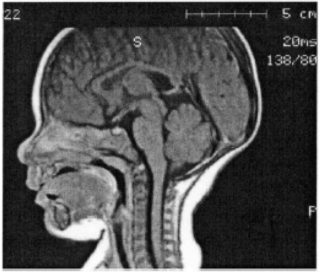
Size effects in lexical access

Kiwako Ito, Kathryn Campbell-Kipler, Elizabeth McCullough, Mary E. Beckman, Eric Ruppe, Tsz-Him Tsui

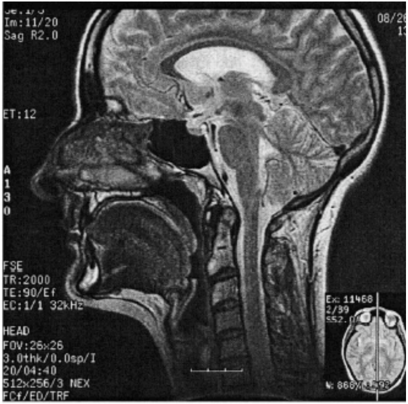
Department of Linguistics
The Ohio State University
http://ling.osu.edu/research/groups/soundsize

Differences between adults and children

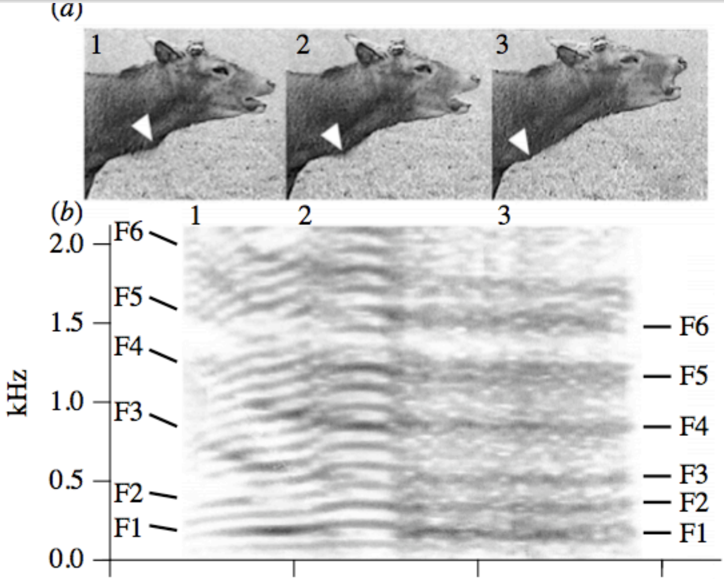
Adults' vocal tracts are longer overall and the pharyngeal cavity is disproportionally longer.



Midsagittal MRIs of 7-mo-old girl (above) and woman (right). (Vorperian, Kent, Lindstrom, Gentry, Yandell, 2005)

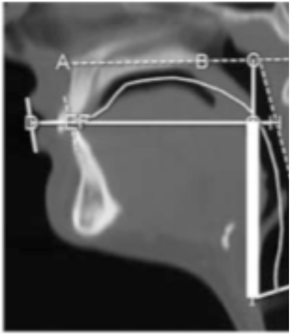


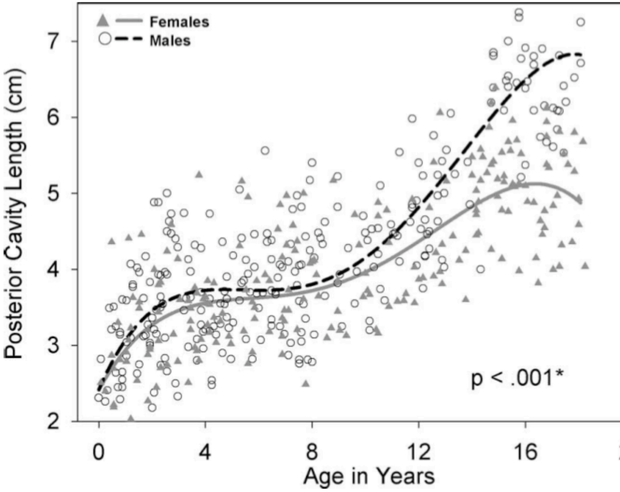
Fitch & Reby (2001) on gendered “roar” of red deer



Physical differences between men and women

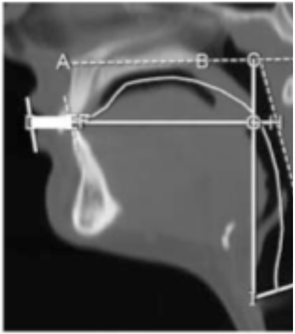
Men's vocal tracts are longer overall and the pharyngeal cavity is disproportionally longer.

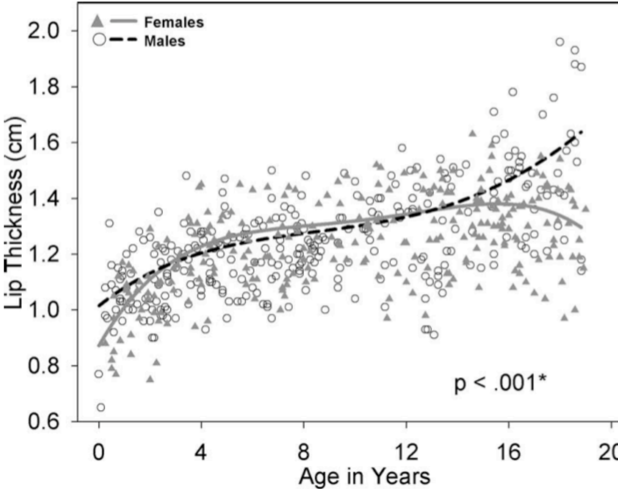




Differences between men and women, cont.

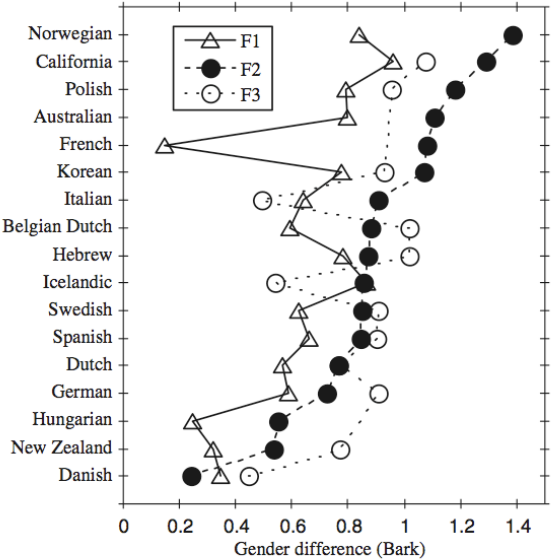
In men, the lip tube also is disproportionally longer. Could this be basis of gendered /s/ ?





Culture-specific performance of talker size effects

Cross-language differences in talker size effects on vowel formant values (Johnson 2005).



The larger research question

Size effects and talker "normalization"

- Different talkers have different sized vocal tracts
- Size effects such as the gendering of /s/ in American English may be rooted in such physical differences, but they are also highly culture-specific
- Phonological contrasts that are cued by spectral differences must always be parsed against this backdrop of "natural" but culture-specific size effects

Size effects and category differentiation

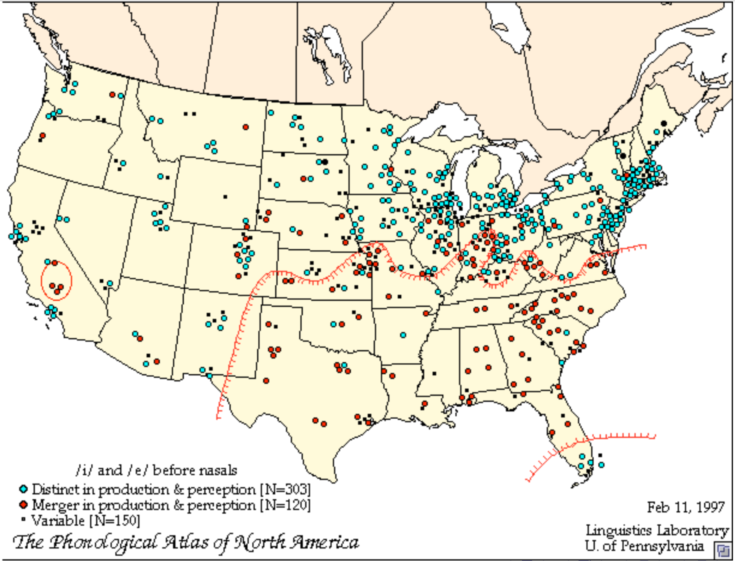
- The gendering of /s/ in American English has the effect of moving /s/ further away from /ʃ/ for some speakers, but reducing the contrast for others
- How can we measure the effects of such reduced contrast on lexical access ?

Strategy for addressing the question

- Look at vowel formants, a better understood phonetic parameter space that shows culture-specific size effects
- Find a contrast that is reduced for some speakers relative to others in this space – i.e., a vowel pair that is merged in some context for some group of speakers
- Develop a measure of degree of merger in productions by listeners who participated in a visual world paradigm study of talker adaptation effects
- Use this measure as a predictor variable in analyzing inter-listener differences in speaker adaptation effects

We will use the *pin-pen* merger – reduction or loss of contrast between /ɪ/ and /ɛ/ before nasals – a feature of some US dialects

http://www.ling.upenn.edu/phono_atlas/maps/



Social stereotypes about the *pin-pen* merger

- The merger is associated with rural and older speakers in the South (Preston, 1989 ; Tillery & Bailey, 2004 ; Gentry, 2006)
- Visually invoked stereotype about older speakers can affect lexical access (Koops et al., 2008)



- The merger is also widely found among African Americans across regions (Labov et al., 2006)
- Evidence of social stereotypes about the merger in, e.g., southern Ohio ... `http://www.ilovesooh.com/2011/05/englewood-not-inglewood-ca.html`

Speaker adaptation and merger

- Listeners store speaker-specific phonetic details in memory and use them to facilitate subsequent lexical processing (Nygaard & Pisoni 1998, Creel et al. 2008)
- Speaker-adaptation may result in lexical re-organization.
- For example, when a listener adapts to a specific speaker who raises the low front vowel /ae/ to /ɛ/ before /g/ ...
  - cohorts in standard pronunciation (e.g., *bag* and *back*) become non-cohorts (Dahan et al. 2008), and ...
  - non-cohorts in standard pronunciation (e.g., *bag* and *baker*) become cohorts (Trude & Brown-Schmidt, 2011)
- Speaker adaptation can be triggered by "phonetic details" inferred from photos suggesting relevant social characteristics of the talker (e.g., Johnson et al., 1999 ; Hay et al., 2006)

Ito & Campbell-Kibler (2012) test Ohio stereotypes

Research question

How do visually evoked stereotypes affect perceptual expectations prior to and during adaption ?

Method

Visual object detection task

- Participant sees 8 pictured objects surrounding a picture of the "speaker", hears voice giving instructions (e.g., *Click on the fence.*), and clicks on picture of perceived word
- On non-filler trials, pictures are of target (e.g., *fence)*, competitor (e.g., *fins*), and 6 distractors
- Fixation locations (i.e., x- & y-coordinates on the screen) measured at 50 Hz using Tobii 1750.

Example trial, with target *fins* and competitor *fence*

subject hears [ɪ] in *fins*

Speaker voice adaptation method

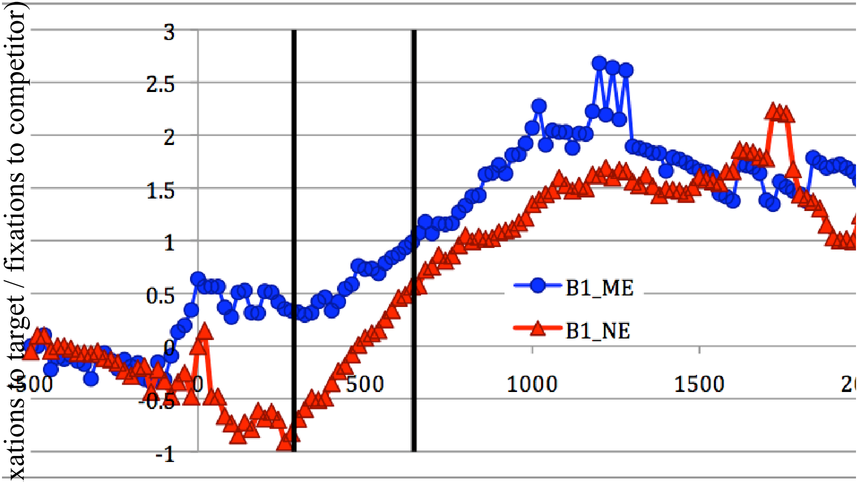
Participants hear four male voices, 2 merged speakers (more [ɛ]-like pronunciations in both *pin* and *pen* & 2 non-merged speakers ([ɛ] only in *pen*-words), in 3 blocks

- Block 1 : familiarization to voice on [ɛ] in *pen*-words : *bench*, *fence*, *tent stake*
- Block 2 : exposure to merger evidence with [ɛ] in *pin* words : (pronounced with "unambiguous" [ɪ] only by non-merged speakers)
  bin, *dinner plate*, *fins*, *mint*, *pins*, *tin-can phone*
- Block 3 : evaluation of adaptation with [ɛ] in *pen*-words : (repeated from Block 1) *bench*, *fence*, *tent stake*
  (and new items) *dentist sign*, *men*, *pencil* ....

Block 1 : Familiarization, unambiguous [ɛ] in *bench*

subject hears [ɛ] in *bench*

Block 1 results : An advantage for merged voices



Group data (n=80): merged voices (ME) led to faster detection of target /ɛn/-objects than non-merged (NE) ($t = 1.74$, $p<.05$)

Block 2 : Exposure to [ɪ] or [ɛ] in *pin* words

subject hears [ɪ] in *mint*

or hears [ɛ] in *mint*

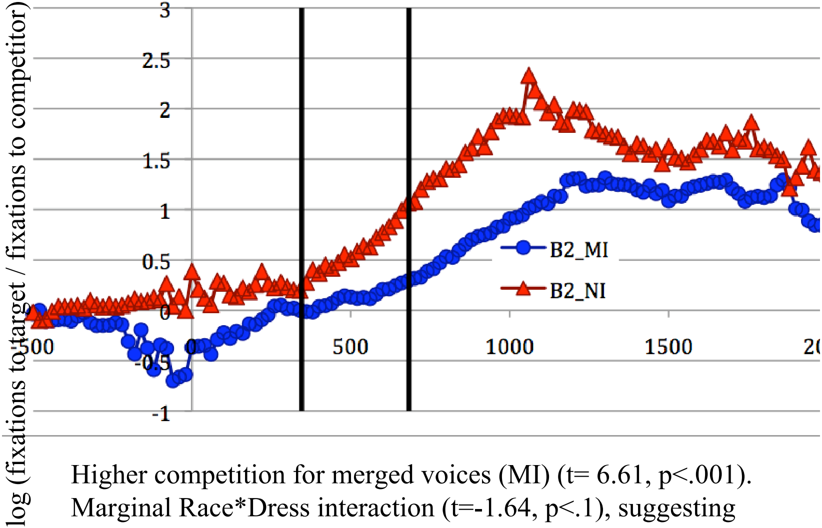








Block 2 results : An advantage for non-merged voices



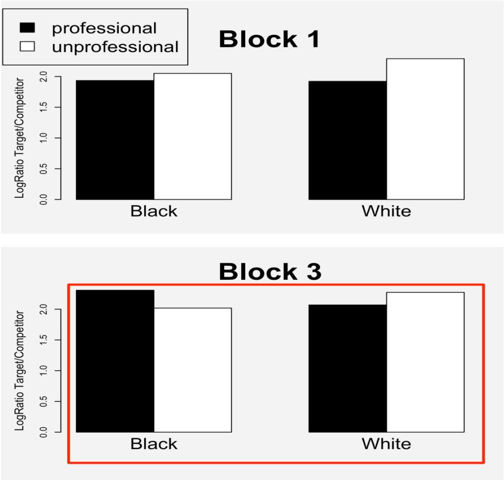
Higher competition for merged voices (MI) (t= 6.61, p<.001).
Marginal Race*Dress interaction (t=-1.64, p<.1), suggesting more looks to competitor for Black "speaker" in casual clothes.

Block 3 : Evaluation of adaptation, target *pencil*

subject hears N voice

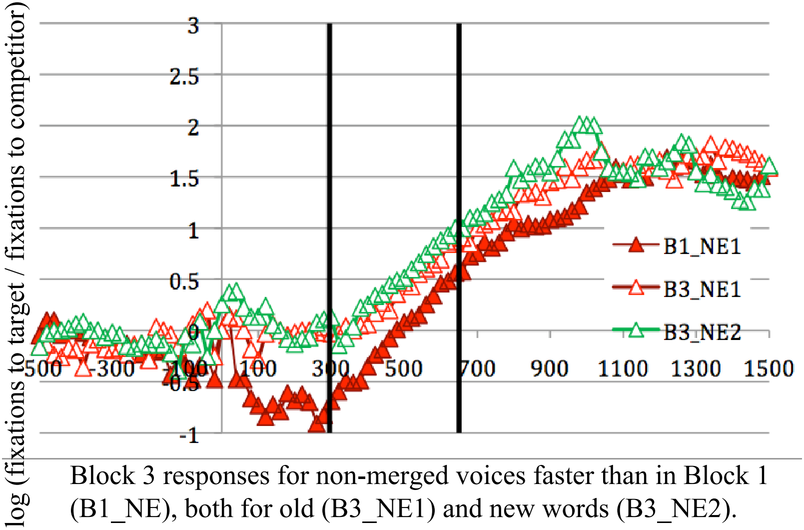
or hears M voice

Block 3 results : Interaction between race and dress



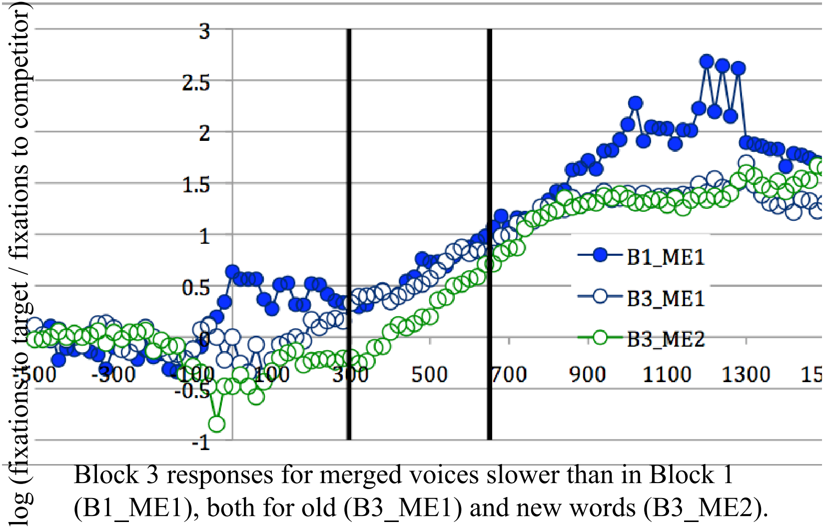
Block 3 results: Significant interaction between Race*Dress (t=-1.63, p<.01)

Voice familiarity an advantage for non-merged voices



Block 3 responses for non-merged voices faster than in Block 1 (B1_NE), both for old (B3_NE1) and new words (B3_NE2).

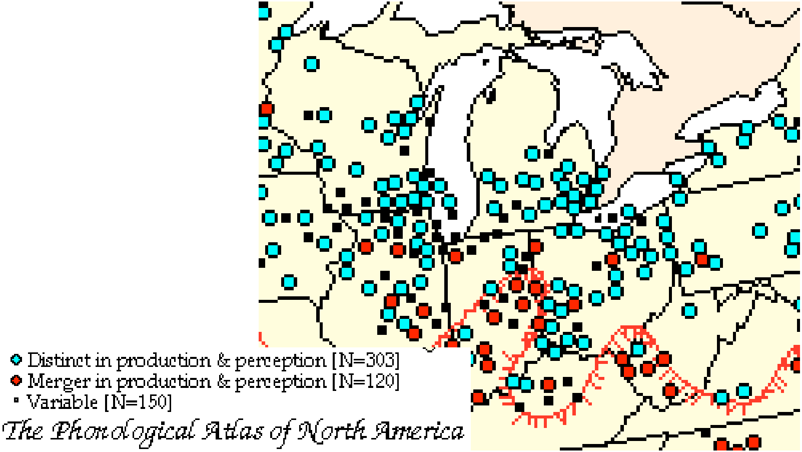
Voice adaptation induces ambiguity for merged voices



Block 3 responses for merged voices slower than in Block 1 (B1_ME1), both for old (B3_ME1) and new words (B3_ME2).

The *pin-pen* merger in Ohio

Ohio is in a border region with much variability. Are we characterizing this variabilty correctly ?



Measuring merger for the stimulus voices

For selection of auditory stimuli, we used VAS task with question "Which of the two words is this syllable part of ?" (see tutorial 2)

- For non-merged voices, ratings clustered around *pen*-word endpoint for *pen* words and around *pin*-word endpoint endpoint for *pin*-words,
- For merged voices, by contrast, more ratings near *pen*-word endpoint for both words, as well as more intermediate ratings

We also measured F1 and F2 at mid point of vowel

- For non-merger voices, F1 values clearly separated
- For merger voices, F1 values completed overlapped

Issues not addressed in Ito & Campbell-Kibler (2012)

What if there are varying degrees of merger ?

- We had to record more than four speakers to find two who clearly merged and two who clearly did not merge.
- There was not just variability in "code-switching" between variants, but also continuous variation in degree of merger
- We also noticed variation in the direction of the merger, with some raising *pen* words to ɪ (as in Brown, 1990) and others lowering pin to ɛ (in our two merged voices)

This raises the following questions about the listeners

- Are the listeners who participated in the eye-tracking experiments people who merge or do not merge ?
- How are the participants' pronunciations patterns linked to their processing of the four voices in the eye-tracking study ?

Productions elicited

Each listener produced two tokens of all words in training on picture names, as in this sample elicitation slide

6 target word pairs

- bin, bench
- fins, fence
- mint, men
- dinner plate, dentist sign,
- pins, pencil
- tin-can phone, tent stake

several words with target vowels before stops or fricatives

- ɪ : lipstick, scissors
- ɛ : drum set, bunk bed

How we got the formant values we are evaluating

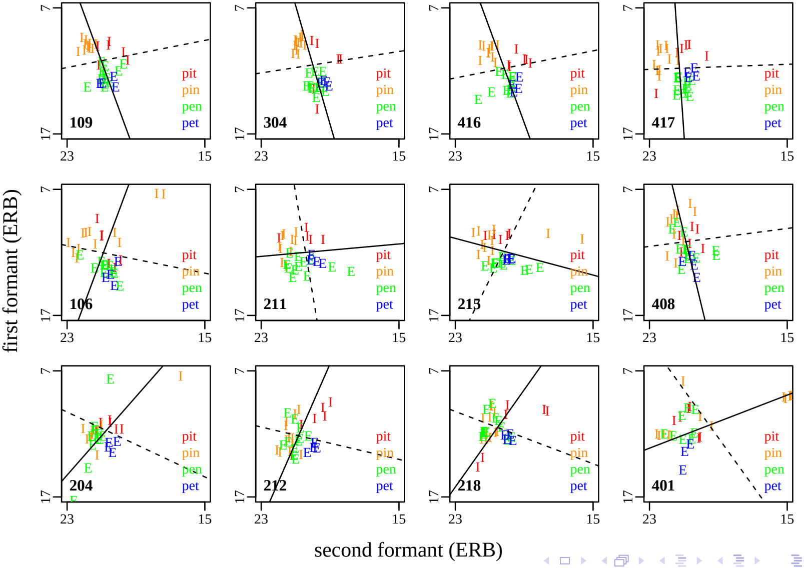
- Segment edges marked from word list using the Penn Phonetics Lab Forced Aligner `http://www.ling.upenn.edu/phonetics/p2fa/`
- Formant values extracted from each ɪ or ɛ token at time point where intensity for the vowel reached its local peak
- This differs to measurement point for stimuli
- Also, we are considering only one point for now, rather than using several measurement points to assess degree and direction of "glide" (cf. Scanlon & Wassink, 2008)
- Choice motivated by idea that peak intensity will reflect "nucleus" if there is any gliding
- (Also because we don't have the manpower to correct aligner errors)

First steps toward a measure of degree of merger

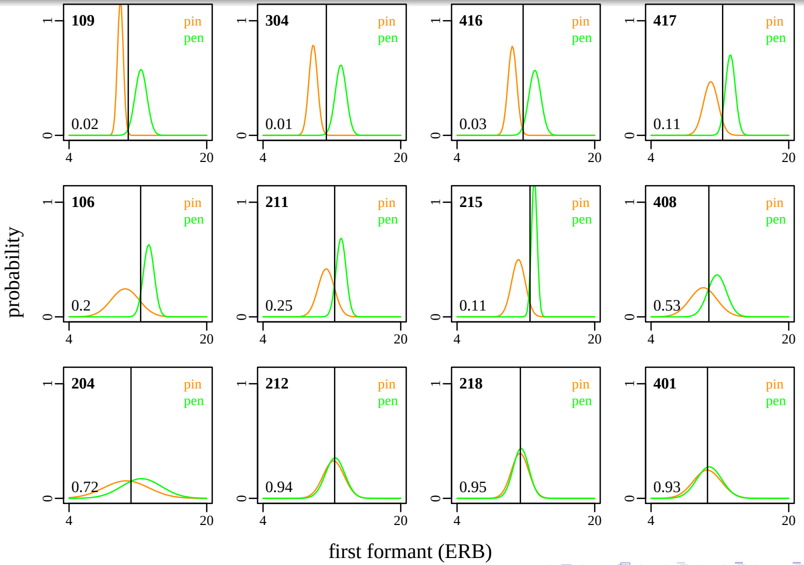
Probability of being on the wrong side of a criterion line

- Considered measures such as Pillai score in MANOVA (e.g., Hay et al., 2006) that evaluate distances between mean values
- Aiming instead for more direct measure of degree of overlap, inspired by sensitivity measures in signal detection theory
- Started by looking for dimension that well separates /ɪ/ from /ɛ/ in non-merger environment for each participant
- Tried principal components analysis, but this did not separate as well as F1 for many participants
- Fit Gaussians to distribution of vowels in merger environment
- Defined a criterion line at intersection of the two Gaussians
- Summed the areas on the “wrong” side of the criterion line and divided by total area under the 2 curves

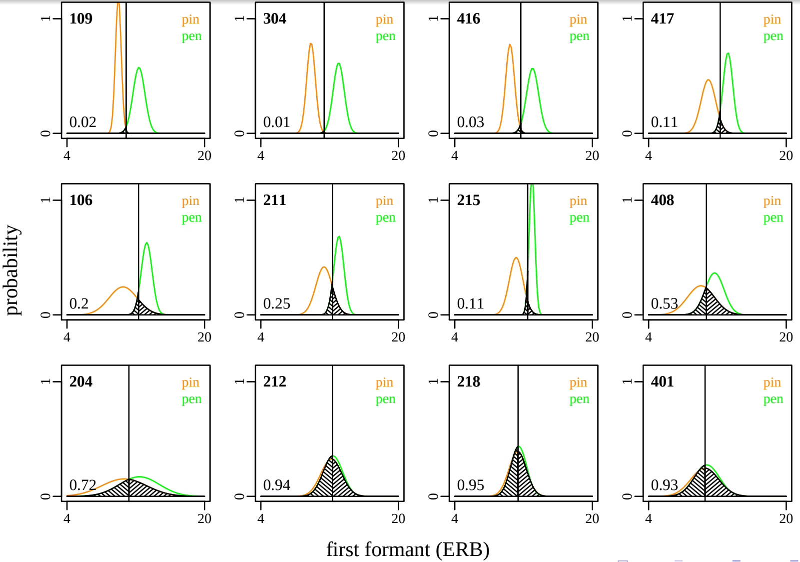
Tried principal components analysis



Gaussians fit to distribution of F1 in merger context



Areas under curve on “wrong” side of criterion line

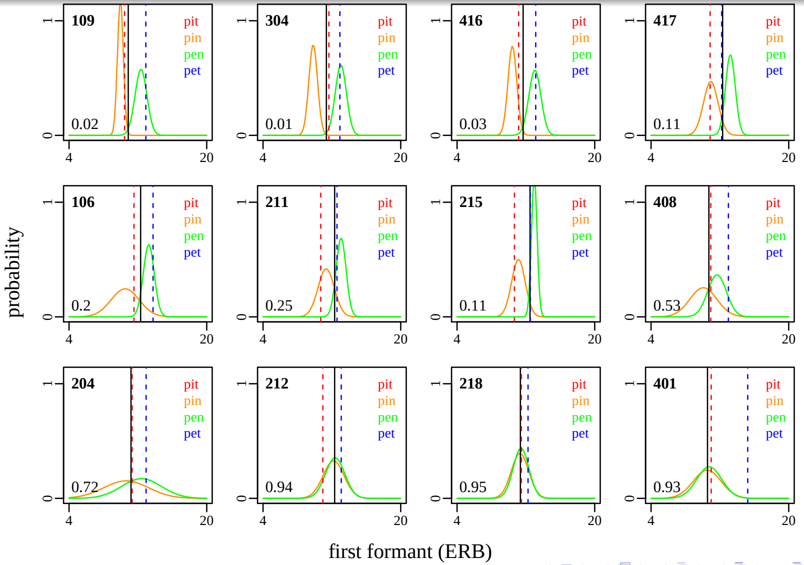


Are we on the right track ?

Problems / questions that we are grappling with include ...

- The first principal component sometimes (often ?) fails to capture a meaningfully differentiating dimension, but using F1 alone won't capture differentiation in F2
- While the measure captures the relative size of overlap for the vowels in the merger context, it does not capture either the degree of absolute dispersion for the vowel contrast or the direction of merger
- Could the (modes of the) distributions of vowel tokens in non-merger context be used as references for interpreting the absolute dispersion of the vowels for each speaker ?
- Could direction of merger be gauged by the log ratio of the distance between the criterion line and the /ɪ/ mode relative to the distance between the criterion line and the /ɛ/ mode ?

Distributions compared to modes in non-merger context



Non-VAS measures of perceived degree of merger

Multiple experimenter judgements

Koops, Gentry & Pantos (2008) by have participants read

- a short passage with embedded *pin* and *pen* words
- a word list with *pin* and *pen* words and an equal number of fillers
- a series of minimal pairs such as *tin-ten* and *pin-pen*

Each independently judges each target as merged or not, to get three mergedness scores

Listener's self-perceived degree of merger

Participants also say for each mimimal pair whether they would pronounce the words "the same", "close", or "different"
Self-perceived score = N "close" + (2 * N "same")

Effect only of self-perceived degree of merger

Degree of self-perceived merger predicted amount of time looking at competitor (e.g., *RINSE* while listening to *rent*)

How can we generalize from this result?

Difference in measure

- Koops et al. (2008) measured degree of merger by counting pair-by-pair judgments by the participants
- We propose to use a formants-based measure

Differnece in stimuli

- Koops et al. (2008) used only one speaker, whose productions they themselves judged to be **not** merged
- Ito & Campbell-Kibler (2012) used four speakers, two of whose *pin*-word productions were often judged to be /ɛ/-like on the VAS

Given these differences, what can we predict?

Tentative predictions for ...

Block 2 : *pin* words only

- Non-merged participants will look momentarily at competitors when listening to the merged speakers' [ɛ] but not when listening to non-merged speakers
- Merged participants will not show such an effect, since both the merged speakers' [ɛ] and the non-merged speaker' [ɪ] should activate both *pin*- and *pen*-words

Block 3 : *pen* words only again

- Non-merged participants will respond to non-merged voices faster than in Block 1 but to merged voices slower than in Block 1.
- Merged participants will not show this Block * Speaker interaction

Are we on the right track ? (again)

Problems / questions that we are grappling with include ...

- The production data analysis suggests continuous variability in degree of overlap ; we cannot categorize participants into Merged vs. Non-Merged groups
- What is an appropriate measure of ease of lexical activation, which could be regressed against the degree of vowel overlap in the participants ? Our current measure is the log ratio of looks to target relative to looks to competitor. Is there a better measure that takes time into account more directly ?
- Also, what kind of generalized linear model can we build in order to asses degree of adaptation ? For example, if we use a mixed effects model, with participant as a random effect, should we include individual-level slopes for Block ?

Are we on the right track ? (cont.)

Other problems / questions that we are grappling with include ...

- Since the target words formed 6 cohort pairs, we could calculate an item-specific measure of vowel merger, but this measure may not be very robust since there were only 2 repetitions
- Given this, is it worth testing whether the effect of merger is word specific ?
- How can we include other information about the participants, such as gender and age and residence history (as reported on the questionnaires that they also filled in) ?
- How can we explore interactions with visually invoked stereotypes of the "speaker" from the photo associated with the voice ?