Trends in Hearing
Volume 24: 1–14
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2331216520915110
journals.sagepub.com/home/tia
SAGE

Contributions of Voice Expectations to Talker Selection in Younger and Older Adults With Normal Hearing

William J. Bologna, Jayne B. Ahlstrom, and Judy R. Dubno

Abstract

Focused attention on expected voice features, such as fundamental frequency (F0) and spectral envelope, may facilitate segregation and selection of a target talker in competing talker backgrounds. Age-related declines in attention may limit these abilities in older adults, resulting in poorer speech understanding in complex environments. To test this hypothesis, younger and older adults with normal hearing listened to sentences with a single competing talker. For most trials, listener attention was directed to the target by a cue phrase that matched the target talker's F0 and spectral envelope. For a small percentage of randomly occurring *probe* trials, the target's voice unexpectedly differed from the cue phrase in terms of F0 and spectral envelope. Overall, keyword recognition for the target talker was poorer for older adults than younger adults. Keyword recognition was poorer on probe trials than standard trials for both groups, and incorrect responses on probe trials contained keywords from the single-talker masker. No interaction was observed between age-group and the decline in keyword recognition on probe trials. Thus, reduced performance by older adults overall could not be attributed to declines in attention to an expected voice. Rather, other cognitive abilities, such as speed of processing and linguistic closure, were predictive of keyword recognition for younger and older adults. Moreover, the effects of age interacted with the sex of the target talker, such that older adults had greater difficulty understanding target keywords from female talkers than male talkers.

Keywords

speech perception, aging, attention, talker identification, talker sex

Received 3 July 2019; Revised 2 March 2020; accepted 3 March 2020

When several sources of speech are present, attention must be directed to the appropriate talker, and competing talkers must be ignored. Whereas this basic role of attention is relatively straightforward, the process(es) by which attention facilitates speech segregation and selection has been debated for 60 years. Early theories of attention invoked the concept of a filter, wherein the unattended signals were either blocked or attenuated prior to higher level processing (Broadbent, 1958; Treisman, 1964). One limitation of these early works is that the specific dimension(s) operated on by the attentional filter is not made explicit (Bronkhorst, 2015). More recent work has described *object selection* as the process of choosing a particular auditory object (i.e., the talker) to be the focus of attention and subsequent higher level processing (Shinn-Cunningham, 2008). This is distinct from the process of separating a mixture of concurrent voices into separate auditory objects, referred to here as *perceptual segregation*. Object selection is often guided by a priori knowledge or expectations about the target such as an expected spatial location, overall level relative to the background, and/or voice features (Brungart, 2001; Kidd et al., 2005; Mackersie et al., 2011). Listener expectations of the target prime low levels of the auditory pathway for

Department of Otolaryngology—Head and Neck Surgery, Medical University of South Carolina

Corresponding Author:
William J. Bologna, Department of Veterans Affairs, National Center for Rehabilitative Auditory Research, Veterans Affairs Portland Health Care System, 3710 Southwest United States Veterans Hospital Road, P5-NCRAR, Portland, OR 97239, United States.
Email: Bolognaw@ohsu.edu

preferential representation of signals matching those expectations, such that perception of the auditory scene is biased toward the salient signals in the foreground, with irrelevant competing signals in the background (Kaya & Elhilali, 2014; Shamma et al., 2011). Thus, certain dimensions of listener expectations may describe the attentional filter. For example, expectations of a signal's pitch, duration, or temporal structure can improve the detection of simple acoustic signals, relative to equally detectable, but unexpected, signals (Dai & Wright, 1995; Dai et al., 1991; Greenberg & Larkin, 1968; Scharf et al., 1987; Schlauch & Hafter, 1991; White & Carlyon, 1997). Here, we investigate more complex acoustic signals (i.e., speech) to determine the extent to which expectations of a talker's voice improve object selection as indicated by improved speech recognition in a two-talker context.

Recently, Bronkhorst (2015) proposed a model of early speech processing that captures the interplay between bottom-up stimulus features and top-down attentional control. In this model, attentional control operates a feedback loop that facilitates *fast top-down selection* where primitive features, such as the voice characteristics of a talker, are compared with an attentional set determined by the listener's task and goals. Bottom-up sensory priming enhances the representation of expected signals prior to object selection. In a typical real-world scenario, this is akin to enhancing the salience of an expected voice in anticipation of that talker's turn in the conversation. Physiological evidence supports this interplay between attention and the sensory system; short-term plasticity of cortical spectrotemporal receptive fields is guided by task demands and functionally increases the contrast between target and competing objects in the auditory scene (see Fritz et al., 2007 for review). This contrast enables listeners to identify and track the target over time as well as avoid unwanted intrusions by irrelevant competing signals (Zion Golumbic et al., 2013). This complex network requires efficient use of processing resources as well as coordination between selective attention and inhibition, all of which may be susceptible to age-related declines.

Numerous studies have demonstrated that older adults are poorer than younger adults at understanding speech in multitalker environments, even after accounting for differences in hearing sensitivity (Bologna et al., 2018; Helfer & Freyman, 2008; Rajan & Cainer, 2008). In many cases, these results have been interpreted as age-related declines in perceptual segregation of speech (e.g., Ben-David et al., 2012; Ezzatian et al., 2015; Lee & Humes, 2012). Others have suggested a more general explanation, that older adults have slower cognitive processing or limited processing resources (Salthouse, 1996). Another possible explanation is that age-related declines in attention reduce the extent to which expected voices are enhanced relative to competing talkers. Best et al. (2018) demonstrated the effects of age and hearing loss on the ability to identify talkers based on their voice features, suggesting that older adults may be poorer than younger adults at learning and/or storing representations of a talker's voice. Studies of speech recognition with competing talkers have noted that responses from older adults are more likely than younger adults to contain words from the masker sentence (Helfer & Freyman, 2008; Lee & Humes, 2012). These *masker-intrusion errors* reflect a form of informational masking and can be used to quantify the extent to which competing signals intrude on the foreground of perception, presumably due to failures in selective attention or inhibition. Taken together, these results suggest that age-related declines in object selection may affect speech recognition with competing talkers for older adults, separately or in conjunction with declines in perceptual segregation.

Most real-world environments allow listeners to generate expectations of a talker's voice based, minimally, on the talker's sex. Talker sex affects primarily two acoustic characteristics: F0, corresponding to the rate of vocal fold vibration and the perception of voice pitch, and the overall shape of the spectral envelope, corresponding to vocal tract length and the perception of voice timbre (Darwin et al., 2003). Extensive research has demonstrated that these acoustic cues provide a basis for perceptual segregation of concurrent talkers (e.g., Darwin et al., 2003; Mackersie et al., 2011). These cues also serve a role in object selection by helping listeners generate expectations of the target voice and direct attention to the appropriate talker (e.g., Johnsrude et al., 2013; Newman & Evers, 2007). While perceptual segregation and object selection can be viewed as distinct processes, few studies have attempted to isolate their respective contributions to speech perception (Ihlefeld & Shinn-Cunningham, 2008). Since these two processes are often driven by the same cues, one of the challenges in studying object selection is to control for potential effects associated with perceptual segregation.

The purpose of this study was to compare the effect of attentional filtering by voice features between younger and older adults. This was achieved by designing a speech recognition task where the burden of perceptual segregation was equivalent across trials, but the benefit of focused attention to an expected voice was variable across trials. Listeners' expectations of the target voice were manipulated by adjusting both F0 and spectral envelope to alter the perceived sex of the talker from relatively more male-like to relatively more female-like or vice versa (i.e., Darwin et al., 2003). We hypothesized that speech recognition would be best when listeners could use their expectations of the target's voice to identify and attend to the target talker. In contrast, speech recognition was expected to decline when the target's voice features differed from the listener's expectation. We predicted that age-related declines in attention would be revealed by an interaction effect, such that older adults would demonstrate less benefit from focused attention to voice features than younger adults. Younger and older adults completed the speech recognition task to test these predictions and determine the extent to which the benefit of voice expectations declines with age. Responses were scored for correct keywords (i.e., from the target talker) as well as for masker-intrusion errors (i.e., from the competing talker) to test the hypothesis that age-related declines in keyword recognition of the target talker can be partially explained by an increase in intrusions from the competing talker.

Methods

Participants

Two groups were tested, 20 younger adults ranging in age from 18 to 29 years (*M*: 24.7 years, *SD*: 2.8 years) and 20 older adults ranging in age from 63 to 84 years (*M*: 69.9 years, *SD*: 5.7 years). Hearing sensitivity was assessed in all the participants by measuring air-conduction thresholds at audiometric frequencies (American National Standard Institute, 2010). Hearing thresholds for younger participants were ≤25 dB HL for .25 to 8.0 kHz. For older participants, thresholds were ≤30 dB HL for .25 to 6.0 kHz. A pure-tone average (PTA; 0.5, 1.0, 2.0, 4.0 kHz) was calculated for each participant so that the effects of hearing sensitivity could be modeled along with other subject- and task-related factors. Each participant reported their level of education in years of formal schooling (i.e., 12 years for high school diploma, 16 years for 4-year bachelor's degree, etc.). The two age groups reported similar levels of education (younger group *M*: 16.8 years, *SD*: 2.3 years; older group *M*: 16.4 years, *SD*: 2.4 years). Normal cognitive functioning of older participants was confirmed by a score of 25 or greater on the Mini Mental Status Examination (Folstein et al., 1975). All participants were native speakers of American English and reported normal or corrected-to-normal vision. Participants were paid for their participation and gave informed consent for the protocol, as approved by the Institutional Review Board at the Medical University of South Carolina (Pro00031785).

Stimuli and Apparatus

Speech stimuli were sentences spoken by male and female talkers from the TIMIT corpus (e.g., "A huge tapestry hung in her hallway"; Garofolo et al., 1993), which were compiled into phonetically balanced lists for the Perceptually Robust English Sentence Test Open-set (PRESTO; Gilbert et al., 2013). These materials include sentences from 630 unique talkers from 8 dialect regions in the United States. Each PRESTO list contains 18 sentences spoken by 18 talkers (9 males and 9 females). The diversity and unpredictability of talkers in this corpus were ideal for this experiment, where the listener's expectation of the target talker's voice features was manipulated. In addition, this allowed for a generalizable analysis of sentences with male versus female target talkers to determine whether talker sex influences the pattern of results or interacts with the age of the listener.

Sentences from the PRESTO lists served as targets and were paired with equal duration, opposite-sex competing talker sentences from the TIMIT corpus. To manipulate voice features, target and competing sentences were processed in Praat with the Pitch Synchronous Overlap and Add algorithm (PSOLA) (Moulines & Charpentier, 1990), such that each sentence pair had an eight-semitone difference in F0. For each sentence pair, the geometric mean between male and female F0 was calculated to serve as a midpoint. Next, pitch contours for each sentence were extracted and shifted higher or lower in frequency such that the mean F0 was exactly four semitones above or below the midpoint. This process resulted in sentences with natural pitch contours and variations in F0 across male and female talkers, while maintaining an equivalent eight-semitone difference in average F0 for each sentence pair.

In addition to shifting the pitch contour, the spectral envelopes of the target and competing talkers were manipulated in Praat based on methods described by Darwin et al. (2003). Linear extrapolation of average male/female ratios for the formant and F0 data reported by Peterson and Barney (1952) were used to obtain spectral envelope shift factors (vt) corresponding to a specific semitone shift of F0. For a given sentence, the semitone shift in F0 applied to the pitch contour was used to find the corresponding vt value to scale the spectral envelope. This method of changing the spectral envelope is similar, but not identical, to a true change in vocal tract length and has been used in previous studies of talker sex differences in speech recognition with competing talkers (Darwin et al., 2003; Mackersie et al., 2011). Example spectrograms are displayed in Figure 1 for a single sentence that has been processed for a downward shift in voice features (*male talker*) or an upward shift in voice features (*female talker*). After processing, the mean F0 was 125.2 Hz (*SD* = 11.1 Hz) for male talkers and 198.3 Hz (*SD* = 17.7 Hz) for female talkers.

For each target talker, a cue phrase ("... greasy wash water all year") was excised from a standard sentence in the TIMIT corpus spoken by that target talker ("She had your dark suit in greasy wash water all year").

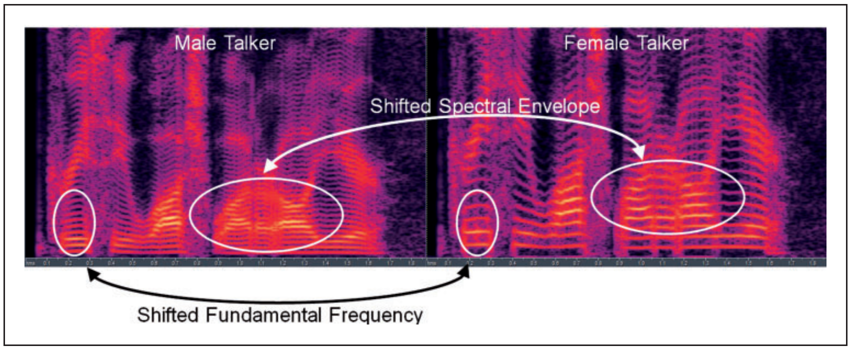

Figure 1. Example spectrograms illustrating a sentence processed for male voice features (left panel) and the same sentence processed for female voice features (right panel). When processing a male talker for female voice features, the fundamental frequency is shifted to a higher voice pitch and spectral envelope is broadened such that formants occur in a higher frequency range.

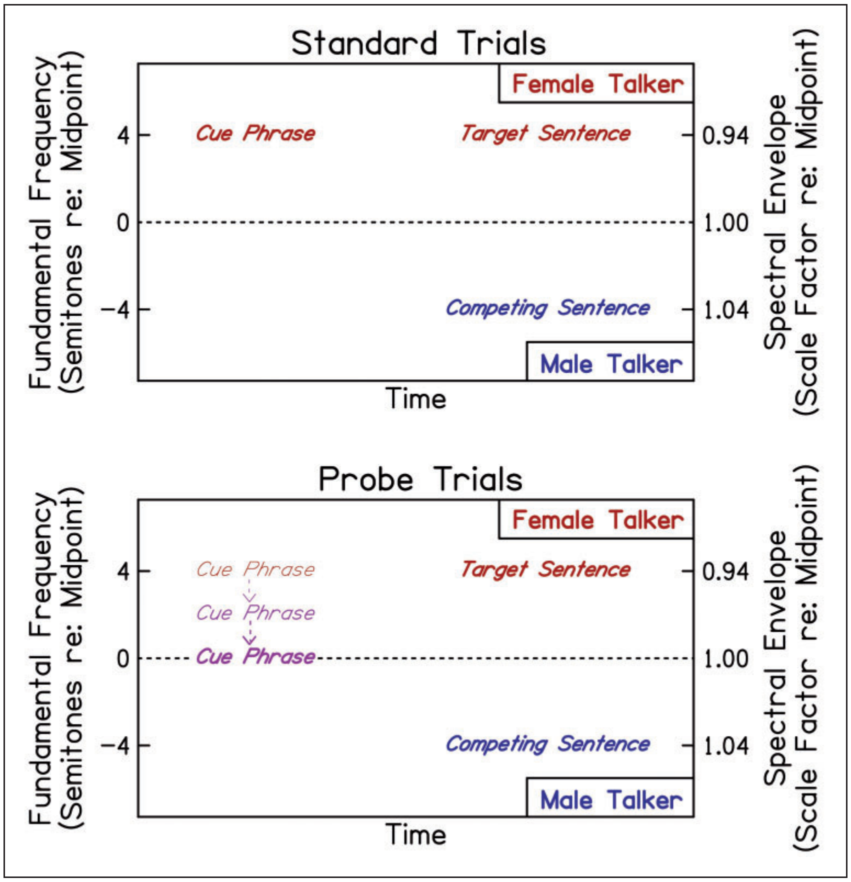

Figure 2. Schematic diagram of standard trials (top panel) and probe trials (bottom panel). All trials began with the cue phrase, followed by 1.5 s of silence, and then the target and competing talker mixture. On standard trials, F0 and spectral envelope of the cue phrase matched the target talker exactly. On probe trials, F0 and spectral envelope of the cue phrase were parametrically shifted toward the midpoint between the target and competing talker.

These cue phrases were processed using the same methods described earlier to create *standard trials* (no change in voice features) and *probe trials* (parametric changes in voice features). Schematic diagrams of the two trial types are shown in Figure 2. The majority of trials were standard trials (top panel), in which the cue phrase was processed such that F0 and spectral envelope matched those of the target talker (i.e., ±4.0 semitones from the midpoint). For a small percentage of randomly occurring probe trials (bottom panel), the cue phrase was processed such that F0 and spectral envelope were shifted 3.0, 3.5, or 4.0 semitones toward the midpoint between the target and competing talker. Note that the four-semitone shift placed the cue phrase at the exact midpoint between the target and competing talker, such that the F0 and spectral envelope of the cue phrase provided ambiguous information on the identity of the target. Whereas the voice shifts only affected the F0 and spectral envelope of the cue phrase, suprasegmental features such as prosody and intonation remained similar between the cue phrase and the target. For all trials, the cue phrase was followed by 1.5 s of silence and then the target/competing talker sentence mixture. Thus, the only difference between standard trials and probe trials was the voice characteristics of the cue phrase, which either matched the target (standard trials) or did not match the target (probe trials).

A total of 16 PRESTO lists were processed as standard trials and 3 PRESTO lists were processed as probe trials, corresponding to the 3 shift conditions (3.0, 3.5, and 4.0 semitones). Each PRESTO list contained 76 preidentified keywords across 18 sentences (3–6 keywords per sentence). To match the number of target keywords, 76 important content words were selected from the competing talker sentences to serve as masker keywords. In some cases, the number of target and masker keywords differed slightly for a specific sentence pair, but each list pair contained a total of 76 target and 76 masker keywords. There were 54 probe trials (3 lists × 18 sentences per list) and 288 standard trials (16 lists × 18 sentences per list); that is, probe trials accounted for 18.75% of all trials. Probe trials were randomly assigned to lists with the restriction that each list contained at least 1 probe trial for each shift condition. The order of sentences within each list was randomized for each participant to ensure that probe trials occurred randomly throughout testing and were not presented in a consistent pattern across participants. The last remaining PRESTO list was reserved to be used as a practice list containing only standard trials.

Stimuli were generated prior to data collection and saved as separate .wav files with 16-bit resolution at a sampling rate of 16000 Hz. The target and competing talker mixtures were presented at 78 dB sound pressure level (SPL) with a 0 dB signal-to-noise ratio (SNR; 75 dB). These levels were verified by acoustic calibration using an acoustic coupler with a Larson Davis model 2559 $^1/_2$-in. microphone and a Larson Davis Model 824 sound level meter with flat weighting.

Procedures

Measurement of speech recognition was completed in a sound-attenuating booth. Participants were instructed to listen to the sentence mixture and repeat the sentence spoken by the target talker and ignore the other voice. They were instructed to use the voice characteristics of the cue phrase preceding the sentence mixture to identify the target talker. They were not informed about the manipulation of voice characteristics or the presence of probe trials. Presentation and keyword scoring were controlled by Token software (Kwon, 2012) with a Lynx Two multichannel audio interface, Tucker–Davis Technologies programmable attenuator (PA4), Tucker–Davis Technologies headphone buffer (HB6), and Sennheiser HDA 200 headphones. Stimuli were presented monaurally to the right ear unless that ear did not meet hearing criteria; four older participants were presented with speech to their left ear. Participant responses were scored for target keywords online by the experimenter using a strict scoring rule (i.e., no missing or additional suffixes) without requiring correct word order. Responses were also recorded using a Realistic Highball Dynamic 33-984C microphone so that they could be subsequently rescored for masker-intrusion errors.

Additional data from a battery of cognitive measures were available for these participants from two other studies completed at the same time (Bologna et al., 2018, 2019). The cognitive measures were included in statistical models described later to assess the contributions of several cognitive abilities to speech recognition and object selection. The test battery included measures of processing speed (Connections, Salthouse et al., 2000; Purdue Peg Board, Tiffin & Asher, 1948), working memory capacity (Reading Span, Daneman & Carpenter, 1980; Rönnberg, 1990), inhibitory control (Stroop Test, Stroop, 1935; Trenerry et al., 1989), and visual linguistic closure (Text Reception Threshold [TRT], Zekveld et al., 2007); descriptions of these measures and scoring can be found in Bologna et al. (2018). Initial modeling results revealed collinearity between age and cognitive factors, which was resolved by residualizing cognitive factors for effects of age. As such, the cognitive factors reflect variance in cognitive abilities within each age-group.

Results

Keyword recognition for the target talker (percent correct) and the masker (percent masker-intrusion errors) are plotted in Figure 3. These data were analyzed using item-level logistic regression implemented in R (R Development Core Team, 2016) using a Generalized Linear Mixed Model (GLMM; R-package: lme4; Bates et al., 2015). Two separate GLMMs were constructed; one for target keywords and one for masker-intrusion errors. Each model specified recognition of each individual keyword as the dependent variable ($W = 0$ if the subject's response did not contain the keyword, 1 if the response included the keyword) and estimated separate β coefficients for each independent variable included in

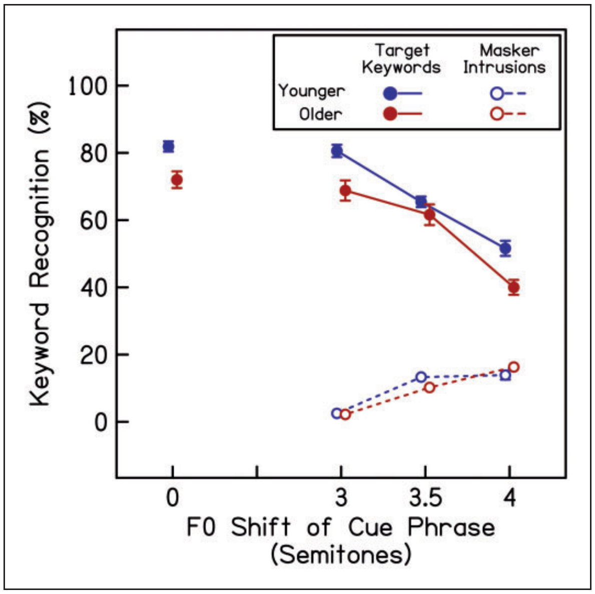

Figure 3. Recognition of keywords in target sentences in percent correct (filled symbols with solid lines) and masker-intrusion errors (open symbols with dotted lines) for younger (blue) and older (red) adults are plotted as a function of semitone shift of the cue phrase. Symbols are offset for clarity. Keyword recognition was higher for standard trials (shift of 0 semitones, left most data points) and declined for probe trials with increasing shift of the cue phrase. Masker-intrusion errors were rare on standard trials and are not plotted. Masker-intrusion errors increased for probe trials with increasing shift of the cue phrase. Keyword recognition for target sentences was higher for younger than older adults for standard and probe trials, but both groups demonstrated a similar number of masker-intrusion errors on probe trials.

the model. GLMM analyses were performed across all keywords independently for all participants with the following factors: age-group of the participant (*Age*), sex of the target talker (*Sex*), voice shift of the cue phrase (*Shift*, treated as a single continuous predictor), number of keywords in the sentence (*nWords*), position of the keyword within the sentence (*Pos*), participant's scores on the cognitive measures, participant's PTA, participant's level of education, interactions between these factors, and random subject effects (*Subj*). A combination of stepwise factor addition and elimination was performed using model testing (Hofmann, 1997) to optimize model fit with all significant factors and interactions.

Data were scored at the keyword level (rather than the sentence level) so that the effect of keyword position could be included in the model. This factor has been shown in previous studies to predict keyword recognition and provides some insight into the time course of speech segregation and selection (e.g., Ben-David et al., 2012; Bologna et al., 2018). Initial attempts to model the data included a sentence identifier as an additional random effect to improve model fit. However, computational overhead associated with attributing variance to the nearly 250 different sentences in the dataset was intractable. Thus, keyword recognition was modeled independently for each keyword without sentence-level nesting. Overall fit of the model was evaluated by binning data points by their predicted probability from the model and calculating the observed probability of a correct response within each bin. A plot of these values revealed close correspondence between the observed data and the model predictions, even without accounting for the interdependence of keyword recognition within a sentence.

Target Keywords

Table 1 shows factor descriptions, coding details, standard estimates, standard errors, and z statistics for each significant factor in the target keyword model, including split factors from post hoc models. Modeling results indicated that younger adults significantly outperformed older adults (Figure 3, blue vs. red; $\beta_{Age} = -0.28$; $z = -4.83$; $p < .001$). Target keyword recognition was higher for sentences with a female target talker than a male target (Figure 4, right vs. left; $\beta_{Sex} = 0.09$; $z = 8.33$; $p < .001$). Keyword recognition declined on probe trials with increasing voice shift of the cue phrase (Figure 3; $\beta_{Shift} = -0.25$; $z = -34.22$; $p < .001$), indicating that the shifted voice characteristics of the cue phrase misled the listeners' expectations of the target voice, leading to poorer keyword recognition. Target keyword recognition was better for sentences with fewer keywords ($\beta_{nWords} = -0.13$; $z = -12.61$; $p < .001$) and was better for keywords at end of sentences compared with the beginning ($\beta_{Pos} = 0.34$; $z = 32.92$; $p < .001$). Participants with faster speed of processing and better linguistic closure, as measured by Connections and TRT, respectively, were more likely to respond with correct target keywords ($\beta_{Connections} = 0.20$; $z = 3.44$; $p < .001$; $\beta_{TRT} = 0.16$; $z = 2.72$; $p < .01$). Three interaction terms significantly contributed to the fit of the model; *Sex* × *Age*, *Shift* × *Sex*, and the three-way interaction *Age* × *Sex* × *Shift*. These interactions were interpreted using separate post hoc models with split factors to describe the effect of a given factor across the two levels of its interacting factor.

The effect of target talker sex interacted with listener age-group ($\beta_{Sex \times Age} = –0.10$; $z = –8.42$; $p < .001$). Post hoc modeling indicated that the effect of talker sex was driven by the younger group, whose keyword recognition was better for female targets than male targets (average of 83.2% vs. 75.4%). In contrast, older adults did not demonstrate this asymmetry in performance (70.0% for female targets, 69.2% for male targets). This suggests that for younger listeners keywords

Table 1. Target keyword GLMM factors, coding, standard estimates, standard error, and z statistics are displayed for each significant fixed effect and interaction term.

Factor	Description and Coding	Standard Estimate (β)	Standard Error	z Statistic
Age	Age-group of listener (older = 1; younger = −1)	−0.28	0.06	−4.83***
Sex	Sex of target talker (female = 1; male = −1)	0.09	0.01	8.33***
Shift	Shift in voice characteristics of cue phrase (0 for standard trials; 3.0, 3.5, or 4.0 for probe trials)	−0.25	0.01	−34.22***
nWords	Sentence length (number of keywords in the sentence, normalized $M = 0$, $SD = 1$)	−0.13	0.01	−12.61***
Pos	Position of keyword within the sentence (serial order position of keyword scaled by total number of keywords in sentence, normalized $M = 0$, $SD = 1$)	0.34	0.01	32.92***
Connections	Listener's score on Connections Test (residualized for effects of age, normalized for $M = 0$, $SD = 1$)	0.20	0.06	3.44***
TRT	Listener's score on Text Reception Threshold Test (residualized for effects of age, normalized for $M = 0$, $SD = 1$)	−0.16	0.06	2.72**
Sex × Age	Interaction between target sex and age-group of listener	−0.10	0.01	−8.42***
Sex_O	Effect of target sex for older listeners	<0.01	0.01	<0.01*ns*
Sex_Y	Effect of target sex for younger listeners	0.19	0.02	11.12***
Shift × Sex	Interaction between voice shift of cue phrase and target sex	0.09	0.01	11.98***
Shift_F	Effect of voice shift on female targets	−0.16	0.01	−15.67***
Shift_M	Effect of voice shift on male targets	−0.33	0.01	−33.61***
Age × Sex × Shift	Interaction between age-group, target sex, and voice shift	−0.03	0.01	−3.54**
Shift_F × Age	Interaction between voice shift and age-group for female targets	<0.01	0.01	−0.27*ns*
Shift_M × Age	Interaction between voice shift and age-group for male targets	0.05	0.01	4.63***
Shift_M_O	Effect of voice shift on male targets for older listeners	−0.26	0.01	−19.23***
Shift_M_Y	Effect of voice shift on male targets for younger listeners	−0.41	0.01	−30.09***

Note. Each interaction was explored with a separate post hoc model with split factors, and statistical results are indented below interactions. Asterisks indicated significance levels for z statistics (***$p < .001$; **$p < .01$). *ns* = not significant; TRT = Text Reception Threshold.

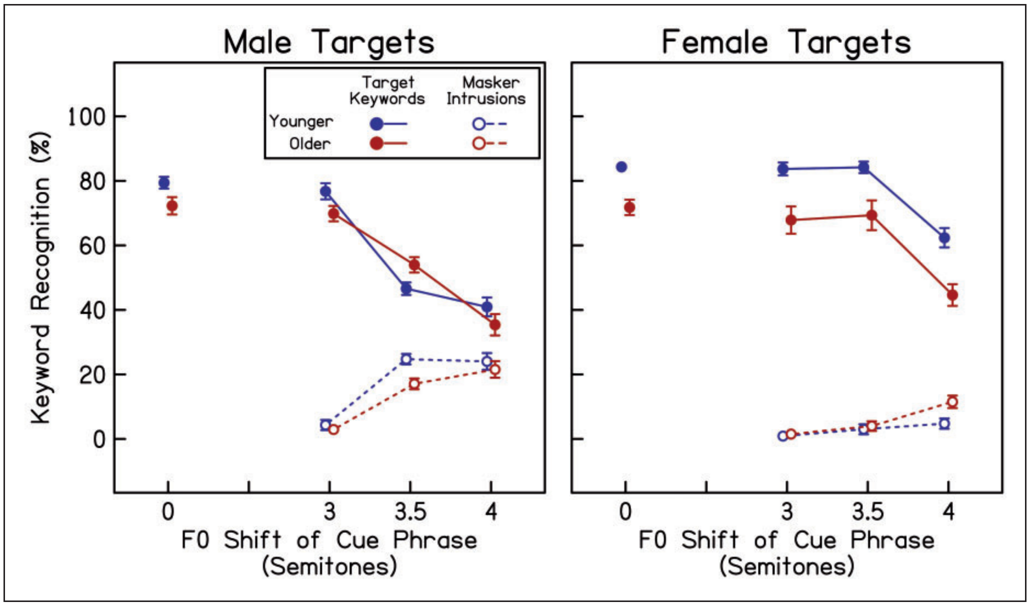

Figure 4. Recognition of keywords in target sentences in percent correct (filled symbols with solid lines) and masker-intrusion errors (open symbols with dotted lines) for younger (blue) and older (red) adults are plotted as a function of semitone shift of the cue phrase for male target talkers (left panel) and female target talkers (right panel). Symbols are offset for clarity. Standard trials are presented on the far left in each panel. Probe trials with male target talkers demonstrate poorer performance and more masker-intrusion errors than trials with female target talkers. Keyword recognition for target sentences with female targets was higher for younger than older adults for standard and probe trials, but the two groups performed more similarly for target sentences with male targets.

spoken by female talkers were easier to recognize than those by male talkers, but older adults recognized keywords equally well for male/female talkers.

The effect of shifting the voice of the cue phrase also interacted with the sex of the talker ($\beta_{Shift \times Sex} = 0.09$; $z = 11.98$; $p < .001$). Post hoc modeling indicated that shifting the voice characteristics of the cue phrase was more disruptive to recognition of male target keywords than female keywords; as the cue phrase was shifted from three semitones to four semitones, keyword recognition decreased from 75.8% to 53.5% for female targets and 73.3% to 38.2% for male targets. This effect can be observed in Figure 4, where keyword recognition declines more precipitously in probe trials with male target talkers (left) than in probe trials with female target talkers (right) for both younger and older adults.

The three-way interaction between age-group, talker sex, and voice shift also improved the fit of the model ($\beta_{Age \times Sex \times Shift} = -0.03$; $z = -3.54$; $p < .01$). To explore this interaction, it was modeled as the interaction between age-group and voice shift for male targets and the (nonsignificant) interaction between age-group and voice shift for female targets (see Table 1). Because the effect of voice shift for female targets did not interact with age-group, this interaction term was replaced with the factor describing the effect of voice shift for female targets described earlier (*Shift_F*). The significant two-way interaction between age-group and voice shift for male targets was split into two factors describing the effect of voice shift for male targets among younger adults and among older adults (see Table 1). These results indicated that shifting the voice characteristics of the cue phrase for male target talkers was more disruptive for younger adults than older adults. Taken together, the pattern of significant effects and interactions indicates that shifting the voice characteristics of the cue phrase was most disruptive to keyword recognition when the target was male and the listener was younger, less disruptive when the target was male and the listener was older, and least disruptive when the target was female for all listeners (similar effect for both age groups).

Many factors and interaction terms were tested for significance using model testing but did not significantly improve the fit and were not included in the final models. Most notable were effects of hearing sensitivity, based on PTA. Whereas all participants were selected to have essentially normal hearing, subtle threshold elevation may have still affected speech recognition. Effects of PTA were modeled to explore this possibility, but model testing revealed that this factor did not significantly improve model fit. However, model results indicated collinearity between the effects of hearing sensitivity and age, which can disrupt reliable estimation of regression coefficients. To address this potential problem, PTA was residualized for effects of age and reentered into the model, but did not significantly improve model fit ($\chi^2 < 2.53$, *ns*). These results indicate that a measure of hearing sensitivity was not a good predictor of keyword recognition for these participants in this two-talker speech recognition task.

Masker-Intrusion Errors

Table 2 shows factor descriptions, coding details, standard estimates, standard errors, and *z* statistics for each significant factor in the masker-intrusion error model,

Table 2. Masker-intrusion error GLMM factors, coding, standard estimates, standard error, and z statistics are displayed for each significant fixed effect and interaction term.

Factor	Description and Coding	Standard Estimate (β)	Standard Error	z Statistic
Sex	Sex of target talker (female = 1; male = −1)	−0.39	0.05	−8.37***
Shift	Shift in voice characteristics of cue phrase (0 for standard trials; 3.0, 3.5, or 4.0 for probe trials)	0.64	0.02	37.45***
Pos	Position of keyword within the sentence (serial order position of keyword scaled by total number of keywords in sentence, normalized *M* = 0, *SD* = 1)	0.22	0.03	7.73***
Connections	Listener's score on Connections Test (residualized for effects of age, normalized for *M* = 0, *SD* = 1)	−0.21	0.08	−2.59**
Sex × Age	Interaction between target sex and age group of listener	0.22	0.03	7.18***
Sex_O	Effect of target sex for older listeners	−0.17	0.05	−3.12**
Sex_Y	Effect of target sex for younger listeners	−0.61	0.06	−10.44***
Shift × Sex	Interaction between voice shift of cue phrase and target sex	−0.10	0.02	−5.77***
Shift_F	Effect of voice shift on female targets	0.55	0.03	19.05***
Shift_M	Effect of voice shift on male targets	0.74	0.02	39.08***

Note. Each interaction was explored with a separate post hoc model with split factors, and statistical results are indented below interactions. Asterisks indicated significance levels for z statistics (***$p < .001$; **$p < .01$).

including split factors from post hoc models. Note that this model is based on keyword responses from the single-talker masker sentence, and so significant effects in the model are interpreted as increasing the likelihood of a masker-intrusion error. Modeling results indicated that masker-intrusion errors were more likely to occur as the voice of the cue phrase was shifted toward the midpoint between the target and masker voice ($\beta_{Shift} = 0.64$; $z = 37.45$; $p < .001$). Masker-intrusion errors were also more likely to occur for sentences with male target talkers than female targets ($\beta_{Sex} = -0.39$; $z = -8.37$; $p < .001$). Masker keywords at the ends of sentences were more likely to be reported by listeners than those at the beginnings of sentences ($\beta_{Pos} = 0.22$; $z = 7.73$; $p < .001$). Participants with faster speed of processing, as measured by Connections, were less likely to respond with masker keywords ($\beta_{Connections} = -0.21$; $z = -2.59$; $p < .01$). Two interaction terms significantly contributed to the fit of the model; *Sex* × *Age*, *Shift* × *Sex*. Post hoc models with split factors were constructed to explore significant interaction terms.

The effect of target talker sex on masker-intrusion errors interacted with listener age-group ($\beta_{Sex \times Age} = 0.22$; $z = 7.18$; $p < .001$). Post hoc modeling indicated that the effect of talker sex on masker-intrusion errors was greater for the younger group than the older group (see Table 2). Masker-intrusion errors were more common for male targets than female targets for both younger and older adults, and the asymmetry in masker-intrusion errors by talker sex was greater for younger adults than older adults.

The effect of shifting the voice characteristics of the cue phrase also interacted with the sex of the talker for masker-intrusion errors ($\beta_{Shift \times Sex} = -0.10$; $z = -5.77$; $p < .001$). Post hoc modeling indicated that shifting the voice characteristics of a male cue phrase was more likely to result in a masker-intrusion error than the same shift applied to a female voice. The asymmetry of this effect can be observed in Figure 4; dotted lines indicate masker-intrusion errors increased from 3.6% to 22.8% for probe trials with male target talkers (left), whereas masker-intrusion errors increased from 0.7% to 8.1% for probe trials with female talkers (right). This asymmetry mirrors the asymmetry in target keyword recognition described earlier, such that masker-intrusion errors were greatest for conditions in which keyword recognition was poorest.

Discussion

This study investigated the effects of age and listener expectations on object selection using a speech recognition task with a competing talker. Sentence mixtures were spoken by male and female talkers with a standard difference in F0 and spectral envelope corresponding to an eight-semitone difference in F0. The eight-semitone difference ensured that the effects of these acoustic characteristics on perceptual segregation were roughly equivalent across trials, and both target and masker sentences were equivalently processed to prevent any processing artifacts from serving as a cue to distinguish the target from masker. Listener expectations of the target talker's voice features were manipulated with a cue phrase that preceded each trial. On most trials (standard trials), F0 and spectral envelope of the cue phrase were identical to the target, which facilitated accurate selection of the target talker in the two-talker mixture. For a small percentage of randomly occurring probe trials, F0 and spectral envelope of the cue phrase were parametrically shifted toward the competing talker's voice. Larger shifts in the voice characteristics of the cue phrase resulted in poorer keyword recognition and more masker-intrusion errors. These results suggest that listeners used the voice characteristics of the cue phrase to prime their attention for selection of an upcoming expected voice, resulting in enhanced recognition on standard trials compared with probe trials.

In the context of the model proposed by Bronkhorst (2015), standard trials would elicit fast top-down selection based on the listener's expectation of hearing specific voice features in the mixture. The unexpected deviation in voice features on probe trials would activate a slow selection process; listeners would compare their memory trace of the cue to the individual voices in the two-talker mixture in sensory memory to find a closest match. This process is slower and more prone to error because the sensory trace is decaying while the listener is selecting a target. We observed the expected decline in performance; probe trials contained fewer correct keywords and more masker-intrusion errors than standard trials. However, the decline in keyword recognition was similar for younger and older adults. This result suggests that advancing age does not affect fast top-down selection or the sensory priming mechanisms that underlie fast top-down selection. Rather, advancing age is better characterized by broad changes and generalized effects on object selection.

The pattern of decline in performance on probe trials suggests fairly broad tuning of attention for voice features in a two-talker context (i.e., Scharf et al., 1987; Schlauch & Hafter, 1991). Probe trials in which the voice of the cue phrase was shifted by three semitones toward the midpoint between talkers resulted in essentially equivalent performance to standard trials. In addition, performance was greater than chance for probe trials in which the voice of the cue phrase was shifted by four semitones (i.e., the midpoint between target and competing talkers). An intuitive hypothesis would propose that when the voice characteristics of the cue phrase was positioned midway between the target and competing talker, listeners would be forced to select a talker at random, resulting in equal numbers of target keywords correct and masker-intrusion errors. However, responses from both younger and older adults for the four-semitone probe trials contained more correct keywords than masker-intrusion errors, suggesting that performance of both groups was greater than chance for this ambiguous condition for selection of the correct target talker. A likely explanation for this effect is that listeners relied on other cues to identify the target talker when F0 and spectral envelope cues were ambiguous. For example, the TIMIT corpus contains talkers from many dialect regions. Because cue phrases were always taken from the target talker's recording of a standard sentence, listeners may have been able to identify the target talker based on dialectal variations, prosody, intonation, and other suprasegmental cues present in both the cue phrase and the target sentence. Thus, broad attentional tuning for voice features may reflect the same multidimensionality of attentional tuning that has been observed with nonspeech sounds (e.g., Dai & Wright, 1995; White & Carlyon, 1997; Wright & Dai, 1994).

The pattern of performance on standard and probe trials can be interpreted in terms of the benefit of voice-feature continuity from the cue phrase to the target. The benefit of voice-feature continuity for attention was proposed by Bressler et al. (2014) and recently supported by Kreitewolf et al. (2018). In these studies, participants listened to digit sequences in a background of competing digit sequences. In separate conditions, the digits in the target sequence were spoken either by the same talker or by different talkers. When the voice of the target talker changed from digit-to-digit, listeners identified fewer correct digits and made more masker-intrusion errors compared with sequences with a consistent target voice (Bressler et al., 2014). Similar results were observed by Kreitewolf et al. (2018) for digit sequences in which F0 and/or spectral envelope were either consistent across the sequence or changed from digit-to-digit. In both previous studies, digit recognition improved over the course of the sequence, similar to the pattern of improvement observed with later keyword position in this study. Importantly, the benefit of voice-feature continuity was greatest when the listener correctly identified the previous digit in the sequence. This was interpreted as evidence that voice continuity automatically primes the listener to organize the auditory scene with the previously selected voice in the foreground. This continuity benefit was described as obligatory or automatic because it was observed even in conditions in which listeners were unable to predict whether the target voice would be consistent or change from digit-to-digit. The assumption in this study was that listeners were *intentionally* focusing attention on the voice features of the cue phrase in anticipation of the need to segregate the two-talker mixture that followed. The task and instructions in this study encouraged this goal-directed behavior, which was expected to result in preferential representation of the expected voice in the perceptual foreground. Regardless of whether this effect is achieved consciously or subconsciously, the benefit to speech recognition would be strongest on standard trials and weaker on probe trials as the voice features of the cue phrase were shifted toward the midpoint between the target talker and competing talker. Thus, the continuity interpretation is consistent with the pattern of results observed in this study. However, we expect that any automatic benefit of voice continuity would be considerably weaker in this study than reported previously, as our stimulus design included a 1.5-s period of silence between the cue phrase and the sentence mixture, as opposed to the 50-ms inter-digit intervals used to demonstrate continuity benefit in previous studies (Kreitewolf et al., 2018). Nevertheless, the current results complement these recent findings by demonstrating that listener expectations of voice features can also guide attention and facilitate organization of an auditory scene.

Effects of Age on Object Selection

We predicted that age-related declines in object selection would result in poorer keyword recognition by older adults compared with younger adults on standard trials, where focused attention on the voice characteristics of the cue phrase would be most beneficial. On probe trials, the benefit of focused attention on the voice characteristics of the cue phrase would be reduced, as these voice characteristics did not provide an accurate method of selecting the target talker from the mixture. Performance of younger and older adults was expected to converge on probe trials as the cue phrase was shifted further toward the midpoint between the target and competing talker. Limited support for this hypothesis can be drawn from the data. Older adults performed more poorly than younger adults overall but shifting the voice characteristics of the cue phrase had a similar effect on performance by younger and older adults (see Figure 3). Thus, the sensory priming afforded by the cue phrase had a similar effect on performance among the two groups. Separating trials based on the sex of the target talker revealed subtle differences in the pattern of decline on probe trials for younger and older adults. When the talker was male (see Figure 4), shifting the cue phrase disrupted performance more for younger adults than older adults, but the effect was modest and no interaction with age was observed for female talkers. Collapsed across talker sex, the effect of shifting the cue phrase was similar for younger and older adults (see Figure 3). Thus, older adults performed more poorly than younger adults regardless of whether focused attention to an expected voice facilitated performance overall.

Better keyword recognition and fewer masker-intrusion errors were predicted by faster speed of processing. That is, age-related declines in speed of processing (e.g., Salthouse, 2000) may have limited the extent to which older adults could quickly attend to an unexpected voice on probe trials. In contrast, variance in working memory capacity did not predict performance, suggesting that overall processing capacity was less critical for keyword recognition than efficient use of available cognitive resources (as measured by Connections). Additional variance in keyword recognition was predicted by the TRT, a measure of the use of partial linguistic information (Zekveld et al., 2007). The TRT has been shown to predict performance on a variety of speech recognition tasks, further supporting the hypothesis that an amodal cognitive ability to use partial linguistic information is important for speech recognition in challenging listening conditions (Bologna et al., 2018, 2019; George et al., 2007; Humes et al., 2013; Krull et al., 2013).

Effects of Talker Sex

The sex of the target talker had a widespread effect on the results, both in terms of the number of correct target keywords and the number of masker-intrusion errors on probe trials. In general, probe trials with a male target talker contained fewer correct keywords and more masker-intrusion errors for both younger and older adults. This suggests that shifting the voice characteristics of a male cue phrase to make it sound more like a female talker was more disruptive to object selection than the opposite shift applied to a female voice. The reason for this asymmetry is unclear. On probe trials, the voice characteristics of the cue phrase were shifted toward a midpoint between talkers that was determined mathematically, rather than a *perceptual* midpoint between voices that sound male-like versus female-like. Characteristics of female voices vary over a larger range than male voices (Peterson & Barney, 1952). Thus, the voice characteristics of the cue phrase on the most difficult probe trials (four-semitone shift) may have fallen more in line with the perception of a low-pitched female voice than a high-pitched male voice. Alternatively, the asymmetry may be related to the effective masking of a female voice by a male voice and vice versa. In this experiment, the two talkers were always opposite sex, and so we cannot disentangle poorer intelligibility of a male target talker (relative to a female target talker) from greater masking effects by a female masker (relative to a male masker). Similar asymmetry in sentence recognition with opposite sex maskers has been noted previously with other speech corpora (Gallun & Diedesch, 2013). Future research using this same task design can distinguish between these alternatives by including same-sex talker pairs.

The sex of the target talker differentially affected the performance of younger and older adults. Across standard and probe trials, younger adults performed better for recognition of female target talkers relative to male target talkers, whereas older adults did not demonstrate this asymmetry. Difficulty understanding female voices is a very common complaint among older adults, even among those with relatively normal hearing thresholds, like the older participants in this study. Investigations of the effects of talker sex and age are sparse but provide some insight. Mackersie et al. (2011) noted that the benefit of increasing F0 difference between talkers was only apparent when the target had the higher F0. Age-related declines have been shown for the ability to distinguish talker sex for noise-band vocoded speech (Schvartz & Chatterjee, 2012). Older adults are also poorer than younger adults at identifying talkers they have heard previously (Best et al., 2018; Yonan & Sommers, 2000). The generalizability of these results, particularly for evaluating asymmetric effects for male/female voices, are limited by the relatively small number of different talkers typically included as stimuli. In contrast, stimuli used in this study come from an extremely diverse sentence corpus (Gilbert et al., 2013), characterized by large numbers of male and female talkers from a broad range of dialect regions around the United States. This provides a means for evaluating talker sex effects that are relatively unaffected by talker-specific findings noted in previous studies. The pairing of opposite sex talkers in this study means that an age-dependent asymmetry in masking effects cannot be ruled out. However, across a wide range of male and female voices, older adults were particularly poor at recognizing speech from female talkers masked by male competing talkers.

Conclusions

Listeners can use expectations of a talker's voice to attend to a sentence spoken by that talker and ignore competing speech. In this study, when the target voice unexpectedly deviated from the listener's expectations, speech recognition declined and masker-intrusion errors increased. However, declines in speech recognition were observed only when the target voice deviated considerably from listener expectations, particularly when the target voice was female. Older adults performed more poorly than younger adults overall, but results did not support the hypothesis that age-related declines in attention underlie age-related differences in performance. Other cognitive factors, such as speed of processing and linguistic closure, may contribute to the overall decline in speech recognition with competing talkers among older adults.

Acknowledgments

This investigation was conducted in a facility constructed with support from Research Facilities Improvement Program Grant Number C06 RR14516 from the National Center for Research Resources, National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was completed as part of the first author's dissertation to fulfill the requirement for PhD at the University of Maryland. The authors would like to thank University of Maryland dissertation committee members Monita Chatterjee, Sandra Gordon-Salant, Rochelle Newman, and Jonathan Simon for their advice and support on this work, Lois Matthews for her assistance with participant recruitment, and Larry Humes for providing the English version of the TRT.

Authors' Note

Portions of this article were previously presented at the *Spring Meeting of the Acoustical Society of America*, Salt Lake City, Utah, United States, May 23 to 27, 2016. Data from this study are available upon request to the corresponding author (Bolognaw@ohsu.edu).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported, in part, by research grants R01 DC000184 and P50 DC000422 from NIH/NIDCD, 017727 from American Academy of Audiology Foundation, and by the National Center for Advancing Translational Sciences of the National Institutes of Health under Grant Numbers UL1 TR001450 and TL1 TR002371.

ORCID iD

William J. Bologna https://orcid.org/0000-0003-1851-6892

References

American National Standard Institute. (2010). *American National Standard Specification for Audiometers (ANSI/ASA S3.6-2010)*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-David, B. M., Tse, V. Y., & Schneider, B. A. (2012). Does it take older adults longer than younger adults to perceptually segregate a speech target from a background masker? *Hearing Research*, *290*(1–2), 55–63. https://doi.org/10.1016/j.heares.2012.04.022

Best, V., Ahlstrom, J. B., Mason, C. R., Roverud, E., Perrachione, T. K., Kidd, G., Jr., & Dubno, J. R. (2018). Talker identification: Effects of masking, hearing loss, and age. *The Journal of the Acoustical Society of America*, *143*(2), 1085–1092. https://doi.org/10.1121/1.5024333

Bologna, W. J., Vaden, K. I., Jr., Ahlstrom, J. B., & Dubno, J. R. (2018). Age effects on perceptual organization of speech: Contributions of glimpsing, phonemic restoration, and speech segregation. *The Journal of the Acoustical Society of America*, *144*(1), 267–281. https://doi.org/10.1121/1.5044397

Bologna, W. J., Vaden, K. I., Jr., Ahlstrom, J. B., & Dubno, J. R. (2019). Age effects on the contributions of envelope and periodicity cues to recognition of interrupted speech in quiet and with a competing talker. *The Journal of the Acoustical Society of America*, *145*(3), EL173–EL178. https://doi.org/10.1121/1.5091664

Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research*, *78*(3), 349–360. https://doi.org/10.1007/s00426-014-0555-7

Broadbent, D. E. (1958). *Perception and communication*. Pergamon Press.

Bronkhorst, A. W. (2015). The cocktail party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Dai, H., & Wright, B. A. (1995). Detecting signals of unexpected or uncertain durations. *The Journal of the Acoustical Society of America*, *98*(2), 798–806. https://doi.org/10.1121/1.413572

Dai, H., Scharf, B., & Buus, S. (1991). Effective attenuation of signals in noise under focused attention. *The Journal of the Acoustical Society of America*, *89*(6), 2837–2842. https://doi.org/10.1121/1.400721

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of Acoustical Society of America*, *114*(5), 2913–2922. https://doi.org/10.1121/1.1616924

Ezzatian, P., Li, L., Pichora-Fuller, K., & Schneider, B. A. (2015). Delayed stream segregation in older adults: More than just informational masking. *Ear & Hearing*, *36*(4), 482–484. https://doi.org/10.1097/AUD.0000000000000139

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of*

Psychiatric Research, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—Focusing the searchlight on sound. *Current Opinion in Neurobiology*, *17*(4), 437–455. https://doi.org/10.1016/j.conb.2007.07.011

Gallun, F. J., & Diedesch, A. C. (2013). Exploring the factors predictive of informational masking in a speech recognition task. *Proceedings of Meetings on Acoustics*, *19*(1), 060145. https://doi.org/10.1121/1.4799107

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). *The DARPA TIMIT acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium.

George, E. L. J., Zekveld, A. A., Kramer, S. E., Goverts, S. T., Festen, J. M., & Houtgast, T. (2007). Auditory and non-auditory factors affecting speech reception in noise by older listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2362–2375. https://doi.org/10.1121/1.2642072

Gilbert, J. L., Tamati, T. N., & Pisoni, D. B. (2013). Development, reliability, and validity of PRESTO: A new high-variability sentence recognition test. *Journal of the American Academy of Audiology*, *24*(1), 26–36. https://doi.org/10.3766/jaaa.24.1.4

Greenberg, G. Z., & Larkin, W. D. (1968). Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *The Journal of the Acoustical Society of America*, *44*(6), 1513–1523. https://doi.org/10.1121/1.1911290

Helfer, K. S., & Freyman, R. L. (2008). Aging and speech-on-speech masking. *Ear & Hearing*, *29*(1), 87–98. https://doi.org/10.1097/AUD.0b013e31815d638b

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, *23*(6), 723–744. https://doi.org/10.1177/014920639702300602

Humes, L. E., Kidd, G. R., & Lentz, J. J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience*, *7*, 55. https://doi.org/10.3389/fnsys.2013.00055

Ihlefeld, A., & Shinn-Cunningham, B. (2008). Disentangling the effects of spatial cues on selection and formation of auditory objects. *The Journal of the Acoustical Society of America*, *124*(4), 2224–2235. https://doi.org/10.1121/1.2973185

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*, 327. https://doi.org/10.3389/fnhum.2014.00327

Kidd, G., Jr., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, *118*(6), 3804–3815. https://doi.org/10.1121/1.2109187

Kreitewolf, J., Mathias, S. R., Trapeau, R., Obleser, J., & Schönwiesner, M. (2018). Perceptual grouping in the cocktail party: Contributions of voice-feature continuity. *The Journal of the Acoustical Society of America*, *144*(4), 2178–2188. https://doi.org/10.1121/1.5058684

Krull, V., Humes, L. E., & Kidd, G. R. (2013). Reconstructing wholes from parts: Effects of modality, age, and hearing loss on word recognition. *Ear & Hearing*, *34*(2), e14–e23. https://doi.org/10.1097/AUD.0b013e31826d0c27

Kwon, B. J. (2012). *Token [Computer program]. Version 1.36*. auditorypro.com/aux/

Lee, J. H., & Humes, L. E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *The Journal of the Acoustical Society of America*, *132*(3), 1700–1717. https://doi.org/10.1121/1.4740482

Mackersie, C. L., Dewey, J., & Guthrie, L. A. (2011). Effects of fundamental frequency and vocal-tract length cues on sentence segregation by listeners with hearing loss. *The Journal of the Acoustical Society of America*, *130*(2), 1006–1019. https://doi.org/10.1121/1.3605548

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5–6), 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85–103. https://doi.org/10.1016/j.wocn.2005.10.004

Peterson, G. H., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

R Development Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org

Rajan, R., & Cainer, K. E. (2008). Ageing without hearing loss or cognitive impairment causes a decrease in speech intelligibility only in informational maskers. *Neuroscience*, *154*(2), 784–795. https://doi.org/10.1016/j.neuroscience.2008.03.067

Rönnberg, J. (1990). Cognitive and communicative function: The effects of chronological age and "handicap age". *European Journal of Cognitive Psychology*, *2*(3), 253–273. https://doi.org/10.1080/09541449008406207

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428. https://doi.org/10.1037/0033-295x.103.3.403

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, *54*(1–3), 35–54. https://doi.org/10.1016/S0301-0511(00)00052-1

Salthouse, T. A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., & Hocking, K. J. (2000). Effects of aging on efficiency of task switching in a variant of the Trail Making Test. *Neuropsychology*, *14*(1), 102–111. https://doi.org/10.1037/0894-4105.14.1.102

Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception & Psychophysics*, *42*(3), 215–223. https://doi.org/10.3758/BF03203073

Schlauch, R. S., & Hafter, E. R. (1991). Listening bandwidths and frequency uncertainty in pure-tone signal detection.

The Journal of the Acoustical Society of America, *90*(3), 1332–1339. https://doi.org/10.1121/1.401925

Schvartz, K. C., & Chatterjee, M. (2012). Gender identification in younger and older adults: Use of spectral and temporal cues in noise-vocoded speech. *Ear & Hearing*, *33*(3), 411–420. https://doi.org/10.1097/AUD.0b013e31823d78dc

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Tiffin, J., & Asher, E. J. (1948). The Purdue Pegboard: Norms and studies of reliability and validity. *Journal of Applied Psychology*, *32*(3), 234–247. https://doi.org/10.1037/h0061266

Treisman, A. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, *3*(6), 449–459. https://doi.org/10.1016/S0022-5371(64)80015-3

Trenerry, M. R., Crosson, B., DeBoe, J., & Leber, W. R. (1989). *The Stroop Neuropsychological Screening Test*. Psychological Assessment Resources.

White, L. J., & Carlyon, R. P. (1997). Detection of signals having expected and unexpected temporal structures. *Hearing Research*, *112*(1–2), 141–146. https://doi.org/10.1016/S0378-5955(97)00115-9

Wright, B. A., & Dai, H. (1994). Detection of unexpected tones with short and long durations. *The Journal of the Acoustical Society of America*, *95*(2), 931–938. https://doi.org/10.1121/1.410010

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, *15*(1), 88–99. https://doi.org/10.1037/0882-7974.15.1.88

Zekveld, A. A., George, E. L., Kramer, S. E., Goverts, S. T., & Houtgast, T. (2007). The development of the Text Reception Threshold test: A visual analogue of the Speech Reception Threshold test. *Journal of Speech, Language, and Hearing Research*, *50*(3), 576–584. https://doi.org/10.1044/1092-4388(2007/040)

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037