Acoustic-phonetic predictors of non-native consonant cluster modification

Colin Wilson, JHU (colin@cogsci.jhu.edu)
work in collaboration with
Lisa Davidson, NYU (lisa.davidson@nyu.edu)

November 17, 2011

Cross-language speech processing

Under a wide variety of natural and experimental conditions, *non-native* sounds and sequences elicit systematic patterns of performance that differ from those evoked by native structures.

- poorer discrimination

  (e.g., **?**; Kuhl et al. 1992; Best 1995; **??**)

- modifications in identification, transcription, **production**

  (e.g., **???**; Bent 2005; **??**; Shaw & Davidson 2011)

  Ex. [bdava] (Russian item) → [bᶤdava] (English production)

- loanword and L2 adaptations (e.g., **????**, Kenstowicz & Uffmann 2006, **??**)

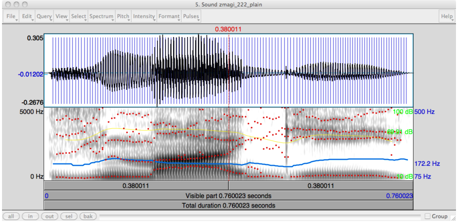
- lower acceptability judgments

  (e.g., **???**, Daland et al. 2011)

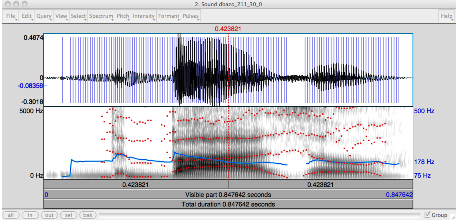
Cross-language speech processing

Examples

[?agi] play



[?azo] play



Cross-language speech processing

Two types of knowledge about the native language that could explain systematic performance on non-native sound structures:

- Knowledge of abstract phonological restrictions on possible sounds and sound combinations
  - sonority sequencing (e.g., **???**)
  - syllable parsing (e.g., **?**)
  - segmental phonotactics (e.g., **??**)
  - restrictions on gestural overlap (eg., **?**)
- Knowledge of how phonological representations are phonetically realized — and their expected acoustic/auditory signatures
  - open vs. close CC transition (e.g., Catford 1977; Zsiga 2003; Davidson 2011)
  - strongly released vs. unreleased C# (e.g., Kang 2004; Peperkamp et al. 2008)
  - durational, laryngeal, and spectral variability of reduced vowels

    (e.g., Keating & Huffman 1984; Tsuchida 1998, 2001; Davidson 2006; Flemming & Johnson 2007)

Cross-language speech processing

Patterns in the perception and production of non-native *sequences* have most often been attributed to abstract phonotactics

- Universality of the Sonority Sequencing Principle

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

- *These results buttress the hypothesis that speech perception is heavily influenced by **phonotactic knowledge**. ... Indeed, not only does phonotactic knowledge influence the classification of individual phonemes, but it **can also induce the perception of "illusory" phonemes that have no acoustic correlates***

  — Dupoux et al. 1999 (emphases added)

  Ex. $*\mathrm{C}_{[\text{-son}]}\mathrm{C}$ alone leads to misperception of [ebzo] as [ebɯzo]

Cross-language speech processing

A purely phonological phonotactic approach to non-native cluster perception and production fails to predict:

- *rate* of modifications made in production across cluster types (e.g., error rate does not decrease with sonority increase)
- *type* of modification made to each cluster type (unless supplemented by knowledge of phonetic realization)
- modifications that result in different, but not phonotactically better, consonant clusters

Phonotactic knowledge may be one predictor but is unlikely to be the only (or even the most important) one . . .

Cross-language speech processing

Analyses of the perception and production of *individual* non-native sounds have long emphasized the role of phonetic realization (e.g., Iverson & Kuhl 1995; Harnsberger 2000; Best et al. 2001, 2003; Escudero & Boersma 2004, Escudero & Vasiliev 2011)

- Ex. Marathi listeners identify the Malayalam retroflex nasal [ɳ] as dental [n̪] up to 54% of the time, even though both languages have a dental-retroflex contrast (Harnsberger 2000)
- Ex. Peruvian listeners identify Canadian French [ɛ] → [e] and [æ] → [a], but Canadian English [ɛ], [æ] → [a] (Escudero & Vasiliev 2011)

How much of non-native consonant cluster processing can be predicted by knowledge of native-language phonetic realization?

(see also Dupoux et al. 2011; Davidson & Shaw, to appear, on non-native cluster perception)

A Bayesian framework

Abstract phonological knowledge and knowledge of phonetic realization can be combined by Bayes' Theorem (for related applications to perception and production see e.g., **???**; Feldman et al. 2009)

Given stimulus $\{z\}$, the probability that an observer will perceive (and attempt to produce) phonological representation $[x]$ is $\propto$

the probability of $[x] \rightarrow \{z\}$ according to the observer's native-language knowledge of phonetic realization

$\times$

the probability of $[x]$ according to the observer's native-language knowledge of phonology

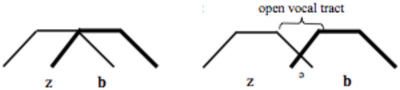
$$P([x] \mid \{z\}) \;\propto\; \underbrace{P([x] \rightarrow \{z\})}_{\textit{likelihood}} \times \underbrace{P([x])}_{\textit{prior}}$$

A Bayesian framework

Motivating example

Part of native English speaker's knowledge of phonetic realization is that consonant clusters C1C2 typically show *close* transition

- traditional phonetic observation (e.g., Catford 1977, 1988)
- modeled as gestural overlap (see Byrd 1992, 1996; Gafos 2002; Zsiga 2003; Davidson 2006)



(close transition) (open transition)

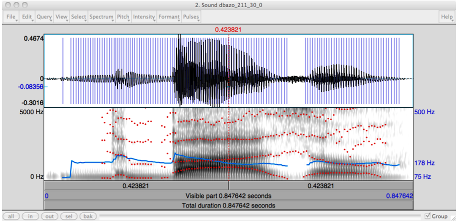
- C1 release is infrequent ($\sim 25\%$) in word- and phrase- medial clusters in spontaneous speech (Davidson 2011)
- expect C1 release to be of short duration when it does occur

A Bayesian framework

Motivating example

How should an English observer be expected to represent (and attempt to reproduce) the following Russian utterance?

[dbazo] play



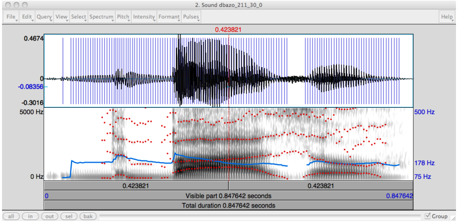
$= \{z\}$ (stimulus)

- Probability of realizing cluster [db] with long open transition should be low — not consistent with English gestural timing
- Probability of realizing sequence [d^{ə}b] or [d^{ɨ}b] this way may be similar or higher: $p([\mathrm{d}^{ə}\mathrm{b}] \rightarrow \{z\}), p([\mathrm{d}^{ɨ}\mathrm{b}] \rightarrow \{z\}) \geq p([\mathrm{db}] \rightarrow \{z\})$

A Bayesian framework

Motivating example

[dbazo] play



$= \{z\}$ (stimulus)

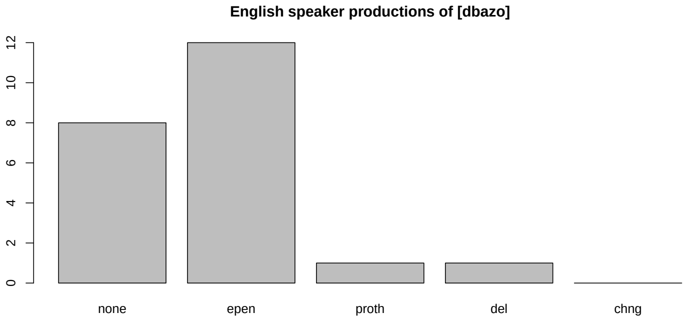
Likelihoods of alternative phonological representations

[dbazo]	open transition unlikely, burst+release too long?
[d^{ə}bazo]	F1 low (407Hz), F2 ok (1775Hz), too short (36ms)?
[d^{ɨ}bazo]	F1 somewhat low, F2 somewhat high, too short?

Flemming & Johnson 2007: [ə] F1=539 (90), F2=1797 (97); [ɨ] F =449 (56), F2=1922 (121)

A Bayesian framework

Motivating example



Likelihoods alone may largely predict this and other patterns of performance, with phonotactics (the prior) playing a more minor role.

Overview

- Cross-language production experiment
  - Systematically manipulate selected acoustic properties of Russian stimuli beginning with C1C2 clusters that are illegal in English.
  - General prediction
    If knowledge of phonetic realization strongly influences patterns of correct perception/production and modification, acoustic manipulations should be mirrored by changes in performance.
- Specifying the likelihood function that reflects language-specific knowledge of phonetic realization.

Cross-language cluster production experiment

English speaking participants (N = 12) heard and repeated critical items of the form [C1C2áCV] produced by a native Russian speaker.

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

- 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
- fillers were [əC1C2áCV] (proth.) and [C1əC2áCV] (epen.) counterparts of the critical items

Cross-language cluster production experiment

Order of events in a trial

- one version of a stimulus item was played twice (with a brief ISI)
- participant repeated the stimulus

Each participant heard and produced 288 total stimuli

- 32 FN, 32 FS
- 64 SN, 64 SS (each approx. half S_{-v}X, half S_{+v}X)
- 48 epenthesis, 48 prothesis fillers

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

Acoustic manipulations

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

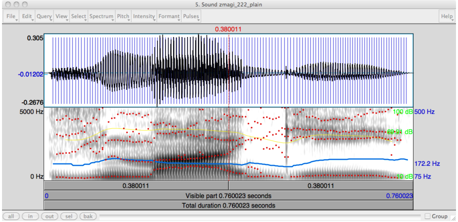
- POV (pre-obstruent voicing): modally-voiced interval that precedes the formation of a voiced obstruent constriction
  POV present → more prothesis
- DUR: duration of the acoustic transition (burst + aspiration) between an oral stop and closure of the following stop
  DUR longer → more epenthesis
-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

Versions of the stimulus items were created by systematically manipulating these acoustic-phonetic properties . . .

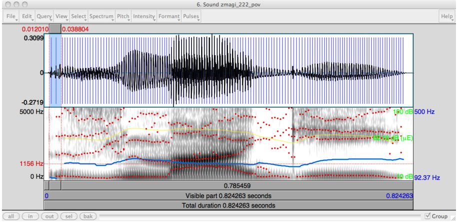
Acoustic manipulations

FX clusters [vm vn zm zn; vd vg zb zg]: POV (pre-obstruent voicing)

[zmagi] play
POV absent



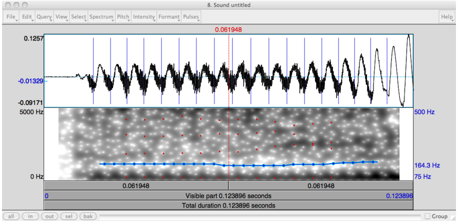
[zmagi] play
POV present



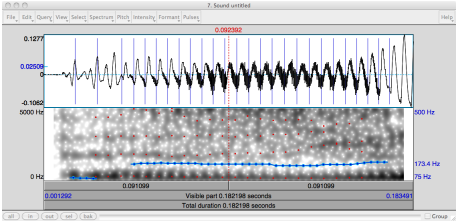
Acoustic manipulations

FX clusters [vm vn zm zn; vd vg zb zg]: POV (pre-obstruent voicing)

[zmagi] play
POV absent



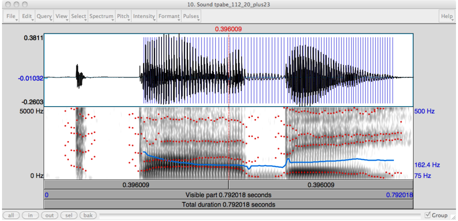
[zmagi] play
POV present



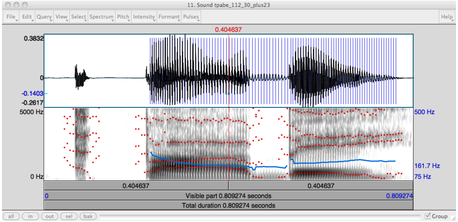
Acoustic manipulations

S$_{-v}$X clusters [pn tm km kn; pt tp kp kt]: DUR (transition duration)

[tpabe] play
DUR = 20ms



[tpabe] play
DUR = 30ms

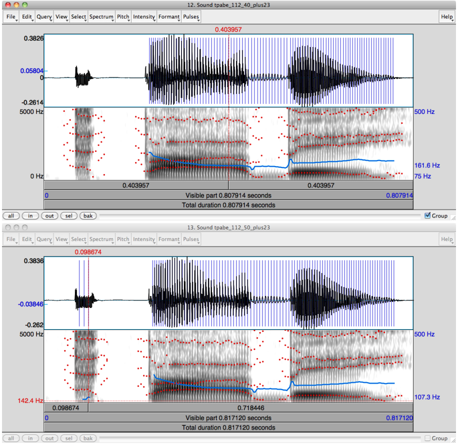


Acoustic manipulations

S$_{-v}$X clusters [pn tm km kn; pt tp kp kt]: DUR (transition duration)

[tpabe] play
DUR = 40ms

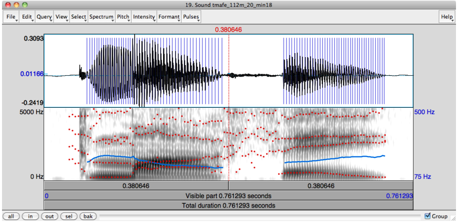
[tpabe] play
DUR = 50ms



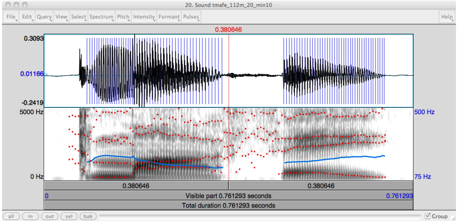
Acoustic manipulations

S$_{-v}$X clusters [pn tm km kn; pt tp kp kt]: AMP (transition amplitude)

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



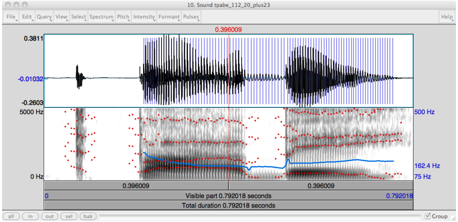
[tmafe] play
DUR = 20ms
AMP = -10dB
(raised)



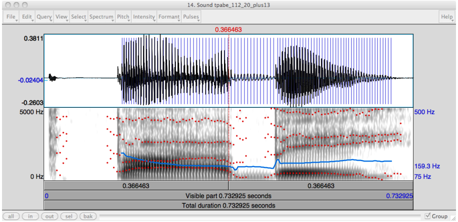
Acoustic manipulations

S$_{-v}$X clusters [pn tm km kn; pt tp kp kt]: AMP (transition amplitude)

[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 (NYU) 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 relative burst intensity see also Stoel-Gammon et al. 1994; Sundara 2005)

Acoustic manipulations: summary

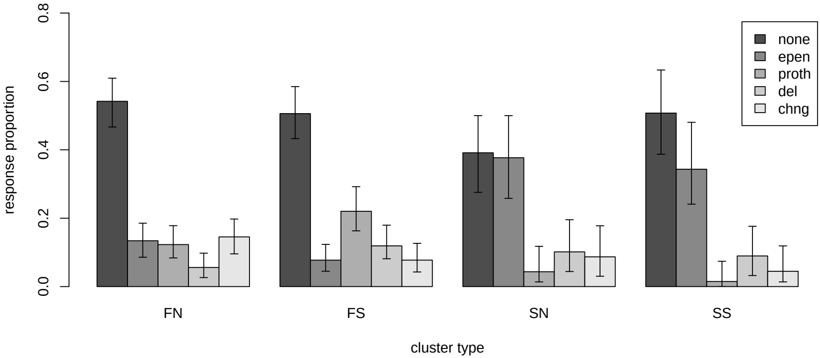
FX clusters
POV (absent vs. present)

S$_{-v}$X clusters
DUR (20ms, 30ms, 40ms, 50ms) $\times$ AMP (base vs. lowered/raised)
(recall AMP denotes raised before N and lowered before S)

S$_{+v}$X clusters
DUR (20ms, 30ms, 40ms, 50ms) $\times$ POV (absent vs. present)
DUR (20ms, 30ms, 40ms, 50ms) $\times$ AMP (base vs. lowered)
(recall AMP denotes raised before N and lowered before S)

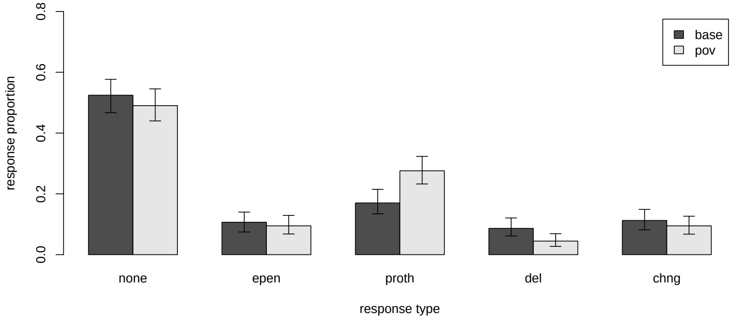
+ fillers = 800 total sound files, distributed across 12 experimental lists so that each critical version occurs equally often across the lists

Response proportions (‘baseline’ versions only)



Note: error bars of all graphs show 95% adjusted bootstrap percentile intervals, $rep = 1000$, as calculated by R functions boot::boot, boot.ci

Effect of POV manipulation on FX production



Effect of POV manipulation on FX production

Significantly more prothesis induced by

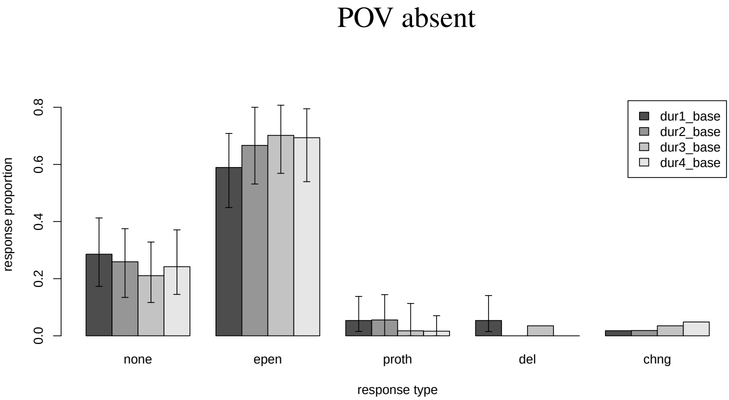
- **variants with POV vs. variants without POV**
- FS clusters vs. FN clusters

No interaction between POV and cluster type, and overall prothesis is rarer than other response types (esp. no modification) for FX clusters.

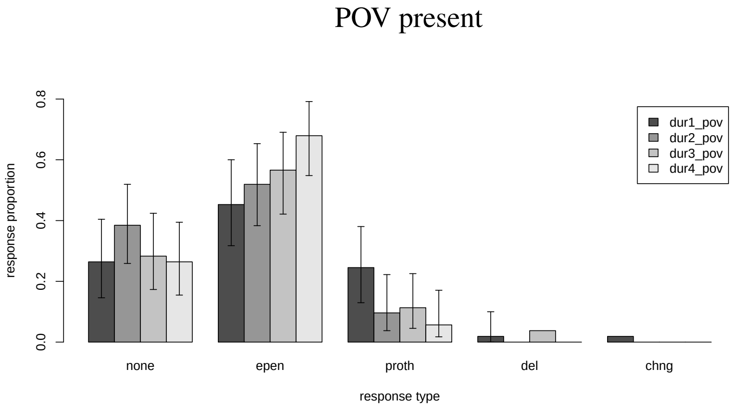
```
Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.74283    0.41108  -4.240 2.24e-05 ***
condition1              -0.31666    0.12051  -2.628   0.0086 **
cluster.type1           -0.36844    0.17027  -2.164   0.0305 *
condition1:cluster.type1 0.01742    0.10787   0.161   0.8717
```

Note: all statistical analysis are mixed-effects logistic regressions with effect (sum-to-zero) coding of fixed factors, as calculated by R lme4::lmer

Effect of POV manipulation on $S_{+v}X$ production



Effect of POV manipulation on $S_{+v}X$ production



Effect of POV manipulation on $S_{+v}X$ production

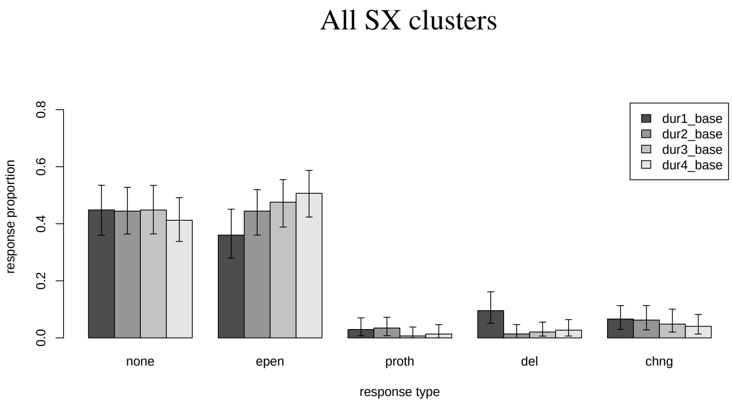
Significantly more prothesis induced by

- **variants with POV vs. variants without POV**
- SN clusters vs. SS clusters (opposite of FN vs. FS effect)

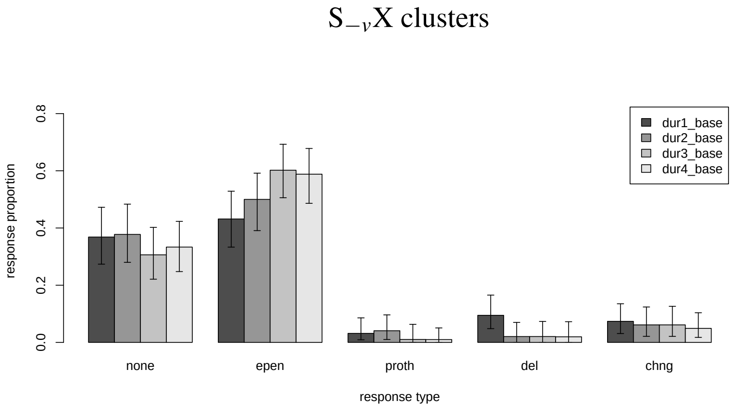
No interaction between POV and cluster type, and overall prothesis is much rarer than other response types for SX clusters.

```
Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.225780   0.393204  -8.204 2.33e-16 ***
pov1                -0.794939   0.257565  -3.086  0.00203 **
cluster.type1        0.618551   0.305260   2.026  0.04273 *
pov1:cluster.type1  -0.004641   0.257723  -0.018  0.98563
```

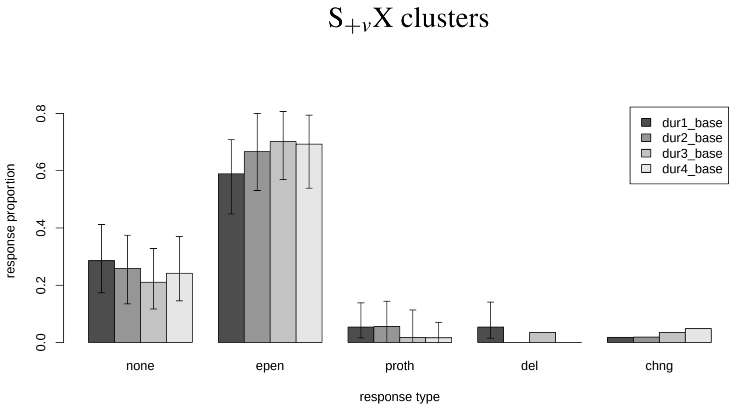
Effect of DUR manipulation on SX production



Effect of DUR manipulation on SX production



Effect of DUR manipulation on SX production



Effect of DUR manipulation on SX production

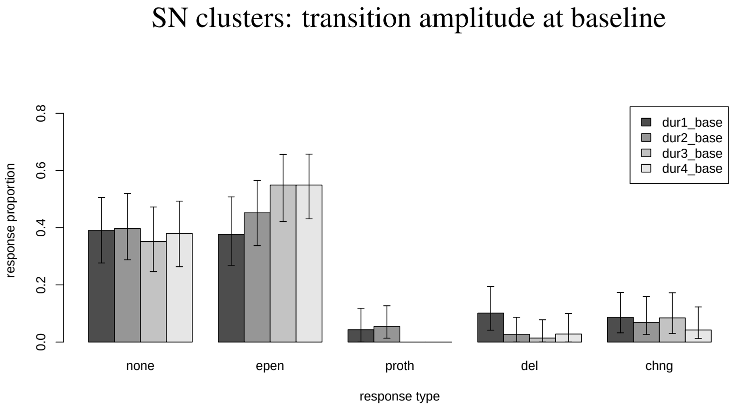
Significantly more epenthesis induced by

- **variants with longer vs. shorter transition duration** specifically, epenthesis rate for dur4 greater than mean rate
- voiced- vs. voiceless- initial clusters

No main effect of cluster type (SS vs. SN) or interaction between interaction DUR and cluster type or cluster voice [just trust me].

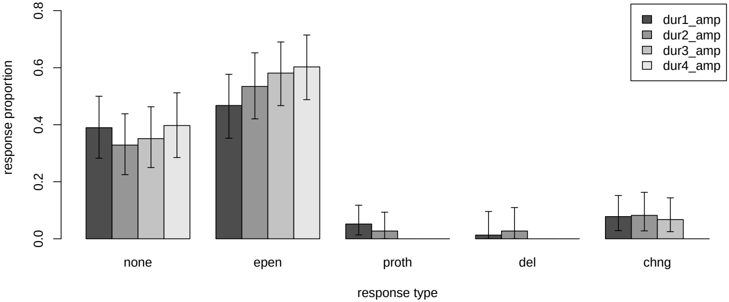
```
Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.03005    0.36404  -0.083   0.9342
condition1     -0.55115    0.26691  -2.065   0.0389 *
condition2      0.01451    0.22502   0.064   0.9486
condition3      0.29406    0.24788   1.186   0.2355
cluster.type1   0.18150    0.18006   1.008   0.3134
cluster.voice1  1.08682    0.14874   7.307 2.73e-13 ***
```

Effect of AMP manipulation on SX production

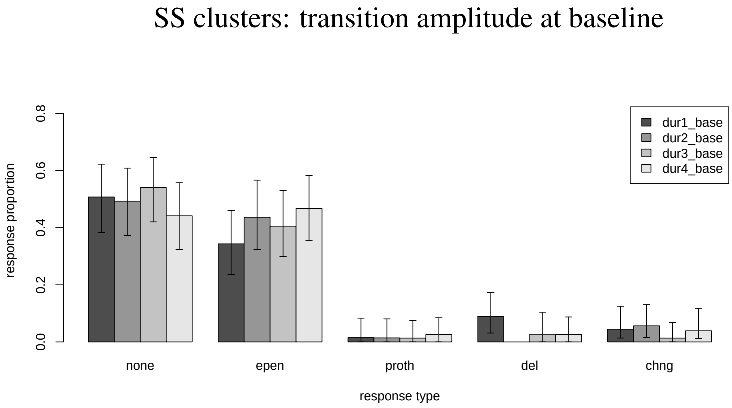


Effect of AMP manipulation on SX production

SN clusters: transition amplitude *raised*

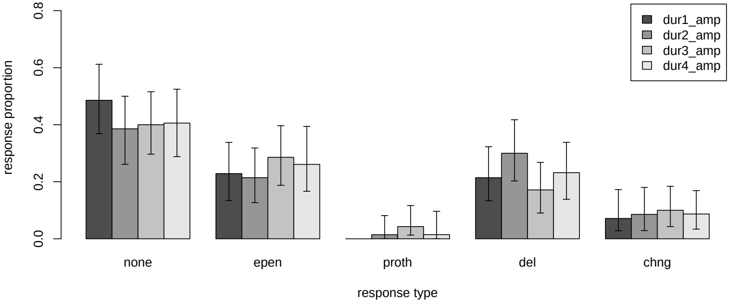


Effect of AMP manipulation on SX production



Effect of AMP manipulation on SX production

SS clusters: transition amplitude *lowered*



Effect of AMP manipulation on SX production

Significantly more deletion induced by

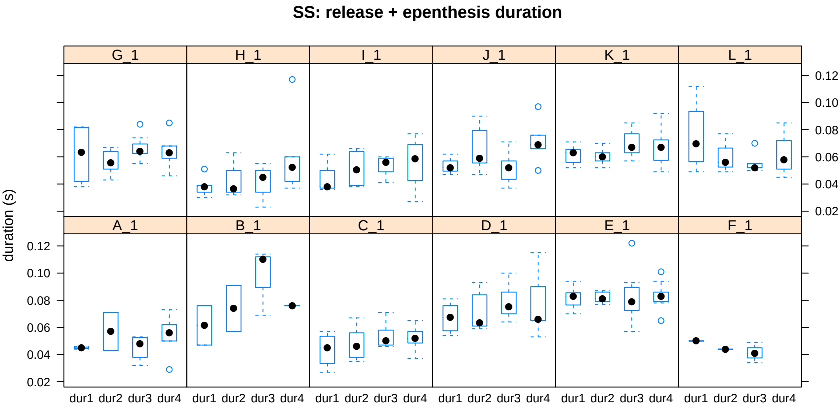
- **variants with lower vs. higher release amplitude**
- SS vs. SN clusters (probably an artifact of lowering vs. raising)

No effect of cluster voice, and no interaction between AMP and cluster type or cluster voice [just trust me].

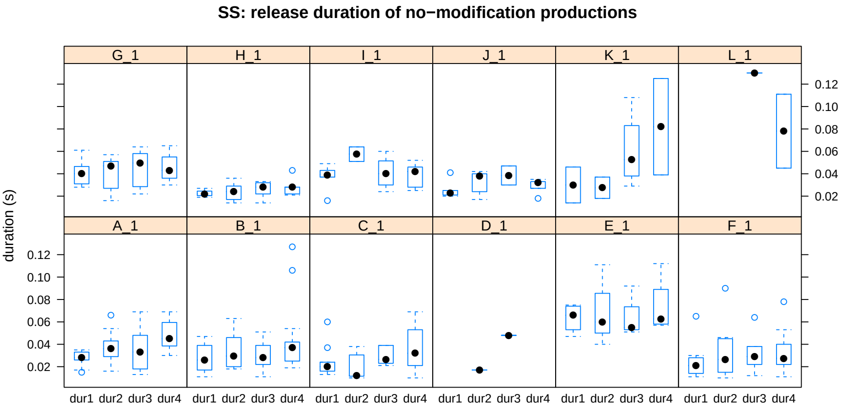
```
Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.7034     0.4939  -9.522  < 2e-16 ***
amp1            -1.7341     0.2995  -5.790 7.04e-09 ***
cluster.type1   -1.0764     0.2601  -4.139 3.49e-05 ***
cluster.voice1  -0.1079     0.3004  -0.359    0.719
```

Note: amp = +1 consistently denotes *higher* amplitude (i.e., modified variants for SN but baseline variants for SS) in this statistical analysis

Are modifications due to phonetic imitation?



Are modifications due to phonetic imitation?



Are modifications due to phonetic imitation?

SS: release + epenthesis duration

```
Fixed effects:
             Estimate Std. Error t value
(Intercept) -2.847787   0.061472  -46.33
dur1        -0.063584   0.027015   -2.35 *
dur2        -0.016802   0.025093   -0.67
dur3         0.005602   0.028078    0.20
```

SS: release duration of no-modification productions

```
Fixed effects:
            Estimate Std. Error t value
(Intercept) -3.35419    0.10031  -33.44
dur1        -0.15347    0.04830   -3.18 *
dur2        -0.04662    0.04913   -0.95
dur3         0.04010    0.04816    0.83
```

Toward a likelihood-based theory of cluster modification

Acoustic-phonetic principles of cluster modification

The observer must *infer* the structure of the phonological representation from noisy measurements of the stimulus.

- Noise an uncertainty are an ineluctable properties of perception in all domains (auditory, visual, tactile, . . .).
- Sources of noise include external transition medium and internal processing by the perceiver.
- Bayes' Theorem provides a rational basis for inference under uncertainty (e.g., as in signal detection theory)
  - strong enough likelihood overwhelms the prior
  - do not posit multiple segments needlessly ('explaining away')
  - do not posit segments in the absence of evidence (parsimony)

Acoustic-phonetic principles of cluster modification

The observer will perceive, and attempt to reproduce, sounds with acoustic signatures ('cues') that cannot reasonably be attributed to other sources.

+ segments with *internal cues* are protected from deletion if these cannot be attributed to neighboring segments
    - frication of F before closure of S or nasal murmur of N
    - voice bar of S_{+v} before burst+release
+ segments with *contextual cues* are protected from deletion if these cannot be attributed to neighboring segments
    - outgoing formant transitions of C2 in all clusters
    - clear burst+release of S_{-v}X in most stimuli
- higher rate of deletion for S_{-v}S and S_{-v}N with *lower-amplitude* burst+release (cf. S_{+v}S with amplitude lowered)
    - not heard at all? too uncertain? not worth effort of reproducing?

Acoustic-phonetic principles of cluster modification

Deletion modifications

C1	FN	FS	SN	SS
voiceless	–	–	12	52
voiced	14	32	5	23

- deletion twice as frequent for $S_{-v}X$ as for $S_{+v}X$
- majority of S_{-v} deletions are in dur1_amp, dur2_amp variants (short burst+release duration with lowered amplitude)

But what about unexpected deletions in FX clusters?

Acoustic-phonetic principles of cluster modification

Deletion modifications

- FN deletions target **near-homorganic** [vm], except two [vn]

	vm	vn	zm	zn
¬del	81	83	94	93
del	12	2	0	0

$\chi^2(3) = 22.5, p < .001$ (Pearson's Chi-square with add-one smoothing)

- FS deletions target **non-strident** [v], except one [zb]

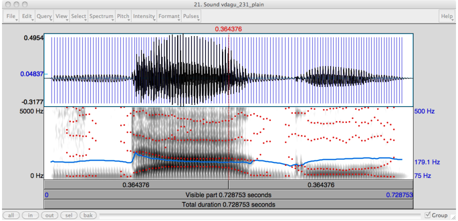
	vd	vg	zb	zg
¬del	75	69	85	80
del	13	18	1	0

$\chi^2(3) = 27.9, p < .001$ (Pearson's Chi-square with add-one smoothing)

It is well-established that the internal cues for strident [s/z] are more robust than those for non-strident [f/v]. (e.g., Miller & Nicely 1955, Steriade 1999, et seq.)

Acoustic-phonetic principles of cluster modification

[vdagu] play
(baseline)



Possible cues for the existence of [v] separate from following [d]:

- frication (higher-frequency energy): but this is characteristically weak in voiced esp. non-strident fricatives (e.g., Ohala 1983, Ohala & Solé 2010).
- voice bar (lower-frequency energy): but this is potentially grouped with voiced bar of following stop (‘explaining away’)*

*We are not sure how sensitive English listeners are to the presence, let alone duration, of voicing during stop closure — ideas?

Acoustic-phonetic principles of cluster modification

The observer does not hallucinate, and will not produce, sounds that have no acoustic source in the stimulus.

+ there should be no acoustic-phonetic evidence for an initial vowel, however short, in stimuli beginning with S_{-v}
  - only 7 instances (.3%) of prothesis modification of $S_{-v}X$ clusters

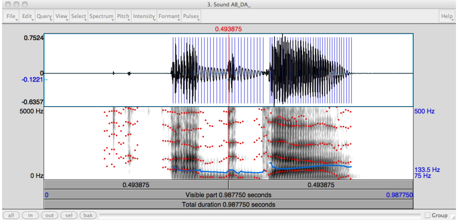
– transitions of $S_{-v}S$ items provide evidence for some acoustic event between the two consonants
  - transition interpretable as a short, devoiced vocoid — likelihood predicted to depend on frequency of reduced vowel devoicing

– transitions of $S_{+v}S$ items provide even clearer evidence for a vocoid between C1 and C2: voicing, (weak) formant structure
  - these are the clusters for which epenthesis is observed most often

Acoustic-phonetic principles of cluster modification

Example of "no vowel" between two consonants in the stimulus set of Dupoux et al. 2011 (thanks to Sharon Peperkamp for making the sound files available)

[abda] play



*"We . . . compared digitally produced clusters (that might have residual coarticulation information) and naturally produced clusters (**that have no coarticulatory information for a vowel**). We found that Japanese participants did not perceive more [u] vowels in digital than in natural clusters (**in fact, there was a nonsignificant trend in the other direction**)."*
— Dupoux et al. 1999 (emphases added)

Factors of the acoustic-phonetic likelihood function

We anticipate a model in which multiple noisy measurements are used to infer ('reconstruct') the most probable phonological representation.

- stops
  - burst+aspiration (and spectrum)
  - VOT and formant transitions
  - perhaps voice bar
- fricatives
  - frication (and spectrum)
  - voicing
- vocoids
  - formant structure and coarticulation
  - duration
  - voicing

(Building heavily on previous research on speech perception and phonetically-based phonology., e.g, Hayes et al. 2004)

Work in progress

- Perception experiments with the same stimuli

  (see also Shaw & Davidson 2010, Davidson & Shaw, to appear)

- Quantification of the likelihood function with English phonetic norms (e.g., phonetic variability of medial reduced vowels)

- Identification and modeling of (residual) effect of the phonotactic prior — we do not anticipate eliminating this!

- Further semi-automatization of data coding

Thanks to everyone who helped us

Graduate R.A.
Sean Martin

Undergraduate R.A.s
Alice Hall, Francesca Himelman, Johnny Mkitarian [yes, for real!]

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)

Thank you!

We gratefully acknowledge NSF grant BCS-1052855 to LD and CW