Modeling talker intelligibility variation in a dialect-controlled corpus

Daniel McCloy • Richard Wright • August McGrath

Linguistic Phonetics Laboratory
Department of Linguistics
University of Washington

4pSC11

Research supported by NIH grant #R01DC006014. Special thanks to Pamela Souza and the Northwestern University Hearing Aid Lab

Background

- Correct identification of vowel phoneme is more difficult the more formant values differ from the regionally appropriate values.[1] Does this effect scale up to sentential stimuli?
- In investigating that question with a cross-dialect study of intelligibility, we found dramatic talker intelligibility differences even for within-dialect listeners.
- To better understand this finding, we modeled the mean intelligibility of each talker against several acoustic measures of their speech.

Methods

Dialect controls (both talkers + listeners):

- Northern Cities (NC) + Pacific Northwest (PNW); lived in region age 5-18; max. 5 years outside region

Corpus

- 3600 RMS-normalized stimuli: 180 sentences × 20 talkers (5 male + 5 female per dialect)
- Subset of IEEE "Harvard" sentences[2] chosen for absence of alliteration, rhyme, or focus/contrast
- Coaching and feedback to ensure consistent, normal declarative prosody; best of 3 readings selected per talker (free of mic overloading, hesitation, etc)

Perception task (15 PNW listeners; 13 NC listeners):

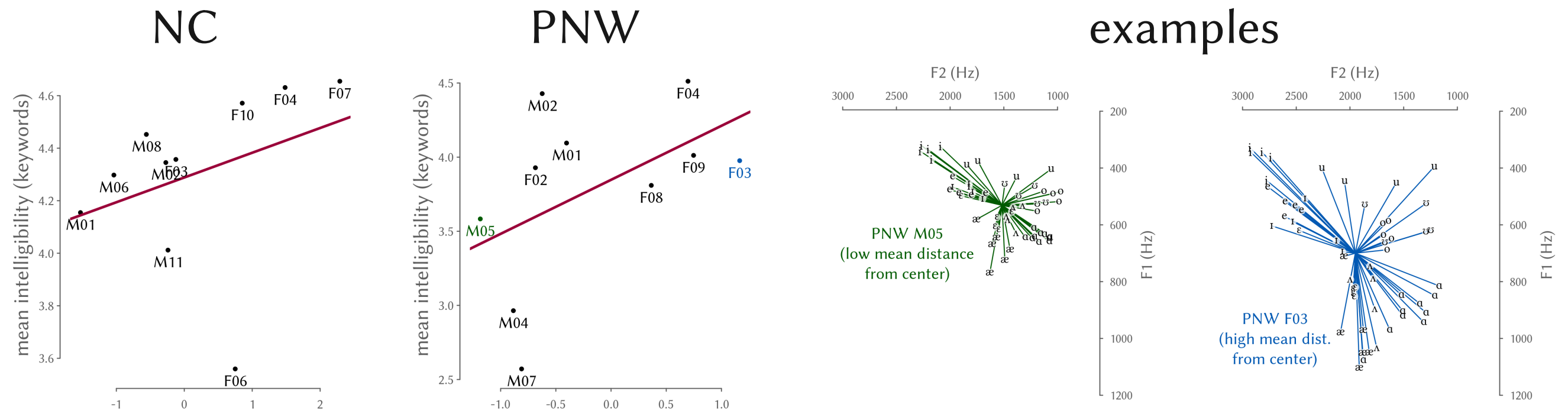
- Unique talker/sentence/SNR randomizations for each listener: 180 unique sentences ÷ 20 talkers ÷ 3 SNRs = 3 sentences per talker-SNR pairing for each listener
- "Repeat what you hear" paradigm scored 0-5 on keywords; converted to binary score (1 = all keywords correct) for statistical models
- Data shown for +2dB SNR only (ceiling effects at higher SNRs)

Acoustics

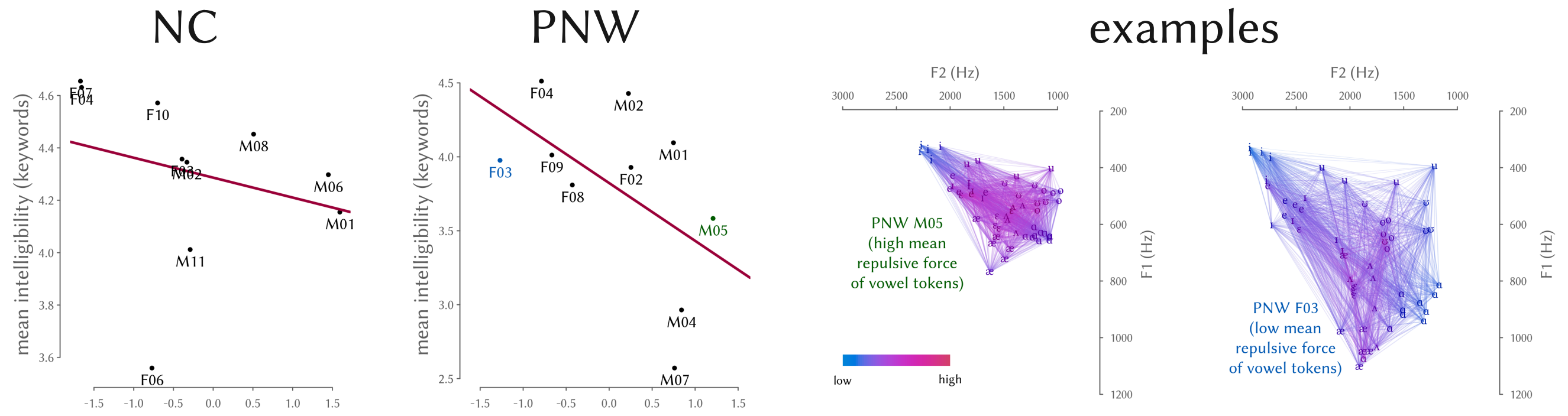
- **Vowels:** 1100 vowel tokens hand-measured (11 vowel phonemes × 5 tokens/vowel × 20 talkers)
- **Pitch:** 300 stimuli (15 sentences × 20 talkers, hand-corrected)
- **Intensity:** all 3600 stimuli (auto-extracted by Praat)[3]

Correlations by talker

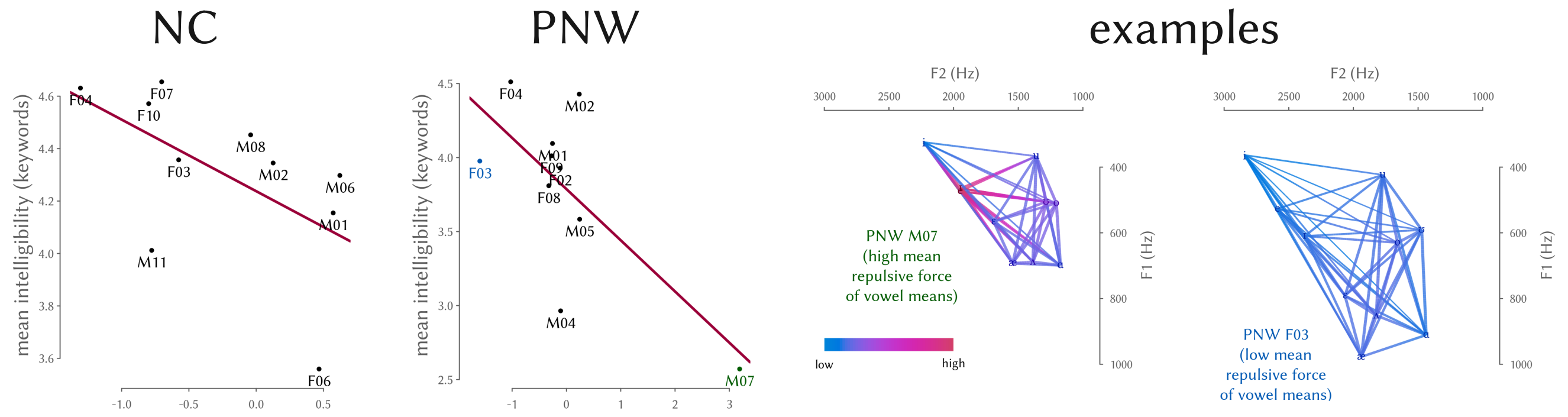
Mean Euclidean distance from center of vowel space



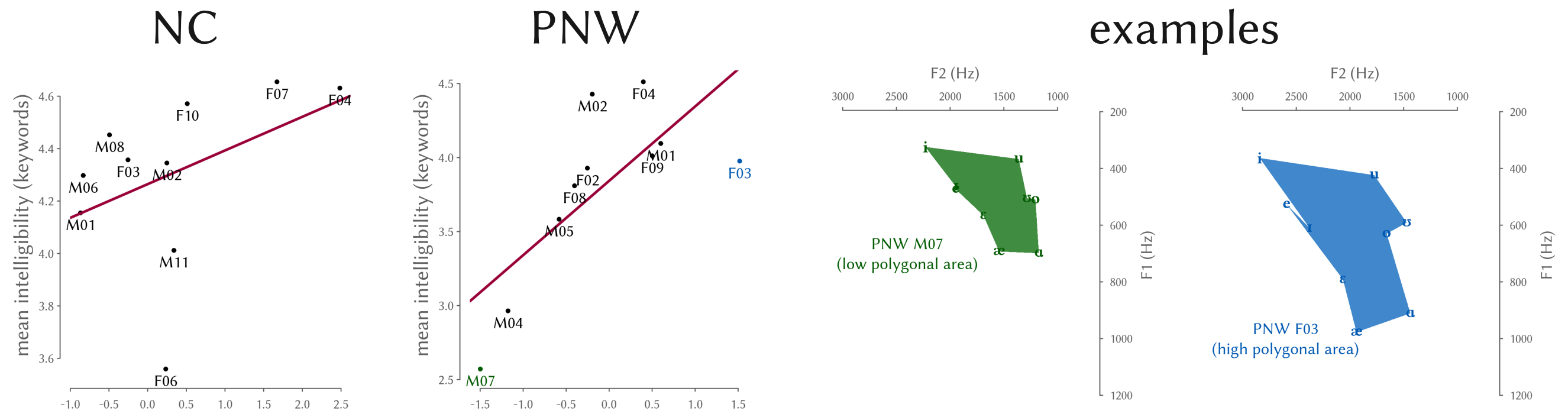
Repulsive force[4,5] of tokens: $\sum(\mathrm{dist}(vowel_i,vowel_j))^{-1}$, $i{\neq}j$



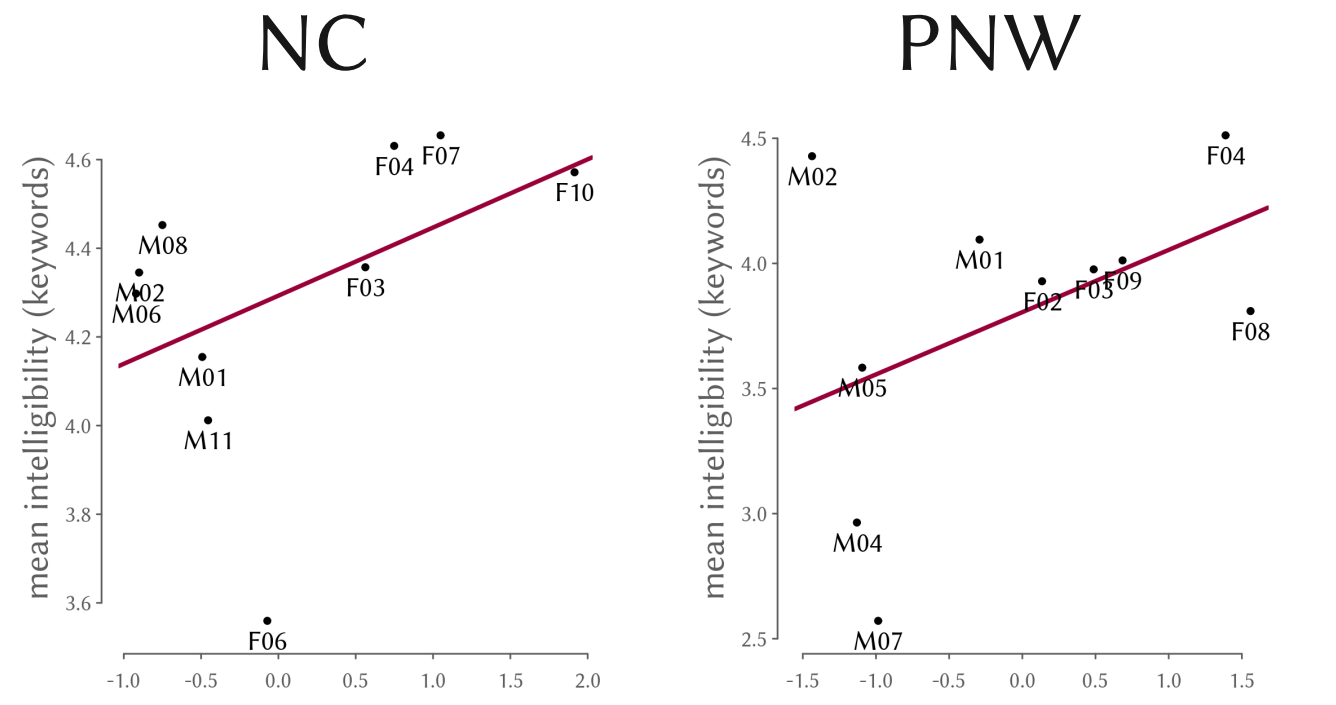
Repulsive force of means: $\sum(\mathrm{dist}(\overline{vowel_i},\overline{vowel_j}))^{-1}$, $i{\neq}j$



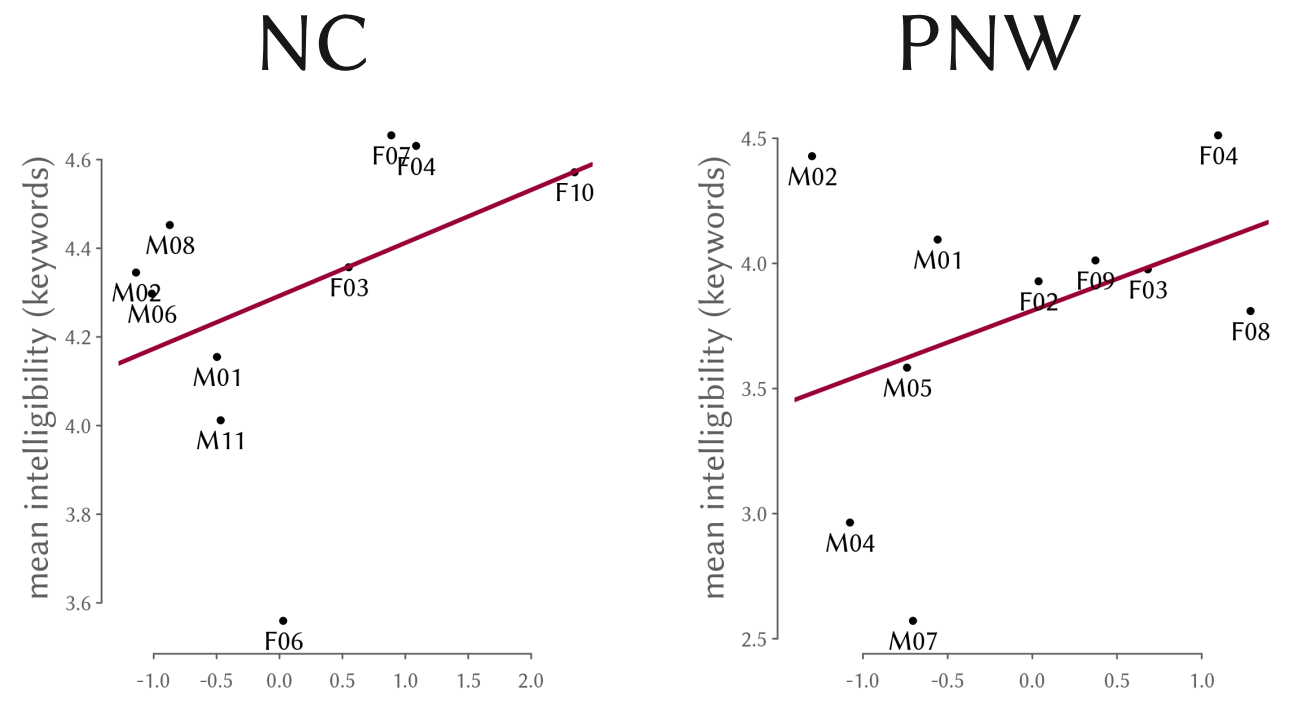
Area of polygon described by vowel means (excluding /ʌ/)



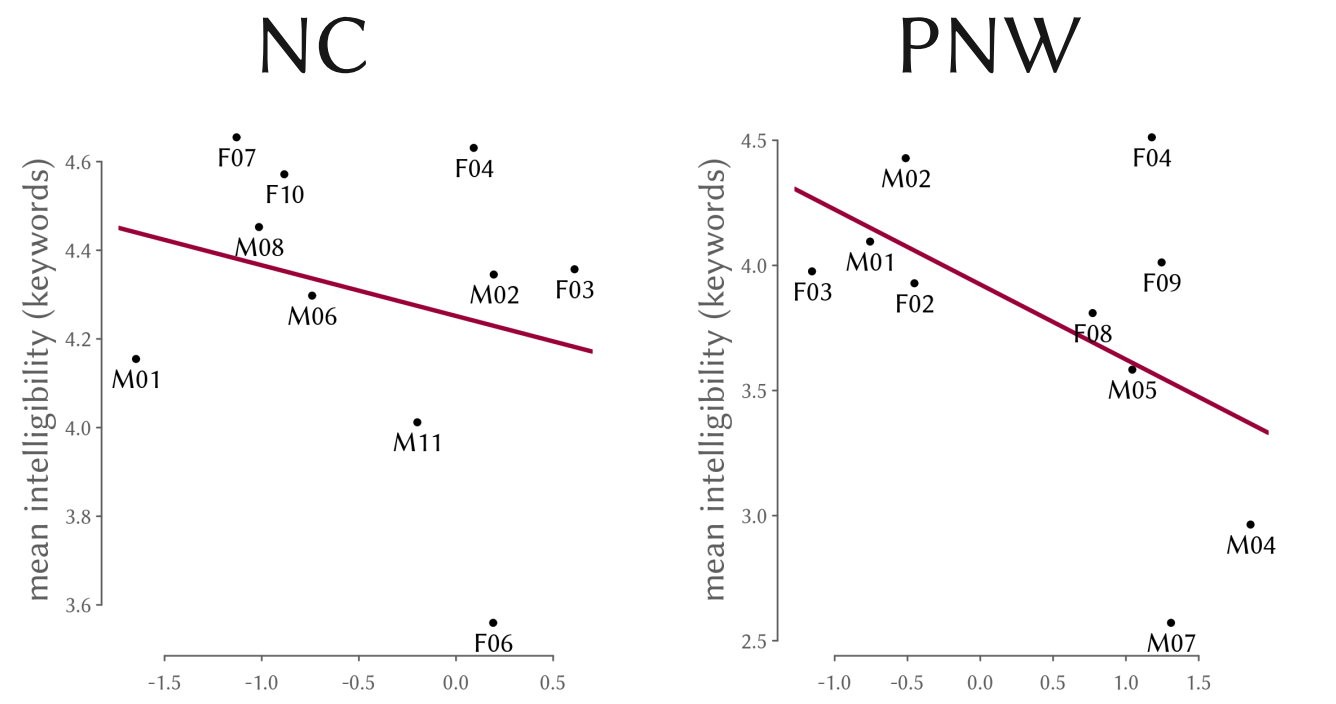
Pitch range



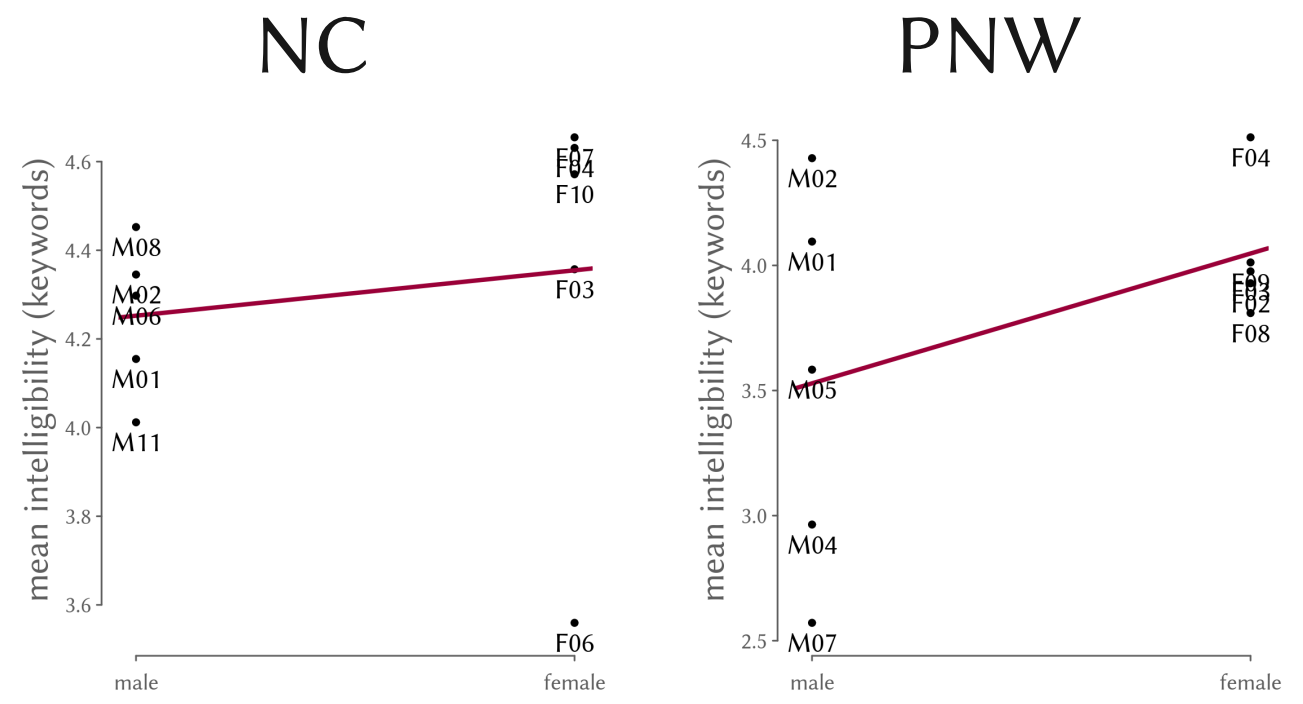
Mean absolute pitch velocity



Mean intensity velocity



Gender



Model construction

- **Linear mixed-effects logistic regression** fit in R[6] using `glmer()`[7]; separate model for each dialect region; all predictors normalized
  - PNW vowel-space predictors (avgDistFromCenter, polygonalArea, repulsiveForceTokens, repulsiveForceMeans) calculated with low-back merger (/a/ and /ɔ/ collapsed to /ɑ/)
  - Full model specification:

```
intel ~ speechRate + avgDistFromCenter + polygonalArea +
repulsiveForceTokens + repulsiveForceMeans + pitchRange +
avgAbsPitchVelocity + avgIntensityVelocity + talkerGender +
(1|talker) + (1|listener) + (1|sent)
```

  - Poor predictors eliminated via likelihood ratio tests, yielding different models for PNW and NC:
    - Mean distance from center of vowel space, repulsive force of vowel tokens, and talker gender significant for both regions
    - Polygonal area, pitch range, pitch velocity, intensity velocity also significant in PNW
    - Repulsive force of vowel means also significant in NC

Summary of fixed effects	*PNW talkers & listeners (N=1350, log-likelihood -750.7)*				*NC talkers & listeners (N=1170, log-likelihood -548.7)*			
	Estimate	**SE**	***z***	***p***	**Estimate**	**SE**	***z***	***p***
intercept	-0.0226	(0.332)	-0.07	> 0.9	1.2060	(0.195)	6.19	$< 10^{-9}$
avgDistFromCenter	-2.5939	(0.733)	-3.54	$< 10^{-3}$	1.2301	(0.437)	2.81	$< 10^{-2}$
repulsiveForceTokens	-2.5047	(0.663)	-3.78	$< 10^{-3}$	1.1521	(0.501)	2.30	< 0.05
repulsiveForceMeans	—	—	—	—	-0.5756	(0.213)	-2.71	$< 10^{-2}$
polygonalArea	1.1706	(0.242)	4.84	$< 10^{-5}$	—	—	—	—
pitchRange	1.8398	(0.380)	4.84	$< 10^{-5}$	0.2086	(0.141)	1.48	$= 0.14$
avgAbsPitchVelocity	-1.4216	(0.423)	-3.36	$< 10^{-3}$	—	—	—	—
avgIntensityVelocity	0.3126	(0.136)	2.30	< 0.05	0.4710	(0.265)	1.78	$= 0.08$
talkerGender	1.7798	(0.544)	3.27	$< 10^{-2}$	1.2567	(0.353)	3.56	$< 10^{-3}$

Discussion

- **Regional differences in models** suggests sample is still too small even with ten talkers / region
- **Lack of significant speech rate finding** disagrees with Sommers et al (1994)[8] and agrees with Bradlow et al (1996).[9] Suggests that the intelligibility cost of fast speech may not be due to speech rate *per se*, but corollary effects (e.g., reduction)
- **Difference in significance pattern of vowel space predictors** possibly due to low back merger in PNW
  - **Polygonal area** disagrees with Bradlow et al (1996),[9] but their polygon based on /i o a/ (ours: /i ɪ e ɛ æ a ɔ o ʊ u/)
  - **Repulsive force** possibly related to Neel (2008)[10] although that study examined vowel identification confusions, not sentential stimuli
- Relation of **prosodic predictors** to intelligibility still unclear; **intensity velocity** may reflect word-by-word SNR differences arising from different phrasal stress habits of talkers

Significance

- **Cross-dialect studies** of intelligibility or speech perception should expect substantial within-group variability and model appropriately; small numbers of talkers may bias results
- **Prosodic patterns** are an important and often overlooked consideration with sentential stimuli: dynamic aspects of intensity and pitch may affect intelligibility

Future directions

- Relation between intensity contour and word-by-word changes in SNR across the duration of a sentence
- Deeper investigation of cross-dialect differences in pitch patterns (esp. creaky voicing) and relation to intelligibility
- Role of acoustic predictors (esp. prosodic ones) in the perceptual benefit of talker familiarity

References

[1] Wright, R., & Souza, P. (2012). Comparing identification of standardized and regionally valid vowels. Journal of speech, language, and hearing research: JSLHR, 55(1), 182–193. doi:10.1044/1092-4388(2011/10-0278)

[2] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., et al. (1969). IEEE recommended practice for speech quality measurements. IEEE Transactions on Audio and Electroacoustics, 17, 225–246. doi:10.1109/TAU.1969.1162058

[3] Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer. http://www.praat.org/

[4] Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. Language, 48(4), 839–862.

[5] Wright, R. A. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), Phonetic interpretation, Papers in Laboratory Phonology (pp. 75–87). Cambridge, UK: Cambridge University Press.

[6] R Development Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

[7] Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. http://cran.r-project.org/package=lme4

[8] Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. The Journal of the Acoustical Society of America, 96(3), 1314–1324.

[9] Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication, 20(3-4), 255–272. doi:10.1016/S0167-6393(96)00063-5

[10] Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. Journal of Speech, Language, and Hearing Research, 51(3), 574–585. doi:10.1044/1092-4388(2008/041)