An Ideal Observer Analysis of Variability in Visual-Only Speech

Brianna Conrey & Jason M. Gold *Indiana University, Bloomington*

1. Introduction.

Normal-hearing observers are typically able to understand speech to some degree when it is presented in the visual-only modality, without an accompanying auditory signal[1]. However, different talkers vary in how easily they can be understood through visual-only speech perception[2]. It has previously been unclear whether this variability in talker intelligibility is due to differences in the amount of physical information available in the visual speech signal or to human perceptual strategies that are more optimally suited to some talkers than others[3-4]. Here, we investigate this issue by comparing human and model observer performance across different talkers in a visual-only word identification task.

2. Human Data.

Observers

- 8 Indiana University students (4 female, 4 male, ages 23-40)
- 7 were naïve to the purpose of the experiment

Materials

- Video recordings of 8 talkers (4 female, 4 male)
- Stimulus set of 8 familiar monosyllabic English words
- Words were the same for each talker
- All stimulus videos were the same length (1.5 seconds)
- An oval-shaped frame around each talker's face prevented the use of cues from features outside the face
  - The frame was the same constant size for each talker
  - The background was also constant and the same for each talker

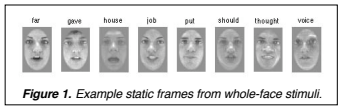

Figure 1. *Example static frames from whole-face stimuli.*

Procedure

- One-of-eight word identification task
- Sessions were blocked by talker
- Talker order was randomized across observers
- Videos of words were presented without sound in dynamic Gaussian white pixel noise ($\sigma = 0.1$)
- Word contrast was varied with an adaptive staircase procedure tracking the 50% correct contrast energy threshold for each talker (chance performance = 12.5%)

Results

- Figure 2 displays contrast energy thresholds for 4 representative observers
- Consistent with previous findings[5], across observers, there were:
  - Differences in overall levels of performance
  - Consistent patterns of variability across talkers

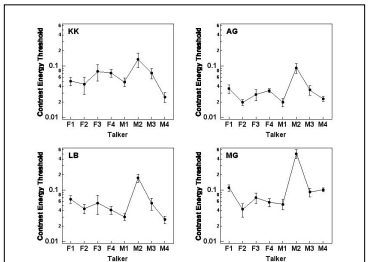

Figure 2. *Contrast energy thresholds for human observers. M = male talker, F = female talker. Female observers are on the left and male observers on the right.*

3. Ideal & Spatially Restricted Models.

- Several simulated observers were used to model possible sources of the variability across talkers seen in human performance
- The performance of these observers was measured using Monte Carlo simulations (1000 trials/talker)
- Decision rule: On each trial, the word template with the highest cross-correlation with the stimulus was chosen[6]

Ideal Observer Model

- The ideal observer used noise-free templates of the whole-face stimuli shown to the human observers (see examples in Figure 1)

Lower Half of Face

- Some previous research on visual-only speech perception has concentrated on the lower half of the face (lips, cheeks, and chin)[2,7]
- Information use for this model observer was restricted to the lower half of the face

Mouth Only

- Eye movement data suggest that humans tend to focus on the area around the mouth when presented with visual-only speech[4]
- Information use for this model observer was restricted to the mouth only

Figure 3. *Example templates for spatially restricted models.*

Results

- Figure 4 shows contrast energy thresholds for the humans on average and for the ideal and spatially restricted observer models
- The model-to-human performance ratios (shown in Figure 5) indicate that the mouth-only model provides the best description of human performance for most of the talkers. The flatter the pattern of performance ratios, the more successful the model is at describing the pattern of human performance.

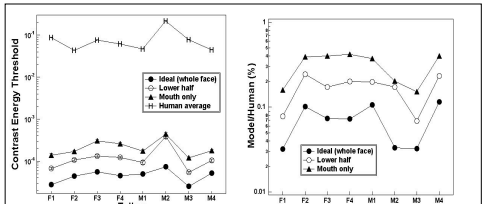

Figure 4. *Contrast energy thresholds.* ***Figure 5.*** *Model-to-human performance ratios.*

4. Models with Spatial & Temporal Uncertainty.

- To attempt to account for remaining differences in cross-talker variability between humans and the mouth-only model, Monte Carlo simulations (200 trials/talker) were conducted to measure the performance of models that used the mouth-only templates and small amounts of spatial uncertainty, temporal uncertainty, or both
- **Spatial Uncertainty (SU):** up to 1, 3, or 5 pixels
- **Temporal Uncertainty (TU):** up to 1, 3, or 5 frames
- **Spatial + Temporal Uncertainty (STU):** up to 1 pixel or frame

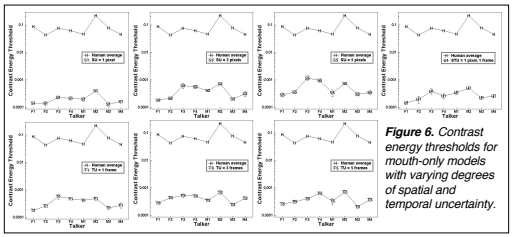

Figure 6. *Contrast energy thresholds for mouth-only models with varying degrees of spatial and temporal uncertainty.*

Results

- Figure 6 shows contrast energy thresholds for the mouth-only model with varying degrees of spatial and temporal uncertainty
- Of these models, the one with SU = 1 pixel provides the best qualitative fit to the pattern of human performance, but still does not improve over the mouth-only model with no uncertainty (see Figure 7)

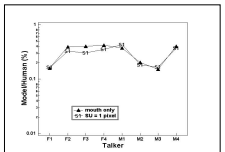

Figure 7. *Comparison of mouth-only models with and without SU = 1 pixel.*

5. Conclusions & Future Directions.

- The ideal observer model revealed some differences in the physical information available across talkers. However, a model restricted to using information from the mouth only produced a better description of cross-talker variability in human performance.
- The addition of spatial or temporal uncertainty to the mouth-only model did not improve its fit to the human pattern of performance. Instead, the addition of spatial or temporal uncertainty tended to make patterns of model and human performance less similar.
- Although these models do not account for all the variability across talkers, the results suggest that both physical information differences across talkers and perceptual strategies such as spatially restricting information use are involved in cross-talker variability in human performance.
- Future projects will include examining human and model observer performance on auditory-only and auditory-visual word identification tasks.

6. References & Acknowledgments.

[1]L.E. Bernstein, M.E. Demorest, & P.E. Tucker, *Percept Psychophys 62:* 233-252 (2000).
[2]P.B. Kricos & S.A. Lesner, *Volta Rev 84*(4): 219-225 (1985).
[3]L.E. Bernstein, J. Jiang, A. Alwan, & E.T. Auer, *Proceedings of the Auditory-Visual Speech Processing Workshop:* 104-109 (2001).
[4]C.R. Lansing & G.W. McConkie, *Percept Psychophys 65*(4): 536-552 (2003).
[5]M.E. Demorest & L.E. Bernstein, *J Speech Hear Res 35*(4): 876-891 (1992).
[6]B.S. Tjan, W.L. Braje, G.E. Legge, & D. Kersten, *Vision Res 35*(21): 3053-5069 (1995).
[7]Campbell, C.S., *Patterns of evidence: Investigating information in visible speech perception.* Unpublished doctoral dissertation, University of California, Santa Cruz.

This research was supported by an NSF Graduate Research Fellowship to the first author and NIH R03 research grant EY015787-01.