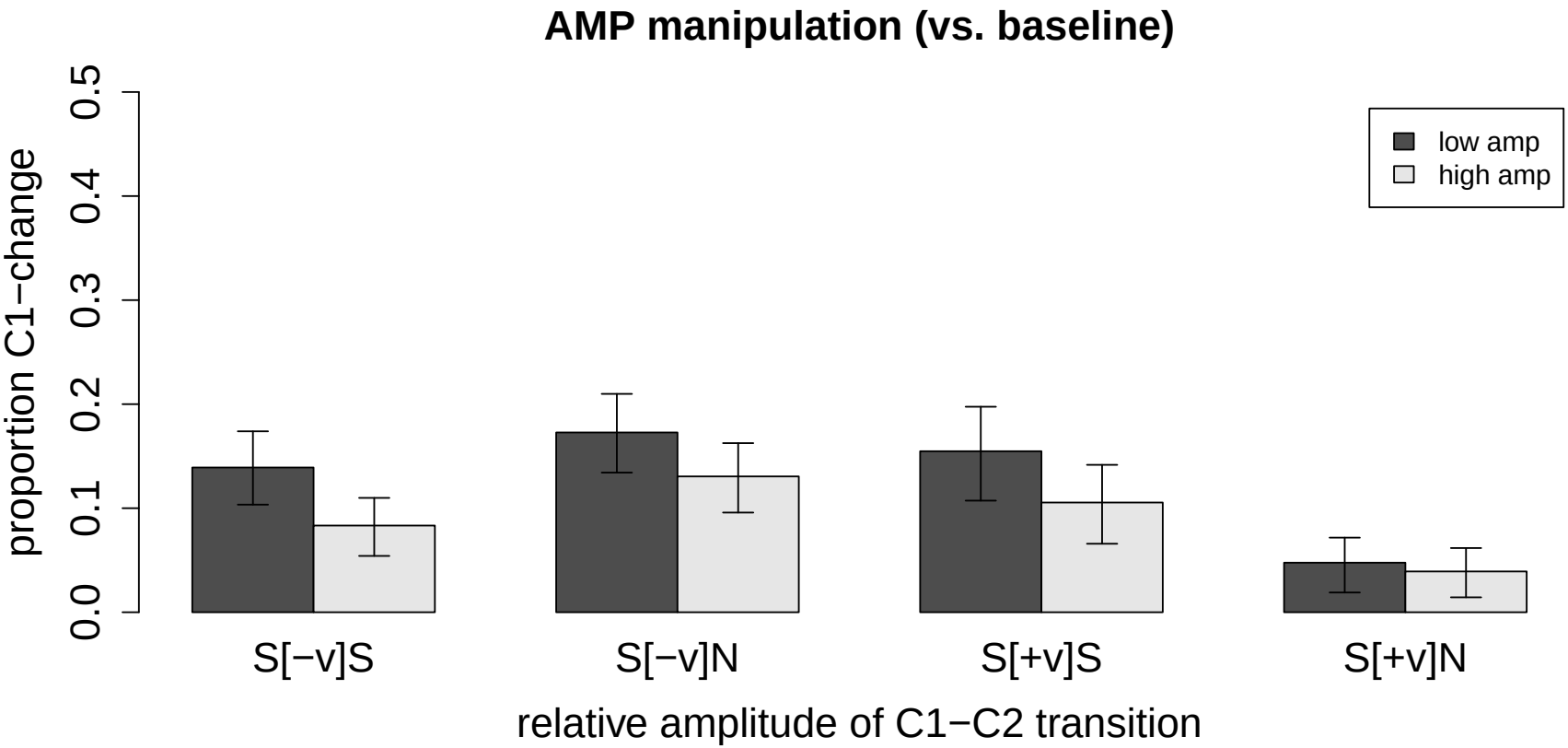
Results and statistical analysis



- **More deletion with lower amplitude** ($\beta = .80, p < .001$)
- Less deletion for voiced clusters ($\beta = -.68, p < .05$)
- Less deletion before nasals ($\beta = -.94, p < .001$)

Bootstrap confidence intervals (`boot`). All statistics based on mixed-effects logistic regression (`lme4`); no sig. interactions.
POV items were excluded from the analysis because this was manipulated orthogonally and was not fully crossed with voicing.

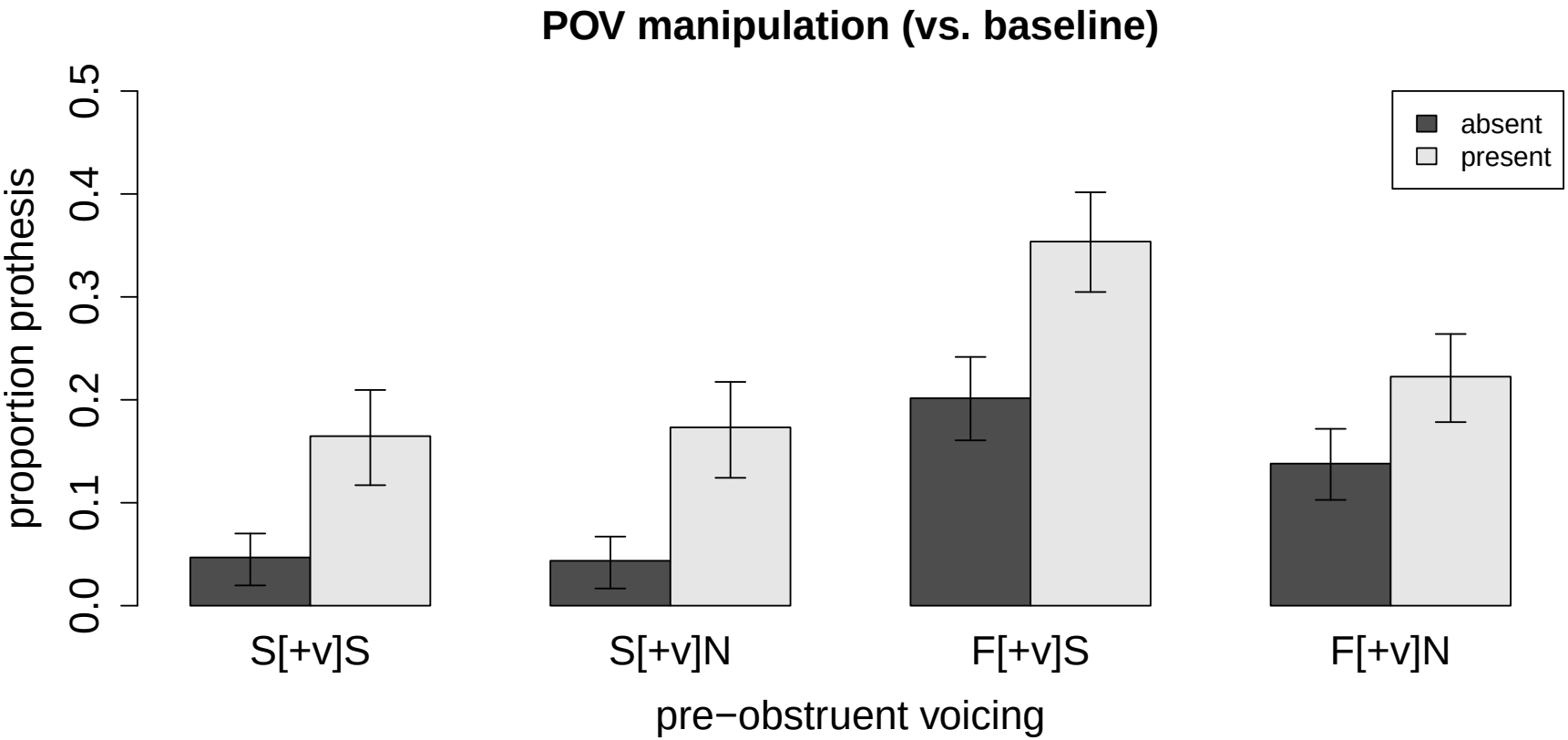
Results and statistical analysis



- **More C1-change with lower amplitude** ($\beta = .22, p < .05$)
- Less C1-change before nasals ($\beta = -.36, p < .05$), esp. S[+v]N

Bootstrap confidence intervals (`boot`). All statistics based on mixed-effects logistic regression (`lme4`); significant cluster.type × cluster.voice interaction ($\beta = .30, p < .05$) reflects consistently low rate of C1-change in S[+v]N clusters.

Results and statistical analysis



- **More prothesis when POV is present** ($\beta = .66, p < .01$)
- Less prothesis for stop-initial clusters ($\beta = -.66, p < .001$)

Bootstrap confidence intervals (`boot`). All statistics based on mixed-effects logistic regression (`lme4`); significant C1 $\times$ C2 interaction ($\beta = -.22, p < .001$) reflects the consistently higher rate of prothesis for F[+v]S clusters.

Bayesian model of cross-language production

Proposal: non-native modifications reflect knowledge of (at least)

- language-specific phonotactics
- **language-specific patterns of phonetic realization**

The stimulus is rich in fine-grained phonetic detail ("raw material") that may be accidental/variable w.r.t. the source language

. . . non-native perceiver attends to these details, interpreting the stimulus as an instance of the 'best-fitting' phon. structure

. . . where the criteria for 'fit' include (gradient) phonotactic well-formedness and **compatibility with the phonetic realization patterns of the perceiver's language**

. . . and the inferred phon. structure then becomes the basis for the speaker's own reproduction of the stimulus

Bayesian model of cross-language production

Perception
infer phon. rep. x given stimulus y
$P(x|y) \propto P(y|x) \times P(x)$

Production
generate reproduction y' given x, y
$P(y'|x, y)$

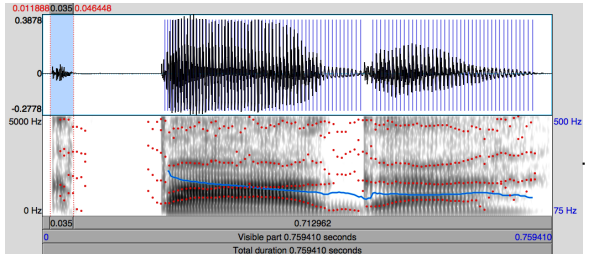
phonetic/phonological prior
$P(x)$

$\times$

phonetic likelihood
$P(y|x)$

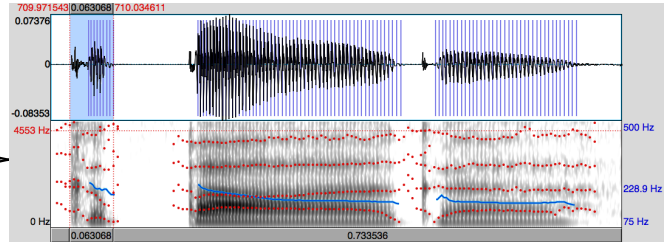
$=$

phonetic realization
$P(y'|x)$



Ex. y = Russian [kpaga]

imitation



Ex. y' = English [kʲpaga]

Bayesian model of cross-language production

Mechanics of the model

- Fine-grained phonetic representation of the stimulus (y) assumed to contain information about duration, formants, voicing, etc.
- Phonetic-phonological representation (x) could be complex, with multiple internal levels (e.g., $x = \langle \text{UR}_x, \text{SR}_x, \text{gestural score}_x \rangle$)
  - minimally a gestural score with prosodic structure
- $P(y|x)$ embodies stochastic knowledge of language- / sound- / context- specific phonetic realization in the native language, $P(x)$ contains knowledge of native phono. well-formedness
- Production y' is generated from the phonetic-phonological representation x — and may further be under pressure to 'imitate' the stimulus — but could differ from y as allowed by $P(y|x)$

Accouting for epenthesis modifications

Experimental results

- Epenthesis much more frequent for SX (.20−.70) than FX (.09)
- **More epenthesis at longer SX transition durations**
- More epenthesis for S[+v]X clusters
- More epenthesis for SN clusters

Phonotactics

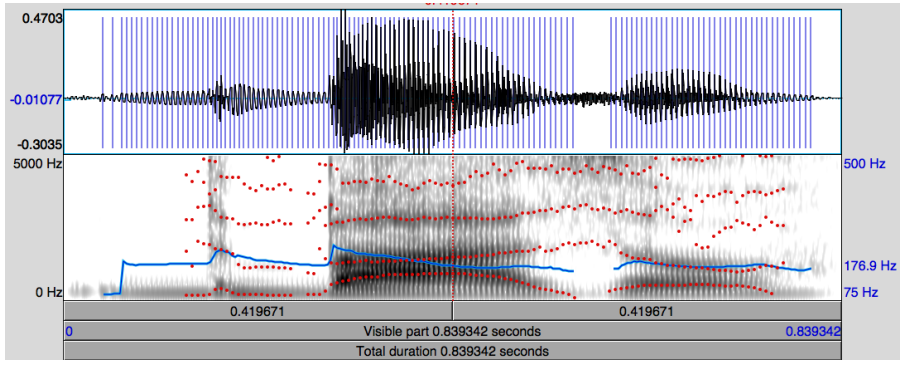
- English lacks *all* of the word-initial SS, SN, FS, and FN clusters tested in the experiment
- English has word-initial C1ɨC2 sequences, where vocoid ɨ
  - plausibly lacks a tongue body target (Flemming & Johnson 2007), gesturally characterized as open transition between C1 and C2 (see Davidson 2003)
  - is variably shortened and overlapped with surrounding consonant place and laryngeal gestures (Davidson 2006)

Accouting for epenthesis modifications

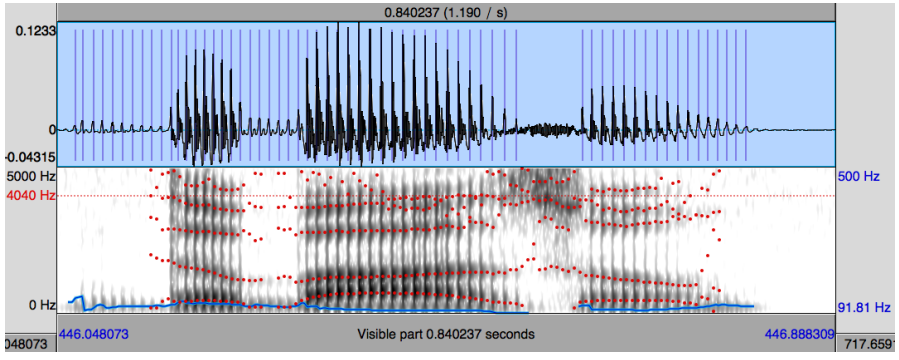
Phonetic realization

The Russian productions of SX clusters have an open transition that differs from English ɨ in having weaker formants / aperiodicity and shorter duration, but which is voiced between voiced consonants

Russian [dbazo]



English [dɨbazo]
(CəCX filler)

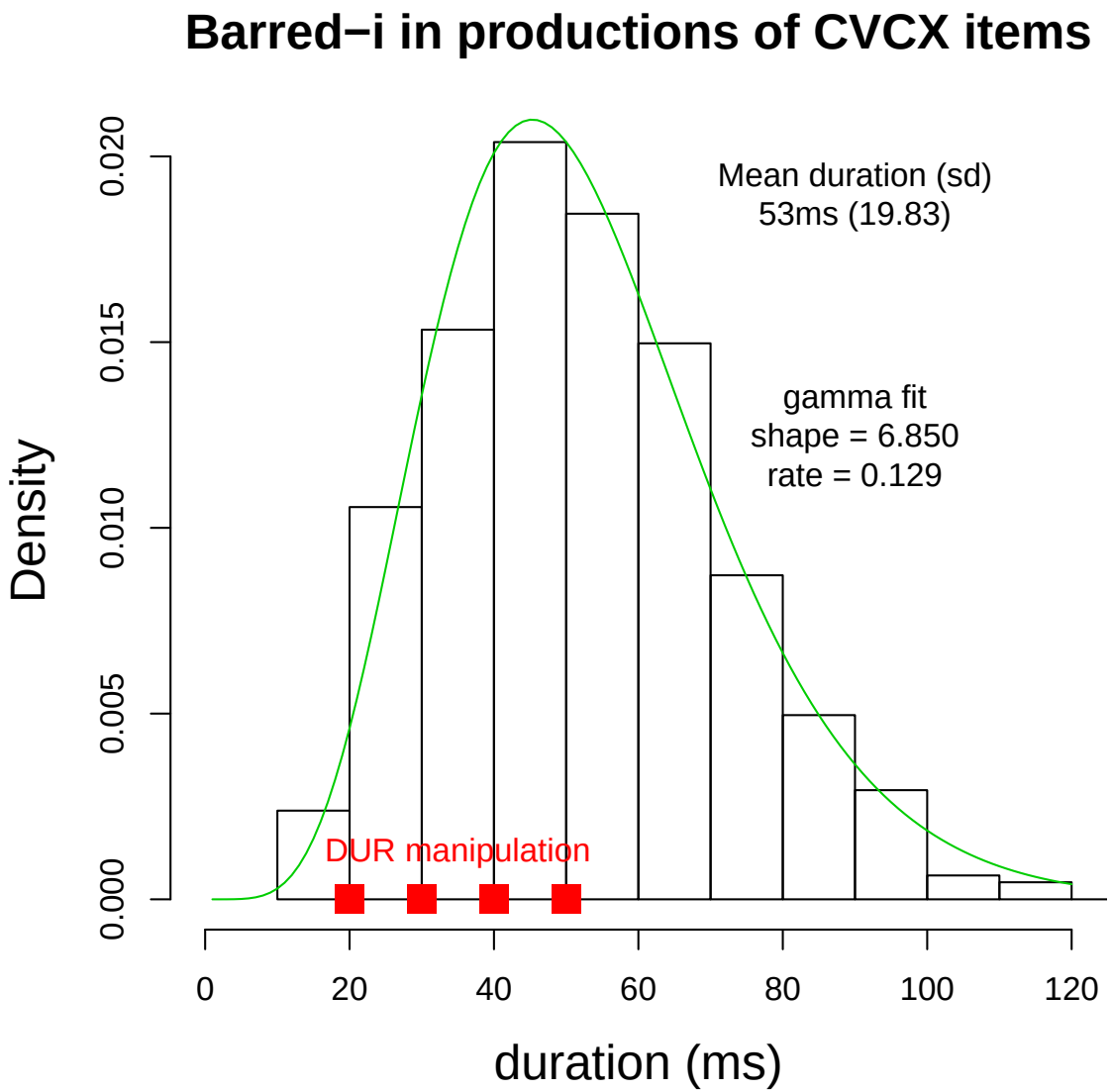


Accouting for epenthesis modifications

Rate of epenthesis modification is higher when the Russian open transition is more similar to typical realizations of English ɨ

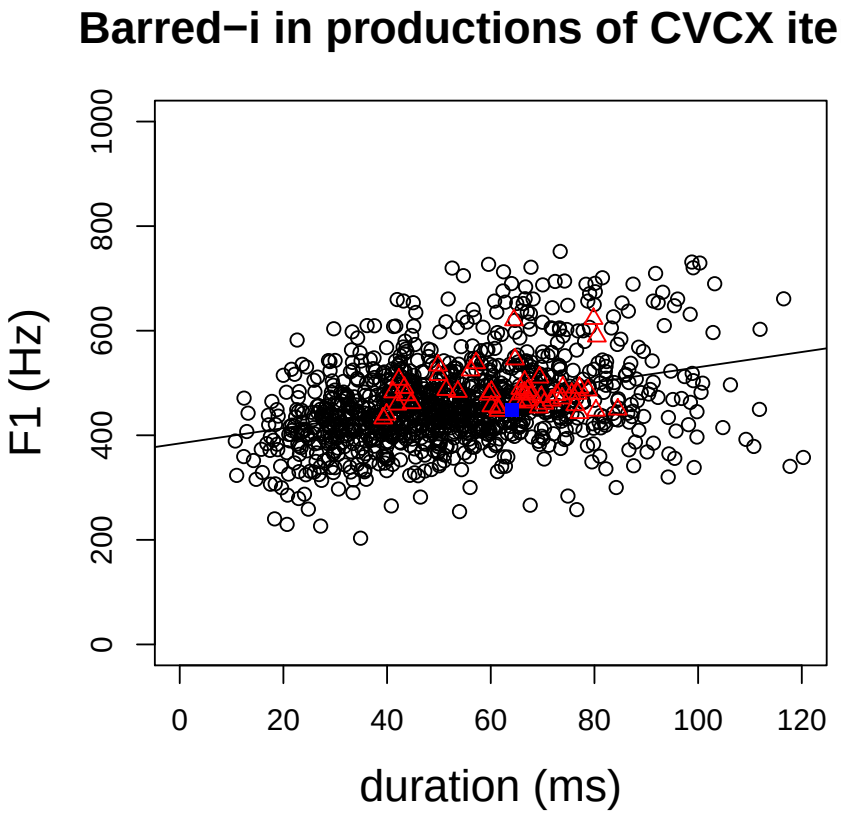
cluster type	voicing	formants	duration	epenthesis rate
S[+v]N	Y	Y (weak)	Y/N (DUR)	higher
S[+v]S	Y	N	Y/N (DUR)	↑
S[-v]N	N	N	Y/N (DUR)	↑
S[-v]S	N	N	Y/N (DUR)	lower

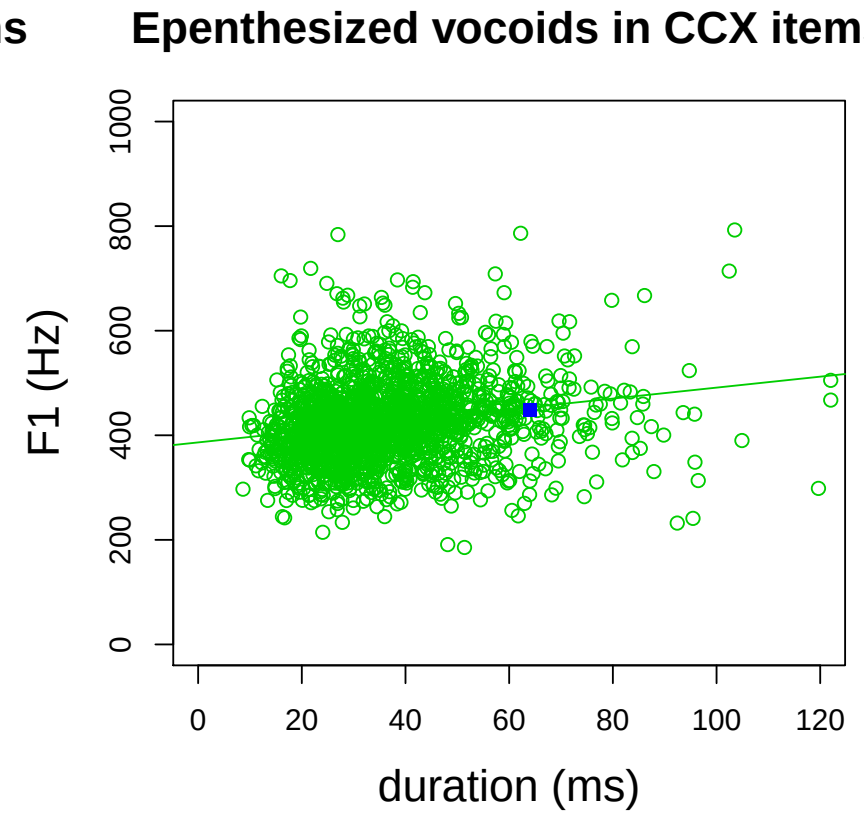
Accouting for epenthesis modifications



Lowest value of DUR (20ms) is unlikely for English ɨ and conditions signif. less epenthesis

Accouting for epenthesis modifications





Both ɨ and epenthesized vocoids show positive correlation between duration and F1, suggesting a similar gestural representation

(red = duration and F1 of Russian schwas in CəCX fillers; blue = mean ɨ duration and F1 from Flemming & Johnson 2007)

Accouting for epenthesis modifications

Epenthesis in voiceless stop-initial clusters

- English speaker may interpret burst+aspiration as containing a short devoiced vocoid: [kpago] → [k^hɨ̥pago]
- . . . and then generate a different realization of the structure, with the vocoid at least partially voiced: [k^hɨ̥pago] → [k^hɨpago]

This scenario is consistent with variable devoicing/laryngeal-overlap of English barred-i after voiceless stops (e.g., Davidson 2006)

Indeed, some 'correct' English productions have transition durations so long (> 100ms) that they may in fact contain instances of devoiced epenthetic vocoids!

Accouting for epenthesis modifications

Relative infrequency of epenthesis in fricative-initial clusters

- Russian productions of FX clusters lack the distinct open transition found in SX clusters
- . . . perhaps there is a relative lack of "raw material" for perceiving a vocoid between F and the following consonant

  (see Fleischhacker 2001, 2005; Minkova 2004; Zuraw 2007 on similar ideas about the importance of distinct 'perceptual breaks' for cluster splitability)

- Alternatively, existence of sN and sT clusters in English may make other FX clusters more phonotactically acceptable

Accounting for deletion and change modifications

Experimental results

- **More C1-deletion and -change found with lower amplitude**
- Less deletion of voiced stops and fricatives
- Less deletion/change before nasals

Perceptual explanation

- Release cues are of particular importance for perceiving the existence and place, other features of stops consonants

  (see Malécot 1958; Winitz et al. 1972; Kingston 1994; Jun 1995, 2004; Steriade 2001)

- . . . and a perceiver that fails to perceive a low-amplitude C1, or misperceives its features, will not reproduce it faithfully
- cf. fricatives and voiced stops have cues to their existence that make them less likely to be perceptually overlooked (internal cues; higher amplitude burst) (see Wright et al. 1996; Steriade 2001; Wright 2004)

Accounting for deletion and change modifications

Unanticipated effect of *raising* amplitude in SN clusters

- Expected to reduce rate of C1-deletion and -change, but rates were low without AMP manipulation
- Instead, raised AMP resulted in increased rate of epenthesis

This suggests another facet of English i that, if shared by the C1-C2 transition, increases prob. of epenthesis: intensity integrated over time

(see also Peperkamp et al. 2008 on **spectral energy** of release of word-final consonants)

Prothesis modifications

Experimental results

- **More prothesis when pre-obstruent voicing is present**
- Less prothesis for stop-initial clusters

Phonetic realization

- Russian word-initial [+v] fricatives have modal voicing throughout their duration
- But English word-initial [+v] fricatives are typically modally voiced only in a post-voiced environments (e.g., Smith 1997)
- . . . which can be provided by prothesizing a schwa
- . . . this is facilitated by initial voicing without frication (POV)
- English word-initial [+v] stops are also typically devoiced, but closure voicing may be unnoticed by English speakers and difficult to distinguish from POV when it is noticed

Summary

Pattern of modifications of non-native clusters largely understood in terms of the phonetic realization of native structures

- How similar is C1-C2 transition to an English reduced vowel w.r.t. voicing, formant structure, duration, total intensity?
  - greater similarity $\rightarrow$ more epenthesis
  - modulated by cluster properties, DUR and AMP manipulations
- What are the perceptual cues to C1 and its features?
  - less perceptible $\rightarrow$ more deletion and change
  - modulated by AMP manipulation
- Is obstruent phonetic voicing expected given the context?
  - less expected $\rightarrow$ more prothesis
  - modulated by POV manipulation

Summary

Results are accounted for by a model in which the perceiver interprets the stimulus as the 'best fitting' phonetic/phonological structure and then produces a phonetic realization of that structure

- Prior: bias against phonotactically illegal structures
- Likelihood: measure of acoustic-phonetic similarity between non-native stimuli and expected realizations of native structures
- Prior $\times$ Likelihood: non-native speaker perceives and produces the phon. structure that best satisfies phonotactic constraints while remaining faithful to the acoustics of the stimulus

Future direction

- Perception experiments with the same manipulated stimuli, to distinguish modifications with perception and production origins

  (see also Davidson & Shaw 2011)

- Quantification of likelihood / phonetic realization distribution (measure of 'similarity' between stimuli and expected phonetic realizations)

- Gradient phonotactic distinctions among non-native clusters? (perhaps C1: F $>$ S, C2: N $>$ S)

- Application of model to perception/reproduction of non-native sounds and native structures

Thank you!

And thanks to everyone who helped us

Undergraduate R.A.s
Alice Hall, Francesca Himelman, Johnny Mkitarian

Tuuli Adams, Adam Albright, Ian Coffman, Edward Flemming, Bruce Hayes, Jeff Heinz, Bill Idsardi, Veronica Monaghan, Brenda Rapp, Jason Shaw, Paul Smolensky, Michael Wolmetz, Julia Yarmolinskaya for their comments, questions, contributions

Penn Phonetics Lab Forced Aligner (Yuan & Liberman 2008)
Praat (Boersma & Weenink 2011)
R (R Development Core Team 2011)

NSF grants BCS-0449560 to LD and BCS-1052855 to LD and CW

References I

*

References

Albright, A. (2009). Feature-based generalization as a source of gradient acceptability. *Phonology*, 26(1):9–41.

Berent, I., Lennertz, T., Jun, J., Moreno, M., and Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences*, 105:5321–5325.

Berent, I., Lennertz, T., Smolensky, P., and Vaknin-Nusbaum, V. (2009). Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology*, 26(75-108).

Berent, I., Steriade, D., Lennertz, T., and Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104:591–630.

Broselow, E. and Finer, D. (1991). Parameter setting in second language phonology and syntax. *Second Language Research*, 7(1):35–59.

Davidson, L. (2003). *The Atoms of Phonological Representation: Gestures, Coordination and Perceptual Features in Consonant Cluster Phonotactics*. PhD thesis, Johns Hopkins University, Baltimore, MD.

Dupoux, E., Kakehi, H., Hirosi, Y., Pallier, C., and Mehler, J. (1999). Epenthetic vowels in Japanese: a perceptual illusion. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6):1568–1578.

Eckman, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning*, 27(2):315–330.

Greenberg, J. H. and Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, 20:157–177.

Hancin-Bhatt, B. (2000). Optimality in second language phonology: codas in thai esl. *Second Language Research*, 16(3):201–232.

Kabak, B. and Idsardi, W. J. (2007). Perceptual distortions in the adaptation of English consonant clusters: Syllable structure of consonantal contact constraints. *Language and Speech*, 50(1):23–52.

References II

Kang, Y. (2004). Perceptual similarity in loanword adaptation: English postvocalic word-final stops in korean. *Phonology*, 20(2):173–218.

Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition*, 84:55–71.

Peperkamp, S., Vendelin, I., and Nakamura, K. (2008). On the perceptual origin of loanword adaptations: experimental evidence from japanese. *Phonology*, 25(129-164).

Pitt, M. A. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & Psychophysics*, 60(6):941–951.

Scholes, R. (1966). *Phonotactic Grammaticality*. Mouton, The Hague.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7:49–63.

Yarmolinskaya, J. (2010). *Perception and Acquisition of Second Language Phonology (in progress)*. PhD thesis, Johns Hopkins University, Baltimore, MD.

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. *Language*, 83:277–316.