Published in final edited form as:

Lang Learn. 2018 June ; 68(Suppl 1): 127–158. doi:10.1111/lang.12281.

Response Errors in Females' and Males' Sentence Lipreading Necessitate Structurally Different Models for Predicting Lipreading Accuracy

Lynne E. Bernstein

Department of Speech, Language, and Hearing Science, George Washington University, 2121 I St NW, Washington, DC 20052

Abstract

Lipreaders recognize words with phonetically impoverished stimuli, an ability that is generally poor in normal-hearing adults. Individual sentence lipreading trials from 341 young adults were modeled to predict words and phonemes correct in terms of measures of phoneme response dissimilarity (PRD), number of inserted incorrect response phonemes, lipreader gender, and a measure of speech perception in noise. Interactions with lipreaders' gender necessitated structurally different models of males' and females' lipreading. Overall, female lipreaders are more accurate, their ability to recognize words with impoverished or degraded input is consistent across visual and auditory modalities, and they amplify their correct responding through top-down insertion of text. Males' responses suggest that individuals with poorer auditory speech perception in noise amplify their responses by shifting towards including text in their response that is more perceptually discrepant from the stimulus. Attention to gender differences merits attention in future studies that use visual speech stimuli.

The biomechanics of speech production generate optical as well as acoustic signals that are correlated with each other (Jiang, Alwan, Keating, Auer, & Bernstein, 2002; Munhall & Vatikiotis-Bateson, 1998; Yehia, Rubin, & Vatikiotis-Bateson, 1998). However, the face occludes some of the articulatory motions that are conveyed in the acoustic signal, causing visual speech stimuli to be reduced in phonetic information relative to clear auditory stimuli. Due to the extent of the reduction in phonetic information, accurate lipreading is a form of spoken language recognition that must engage top-down processing. Visual speech perception exerts significant and diverse effects on auditory speech perception. Among the prominent examples are audiovisual speech perception in noise and the McGurk effect. Estimates vary, but the functional increase in signal-to-noise ratio (SNR) contributed by seeing the talker while listening to acoustic speech in noise is on the order of ~12 dB (e.g., Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). The McGurk effect demonstrates that when auditory and visual stimulus phonemes are mismatched, the phoneme category that is perceived is frequently neither the auditory nor the visual one (e.g., Jiang & Bernstein, 2011; McGurk & MacDonald, 1976).

The phonetically impoverished visual stimulus has been characterized through numerous studies of consonant identification in nonsense syllables (e.g., Massaro, 1998; Owens & Blazek, 1985; Walden, Prosek, Montgomery, Scherr, & Jones, 1977). A commonly accepted inference from phonemic perceptual errors is that the phonetic information in visual speech is so reduced that only a few categories, referred to as *visemes*, can be perceived (Fisher, 1968; Massaro, Cohen, Tabain, & Beskow, 2012; Mattheyses & Verhelst, 2014; Walden et al., 1977). Fisher (1968) defined viseme groups to be /p, b, m, d/, /f, v/, /k, g/, /hw, w, r/, and, /ʃ, t, n, l, s, z, h, dʒ, tʃ/. Phonemes within visemes are sometimes regarded as, “virtually indistinguishable by sight, even from a natural face, and are expected to be easily confused”, and “a difference between visemes is significant, informative, and categorical to the perceiver; a difference within a viseme class is not” (Massaro et al., 2012, p. 316). However, there is evidence that the viseme is not the limit of visual phoneme perception. Significant discrimination has been demonstrated for phonemes within visemes in natural and synthetic speech (Files, Tjan, Jiang, & Bernstein, 2015). Furthermore, when words that differed within viseme groups were presented in a target identification task, highly accurate identification performance was obtained (Bernstein, 2012). The groupings that comprise visemes are computationally derived groupings at some threshold applied to phoneme identification confusions (Walden et al., 1977). They capture perceptual distinctiveness that is higher across certain groups of phonemes and lower within those groups. Thus, the extent of phonetic impoverishment is not as extreme as what is implied by the viseme notion. This study examined whether a continuous measure of phonemic perception is predictive of lipreading accuracy.

Conventionally, lipreading is scored as words correct in sentences, and mean scores are typically low (e.g., Sommers, Tye-Murray, & Spehar, 2005; Stevenson et al., 2015). For example in Auer and Bernstein (2007), across 220 normal-hearing adults the mean score was 19 % words correct. However, among adults with normal hearing, there are very large individual differences in lipreading ability. In Bernstein, Demorest, and Tucker (2000), the performance range for open set recognition of isolated monosyllabic words was 0%–58% in 96 normal-hearing adults. The range in Auer and Bernstein (2007) was ~0 % to ~65 % words correct. Individual variation may be attributable to bottom-up perception, top-down processing, and response biases. There are some indications that female lipreaders are more accurate than males (Auer & Bernstein, 2007; Johnson, Hicks, Goldberg, & Myslobodsky, 1988; Tye-Murray, Sommers, & Spehar, 2007; Watson, Qui, Chamberlain, & Li, 1996). Explanations have been sought in terms of cognitive or visual perceptual abilities (e.g., intelligence, lower-level visual skills, and cloze abilities) (Jeffers & Barley, 1971; Mohammed et al., 2005; Watson et al., 1996).

Lipreading is a process that necessarily engages word recognition processes. The structure of the lipreader’s lexicon does not require that all phonemes always be unambiguous for accurate word recognition. For example, although the phonemes /b/, /p/, and /m/ in prevocalic position are perceptually near, in English the word “bought” ordinarily does not have to be discriminated from the non-words “mought” and “pought.” Indeed, isolated word recognition is a better predictor of sentence lipreading scores than is phoneme identification (Bernstein et al., 2000). Visual spoken word neighborhood characteristics, that is their visual confusability, are predictive of lipreading accuracy (Auer, 2002; Mattys, Bernstein, & Auer,

2002). However, increments in phoneme category discriminability predict relatively large decreases in the number of words with which a particular word might be confused (Auer & Bernstein, 1997). More accurate lipreading (words correct) is expected to arise in the context of better perceptual errors. That is, better lipreaders should make errors involving perceptually closer phonemes. Worse lipreaders should make errors involving phonemes that are perceptually far as well as near.

To test this prediction requires that lipreaders actually produce enough perceptual errors to study, as opposed to merely responding with the words that are correctly identified. The low mean scores on lipreading tests might suggest that responses are sparse. In the current study, the mean words correct was 13%. However, Figure 1, which is a histogram of the number of words in 10,232 responses, suggests that there is a very large volume of unanalyzed incorrect responding relative to the 13% words correct. Here, the hypothesis was that the incorrect responses are significantly driven by visual perception, and accuracy is predicted to increase with less extreme perceptual errors. That is, more accurate lipreading is carried out by individuals who capitalize on more accurate visual speech perception. An alternate hypothesis is that the vast amount of errors in lipreading data are generated through mostly higher-level constructive processes such as guessing, in which case measures of perceptual errors should not be predictive of response accuracy.

Sequence alignment.

In order to test the foregoing hypotheses, a systematic method is needed to score incorrect responses. For example, consider the following stimulus and observed response:

Stimulus: “Fishing in a mountain stream is my idea of having a good time.” (Example 1)

Response: “Virginia mouse cream is having a good time.”

The correct part, “Having a good time,” has an obvious mapping with the stimulus. Scoring words correct is easy, involving little more than direct matching after checking for spelling errors, word order, and possible homophones (e.g., “too”, “to,” “two”). The incorrect part of the response presents a scoring challenge, because word boundaries do not align, and the perceptual confusability of the stimulus phonemes may or may not have driven the selection of words in the response.

We developed software that aligns the stimulus and response based on modeled phonemic perceptual errors and the technique of sequence alignment. Detailed description of the phoneme-to-phoneme alignment software has been published (Bernstein, Demorest, & Eberhardt, 1994). The approach was applied successfully in previous studies (e.g., Bernstein, Auer, & Tucker, 2001; Bernstein et al., 1994; Mattys et al., 2002). Stimulus and response text is first phonemically transcribed online. Then, the perceptual distance between the two strings of phonemes is minimized via a sequence aligner that implements a modification (Bernstein et al., 1994; Kruskal & Sankoff, 1983a, p. 296–298) of the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970; Schlosshauer & Ohlsson, 2002), which guarantees minimization of the overall distance between two strings of elements, given a set of costs. Sequence alignment uses the following definitions: A

substitution is any (correct or incorrect) single stimulus-to-response alignment (one stimulus phoneme to one response phoneme). An insertion is a phoneme in a response string that is aligned with a “-” in the stimulus string. A deletion is a stimulus phoneme aligned with a “-” in the response string. The costs are estimates of the visual perceptual distances between any two phonemes in the language. Costs derive from phoneme perception confusion matrices that are compiled from identification tasks with nonsense syllable stimuli.

Example 1 above was submitted to the sequence aligner. The alignment for the incorrect part of the response to “fishing in a mountain stream is” appears in this example:

```
Stimulus: fISIG In e mWntxn strim Is (Example 2)
Response: vRJIn y- x mW---- skrim Is
```

The vertically aligned phonemes in Example 2 are given in ARPABET (Shoup, 1980), an ASCII-based transcription system. The example shows the systematic relationship between the stimulus and response. Perceptual errors such as /v/ for /f/ and /d/ for /t/ are common in lipreading. The type of word boundary errors involving “fishing in a” and “Virginia” is commonly observed in lipreading data. Table 1 displays additional examples.

After alignments are computed, various metrics can be derived. In order to evaluate the hypothesis that responses are driven by perceptual errors, phoneme response dissimilarity (PRD) was computed for every stimulus-response sentence pair in the current study. PRD is the sum of the costs of incorrect phoneme-to-phoneme substitutions weighted by the number of such phonemic alignments for each trial. The weighted sum of costs goes down as incorrect phonemes in a response are perceptually closer to the stimulus. Higher PRD is predicted when lipreading is less accurate.

A second measure that was examined was the number of response phonemes per trial that could *not* be aligned with stimulus phonemes (see Table 1, Example A, dashes in the stimulus string), referred to as *Insertions*. Because Insertions are phonemes in the response for which the cost of alignment with any stimulus phoneme is too high, they are interpreted as a function of top-down constructive processes. For example, if errors comprised extensive guessing, Insertions would be expected to be high and unrelated to lipreading accuracy. Alternatively, Insertions could be used to generate response text when perceptual processing fails in the context of correct word recognition. That is, Insertions could be an indication of successful interactions between top-down and bottom-up processes.

Individual participant differences.

Several studies have reported that normal-hearing females are better lipreaders than normal-hearing males. Johnson, Hicks, Goldberg, and Myslobodsky (1988) reported on 36 females who recognized common phrases more accurately than 17 males. Watson, Ki, Chamberlain, and Li (1996) suggest that sex differences may be found primarily with sentence stimuli. In their separate samples of 40 and 50 lipreaders, females were more accurate. However, in a previous study (Auer & Bernstein, 2007) with 220 normal-hearing and 112 congenitally deaf

adults, neither gender nor interaction of gender with hearing status differences were obtained. Likewise, Tye-Murray, Sommers, and Spehar (2007) tested sentence lipreading in 122 adults (54 males) and reported that males and females scored similarly. The current study was an opportunity to examine the effect of gender and its potential interactions with PRD and Insertions. If there are gender differences among lipreaders, a question is whether accuracy differences are explained by the same or different statistical models.

Even with normal hearing, adults vary in their abilities to recognize words accurately when auditory speech stimuli are embedded in noise (Fullgrabe & Rosen, 2016). Watson et al. (1996) reported significant positive correlations (0.43 and 0.47) between lipreading sentences and perceiving auditory speech in noise. They argued that these correlations imply modality-independent processes whose efficacy vary across individuals. On the other hand, research in the literature on audiovisual speech processing has shown that there are shifts to the more informative modality when the information in the visual or auditory stimulus is degraded or noisy (e.g., Nath & Beauchamp, 2011; Stevenson et al., 2012; Stevenson & James, 2009). In the extreme case of prelingually deafened adults who do not use cochlear implants and rely on vision for speech communication, lipreading accuracy is greatly enhanced. The effect size of auditory deprivation on lipreading was estimated as 1.69 (Cohen's *d*), suggesting that the average prelingually deaf lipreader would score above 95% of normal-hearing adults (Auer & Bernstein, 2007). Recent studies with cochlear implant patients have also been interpreted as evidence that with the degraded auditory stimuli, visual speech perception improves (Song et al., 2014; Strelnikov et al., 2015). If measures of auditory speech perception in noise are sensitive to the daily quality of information received in face-to-face communication, then it may be that poorer speech perception in noise should predict more accurate lipreading. In this study, participants carried out a clinical speech in noise screening test, which measured their threshold in dB signal-to-noise ratio (SNR) at which 50% of key words in sentences could be recognized (Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004).

The current study.

A large body of lipreading responses was obtained from 341 normal-hearing young adults who lipread sentences spoken by two talkers. It was hypothesized that the errors in their responses would be informative about how lipreading is carried out. Sequence alignment software was used to compute phoneme-to-phoneme stimulus-response alignments to score words and phonemes correct in each sentence and to compute two error scores, PRD and Insertions. Along with the error scores, two individual differences measures, gender and speech perception in noise thresholds, were used in modeling the results. Previous studies of lipreading have typically failed to deal with generalization problems across individuals and stimulus materials raised by Clark (1973; c.f., Demorest & Bernstein, 1992; Demorest, Bernstein, & DeHaven, 1996). Here, modeling was carried out with generalized linear mixed models (GLMMs) (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013; Bates, Maechler, Bolker, & Walker, 2015; West, Welch, & Gatecki, 2015).

Methods

Participants.

Four-hundred sixty-one (112 males) young adults were tested in the years 2011–2016. Their mean age was 22 years (range 18–39 years). Most received tests that evaluated pure tone hearing thresholds, speech perception in noise [Quick Speech in Noise Test, QSIN (Killion et al., 2004)], vocabulary knowledge [the Peabody Picture Vocabulary Test, PPVT (Dunn & Dunn, 1981)], and non-verbal intelligence [Test of Non-verbal Intelligence (TONI) (Hammill, Pearson, & Wiederholt, 1996)]. Three-hundred-thirty received right- and left-ear pure tone audiometric testing at six frequencies (.25, .5, 1, 2, 4, 8 kHz). The PPVT and the TONI were administered to 284 participants. The QSIN test, which estimates the SNR at which 50% of words in sentences can be recognized correctly in 4-talker babble noise, was administered to 339 participants who had normal hearing determined through pure tone testing or through self-report. Their QSIN scores were in the normal range of < 4.0 dB SNR. These 339 individuals were accepted into the sample that was studied here. Among them, 262 were females. Ages were mean 22.5 years (range 18 to 32). All participants gave written informed consent as approved by the George Washington University Institutional Review Board.

Materials.

The test of lipreading comprised 15 different sentences spoken by a male and another 15 entirely different sentences spoken by a female. The sentences derived from video recordings (Bernstein et al., 2000) of the Central Institute for the Deaf Everyday Sentences (Davis & Silverman, 1970), a list that has been widely used in lipreading studies. There were 3 to 15 stimulus words per sentence. The sentences with the highest coefficient *alphas* (Cronbach, 1951) were selected from the full list of 100 sentences spoken by both talkers (Auer & Bernstein, 2007). The talkers' faces filled the screen and were in color video.

Procedures.

All of the testing took place in a sound-treated booth under computer control, except for the pure-tone auditory threshold tested with a manual audiometer. During lipreading testing, participants were instructed to view each sentence as it was presented by a computer and then type at the computer terminal whatever they thought the talker had said. Participants were told that it was all right to type responses if they were unsure but not to fabricate their responses. Sentences were presented in random order blocked by talker, whose order was also randomly assigned. There was no feedback, and the test was self-paced.

Response scoring using sequence alignment.

Sequence alignment software was used for scoring words correct, phonemes correct, phoneme response dissimilarity (PRD), and Insertions in lipread sentences. Typed responses were all manually reviewed, and only obvious spelling errors were corrected. Individual letters and partial words were left in the response. Then phonemic transcription was performed automatically with an online database (Seitz, Bernstein, Auer, & MacEachern, 1998). The phoneme-to-phoneme stimulus-response alignment process minimized the

overall distance computed based on perceptual distance between the stimulus and the response. The elements in the alignment comprised only substitutions, insertions, and deletions. The substitution costs were quantified for all possible elementary alignments through a cost lookup table. Costs were set so that a correct elementary substitution cost zero. Non-zero costs were charged for elementary incorrect phoneme-to-phoneme substitutions and for insertions and deletions. Substitution costs were derived empirically from perceptual phoneme identification confusions obtained from previously compiled data from normal-hearing participants who carried out visual-only phoneme identification with monosyllabic nonsense stimuli. Consonant and vowel phoneme confusion matrices were transformed through multidimensional scaling (MDS) (Kruskal & Wish, 1978; Young & Hamer, 1987) into Euclidean distances between phonemes in MDS space. Sequence alignment requires insertions and deletions in order to minimize the overall alignment cost (Kruskal & Sankoff, 1983b): They are placed in the alignment solution when substitutions are too costly. In the current application, insertion and deletion costs were fixed so that elementary substitutions that are unlikely to be perceptually driven, such as alignments of a vowel with a consonant, or a consonant with a consonant that is typically not confused.

In sequence alignment, all of the possible alignments are computed for stimulus-response pairs. If stimulus and response sentences are randomly paired, very few correct phoneme substitutions are obtained, mostly in words randomly shared across the pairs (Bernstein et al., 1994). Additional alignment examples are in Table 1. The examples demonstrate close correspondence between the stimulus and response despite numerous word errors. Most of the substitutions demonstrate that vowels and consonants are not aligned with each other. Example G shows the occasional occurrence of a consonant aligned with a vowel (/z/ with /[^]/). This shows that minimization of the overall alignment cost can overcome the high costs of this type of substitution in the context of an alignment with otherwise close correspondence between strings.

In addition to aligning the transcribed stimulus and response, the aligner outputs the sum of the costs of the incorrect substitutions per stimulus-response pair weighted by the number of those substitutions. This measure is referred to as *phoneme response dissimilarity* (PRD). The higher the PRD, the more severe the perceptual error. However when PRD equals zero, there may be other errors that are scored as Insertions. Insertions were used whenever a phoneme in the response could not be aligned with a phoneme in the stimulus. Although the number of deleted phonemes per sentence was output by the sequence aligner, they were not used in modeling, because they are highly correlated with words correct. The aligner output the number of phonemes correct, which was used in confirming models.

Analyses.

Trial-level proportion words correct and proportion phonemes correct data were modeled using generalized linear mixed models with nested model testing (Baayen et al., 2008; Barr et al., 2013; Bates et al., 2015; West et al., 2015) via maximum likelihood (Laplace approximation) using the lme4 package (Bates et al., 2015) in the R language for statistical computing.

GLMMs model fixed and random effects. The fixed effects describe the relationships between the dependent variable and predictor variables (i.e., fixed factors or continuous covariates) for a population. Participants and stimulus sentences were modeled as crossed random effects with sentences nested within talkers, because each talker spoke a different sentence set. Random effects (intercepts in the foregoing models) are deviations from the relationships involving the fixed factors (West et al., 2015). This modeling approach allows for generalization beyond the specific participants and stimulus materials (Clark, 1973). The GLMM logit link function was used, because the lipreading scores were proportion words or phonemes correct (Jaeger, 2008). The coefficients in the models that appear in tables here are log odds (i.e., the logit transformation). The magnitude of the effects in odds is the exponentiated log odds of the coefficient. The reference in modeling was lipreading scores for male participants. Positive coefficients indicate that the tested level is likely to contribute to a more accurate response and negative coefficients that the tested level is likely to contribute to a less accurate response. Interactions imply the need to examine simultaneously change along the interacting dimensions. The Effects package (Fox & Hong, 2009) in R was used here to plot interactions. Nested model testing used *chi-square* tests with maximum likelihood estimates to determine whether models were significantly different. The Bayesian Information Criterion (BIC) and the Akaike Information Criterion (Burnham & Anderson, 2004; Howardson, Karim, & Horn, 2016; Liu, Rovine, & Molenaar, 2012) (AIC) were consulted along with confidence intervals.

Results and Discussion

Descriptive statistics.

Before modeling the lipreading results, correlations were computed using participants' PPVT, TONI, QSIN, and individual mean proportion correct lipreading scores. Pearson correlations between lipreading scores and PPVT or TONI scores were not significant ($p > .2$). The PPVT and TONI scores were not considered further. QSIN scores were in a normal range between -1.67 and 3.83 dB SNR with a mean of 1.63 dB SNR, and there was a small significant correlation between QSIN and lipreading ($r = -0.19$, $p = 0.0005$). This negative correlation indicates that lipreading was worse in individuals who had poorer auditory spoken word recognition in noise. Figure 2 shows boxplots of QSIN scores as a function of gender. Although the range of QSIN scores was normal, the male group had significantly worse (higher dB SNR) scores ($t(339) = -2.071$, $p = .0391$). Figure 3 shows boxplots of the mean proportion words correct scores as a function of gender. A conventional *t-test* showed that there was a gender difference in lipreading accuracy ($t(339) = 3.175$, $p = .00164$). Figures 4 and 5 show, respectively, boxplots of mean PRD and Insertions scores as a function of gender. These measures did not differ across gender ($p > .652$).

Modeling results.

GLMMs were applied to determine whether PRD, Insertions, gender, and QSIN are significant predictors of lipreading accuracy measured in terms of words and phonemes correct. Models were computed with individual response-level data. Lipreader gender was investigated to determine whether it predicts lipreading accuracy. The QSIN measure of speech perception in noise was investigated to determine whether there might be a common

basis for auditory and visual recognition of spoken words with degraded or impoverished stimuli. The framework for modeling was to first test a full model with all interactions and then to reduce models successively by removing non-significant factors in order to obtain the best-fitting, most parsimonious model/s.

An initial omnibus model had the fixed effects of gender, and the continuous effects of PRD, Insertions, and QSIN, and all of the interactions among fixed effects at the two-way, three-way, and four-way levels. The random effects were the intercepts for subject and the individual sentences (15) nested within talker (2). The four-way interaction was not significant, and there were only two significant three-way interactions. A reduced model was generated with the interactions of gender x PRD X Insertions and gender x PRD x QSIN that were significant in the omnibus model, along with all of the main effects and two-way interactions implied by the retained three-way interactions. The summary output in Table 2 for this reduced omnibus model provides *p*-values based on asymptotic Wald tests (Bates et al., 2015).

Table 2 shows that the gender x PRD, gender x QSIN, and QSIN x Insertions interaction were not significant; but gender x PRD x QSIN and gender x PRD x Insertions were significant ($p < .001$). These interactions suggest that males and females do lipread differently, beyond their overall difference in accuracy levels. Therefore, the results for males and females were modeled separately to determine whether there are structural differences in models that predict males' versus females' lipreading.

Separate male and female models of lipreading.

Initially, male and female models were separately tested with the fixed factors of QSIN, PRD, and Insertions, and all of the two-way and one three-way interactions. The random factors were intercepts for subject and sentences (15) nested in talker (2).

Lipreading in males.

The initially tested model of males' proportion words correct lipreading had significant PRD and QSIN main effects, PRD x QSIN interaction effects, and no main effect or interactions involving Insertions. The full model was a significantly better fit of the data than a model without the non-significant factors in terms of *chi-square* ($chi-square(4) = 12.058, p = 0.02$) and BIC, but not AIC, which was larger in the reduced model. The more parsimonious model in Table 3a was accepted, because confidence intervals for the full model's non-significant effects included 0.0.

Both fixed factor coefficients in Table 3a are negative, suggesting that lipreading accuracy is higher with a lower QSIN (i.e., a lower SNR threshold for speech perception) and a lower PRD score (closer perceptual proximity between incorrect response phonemes and stimulus phonemes). The significant interaction of QSIN and PRD ($p = .0000$) indicates that the QSIN and PRD slopes interact, and so interpretation of the main effects separately from their interaction is not possible.

Figure 6a shows plots of proportion words correct as a function of QSIN at fixed PRD values. It shows that as these predictor variables increase, the slope of the regression line

shifts from negative to positive: Higher values of QSIN and PRD together predict a more accurate lipreading response. This is a paradoxical result, because it suggests that lipreading is more accurate with more extreme perceptual errors when the lipreader has a poorer QSIN score. Because Figure 6a shows that words correct scores were generally low, there was concern about the importance of the slope change. As a check on the results, the proportion phonemes correct scores were used in the same model. Because there are generally more phonemes correct than words correct in sentence lipreading responses, as well as more phonemes than words per stimulus, phoneme-level accuracy scores have higher resolution and wider variation. The log likelihood of the phonemes correct model (-6563.5) was worse than the words correct model (-4926.4), and there was a change in sign across models for the PRD main effect. However, the significance and sign of the PRD x QSIN interaction was also obtained with phoneme scoring. Notably, as Figure 6b shows, proportion phonemes correct scores were generally higher than words correct scores, and the same shift from negative to positive slopes was obtained when predicting lipreading as a function of QSIN at different values of PRD. These results are discussed in more detail below.

Lipreading in females.

The same analysis path was followed with female lipreaders' data. A full model predicting proportion words correct was run with the fixed factors of QSIN, PRD, and Insertions, and all the two-way and one three-way interactions. The random factors were intercepts for subject and sentences (15) nested in talker (2). QSIN and Insertions were significant main effects, and only the interaction of PRD x Insertions was significant. Examination of 95% confidence intervals showed that only those significant effects had confidence intervals that did not include zero. In addition, the full model did not converge with alternative optimizers. Removal of the non-significant factors resulted in the model in Table 4a, which did converge.

The Table-4a model has significant negative coefficients for the main effects of QSIN and PRD, and for the interaction between PRD and Insertions. A large positive coefficient was obtained for Insertions. Exponentiation of the logit scale QSIN coefficient, resulting in the odds ratio of .795, which suggests that for every 1-dB SNR increase in QSIN score, there is a 20% decrease in the odds of a correct lipreading response $((1-.795) \times 100 \%)$ (Figure 7a). Unlike the model for males, the QSIN effect is not modulated by any other variable. Contrary to the possibility that lipreading improves with poorer speech perception in noise, in females there is a straightforward indication that both lipreading and speech perception in noise share common amodal mechanisms that determine accuracy in recognizing words in sentences with reduced phonetic information.

The significant PRD and Insertions main effects in the females' Table 4a model have different signs, and their interaction has a negative coefficient. The interpretation of the main effects is not permitted in the presence of the interaction. Increased PRD (phoneme errors that are perceptually more distant) predicts less accurate lipreading, and the negative slope becomes steeper with increased numbers of Insertions. Figure 7b shows that the most accurate lipreading is predicted by a low PRD score and a high Insertions score. Insertions

are response phonemes that cannot be aligned with any phonemes in the stimulus, because they are too perceptually distant.

Table 4b shows the same model predicting proportion phonemes correct scores. The model's log likelihood (−26726.3) was worse than the log likelihood of the Table-4a model (−10492.4). The sign of the PRD coefficient was reversed across models, but the significant interaction precluded interpreting the two main effects. Figures 7b and 7c show the same pattern of interactions across different fixed Insertions values. The PRD x Insertions interaction is the more dramatic in Figure 7c, as phonemes correct scores vary from zero to one when Insertions were as high as 9. The interaction between PRD and Insertions suggests that female lipreaders couple their more accurate phonemic perception with increased use of top-down processes that involve inserting additional text into their typed responses.

Discussion.

In the females' lipreading model (Table 4a), the main effect of QSIN threshold was significant: Lower QSIN score (i.e., better ability in noise) predicted higher lipreading accuracy, a result that is consistent with the hypothesis that lipreading and speech perception in noise share common higher-level, amodal processing mechanisms. This implies that an alternative hypothesis is not true: That is, increased difficulty perceiving speech in noise does not result in more accurate lipreading in normal-hearing females. Therefore, demonstrations of short-term shifts to visual speech when auditory speech is degraded, as reported in some neuroimaging studies (e.g., Nath & Beauchamp, 2011; Stevenson et al., 2012; Stevenson & James, 2009), appear to be irrelevant to lipreading ability; and similarly, better lipreading accuracy in individuals with hearing loss seems to be irrelevant to causes for lipreading accuracy in normal-hearing adult females.

In females, lipreading accuracy was higher when the PDR score was lower—that is, when phonemic errors implied finer differentiation among visual spoken phonemes, *and* when there were more Insertions in the typed response. Figure 7b-c shows lipreading accuracy as a function of PDR for the range of Insertions values. Insertions are evidence of the lipreader generating parts of her response through top-down processes. Inserted text is estimated to be too distant from the stimulus to be aligned with it during the minimization procedure. The interaction is consistent with the negatively sloped QSIN effect: Good female lipreaders not only have more accurate visual speech perception, they have lower thresholds for speech perception in noise, and they amplify their more accurate lipreading with insertion of text based on top-down processing.

In the males' model (Table 3a), the main effect of QSIN threshold was significant, but so was its interaction with PRD. This interaction (Figures 6a-b) shows that lipreading accuracy as a function of QSIN score changed slope from negative to positive as PRD scores increased. Phonemic perceptual errors represented poorer differentiation among visual spoken phonemes in responses that were otherwise more accurate in those responses emitted by males with poorer QSIN scores. A possible explanation is that response bias shifted with QSIN score: Males with poorer speech perception in noise thresholds were more willing to type perceptually dissimilar text in the context of more accurate responding. Additionally,

although the mean Insertions scores did not differ across gender, the use of Insertions was not a predictor in the males' model although it was in a significant interaction in the females'. Furthermore, mean PRD did not differ across males and females but was a significant factor in the interactions that predict lipreading in the two groups.

General Discussion

This study investigated lipreading in normal-hearing young adults. In contrast with previous studies that analyzed participants' mean scores pooled across all of their responses, here individual trials were analyzed. A phoneme-to-phoneme stimulus-response sequence aligner scored the individual responses in terms of proportion words or proportion phonemes correct, and two error measures, the number of unaligned response phonemes (Insertions), and the phoneme response dissimilarity (PRD), a measure of the weighted visual perceptual dissimilarity between phonemes in the stimulus and phonemes that were substituted in the response. Generalized linear mixed modeling tested the hypothesis that the perceptual error scores predict the likelihood of words (or phonemes) correct in the response. That is, accuracy increases when the lipreader's perceptual errors represent finer differentiation among visual spoken phonemes. This prediction followed from observations that better resolution of phonemic perceptual differences should lead to fewer confusions among lipread words (Auer, Bernstein, Waldstein, & Tucker, 1997; Iverson, Auer, & Bernstein, 1997; Mattys et al., 2002). Gender and speech perception in noise threshold (QSIN score; Killion et al., 2004) were additional predictors. Based on accounts in the literature, females were predicted to be better lipreaders. The speech in noise score was used to test alternate hypotheses that either difficulty perceiving speech in noise leads to more accurate lipreading, or that lipreading accuracy and auditory thresholds in noise are supported by some common abilities to recognize words based on impoverished or degraded phonetic information. All of the foregoing factors and several of their interactions were significant predictors of lipreading accuracy. Notably, interactions with gender required separate models for females' and males' lipreading. The separate models were structured differently. For females, there was a main effect of QSIN score supporting the hypothesis that common processes support lipreading and auditory speech perception in noise. Additionally, females appear to amplify their correct responding by inserting additional text into their typed responses. In contrast, the model of males' lipreading did not include the Insertions factor. Only the main effects of PRD and QSIN and their interaction were significant. The interaction prevents interpretation of the main effects and shows that males with poorer speech perception in noise thresholds typed responses that were more perceptually dissimilar to the stimulus in the context of their otherwise more accurate responses.

Gender differences.

The gender differences in this study imply top-down and bottom-up processing differences. The interactions involving the QSIN factor suggest that top-down processing of degraded or impoverished speech varies across females and males. This study was not designed for detailed investigation of investigate top-down mechanisms that may support auditory and visual spoken word recognition. Currently, there is considerable interest in how top-down mechanisms support individual differences in auditory speech perception in noise, with

particular interest in working memory (e.g., Millman & Mattys, 2017; Ronnberg et al., 2013). The results here suggest that further insights may be gained by increasing the scope of such studies to include audiovisual and visual-only experiments. What appears to be amodal processes may instead represent modality-specific mechanisms that vary together within individuals. For example, Author (xxxx) suggests that there are parallel auditory and visual speech processing pathways, possibly to the level of representing the forms of spoken words. If so, the individual variation may be due to modality-specific variation in bottom-up processing of phonological input regardless of the modality-specific processing pathway.

The males' model of lipreading included the QSIN and PRD main effects and their interaction. These results seem somewhat odd: Poorer QSIN and poorer PRD scores somehow led to more accurate responding. A possible explanation is that males with poorer QSIN scores had different response biases than ones with better scores. They were more willing to type response text that was perceptually discrepant from the stimulus. However, in order for this interpretation to be credible, lipreaders would have to demonstrate evidence for internal measures of their own perceptual accuracy.

In fact, there is evidence for internal accuracy signals. An early study demonstrated that lipreading scores improved when participants were encouraged to guess (Van Tasell & Hawkins, 1981), supporting the availability of internal feedback signals. In another study, both normal-hearing and deaf lipreaders were shown to be able to validly judge their own lipreading accuracy (Demorest & Bernstein, 1997). In that study, regression analyses were carried out to model subjective confidence ratings. Models showed that confidence was predicted by the number of correct response words, even though no feedback was given to the participants, supporting the conclusion that lipreaders have access to internal accuracy signals. Importantly, when phoneme-level measures were used, a measure similar to PRD was shown to predict subjective confidence levels, suggesting that lipreaders also have access to perceptual error signals that increase as the stimulus-to-response dissimilarity increases. In a recent unpublished training study by the author and colleagues, subjective accuracy was also shown to predict perceptual errors in lipread sentences. Therefore, the results in the current study could be the result of processes that lead males with poorer speech perception in noise to amplify their more accurate lipreading responses with additional text that includes worse perceptual errors.

Overall, the current results suggest that female lipreaders achieve better visual perception of speech, their ability to recognize words with impoverished or degraded input is consistent across modalities, and they amplify their correct responding through top-down insertion of text. In contrast, the pattern of responding in males suggests that those with the poorer auditory speech perception amplify their more accurate lipreading by shifting towards typing text that is more perceptually discrepant from the stimulus but nonetheless perceptually driven. These results lead to questions about generalizability across males and females.

Nomothetic versus idiographic models of visual speech perception and neural processing.

Often studies that use visual speech (in lipreading and audiovisual research) pool results across males and females. This approach is consistent with nomothetic science whose aim is, "to make *general predictions about the population* ... it permits pooling across

participants...for data collected in both cross-sectional and longitudinal designs” (Beltz, Wright, Sprague, & Molenaar, 2016, p. 447). A scientific view that is opposed to the nomothetic one is sometimes referred to as *ideographic*, the search for accounts of individual differences. Malenaar and Campbell (2009, p. 113) state that the conditions that must hold in order to justify pooling results across groups are: “First, the same statistical model should apply to the data of all subjects in the population (homogeneity of the population). Second, the data must be stationary—that is, the data must have invariant statistical characteristics across time, for instance, having constant mean and variance.” The concern here is with the first condition (testing occurred only once). The models that were obtained are structurally different for males and females, suggesting that a nomothetic approach that pools lipreading data across gender is not be valid.

However, it would be a mistake to conclude that gender differences imply different lipreading mechanisms (Hommel & Colzato, 2017) just as similar performance across groups does not necessitate the same neural mechanisms (Cahill, 2006). An interesting example of different effects arising from the same mechanisms is presented by Ramscar and colleagues (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014), who argue that age differences in spoken word processing may arise from common word learning effects across the lifespan. It is possible that the same mechanisms result in different lipreading effects in males and females. For example, better visual speech perception could encourage insertion of text that was not perceived in the stimulus. This mechanism may not be significant in males, because their perceptually accuracy is generally too low.

However, research on face processing suggests that there may be fundamental differences in how males and females process faces, and by extension, those differences could lead to fundamental differences in lipreading. Interactions between sex, cortical, and/or subcortical activations to emotional faces have been reported (Cahill, 2006; Filkowski, Olsen, Duda, Wanger, & Sabatinelli, 2016). When asked to rate the valence and arousal levels to pictured faces, regardless of the sex of the pictured faces, women’s ratings were higher than men’s, suggesting that women find faces more interesting and arousing (Proverbio, 2017). Visual stimuli that are more arousing may have higher priority of processing or may become the focus of attention (Harmon-Jones, Gable, & Price, 2011). Women may be more accurate in deciphering facial emotions (Thompson & Voyer, 2014). They may have greater functional connectivity between the right amygdala and the bilateral posterior superior temporal sulcus (pSTS) during perception of biological motion stimuli (Anderson et al., 2013). The left pSTS is typically activated when visual speech stimuli are presented (Author, xxxx), and the amygdalae are highly implicated in emotion processing (Haxby, Hoffman, & Gobbini, 2002). Differences in the arousal or interest value of faces could lead to better lipreading through spending more time looking at faces, affording more opportunities to learn visual speech features, although visual speech and emotion features are different. Two small-sample size studies make claims for different neural lipreading networks in males and females (Ruytjens, Albers, van Dijk, Wit, & Willemsen, 2006, 2007). However, considerable additional work will be needed to determine whether in fact there are structural differences in underlying neural mechanisms across males and females, and whether they are the outcome of lifespan differences in perceiving emotional and talking faces. Alternatively, behavioral outcomes may be the result of the same mechanisms with different drivers. At

this time, a simple recommendation is that researchers who use visual speech stimuli should also investigate whether their results significantly vary across gender.

Lipreading scoring methods.

Because lipreading scores tend to be low, scoring is sometimes carried out in terms of visemes, thereby increasing the upward range of scores (e.g., Tye-Murray et al., 2007). However, viseme scoring implies that only a small number of visual phonological categories are perceived by the lipreader (Massaro et al., 2012), and importantly, the scoring method does not differentiate between exactly correct and incorrect phonemic perception. The sequence aligner used here minimizes the overall dissimilarity (cost) of stimulus and response phoneme-to-phoneme alignment in terms of a continuum of perceptual dissimilarities, which are shown here to be sensitive to lipreading accuracy and to response modes. The current approach to scoring lipreading responses, which are mostly incorrect, is encouraged.

Implications for effects of aging.

Hearing difficulties in older adults might predict increased reliance on visual information. However, there are reports (based on small samples with low power) that lipreading declines with age (Dancer, Krain, Thompson, Davis, & Glenn, 1994; Honnell, 1991; Middleweerd & Plomp, 1987; Shoop & Binnie, 1979; Sommers et al., 2005). Previous reports used mean scores focused on correct responding. Thus, if older adults behave similarly to younger adults, emitting more incorrect word responses than correct ones, much of the data were unscored. In those unscored responses may be evidence of how lipreading changes with age. For example, it may be that word recognition accuracy declines even though phonemic perceptual errors do not. Another possibility is that because previous statistical analyses used methods such as ANOVA with proportional data that the reports about age effects across the lifespan may include spurious significant as well as non-significant results (Jaeger, 2008). Statistical modeling that uses individual trial and error data may improve understanding of older lipreaders.

Summary and conclusions.

The generally low accuracy of lipreading in normal-hearing adults is expected to increase their deployment of higher-level lexical, semantic, and syntactic processing resources in addition to guessing, all of which may introduce diverse errors as these lipreaders attempt to produce responses based on impoverished visual phonetic input. The results here suggest that lipreading accuracy and top-down processing differ across males and females. The results also show that responses in the sentence lipreading task are highly coupled to perceptual processing as evidenced by the predictive significance of a measure of phonemic response errors. While the majority of the typed responses were incorrect, phoneme-to-phoneme stimulus-response alignment suggest that they were driven by ongoing bottom-up perceptual processes interacting with top-down ones. Differently structured models were required by the differences in the data from males and females. These differences could be related to differences in face processing and/or to differences that may be more relevant to language processing.

Acknowledgments

I think Silvio P. Eberhardt for his comments on this paper as well as his contribution during data collection. The research was supported by the US NIH (DCxxxxx and DCxxxxx).

References

- Anderson LC, Bolling DZ, Schelinski S, Coffman MC, Pelphrey KA, & Kaiser MD (2013). Sex differences in the development of brain mechanisms for processing biological motion. *Neuroimage*, 83, 751–760. doi:10.1016/j.neuroimage.2013.07.040 [PubMed: 23876243]
- Auer ET Jr. (2002). The influence of the lexicon on speech read word recognition: contrasting segmental and lexical distinctiveness. *Psychon Bull Rev*, 9(2), 341–347. [PubMed: 12120798]
- Auer ET Jr., & Bernstein LE (1997). Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America*, 102(6), 3704–3710. [PubMed: 9407662]
- Auer ET Jr., & Bernstein LE (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research*, 50(5), 1157–1165.
- Auer ET, Bernstein LE, Waldstein RS, & Tucker PE (1997, September 26–27, 1997). Effects of phonetic variation and the structure of the lexicon on the uniqueness of words. Paper presented at the ESCA/ESCOP Workshop on Audio-Visual Speech Processing, Rhodes, Greece.
- Baayen RH, Davidson DJ, & Bates DM (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Barr DJ, Levy R, Scheepers C, & Tily HJ (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang*, 68(3). doi:10.1016/j.jml.2012.11.001
- Bates D, Mächler M, Bolker B, & Walker S (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Bates D, Maechler M, Bolker B, & Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:doi:10.18637/jss.v067.i01
- Bernstein LE (2012). Visual speech perception In Vatikiotis-Bateson E, Bailly G, & Perrier P (Eds.), *AudioVisual Speech Processing* (pp. 21–39). Cambridge: Cambridge University.
- Bernstein LE, Auer ET Jr., & Tucker PE (2001). Enhanced speechreading in deaf adults: can short-term training/practice close the gap for hearing adults? *Journal of Speech, Language, and Hearing Research*, 44(1), 5–18.
- Bernstein LE, Demorest ME, & Eberhardt SP (1994). A computational approach to analyzing sentential speech perception: phoneme-to-phoneme stimulus-response alignment. *Journal of the Acoustical Society of America*, 95(6), 3617–3622. [PubMed: 8046151]
- Bernstein LE, Demorest ME, & Tucker PE (2000). Speech perception without hearing. *Percept Psychophys*, 62(2), 233–252. [PubMed: 10723205]
- Burnham KP, & Anderson DR (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2), 261–304.
- Clark HH (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359.
- Cronbach LJ (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297–334.
- Dancer JE, Krain M, Thompson C, Davis PN, & Glenn J (1994). A cross sectional investigation of speechreading in adults: Effects of age, gender, practice, and education. *Volta Review*, 96, 31–40.
- Demorest ME, & Bernstein LE (1992). Sources of variability in speechreading sentences: a generalizability analysis. *J Speech Hear Res*, 35(4), 876–891. [PubMed: 1405543]
- Demorest ME, & Bernstein LE (1997). Relationships between subjective ratings and objective measures of performance in speechreading sentences. *Journal of Speech, Language, and Hearing Research*, 40(4), 900–911.

- Demorest ME, Bernstein LE, & DeHaven GP (1996). Generalizability of speechreading performance on nonsense syllables, words, and sentences: subjects with normal hearing. *J Speech Hear Res*, 39(4), 697–713. [PubMed: 8844551]
- Dunn LM, & Dunn LM (1981). *Peabody Picture Vocabulary Test-Revised*. Pines, MN: American Guidance Service.
- Files BT, Tjan B, Jiang J, & Bernstein LE (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Front Psychol*, 6, 878. doi:doi: 10.3389/fpsyg.2015.00878 [PubMed: 26217249]
- Fisher CG (1968). Confusions among visually perceived consonants. *J Speech Hear Res*, 11(4), 796–804. [PubMed: 5719234]
- Fox F, & Hong J (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, 32(1), 1–24.
- Fullgrabe C, & Rosen S (2016). On The (Un)importance of Working Memory in Speech-in-Noise Processing for Listeners with Normal Hearing Thresholds. *Front Psychol*, 7, 1268. doi:10.3389/fpsyg.2016.01268 [PubMed: 27625615]
- Hamill DD, Pearson NA, & Wiederholt JL (1996). *Comprehensive Test of Nonverbal Intelligence*. Austin, TX: Pro-Ed.
- Honnell S (1991). Age and Speechreading Performance in Relation to Percent Correct, Eyeblinks, and Written Responses. *Volta Review*, 93(4), 207–213.
- Howardson GN, Karim MN, & Horn RG (2016). The latent change score model: A more flexible approach to modeling time in self-regulated learning. *Journal of Business and Psychology*, advanced online publication, 1–18.
- Iverson P, Auer ET Jr., & Bernstein LE (1997). A comparison of perceptual word similarity metrics. *Journal of the Acoustical Society of America*, 102, 3189.
- Jaeger TF (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434–446. [PubMed: 19884961]
- Jeffers J, & Barley M (1971). *Speechreading (Lipreading)*. Springfield, IL: Charles C. Thomas.
- Jiang J, Alwan A, Keating P, Auer ET Jr., & Bernstein LE (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing: Special issue on Joint Audio-Visual Speech Processing*, 11, 1174–1188.
- Jiang J, & Bernstein LE (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Performance and Perception*, 37(4), 1193–1209.
- Johnson FM, Hicks LH, Goldberg TE, & Myslobodsky MS (1988). Sex differences in lipreading. *Bulletin of the Psychonomic Society*, 26(2), 106–108.
- Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, & Banerjee S (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*, 116(4 Pt 1), 2395–2405. [PubMed: 15532670]
- Kruskal JB, & Sankoff D (1983a). An anthology of algorithms and concepts for sequence comparison. In Sankoff D & Kruskal JB (Eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (pp. 265–310). Reading, MA: Addison-Wesley.
- Kruskal JB, & Sankoff D (1983b). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley.
- Kruskal JB, & Wish M (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.
- Liu S, Rovine MJ, & Molenaar P (2012). Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological methods*, 17(1), 15. [PubMed: 22251268]
- Massaro DW (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.
- Massaro DW, Cohen MM, Tabain M, & Beskow J (2012). Animated speech: Research progress and applications. In Clark RB, Perrier J, & Vatikiotis-Bateson E (Eds.), *Audiovisual Speech Processing* (pp. 246–272). Cambridge: Cambridge University.

- Mattheyses W, & Verhelst W (2014). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66, 182–217.
- Mattys SL, Bernstein LE, & Auer ET Jr. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Percept Psychophys*, 64(4), 667–679. [PubMed: 12132766]
- McGurk H, & MacDonald J (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748. [PubMed: 1012311]
- Middleweerd MJ, & Plomp R (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America*, 82(6), 2145–2147. [PubMed: 3429736]
- Millman RE, & Mattys SL (2017). Auditory Verbal Working Memory as a Predictor of Speech Perception in Modulated Maskers in Listeners With Normal Hearing. *J Speech Lang Hear Res*, 60(5), 1236–1245. doi:10.1044/2017_jslhr-s-16-0105 [PubMed: 28492912]
- Mohammed T, Campbell R, MacSweeney M, Milne E, Hansen P, & Coleman M (2005). Speechreading skill and visual movement sensitivity are related in deaf speechreaders. *Perception*, 34(2), 205–216. [PubMed: 15832570]
- Munhall KG, & Vatikiotis-Bateson E (1998). The moving face during speech communication In Campbell R, Dodd B, & Burnham D (Eds.), *Hearing by Eye II. Advances in the Psychology of Speechreading and Auditory-visual Speech* (pp. 123–142). Hove: Psychology Press.
- Nath AR, & Beauchamp MS (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, 31(5), 1704–1714. doi:10.1523/jneurosci.4853-10.2011 [PubMed: 21289179]
- Needleman SB, & Wunsch CD (1970). A general method applicable to the search for similarities in the amino-acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443–453. [PubMed: 5420325]
- Owens E, & Blazek B (1985). Visemes observed by hearing-impaired and normal hearing adult viewers. *J Speech Hear Res*, 28(3), 381–393. [PubMed: 4046579]
- Proverbio AM (2017). Sex differences in social cognition: The case of face processing. *J Neurosci Res*, 95(1–2), 222–234. doi:10.1002/jnr.23817 [PubMed: 27870403]
- Ramscar M, Hendrix P, Shaoul C, Milin P, & Baayen H (2014). The myth of cognitive decline: non-linear dynamics of lifelong learning. *Top Cogn Sci*, 6(1), 5–42. doi:10.1111/tops.12078 [PubMed: 24421073]
- Ronnberg J, Lunner T, Zekveld A, Sorqvist P, Danielsson H, Lyxell B, . . . Rudner M (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front Syst Neurosci*, 7, 31. doi:10.3389/fnsys.2013.00031 [PubMed: 23874273]
- Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, & Foxe JJ (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17(5), 1147–1153. doi:10.1093/cercor/bhl024 [PubMed: 16785256]
- Ruytjens L, Albers F, van Dijk P, Wit H, & Willemsen A (2006). Neural responses to silent lipreading in normal hearing male and female subjects. *Eur J Neurosci*, 24(6), 1835–1844. doi:10.1111/j.1460-9568.2006.05072.x [PubMed: 17004947]
- Ruytjens L, Albers F, van Dijk P, Wit H, & Willemsen A (2007). Activation in primary auditory cortex during silent lipreading is determined by sex. *Audiol Neurotol*, 12(6), 371–377. doi:10.1159/000106480 [PubMed: 17664868]
- Schlosshauer M, & Ohlsson M (2002). A novel approach to local reliability of sequence alignments. *Bioinformatics*, 18(6), 847–854. [PubMed: 12075020]
- Seitz PF, Bernstein LE, Auer ET Jr., & MacEachern M (1998). PhLex (Phonologically Transformable Lexicon): A 35,000-word computer readable pronouncing American English lexicon on structural principles, with accompanying phonological transformations, and word frequencies.
- Shoop C, & Binnie CA (1979). The effects of age upon the visual perception of speech. *Scand Audiol*, 8(1), 3–8. [PubMed: 515681]
- Shoup JE (1980). Phonological aspects of speech recognition In Lea WA (Ed.), *Trends in Speech Recognition* (pp. 125–138). NY: Prentice-Hall.

- Sommers MS, Tye-Murray N, & Spehar B (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear*, 26(3), 263–275. [PubMed: 15937408]
- Song JJ, Lee HJ, Kang H, Lee DS, Chang SO, & Oh SH (2014). Effects of congruent and incongruent visual cues on speech perception and brain activity in cochlear implant users. *Brain Structure & Function*. doi:10.1007/s00429-013-0704-6
- Stevenson RA, Bushmakina M, Kim S, Wallace MT, Puce A, & James TW (2012). Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech. *Brain Topography*, 25(3), 308–326. doi:10.1007/s10548-012-0220-7 [PubMed: 22367585]
- Stevenson RA, & James TW (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage*, 44(3), 1210–1223. [PubMed: 18973818]
- Stevenson RA, Nelms CE, Baum SH, Zurkovsky L, Barense MD, Newhouse PA, & Wallace MT (2015). Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. *Neurobiol Aging*, 36(1), 283–291. doi:10.1016/j.neurobiolaging.2014.08.003 [PubMed: 25282337]
- Strelnikov K, Rouger J, Lagleyre S, Fraysse B, Demonet JF, Deguine O, & Barone P (2015). Increased audiovisual integration in cochlear-implanted deaf patients: independent components analysis of longitudinal positron emission tomography data. *Eur J Neurosci*, 41(5), 677–685. doi:10.1111/ejn.12827 [PubMed: 25728184]
- Sumbly WH, & Pollack I (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2), 212–215.
- Tye-Murray N, Sommers MS, & Spehar B (2007). The effects of age and gender on lipreading abilities. *J Am Acad Audiol*, 18(10), 883–892. [PubMed: 18496997]
- Van Tasell DJ, & Hawkins DB (1981). Effects of guessing strategy on speechreading test scores. *American Annals of the Deaf*, 126, 840–844. [PubMed: 7304382]
- Walden BE, Prosek RA, Montgomery AA, Scherr CK, & Jones CJ (1977). Effects of training on the visual recognition of consonants. *J Speech Hear Res*, 20, 130–145. [PubMed: 846196]
- Watson CS, Qui WW, Chamberlain MM, & Li X (1996). Auditory and visual speech perception: confirmation of a modality-independent source of individual differences in speech recognition. *Journal of the Acoustical Society of America*, 100(2 Pt 1), 1153–1162. [PubMed: 8759968]
- West BT, Welch KB, & Gatecki AT (2015). *Linear Mixed Models: A Practical Guide Using Statistical Software* (Second Edition ed.).
- Yehia H, Rubin P, & Vatikiotis-Bateson E (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1–2), 23–43.
- Young FW, & Hamer RM (1987). *Multidimensional Scaling. History, Theory, and Applications*. Hillsdale, NJ: Erlbaum.

Figure 1. Histogram of the number of words per response.

The plot shows that a small proportion of the trials resulted in no response words. Stimulus sentences varied from 3 to 15 words in length, as did response length.

Figure 2. Boxplot of QSIN Scores for female versus male lipreaders.

Boxplots show the distribution of QSIN scores as a function of gender. Medians are the horizontal lines within boxes. Whiskers are the first and fourth quartiles. Outliers are dots. Larger QSIN scores indicate a higher signal-to-noise ratio is needed to recognize 50% of words in babble noise.

Figure 3. Boxplots of mean proportion words correct in sentences lipread by females versus males.

The figure shows the lipreading advantage held by females, both in terms of the quartiles and in terms of outliers.

Figure 4. Boxplots of mean phoneme response dissimilarity (PRD) scores for female versus male lipreaders.

The PRD was averaged across 30 sentences within participant in this figure.

Figure 5. Boxplot of mean number of Insertions for females versus males.
The number of insertions was averaged across 30 sentences within participants.

QSIN x PRD Interaction Effect Plot

Figure 6a-b. Figures illustrating males' lipreading model with QSIN x PRD interaction.

Figure 6a illustrates proportion words correct as a function of QSIN for the range of males' PRD scores. Figure 6b illustrates the same responses scores with proportion phonemes correct as a function of QSIN for the range of PRD scores.

Note: The non-linear *y-axis* is due to linear modeling in the logit transformation. The different *y-axis* ranges in the 6a and 6b figures are due to higher scores when sentences are evaluated in terms of phonemes correct (6b). The shaded areas represent loess smoothing.

Figure 7a-c. Figures illustrating females' lipreading model as a function of QSIN score, with Insertions x PRD interaction.

Figure 7a illustrates proportion words correct as a function of QSIN score. Figure 7b illustrates proportion words correct as a function of PRD across the range of Insertion scores. Figure 7c illustrates proportion phonemes correct across the range of Insertion scores.

Note: The non-linear *y-axis* is due to linear modeling in the logit transformation. The *y-axis* in the 7b and 7c figures is different due to higher scores when sentences are evaluated in terms of phonemes correct (7c). The shaded areas represent loess smoothing.

Table 1.

Examples of sentence lipreading.

The center column lists the stimulus and response. The rightmost column shows the stimulus and response phoneme strings aligned vertically phoneme-by-phoneme by the sequence aligner. Transcriptions are in ARPABET notation. Dashes in the stimulus are aligned with phonemes inserted into the response. Dashes in the response are aligned with phonemes in the stimulus that were deleted in the response.

Example	Stimulus: Response:	Phonemic transcription of the stimulus: Phonemic transcription of the response:
A	S: Come here when I call you. R: let me have your number to call you	k^-m-hirwE-nA-----kclyu lEtmiH@vyUr^n^mbRtukclyu
B	S: There's not enough room in the kitchen for a new table. R: There's not enough room in the kitchen for you and Samuel.	DErznatxn^frumInDxkICxnfcR^nu---teb--xl DErznatxn^frumInDxkICxnfcR-yu@nds@myUxl
C	S: Wait for me at the corner in front of the drug store. R: Wait for me in the corner of the grocery store.	wetfcrmi@tDxkcrnRInfr^nt^vDxdr--^g-stcr wetfcrmiInDxkcrnR-----^vDxgrosxrstcr
D	S: It's a real dark night so watch your driving R: that's a really long you're having	-Its^xildarknAtsow@CyUrdrAvIG D@ts^xil-i--lcG-----yUrH-@vIG
E	S: People ought to see a doctor once a year. R: people always want to hear	pipxlcttusi^daktRw^N-s^yir pipxlcl--wi-----z-w^nttuHir
F	S: If we don't get rain soon we'll have no grass. R: if we go down to school we'll have no class	IfwidontgEtrensunwilh@vnogr@s Ifwigo--dWntuskulwilh@vnokl@s
G	S: Music always cheers me up. R: You always need a map	myuzxklwizCirzmi^p -yu---clwiznid^m-@p
H	S: Don't let the dog out of the house. R: so is there anything else	dontlEtDxcdcgWt^vDxhWs so---IzDErIniTIG-El-s

Table 2.

Results of the most parsimonious omnibus GLMM. The reference category was male lipreaders.

Formula: Proportion words correct =
gender + QSIN + PRD+ Insertions + Insertions x PRD +
Insertions x gender + Insertions x QSIN + QSIN x PRD +
QSIN x gender + PRD x gender + gender x QSIN x PRD +
Insertions x PRD x gender + (1 | subject) + (1 | talker/index)

Number of responses = 10172; Subjects = 339; Stimuli = 30;
Talkers = 2.

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	1.0114	1.0057
index:talker	(Intercept)	0.5889	0.7674
talker	(Intercept)	0.1864	0.4318

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.19274	0.42869	-5.115	3.14e-07	***
Female	0.34595	0.29539	1.171	0.241527	
QSIN	-0.36291	0.12997	-2.792	0.005236	**
PRD	-0.23496	0.04485	-5.239	1.62e-07	***
Insertions	-0.23959	0.12224	-1.960	0.049998	*
PRD x Insertions	0.09460	0.04015	2.356	0.018459	*
Female x Insertions	0.54686	0.12790	4.276	1.91e-05	***
QSIN x PRD	0.09749	0.02279	4.278	1.89e-05	***
Female x QSIN	0.10212	0.14809	0.690	0.490440	
Female x PRD	0.08782	0.04836	1.816	0.069350	.
QSIN x Insertions	0.02317	0.01197	1.935	0.052960	.
Female x QSIN x PRD	-0.08356	0.02510	-3.329	0.000873	***
Female x PRD x Insertions	-0.15796	0.04271	-3.699	0.000217	***

Notes: Convergence was obtained with the bobyqa optimizer (Bates, Mächler, Bolker, & Walker, 2014).

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.

Models predicting male lipreaders' proportion words correct in sentences (Table 3a) and proportion phonemes correct in sentences (Table 3b).

Table 3a					
Formula:					
Proportion words correct = QSIN + PRD+ QSIN x PRD+ (1 subject) + (1 talker/sentence)					
Number of responses=2310, subjects = 77; Stimuli = 30; Talkers = 2.					
Random effects:					
Groups	Name	Variance	Std.Dev.		
subject	(Intercept)	0.9217	0.9601		
index:talker	(Intercept)	0.5966	0.7724		
talker	(Intercept)	0.2448	0.4948		
Fixed effects:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.22601	0.45583	-4.883	1.04e-06	***
QSIN	-0.34778	0.12507	-2.781	0.00542	**
PRD	-0.20898	0.04480	-4.665	3.09e-06	***
QSIN x PRD	0.09539	0.02276	4.192	2.77e-05	***

Table 3b					
Formula:					
Proportion phonemes correct = QSIN + PRD+ QSIN x PRD+ (1 subject) + (1 talker/sentence)					
Number of responses=2310, subjects = 77; Stimuli = 30; Talkers = 2.					
Random effects:					
Groups	Name	Variance	Std.Dev.		
subject	(Intercept)	0.5097	0.7127		
index:talker	(Intercept)	0.3163	0.5624		
talker	(Intercept)	0.0530	0.2297		
Fixed effects:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.62575	0.26362	-6.167	6.96e-10	***
QSIN	-0.31719	0.08771	-3.616	0.000299	***
PRD	0.04381	0.02055	2.132	0.033024	*
QSIN x PRD	0.08110	0.01050	7.725	1.12e-14	***

Notes: Significance codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.

Model predicting female lipreaders' proportion words correct in sentences (Table 4a) and proportion phonemes correct in sentences (Table 4b).

Table 4a				
Formula: Proportion Words Correct ~ QSIN + PRD+ Insertions + PRD x Insertions + (1 subject) + (1 talker/sentence)				
Number of responses=7862, subjects = 262; Stimuli = 30; Talkers = 2.				
Random effects:				
Groups	Name	Variance	Std.Dev.	
subject	(Intercept)	1.0475	1.0235	
index:talker	(Intercept)	0.6022	0.7760	
talker	(Intercept)	0.1749	0.4182	
Fixed effects:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.90110	0.35223	-5.397	6.76e-08 ***
QSIN	-0.22951	0.07007	-3.275	0.00106 **
PRD	-0.12743	0.01068	-11.935	< 2e-16 ***
Insertions	0.33419	0.04319	7.738	1.01e-14 ***
PRD x Insertions	-0.06089	0.01498	-4.064	4.83e-05 ***

Table 4b				
Formula: Proportion Words Correct ~ QSIN + PRD+ Insertions + PRD x Insertions + (1 subject) + (1 talker/sentence)				
Number of responses=7862, subjects = 262; Stimuli = 30; Talkers = 2.				
Random effects:				
Groups	Name	Variance	Std.Dev.	
subject	(Intercept)	1.6012	0.7754	
index:talker	(Intercept)	0.3389	0.5821	
talker	(Intercept)	0.0623	0.2497	
Fixed effects:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.463427	0.227175	-6.442	1.18e-10 ***
QSIN	-0.154290	0.052023	-2.966	0.00302 **
PRD	0.064814	0.005198	12.470	< 2e-16 ***
Insertions	0.770985	0.026406	29.197	< 2e-16 ***
PRD x Insertions	-0.198308	0.009046	-21.921	< 2e-16 ***

Notes:Significance codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1