1pSCb28

Gender, the individual, and intelligibility

Daniel McCloy
U. Washington I-LABS

Laura Panfili
Amazon Seattle

Cornelia John
U. Washington Speech & Hearing Sciences

Matthew Winn
U. Minnesota Speech-Language-Hearing Sciences

Richard Wright
U. Washington Linguistics

INTRODUCTION

There are contradictory claims about gender effects in speech intelligibility

- **Clinic:** men's speech is more intelligible than women's
  - Caveat: anecdotal
- **Behavioral research:** women more intelligible than men (e.g., Bradlow et al., 1996)
  - Caveat: gender effects not tested directly – incidentally observed
- **Received wisdom:**
  Men > Women "because women have higher frequencies in their speech"
  - Explanation not backed up by acoustics

  Women > Men: "because women have more precise articulations"
  - Asserted from few observations, not reliable when statistically controlled (McCloy et al., 2015)

PURPOSE OF THIS STUDY

To test whether there are general trends of gender affecting talker intelligibility...
or whether observed effects are particular to individual talkers

METHODS

Stimuli

- 198 IEEE ("Harvard") sentences; 18 for familiarization, 180 for testing

Talkers

- Native English; from WA, OR, and ID
- 15 women, 15 men
- Natural, relaxed speaking style; picked from multiple recordings of each sentence

Masking Noise

- Corpus-shaped noise (steady)
- Four SNRs: -4, -2, 0, and 2 dB

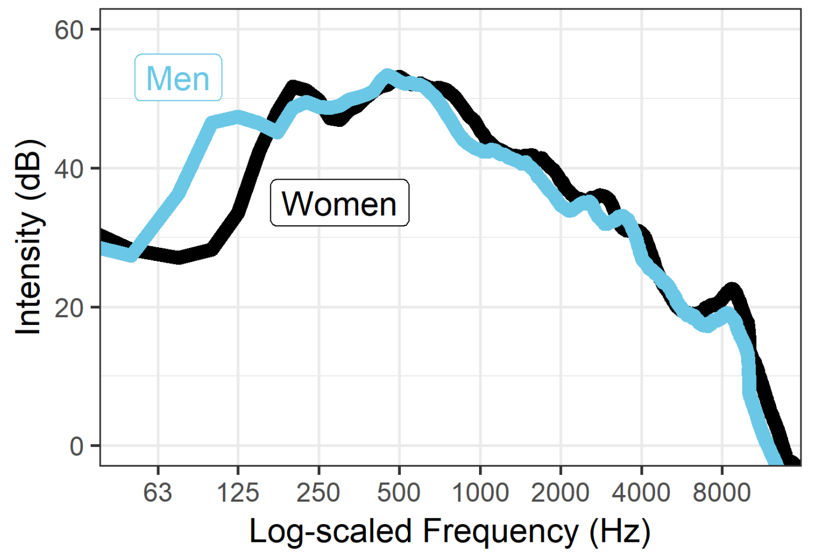

Figure 1. Long Term Average Speech Spectrum dB SPL by log frequency (Hz) of men's (blue) and women's (black) speech. Only notable difference (below 500 Hz) is commonly observed (e.g., Byrne et al., 1994)

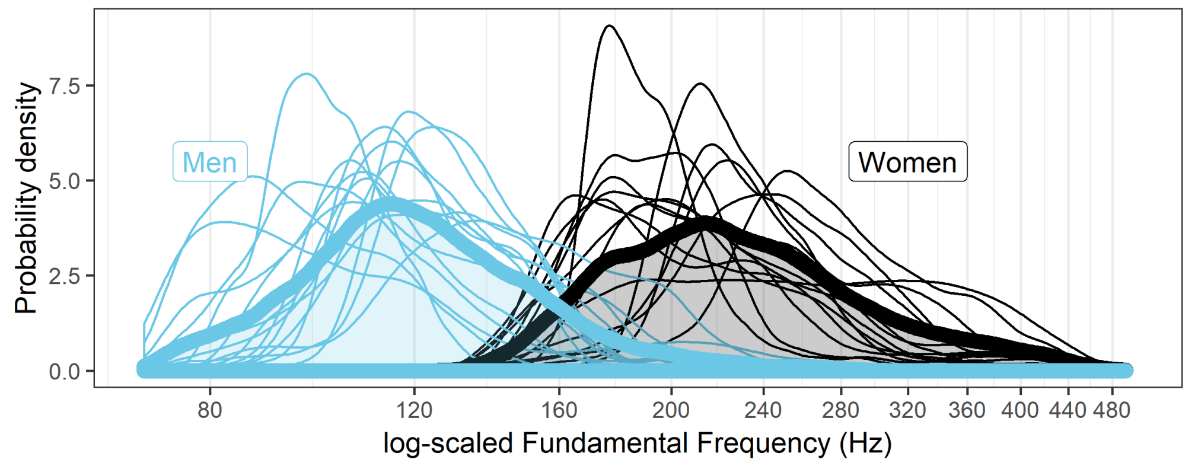

Figure 2. Density plot of f0 of men (blue) and women (black). Thin line f0 of all timepoints in all sentences for individuals. The thick line is the group average.

Listeners

- 32 native English-speaking adults (13 men and 19 women)
- Normal hearing (thresholds ≤20 dB HL at octave frequencies 125 – 8000 Hz)

Perception Task

- Listeners in a sound-attenuated booth in UW Linguistic Phonetics Lab
- Circumaural headphones
- 18 training sentences; 180 test sentences (6 blocks of 30)
- Listeners instructed to repeat each sentence to the best of their ability; if unsure, instructed to give partial answers or guess.

Scoring

- Each keyword in each sentence scored 0 (incorrect) or 1 (correct)
- Two researchers each scored 60% of all sentences (20% overlap).
- Cohen's kappa was calculated at 0.984 showing good agreement.
- A third researcher resolved any scoring disagreements.

RESULTS

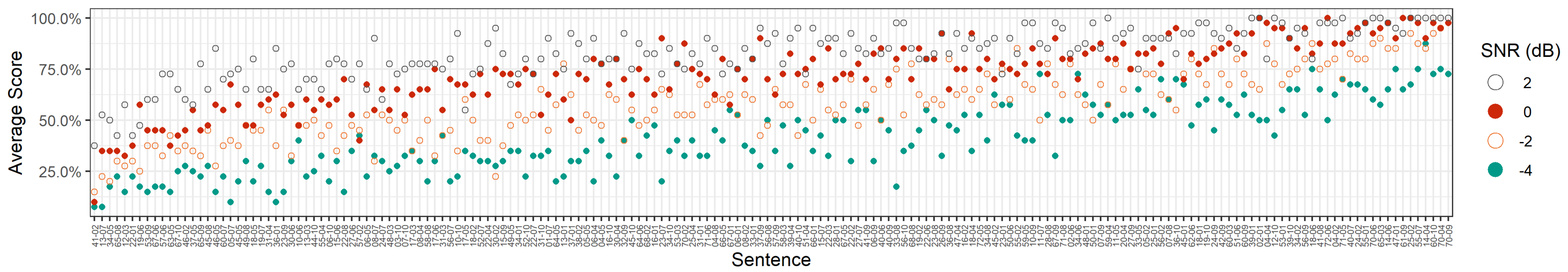

Figure 3. Scatterplot of mean intelligibility score for each sentence, broken down by SNR, ordered by average score.
Variability of sentence intelligibility. Consistent effect of SNR across sentences. Consistent effect of sentence on intelligibility. No interaction.

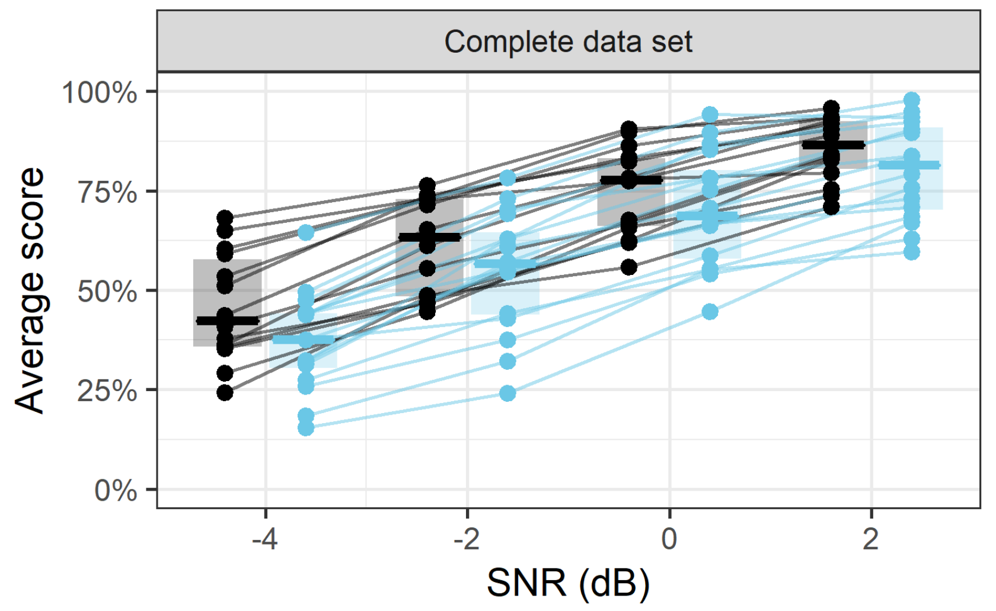

Fig. 4: Intelligibility by SNR, talker, and gender
Each dot indicates the mean across all trials.
Boxes show median and quartiles.

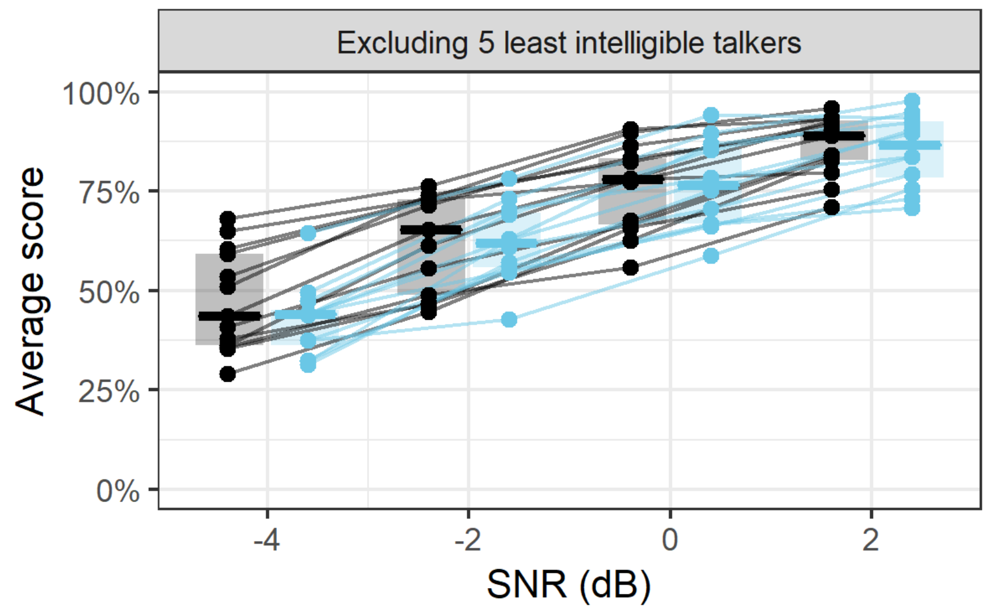

Fig. 5: Same as Fig. 4, but excluding the 5 least intelligible talkers

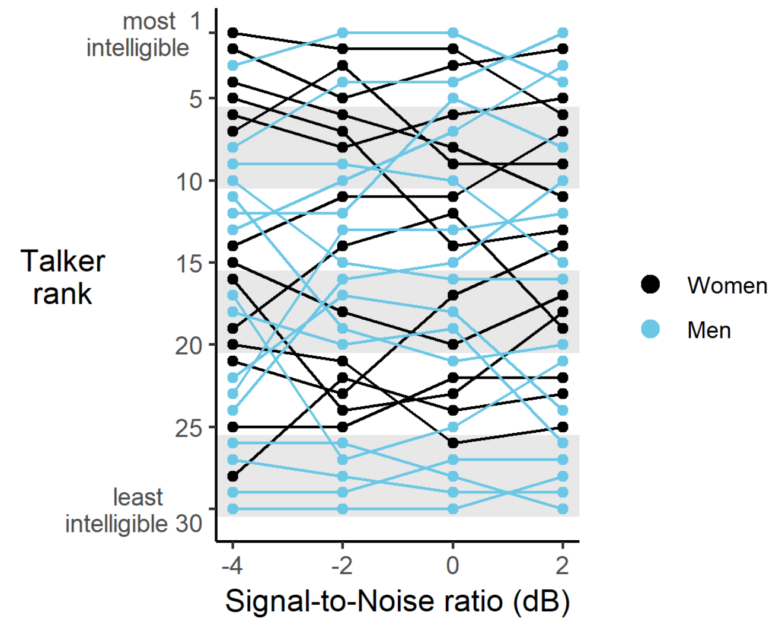

Fig. 6: Ranking of talker intelligibility
Bottom 5 are mostly males;
top 20 are mixed

- All talkers included, mixed model with random intercepts for sentence & listener → *apparent gender effect (women > men)*
- Add a random effect for talker → *gender effect goes away (cf. Fig. 9)*

Suggests that the difference is driven by individual variability (i.e., a few unintelligible men) rather than a true group difference

How Sample Size Affects the Gender-Intelligibility Finding

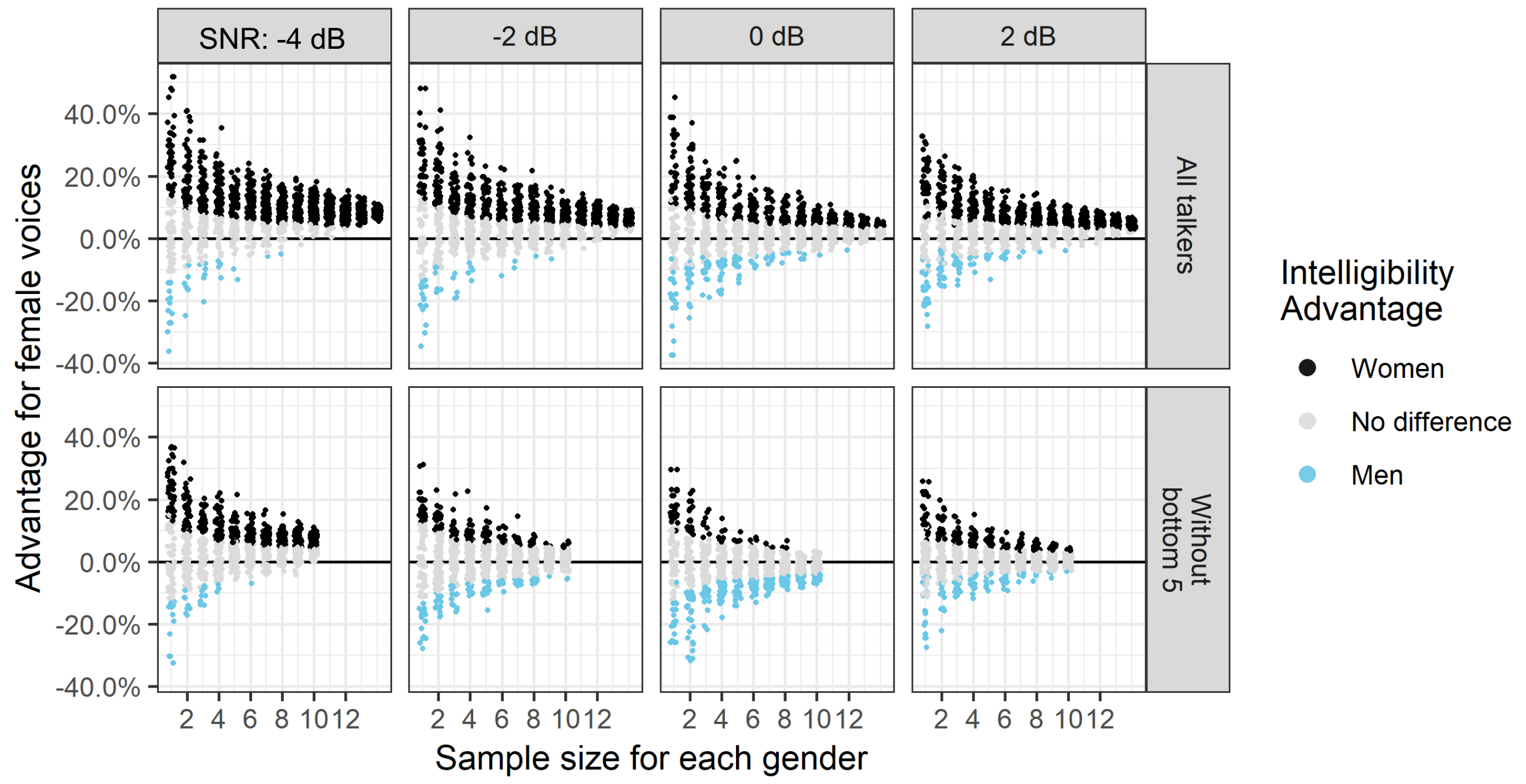

Fig. 7: Gender-related intelligibility disparity for sample sizes from 1 to 14. Each point is 1 of 100 comparisons, where equal numbers of women and men are randomly sampled from the available pool. "No difference" indicates means comparison that did not reach significance criterion of 0.05. The full data set converges on a small but significant advantage for women. The filtered data set (excluding bottom 5 talkers) converges on a null result.

Simulating a Completely Random Effect of Gender

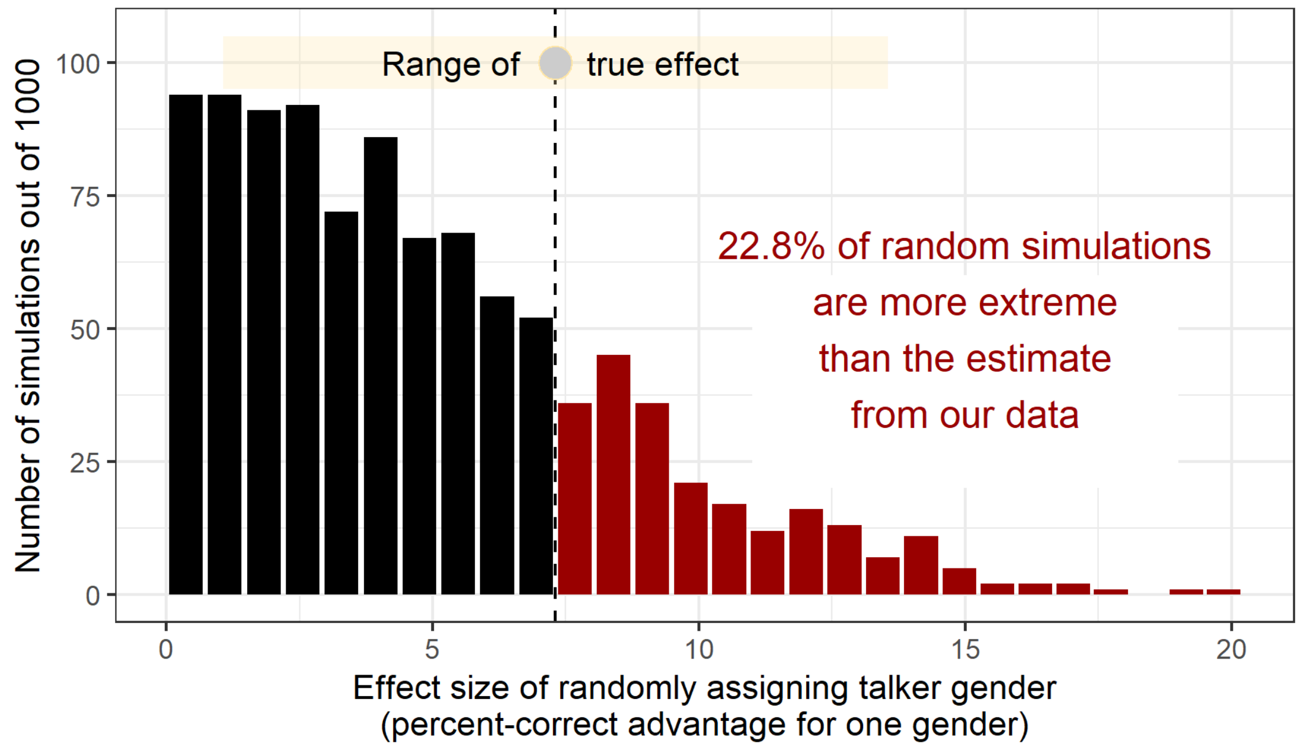

Fig. 8: Estimation of group effect when "gender" is randomly assigned to talkers (thus estimating a completely random effect). On 228 of 1000 simulations (23%), the estimated phantom effect exceeds the true effect of talker gender from the real data. This suggests that the "true effect" is spurious.

MODEL DETAILS

- Model fitting with R AFEX library (Singmann et al 2018)
- Keyword-level mixed models with/without talker random effect:
  `word_correct ~ SNR * talker_gender + (1|sentence) + (1|listener) + (1|talker)`
- Model including talker RE preferred (likelihood ratio test, $\chi^2(1)=1892.2$, $p<10^{-15}$)

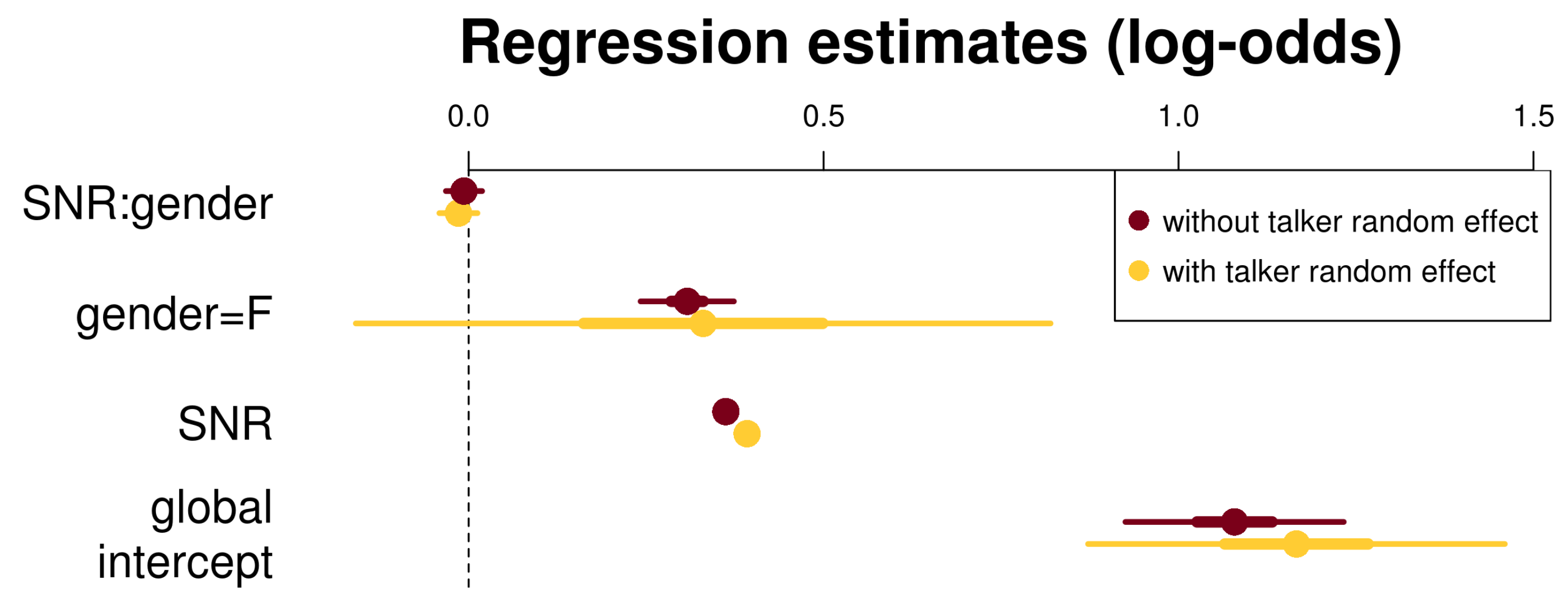

Fig. 9: Model estimates with and without a random effect for talker
Estimates for the effect of gender are quite similar in magnitude, but differ dramatically in the standard error of the estimates.

DISCUSSION

- **Greater variability within gender than across gender.** In our sample:
  - 4 of the 5 least intelligible happened to be men
  - Aside from those, any one man might be more or less intelligible than any one woman
- **Statistical modeling choices matter a lot**
  - The *size of the estimated intelligibility difference between genders* is roughly consistent across models, but…
  - …the *confidence that it is non-zero* varies a lot across models
- **Past findings probably artefactual**
  - Reports that women are more intelligible than men probably a result of small sample sizes

FUTURE WORK

- Stimuli recorded in quiet; gender effect in Lombard speech is unknown.
- Stimuli were read, not spontaneous; gender effect may emerge under more natural communicative conditions.
- Research is underway to record, measure, and test communicative tasks in noise using collaborative tasks in quiet and with noise presented over headphones.

REFERENCES

- Bradlow, Torretta & Pisoni (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
- Byrne (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America* 96(4), 2108–2120. doi:10.1121/1.410152
- McCloy, Wright & Souza (2015). Talker versus dialect effects on speech intelligibility: A symmetrical study. *Language and Speech* 58(3), 371–386. doi:10.1177/0023830914559234
- Singmann, Bolker, Westfall & Aust (2018). afex: Analysis of Factorial Experiments. R package version 0.22-1. https://CRAN.R-project.org/package=afex

ACKNOWLEDGMENTS

Special thanks to Courtney Mansfield, for dedicated work on this project.
This work was supported by NIH NIDCD R01 DC006014