

## Research Article

# Enhancing Speech Intelligibility: Interactions Among Context, Modality, Speech Style, and Masker

Kristin J. Van Engen,<sup>a,b</sup> Jasmine E. B. Phelps,<sup>a</sup> Rajka Smiljanic,<sup>a</sup> and Bharath Chandrasekaran<sup>a</sup>

**Purpose:** The authors sought to investigate interactions among intelligibility-enhancing speech cues (i.e., semantic context, clearly produced speech, and visual information) across a range of masking conditions.

**Method:** Sentence recognition in noise was assessed for 29 normal-hearing listeners. Testing included semantically normal and anomalous sentences, conversational and clear speaking styles, auditory-only (AO) and audiovisual (AV) presentation modalities, and 4 different maskers (2-talker babble, 4-talker babble, 8-talker babble, and speech-shaped noise).

**Results:** Semantic context, clear speech, and visual input all improved intelligibility but also interacted with one another and with masking condition. Semantic context was beneficial across all maskers in AV conditions but only in

speech-shaped noise in AO conditions. Clear speech provided the most benefit for AV speech with semantically anomalous targets. Finally, listeners were better able to take advantage of visual information for meaningful versus anomalous sentences and for clear versus conversational speech.

**Conclusion:** Because intelligibility-enhancing cues influence each other and depend on masking condition, multiple maskers and enhancement cues should be used to accurately assess individuals' speech-in-noise perception.

**Key Words:** speech perception in noise, audiovisual speech, clear speech, visual enhancement, semantic context

During everyday speech communication, the cues available to listeners for understanding speech vary widely. Background noise interferes with access to speech signals, and signals themselves vary with respect to how clearly they are produced by speakers and how much semantic contextual information they contain. In addition to this variability within the auditory domain, listeners may or may not have the benefit of being able to see the person to whom they are listening. Each of these factors plays a significant role in determining the intelligibility of speech in challenging listening conditions: Listeners perform better on speech-in-noise tasks when target speech (i.e., the speech a person intends to listen to) has been produced clearly (see review in Smiljanic & Bradlow, 2009), when it contains semantic contextual information (Bradlow & Alexander, 2007; Kalikow, Stevens, & Elliott, 1977; Miller, Heise, & Lichten, 1951; Smiljanic

& Sladen, 2013), and when listeners can see the speaker (Fraser, Gagne, Alepins, & Dubois, 2010; Grant & Seitz, 2000; Helfer, 1997; Schwartz, Berthommier, & Savariaux, 2004; Sumbly & Pollack, 1954). In the present study, we aimed to extend our understanding of these intelligibility-enhancing cues by examining the interactions among them and with the noise environment in which they are presented. A few studies, reviewed below, have shown significant interactions between two such factors, but no study, to our knowledge, has addressed speech clarity, semantic context, presentation modality (audio vs. audiovisual), and masker type within a single experiment. In typical listening environments (e.g., noisy restaurants), however, listeners are able to make use of multiple simultaneous cues that aid speech intelligibility. Understanding how these cues interact is therefore important for understanding speech perception in naturalistic conditions.

Such investigation is especially important given the many listener populations for whom understanding speech in noisy conditions remains particularly problematic. These include individuals with hearing loss and auditory processing disorders, people communicating in nonnative languages, and individuals with learning disabilities and language impairments. Most clinical measures of speech perception in noise test listeners in very a limited range of conditions,

<sup>a</sup>The University of Texas at Austin

<sup>b</sup>Washington University in St. Louis, St. Louis, MO

Correspondence to Bharath Chandrasekaran:  
bchandra@mail.utexas.edu

Editor: Craig Champlin

Associate Editor: Karen Helfer

Received March 29, 2013

Revision received October 2, 2013

Accepted February 20, 2014

DOI: 10.1044/JSLHR-H-13-0076

**Disclosure:** The authors have declared that no competing interests existed at the time of publication.

usually including a single level of semantic context (n.b., Revised Speech Perception in Noise Test [R-SPIN]; Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984); one type of target speech (produced in a careful speaking style by a voice-over professional); one type of noise; and only auditory presentation. By omitting several of the features present in real-world speech communication, such tests provide only a partial evaluation of listeners' speech-in-noise abilities. It may be the case, for example, that a listener who has tremendous difficulty coping with interfering talkers in an auditory-only situation would show very little impairment when he or she can see the speaker. The goal of the present study was to assess the interactions among such cues during speech perception in noise for native-speaking listeners with normal language and hearing abilities. Extending basic knowledge of the factors modulating speech intelligibility in noise will help professionals continue to improve clinical assessment and intervention.

The clarity of speech signals themselves, first of all, plays an important role in determining the intelligibility of speech in noise. *Clear speech*—a speaking style that is naturally and spontaneously adopted by speakers when they are aware their interlocutors are having difficulty understanding them—is more intelligible than conversational speech for a wide range of listeners in noisy conditions and, further, improves recognition memory for speech in quiet and in noise (Gilbert, Chandrasekaran, & Smiljanic, 2014; Van Engen, Chandrasekaran, & Smiljanic, 2012). Among those for whom intelligibility benefits have been documented are adults with normal and impaired hearing (Payton, Uchanski, & Braida, 1994; Picheny, Durlach, & Braida, 1985), older adults with hearing loss (Ferguson & Kewley-Port, 2002; Schum, 1996), nonnative speakers (Bradlow & Bent, 2002), and children with and without learning disabilities (Bradlow, Kraus, & Hayes, 2003). The beneficial effects of clear speech, however, have also been shown to be modulated by the presence of semantic contextual information (Bradlow & Alexander, 2007; Smiljanic & Sladen, 2013), by the masker in which the signals are presented (Calandruccio, Van Engen, Dhar, & Bradlow, 2010), and by whether the listener can see the speaker (Gagné, Rochette, & Charest, 2002; Helfer, 1997).

Bradlow and Alexander (2007), for example, examined word recognition in noise for conversational and clear speech in high- and low-predictability sentences, showing that native speakers of English derived significant benefits from semantic context and clear speech, and, crucially, that these two cues were mutually enhancing in their effects on speech recognition. In contrast, nonnative speakers of English required both semantic context and clear speech to obtain any improvement in their speech-in-noise perception. In a study of normal hearing children and children who use cochlear implants, Smiljanic and Sladen (2013) similarly showed significant gains in speech intelligibility in noise in the presence of both semantic context and clear speech, but little to no improvement with only one of these two cues. Clear speech, therefore, appears to be of greater benefit to listeners when semantic context is also available.

Calandruccio et al. (2010) investigated clear and conversational speech intelligibility in the presence of different maskers: two-talker English babble and two-talker Croatian babble. They observed a larger benefit of clear speech in the presence of the English babble for native English-speaking participants, suggesting that properties of the masker can also modulate the benefits listeners obtain from clearly produced speech. It must be noted that performance was also generally lower in English than in Croatian (see also Garcia Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007, for studies showing greater masking by native versus foreign-language maskers). Listeners may experience more interference from native-language maskers because they know the language or because the masker is more similar to the target speech and therefore more difficult to segregate from that target (see, e.g., Van Engen, 2012). It cannot be determined conclusively whether differences in the clear-speech benefit across the English and Croatian maskers were due to differences in overall performance level (i.e., clear speech is simply more beneficial in more difficult conditions) or to the linguistic content of the maskers. In other studies (e.g., Payton et al., 1994), researchers have shown, for example, that the benefit of clear speech increased as listening conditions were degraded with increasing noise and reverberation. Follow-up studies are required to clarify the relationship between clear speech and masker content, but it does appear that the effectiveness of clear speech in enhancing intelligibility in noise can be affected not only by semantic context, but also by properties of the noise in which it is presented.

Along with semantic context and noise environment, the effectiveness of clear speech in enhancing intelligibility has also been investigated across presentation modalities (i.e. audio-only, visual-only, audiovisual presentation; Gagné, Masterson, Munhall, Bilida, & Querengesser, 1994; Gagné, Querengesser, Folkeard, Munhall, & Masterson, 1995; Gagné et al., 2002; Helfer, 1997). Gagné and colleagues, in a series of articles, have shown intelligibility benefits of clear speech for words, sentences, and syllables in audio-only, visual-only, and audiovisual speech, although these benefits vary across individual talkers. Helfer (1997) showed that, for nonsense sentences in noise, the total benefit derived by listeners from clear speech and visual input was the sum of their effects. Enhanced acoustic-phonetic cues and visual information, therefore, are argued to provide complementary (as opposed to redundant) information to listeners.

In addition to these studies of the interplay of clear speech with other factors affecting speech intelligibility, one study by Helfer and Freyman (2005) examined the interaction between visual information and masking environment. This experiment tested sentence intelligibility in the presence of steady-state noise and a two-talker masker, showing that visual information was more beneficial to speech intelligibility in the presence of the speech masker than the steady-state noise. The authors argued that their data were consistent with the hypothesis that lipreading cues provide supplementary information for the recovery of masked

phonetic information and help listeners segregate target and competing voices.

There is ample evidence, therefore, that cues that enhance speech intelligibility in adverse conditions interact with one another. The current study investigated the additional interaction of semantic context and visual information, and, importantly, compares these various cue enhancements and their interactions across a range of masking environments. The factors included here (i.e., semantic context, clear speech, visual information) are typical of real-world speech perception in challenging listening environments.

The difficulty associated with understanding speech in noisy environments arises via two general mechanisms: energetic masking and informational masking (Brungart, Simpson, Ericson, & Scott, 2001; Freyman, Balakrishnan, & Helfer, 2001; Freyman, Helfer, McCall, & Clifton, 1999). *Energetic masking* (EM) refers to masking that occurs in the auditory periphery, rendering portions of the target speech inaudible to the listener. *Steady-state speech-shaped noise* (SSN)—white noise filtered to match the long-term average spectrum of speech, which is frequently utilized in laboratory and clinical testing—exerts only this type of masking on target speech. In contrast, *informational masking* (IM) refers to interference at higher levels of auditory and cognitive processing; it arises when a listener has difficulty, for example, in extracting an audible speech signal from one or several simultaneous speech signals. The possible sources of IM in such situations are numerous: misattribution of components of the noise to the target (and vice versa), competing attention from the masker, increased cognitive load, and linguistic interference (Cooke, Garcia Lecumberri, & Barker, 2008).

Because the noise listeners contend with on a daily basis varies with respect to the degree of EM and IM imposed on target speech, the current study used maskers that varied in this respect as well. In particular, we used two-talker babble (2T), four-talker babble (4T), eight-talker babble (8T), and SSN to investigate how the benefits of intelligibility-enhancing cues are affected by the masking environment in which they are presented. These maskers approximate a range of frequently encountered challenging listening environments: situations in which listeners must ignore one competing conversation, busier social gatherings like restaurants and parties, and environments where other nonspeech noise (e.g., a loud air conditioning system) interferes with speech communication. These maskers also vary with respect to their EM and IM components. As the number of talkers in a multitalker babble increases, EM also increases, because the masker becomes denser in its spectral and temporal structure, offering fewer dips in energy where listeners may “glimpse” the target speech. (During real-world communication, additional talkers also increase the overall energy in the masker relative to the target. This factor was controlled in the current experiment.) By contrast, IM generally makes a greater contribution to the overall masking imposed by babble when there are fewer talkers in it—that is, when

distracting information in the maskers is more accessible to listeners (Freyman, Balakrishnan, & Helfer, 2004; Rosen, Souza, Ekelund, & Majeed, 2013; Simpson & Cooke, 2005). Where there is a large number of talkers, much of the informational content of the masker is eliminated by EM within the babble itself. As mentioned above, there is some evidence that masker type—that is, maskers that vary in the degree of EM and IM they impose—can affect the intelligibility benefits listeners obtain from clear speech (Calandruccio et al., 2010) and visual information (Helfer & Freyman, 2005). Here we additionally investigated the effects of different maskers on the benefit listeners obtain from semantic context.

## Method

### Participants

Twenty-nine adults (9 men, 20 women) were recruited from the University of Texas community and paid for their participation. All participants were between the ages of 18 and 38 (average age = 21.3 years) and reported normal or corrected-to-normal vision. Participants completed the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian, Blumenfeld, & Kaushanskaya, 2007) prior to testing to verify they were monolingual English speakers with no second-language exposure before the age of 12 years. Hearing thresholds in both ears were less than 25 dB HL at 1000 Hz, 2000 Hz, and 4000 Hz. All materials and procedures were approved by the Institutional Review Board at The University of Texas at Austin.

### Materials

*Target sentences.* One 33-year-old female native speaker of American English was video-recorded producing two sets of sentences on a sound-attenuated sound stage at The University of Texas at Austin. The first set consisted of 80 semantically anomalous sentences taken from the Syntactically Normal Sentence Test (SNST; Nye & Gaitenby, 1974). The second set was composed of 80 semantically meaningful sentences based on sentences from the Basic English Lexicon (BEL; Calandruccio & Smiljanic, 2012) and adjusted to conform to the sentence length and syntax of the SNST set. Sentences from both sets contained four key words each (e.g., *The green week did the page; The hot sun warmed the ground*). All sentences were produced in both clear and conversational speaking styles. To produce conversational speech, the speaker was instructed to speak as if she were talking to someone familiar. To elicit clear speech, the speaker was asked to talk as if she was speaking to someone who was having trouble understanding her, whether due to hearing impairment or because the listener was a nonnative speaker of English. The video recording was captured using a Sony PMW-EX3 studio camera with target sentences presented to the speaker on a teleprompter. Camera output was processed through a Ross

crosspoint video switcher and recorded on an AJA Pro video recorder. Audio was recorded at a sampling rate of 48000 Hz with an Audio Technica AT835b shotgun microphone placed on a floor stand in front of the speaker. The long video recording was segmented into individual sentences, and the audio from each sentence was detached from the video using Final Cut Pro. All audio tracks were equalized for RMS amplitude using Praat software (Boersma & Weenink, 2009). The leveled audio clips served as the stimuli for the audio-only (AO) condition. For the audiovisual (AV) condition, the leveled audio was reattached to the corresponding videos using Final Cut Pro. Table 1 provides a summary of the experimental factors.

**Maskers.** Multiple-talker babble tracks were created as follows: eight female speakers of American English (different from the talker who produced the target sentences) were recorded in a sound-attenuated booth at Northwestern University as part of the Wildecat Corpus project (Van Engen et al., 2010). Each participant produced a set of 30 simple, meaningful English sentences (from Bradlow & Alexander, 2007). All sentences were segmented from the recording files and equalized for RMS amplitude in Praat. To create *N*-talker babbles, the sentences from each talker were concatenated in random order to create 30-sentence strings without silence between sentences. Two of these strings were mixed together using the mix paste function in Audacity (Version 1.2.5; www.audacity.sourceforge.net) to generate 2T babble. Two more talkers were added to create 4T babble, and four more for 8T babble. SSN was generated by obtaining the long-term average spectrum from the full set of 240 sentences (30 sentences × 8 talkers) and shaping white noise to match that spectrum. All masker tracks were truncated to 50 s and then equated for RMS amplitude.

**Mixing targets and maskers.** Targets were mixed with maskers in real time on each trial. For each stimulus, the noise began 500 ms before the onset of the target and ended 500 ms after the target's offset. Target sentences were presented at a constant RMS level of 70 dB SPL, and the



maskers were presented at a constant 78 dB SPL, so that all stimuli were presented at -8 dB signal-to-noise ratio (SNR).

### Procedure

Before the speech-in-noise test, participants received an otoscopic evaluation. Testing then took place in a sound-attenuated room. ER-1 earphones (Etymotic, www.etymotic.com) were inserted into both ears of the participant, and all auditory stimuli were presented diotically. A MOTU UltraLite external audio interface was used for digital-to-analog conversion (24 bit), and audio signals were passed through an Aphex Headpod 4 headphone amplifier. Video signals were presented at a rate of 29.97 fps on a Dell 2007WFPb 20-inch computer monitor with a resolution of 1280 × 720.

The experiment was run using custom software created with Max/MSP (Cycling '74, www.cycling74.com), and the experimenter regulated the presentation of the stimuli from an administrator computer. Participants sat across from the experimenter facing their own computer monitor. They were positioned approximately 90 cm from the screen. Listeners were presented with a total of 160 target sentences in two blocks. For all participants, the first block (80 sentences) was composed of meaningful sentences and the second block consisted of anomalous sentences. Any observed benefit of semantic context, then, would arise despite the fact that listeners had had more practice with the task when they heard the anomalous sentences. Each target sentence was presented in one of four maskers (2T, 4T, 8T, SSN), one of two speaking styles (clear speech, conversational speech), and one of two presentation modalities (AO, AV); that is, five sentences with 20 key words for scoring were presented in each combination of masker, speaking style, and presentation modality. This number of trials allowed us to complete the experiment in a single testing session while minimizing the effects of fatigue and learning (see, e.g., Van Engen & Bradlow, 2007). The

**Table 1.** Examples of semantic context, speaking style elicitation instructions, and modality of stimuli presentation.

Context	Style	Modality
<u>Meaningful:</u> The HOT SUN WARMED the GROUND.	<u>Clear:</u> Speak as if you are talking to someone who is having a hard time understanding you.	Audiovisual 
<u>Anomalous:</u> The WRONG SHOT LED the FARM.	<u>Conversational:</u> Speak as if you are talking to someone familiar with your speaking style.	Audio only 

assignment of each sentence to a particular condition was randomized for each participant, as was the order of presentation. No target sentence was presented more than once. AV presentation displayed a full-screen video of the speaker, whereas AO presented listeners with a visually centered black crosshair on a white background. Listeners were instructed that they would be listening to sentences mixed with noise and each sentence would either be audio only or accompanied by the video of the speaker. Listeners were also informed that the target sentences would always begin one half second after the noise. They were asked to repeat the target sentence aloud to the experimenter. If they were unable to understand the entire target sentence, they were asked to report any intelligible words or make their best guess. Each stimulus was presented once. After each trial, the experimenter immediately scored the participant's response for correct key word identification. Responses with added or omitted morphemes were scored as incorrect.

## Results

The raw results of the experiment are presented in Figure 1. The overall trends are as follows: Performance was generally higher for the semantically meaningful targets (right panels), for clear speech (right two bars in each panel), for audiovisual presentation (gray bars), and in 2T babble and SSN as opposed to 4T or 8T babble (top and bottom panels).<sup>1</sup> In order to visualize the results with respect to how much each cue or cue combination enhanced intelligibility, the raw data are presented again in Figure 2 as average improvement scores with respect to baseline performance in the most difficult condition: AO, conversational, semantically anomalous targets. For each listener, the proportion of key words identified in the baseline condition was subtracted from the proportion of key words identified in each other condition; the figure presents the average of these difference scores across all participants.

### Analysis 1: Key Word Identification Across Listening Conditions

The proportion of key words correctly identified in the various conditions for each listener was transformed into rationalized arcsine units (Studebaker, 1985) and analyzed using a repeated-measures analysis of variance (ANOVA) with four within-subject factors: noise type (2T vs. 4T vs. 8T vs. SSN), context (meaningful vs. anomalous), speaking style (clear vs. conversational), and modality (AO vs. AV). Results from this analysis revealed significant main effects

of noise type,  $F(3, 26) = 75.821, p < .001, \eta^2 = .897$ ; context,  $F(1, 28) = 16.901, p < .001, \eta^2 = .376$ ; speaking style,  $F(1, 28) = 44.419, p < .001, \eta^2 = .613$ ; and modality,  $F(1, 28) = 305.345, p < .001, \eta^2 = .916$ . Further, there were significant two-way interactions between context and noise,  $F(3, 26) = 10.614, p < .001, \eta^2 = .551$ ; modality and noise,  $F(3, 26) = 7.139, p < .001, \eta^2 = .452$ ; context and speaking style,  $F(1, 28) = 5.595, p = .025, \eta^2 = .167$ ; and modality and speaking style,  $F(1, 28) = 6.245, p < .019, \eta^2 = .182$ . Finally, there was a significant three-way interaction among sentence context, modality, and noise type,  $F(3, 26) = 4.771, p = .009, \eta^2 = .355$ . All other two- and three-way interactions were nonsignificant.

The two-way interactions that were not involved in the three-way interaction were analyzed using a Bonferroni correction for multiple comparisons. The interaction between context and speaking style, first of all, was driven by the fact that clear speech improved intelligibility over conversational speech more for anomalous,  $F(1, 28) = 68.915, p < .0001, \eta^2 = .711$ , than for meaningful,  $F(1, 28) = 4.597, p = .041, \eta^2 = .141$ , sentences. The Modality  $\times$  Speaking Style interaction showed that the benefit of clear speech was larger in the AV condition,  $F(1, 28) = 54.56, p < .0001, \eta^2 = .661$ , than in the AO condition,  $F(1, 28) = 6.888, p < .014, \eta^2 = .197$ .

The three-way interaction among sentence context, modality, and noise type (Figure 3) was driven by the fact that the benefit of semantic context was present in the AO condition for the SSN masker only ( $p < .0001$ ; Figure 3a). In the AV condition, by contrast, the effect of semantic context (meaningful > anomalous) was significant in all noise conditions (2T,  $p < .001$ ; 4T,  $p < .046$ ; 8T,  $p < .004$ ; SSN,  $p < .0001$ ; Figure 3b).

### Analysis 2: Visual Enhancement

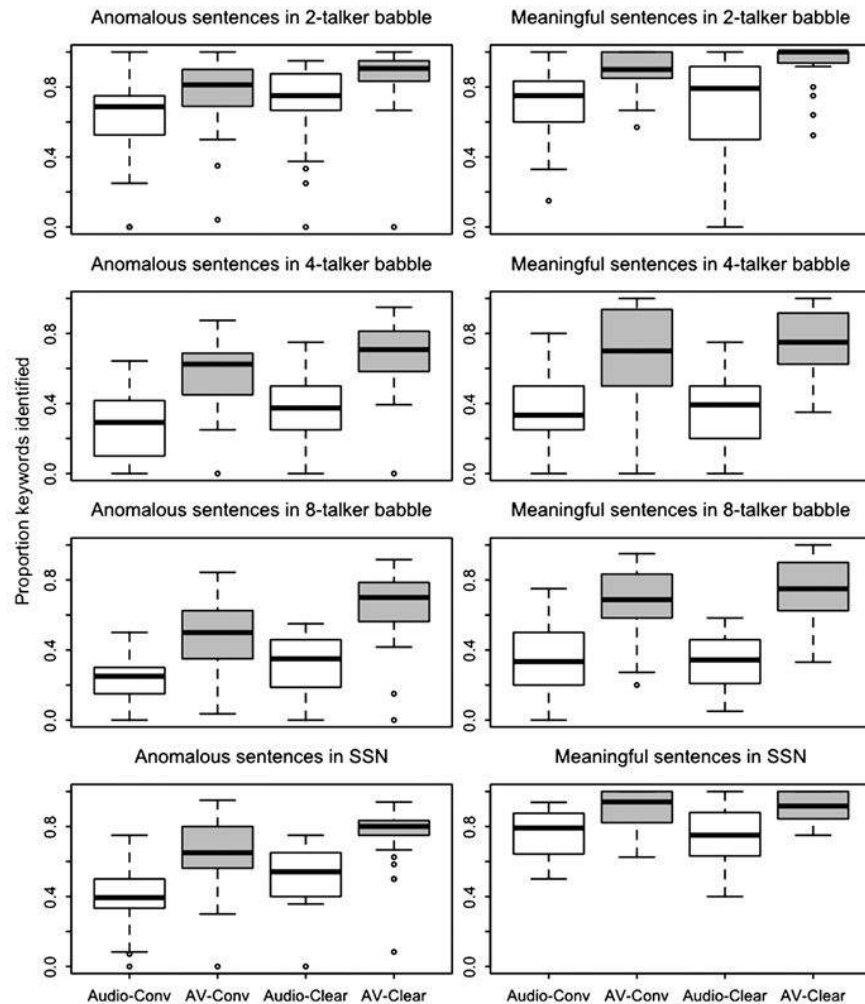
To assess the effects of clear speech and semantic context on listeners' ability to take advantage of visual cues, visual enhancement (VE) scores were calculated by taking the difference between a listener's performance in the AV and AO versions of a given condition and dividing it by the amount of improvement available to the listener due to the addition of visual information (Grant & Seitz, 1998; Grant, Walden, & Seitz, 1998; Sommers, Tye-Murray, & Spehar, 2005):

$$VE = (AV - AO)/(1 - AO). \quad (1)$$

An analysis of VE scores was performed for the 4T and 8T conditions, in which each listener's AO performance was below ceiling (VE cannot be computed for an AO proportion score of 1, and nine subjects identified all of the key words in either 2T or SSN in the AO condition). A repeated-measures ANOVA was then performed on the calculated VE scores with three within-subjects factors: noise type (4T vs. 8T), context (meaningful vs. anomalous), and speaking style (clear vs. conversational).

<sup>1</sup>The wide range of scores for this group of participants is particularly interesting given the relative homogeneity of the group: They were all monolingual young adults with no hearing loss, little musical experience, and similar educational backgrounds. This variability in performance likely reflects natural individual differences in sensory or cognitive function (Chandrasekaran, Hornickel, Skoe, Nicol, & Kraus, 2009; Vaden et al., 2013; Wong, Uppunda, Parrish, & Dhar, 2008).

**Figure 1.** Proportions of key words identified in all listening conditions (SNR = -8 dB). Left panels: Data for semantically anomalous target sentences. Right panels: Data for semantically meaningful sentences. Each row presents data for a given masker. Audio-only (AO) conditions are presented in white; audiovisual (AV) conditions are shaded. The center line on each box plot denotes the median score, the edges of the box denote the 25th and 75th percentiles, and the whiskers extend to data points that lie within 1.5 times the interquartile range. Points outside this range appear as outliers.



Results revealed significant main effects of context,  $F(1, 28) = 6.37, p = .018$ , and speaking style,  $F(1, 28) = 27.09, p < .001$ , but not of noise type ( $p = .371$ ). There were no significant interactions. The results are illustrated in Figure 4.

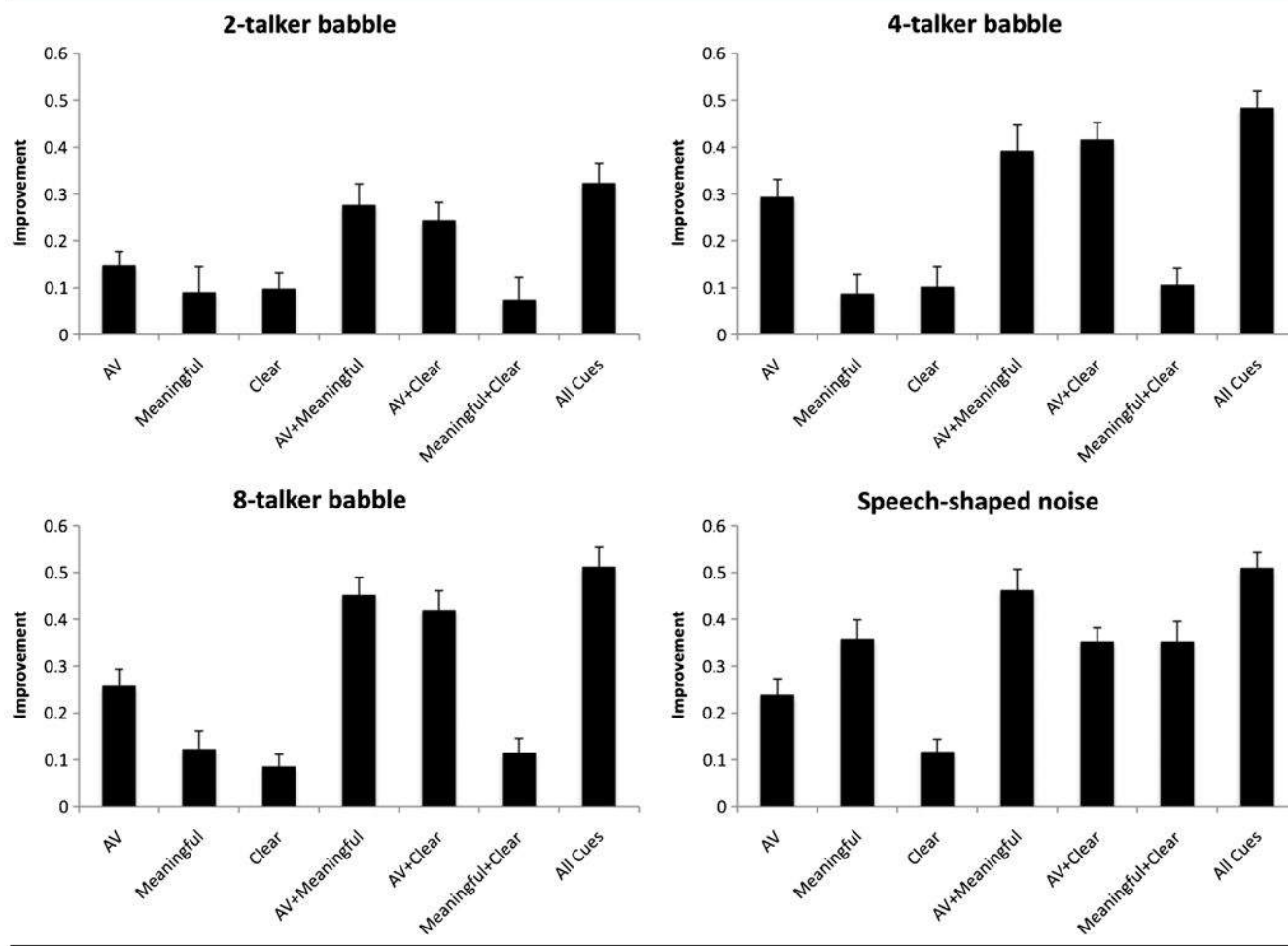
## Discussion

The investigation of semantic context, speech clarity, presentation modality, and masker type in combination revealed a number of significant interactions, showing that the contributions of various intelligibility-enhancing cues to target speech intelligibility influence one another and depend on the masking environment in which they are presented. The main effects of the tested factors are in line with previous work: Some masker types are more

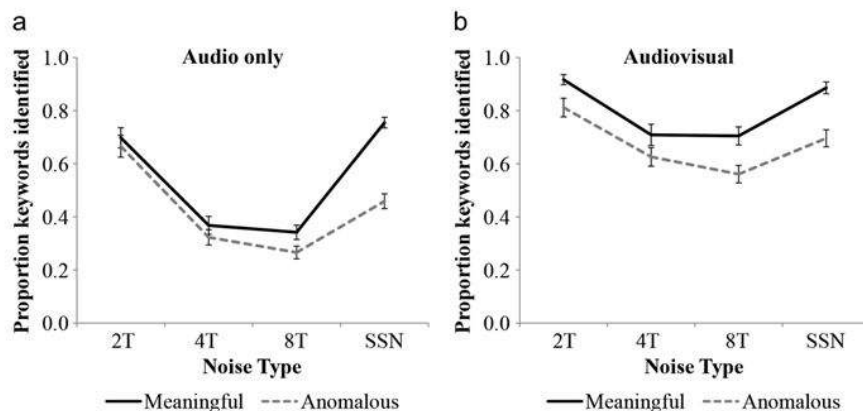
challenging than others (4T and 8T babble, with their combination of EM and IM, led to more overall masking than 2T babble, with relatively less EM,<sup>2</sup> or SSN, with no IM); meaningful sentences were generally more intelligible than anomalous sentences (Bradlow & Alexander, 2007; Kalikow et al., 1977); clear speech was more intelligible than conversational speech (Ferguson & Kewley-Port, 2002; Payton et al., 1994; Picheny et al., 1985; Smiljanic & Sladen, 2013; Uchanski, Choi, Braida, Reed, & Durlach,

<sup>2</sup>The 2T babble here exerts relatively less EM because there are more/larger spectro-temporal dips in it than in 4T or 8T babble; the overall amplitudes of the maskers were held constant. It should be noted that because the overall energy was held constant, each individual masking message in the 2T masker was louder than the individual voices in the maskers with more talkers.

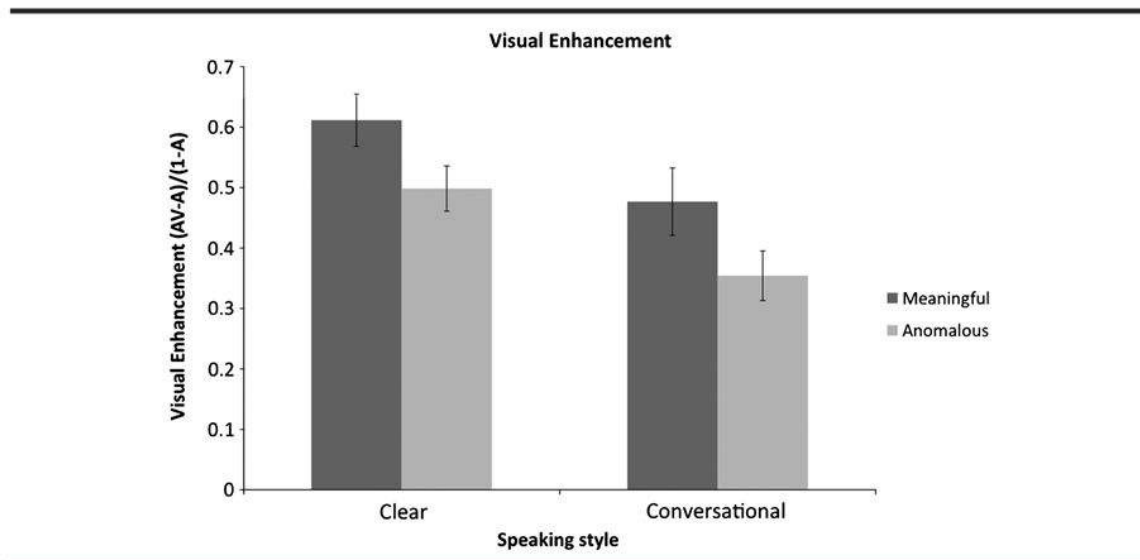
**Figure 2.** Average improvement scores with respect to performance on anomalous sentences produced in a conversational style and presented in the audio-only modality at an SNR of  $-8$  dB. The first three bars represent performance improvement with the addition of one cue (i.e., visual input, semantic content, or clear speech); the fourth through sixth bars represent improvement with the addition of two cues; and the final bar shows improvement when all three cues are available. Error bars represent standard error.



**Figure 3.** A three-way interaction among context, modality, and noise. Average proportions of key words identified in each condition, collapsed over clear/conversational targets. Error bars represent standard error. (a) In the audio-only (AO) condition, semantic contextual cues improve intelligibility in speech-shaped noise only. (b) In the AV condition, semantic contextual cues increase intelligibility across all noise conditions.



**Figure 4.** Average visual enhancement collapsed across four-talker and eight-talker maskers. Error bars represent standard error.



1996; Van Engen et al., 2012); and AV speech was more intelligible than AO speech (Fraser et al., 2010; Gagné et al., 2002; Grant & Seitz, 2000; Schwartz et al., 2004; Sumbly & Pollack, 1954). Crucially, however, these factors showed several significant interactions with one another.

First, the two-way interactions involving speaking style showed that (for this female test talker at an SNR of  $-8$  dB) clear speech was more beneficial to listeners when they listened to anomalous sentences and was more helpful for AV speech than AO speech. The finding that the clear-speech benefit was greater for the anomalous sentences suggests that the acoustic-phonetic enhancements of clear speech may be particularly useful when semantic contextual cues are lacking. It could also be the case that clear speech simply makes a greater contribution to intelligibility in relatively more difficult conditions (in this case, with anomalous vs. meaningful sentences). If this were the explanation, however, we would also expect a greater clear-speech effect in AO speech rather than AV. This is the opposite of what was found. Instead, clear speech improved intelligibility to a greater extent in AV speech, which is more intelligible than AO speech. There may be multiple reasons for this. First of all, the articulatory gestures involved in producing clear speech may enhance the visual information available for identifying speech. Clear speech may therefore improve lipreading itself (i.e., speech identification in visual-only presentation; Gagné et al., 2002) and/or it may provide better visual information to aid listeners in recovering auditory information lost due to signal degradation (i.e., to cope with EM). Enhanced visual cues may also help listeners take advantage of temporal cues in the visual signal that allow them to attend to the correct auditory stream in the presence of multiple talkers (i.e., to cope with IM). Alternatively, the presence of a visual cue

(regardless of its “clarity”) may help listeners make better use of the enhanced acoustic attributes of clear speech (e.g., slower speaking rate and reduced coarticulation). That is, once stream segregation is improved by the presence of a visual temporal cue, listeners are more able to benefit from the acoustic enhancement of the target speech.

The three-way interaction among masker type, semantic context, and presentation modality showed an intelligibility enhancement for this talker with the presence of semantic context across all masker types for AV speech, but only in SSN for AO speech (see Figure 3). This novel finding—that the benefit of semantic context was present only in the pure EM condition for AO perception—is of particular interest, because it demonstrates that different types of noise interfere differently with the listener’s ability to make use of the semantic context in target speech. This result suggests that, in the presence of several speech streams (as in the multitalker babbles), inhibiting the linguistic information in the maskers themselves may also cause listeners to inhibit (at least to some extent) the linguistic content of the target speech. When visual information is available, by contrast, listeners’ ability to select the target sentence is presumably improved by the provision of visual temporal cues that aid auditory stream segregation and talker selection. That is, seeing the speaker produce the target sentence furnished key information (e.g., the exact start of the target sentence) that could help the listener identify the target speech stream and hold it together over time, and therefore appropriately inhibit informational interference from speakers in the masker. The temporal cue available in the visual domain, therefore, increased listeners’ ability to use semantic contextual cues in all masking conditions. These results accord with those of Helfer and Freyman (2005) in showing that visual information aids speech perception in noise not only by



providing an alternative source of information about cues that may be masked in the auditory domain, but also by aiding the segregation of multiple talkers.

Finally, the analysis of visual enhancement (i.e., improvement in AV versus AO normalized by the amount of improvement available to each listener) in 4T and 8T babble showed that semantic context and clear speech both significantly improved listeners' ability to benefit from visual input. That is, listeners were able to take greater relative advantage of visual input when the speaker was producing clear speech and when the targets were semantically meaningful. In general, these results run counter to the principle of *inverse effectiveness*, which asserts that multisensory integration is enhanced as unimodal performance declines (Stein & Meredith, 1993). Here, unimodal performance was varied by changing the clarity of the target speech (i.e., conversational vs. clear) and by varying the presence of semantic contextual cues. The principle of inverse effectiveness would predict enhanced multisensory integration in the more difficult conditions: conversational speech and semantically anomalous sentences. Contrary to this prediction, we found greater visual enhancement in the less difficult conditions. This finding calls into question the pervasiveness of inverse effectiveness in multisensory processing. (See also Tye-Murray, Sommers, Spehar, Myerson, & Hale, 2010; and Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007, for other cases in which the principle of inverse effectiveness does not appear to capture behavior in AV speech perception.)

This benefit of clear versus conversational speech on visual enhancement suggests, first of all, that the visual cues associated with the more extreme articulatory gestures and/or slower rate of clear speech may be more robust than those associated with conversational speech for augmenting auditory speech recognition in noise. This result differs notably from that of Helfer (1997), who found equal visual enhancement for clear and conversational sentences and less visual enhancement overall: We found VE proportions of 61 (clear) and 48 (conversational) for meaningful targets and 50 (clear) and 35 (conversational) for anomalous targets, whereas Helfer found VE proportions of 18.2 (clear) and 16.8 (conversational) for her anomalous targets.) Several notable differences between the listening conditions in the two studies may account for these different outcomes. In particular, Helfer (1997) used 12-talker babble and an SNR of +2 dB, whereas the present analysis includes 4T and 8T babble presented at -8 dB. The masking conditions across the two studies, therefore, differed both quantitatively and qualitatively, and both factors may modulate the relative effects of clear and conversational speech on audiovisual benefit. The larger visual enhancement scores in the current study may, therefore, result from one or both of these facts: The participants in this study performed their task in higher levels of noise (lower SNR) and in babble that, by virtue of containing fewer talkers, likely had a more significant IM component. Both of these explanations for increased visual enhancement find corroboration in previous research: First, greater enhancement

in more degraded listening conditions is in line with studies in which AV perception does appear to follow the principle of inverse effectiveness (i.e., greater integration in conditions where unimodal performance is lower); second, greater visual enhancement in a more IM-weighted masker was also observed in Helfer and Freyman (2005), where visual information was more beneficial in a speech masker than in steady-state noise. As for the observed difference between visual enhancement for clear versus conversational speech (not found in Helfer, 1997), it may simply be the case that clear speech enhances the usefulness of the visual signal only in relatively more difficult listening conditions and/or that these cues are of greater benefit in maskers imposing a greater degree of IM (i.e., in multitalker babbles with fewer talkers). Finally, differences in the production of clear and conversational speech by the target talkers in Helfer (1997) and the present study may account for at least some of the differences in the studies' results. Additional research is needed to clarify this relationship.

Along with the clear-speech benefit for visual enhancement, the current study also found a benefit of contextual information for visual enhancement: Listeners were better able to take advantage of visual input for semantically meaningful versus semantically anomalous targets. This effect may arise as follows: In a noisy listening environment, semantic cues present in the preceding context reduce a listener's uncertainty about speech targets by constraining the set of possible speech targets. Visual input then serves to reduce an already-constrained set of reasonable targets. In semantically anomalous target speech, the unconstrained set of possible auditory targets remains so large that visual information cannot narrow the search with as much precision as it can when semantic cues are present. Further studies are required to test this hypothesis.

The current set of results demonstrates that intelligibility-enhancing cues do interact both with one another and with the masking conditions in which they are presented. One limitation, of course, is that a single target talker was used throughout testing. Intertalker variability has been demonstrated with respect to the enhancements associated with clear speech in all presentation modalities: AO, AV, and VO (Gagné et al., 1994; Gagné et al., 2002; Picheny et al., 1985), so additional research is required to further clarify the relationships among intelligibility-enhancing cues across talkers. It is also unclear whether these results will generalize to listening conditions with mixed-gender or opposite-gender babble. Such conditions could be explored in future studies. What we have seen in the present experiment, however, is that the effectiveness of clear speech depends not only on the acoustic-phonetic modifications made by a particular speaker, but also on both the semantic content of the target and the availability of visual information. Further, listeners' ability to take advantage of semantic context is more detrimentally affected by multitalker babbles when visual information is unavailable than when it is present. Finally, listeners are better able to make use of visual input when the speech signal contains semantic context and when speech is produced clearly.

These results have several implications for the testing of human speech processing in noise in experimental and clinical settings. First, they suggest that by employing a single type of noise, clinical tests are providing incomplete pictures of listeners' ability to understand speech in the range of listening environments encountered in everyday communication. The Hearing in Noise Test (HINT; Nilsson, Soli, & Sullivan, 1994), for example, measures full-sentence intelligibility in SSN, whereas the Quick Speech in Noise Test (QuickSIN; Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004) measures key-word identification in sentences in 4T babble (composed of one male and three female voices). These two tests, therefore, provide information about speech identification in very different noise conditions. However, it is entirely possible for an individual to be particularly good at coping with maskers that impose primarily energetic masking (e.g., SSN), but have difficulty coping with maskers that exert a large amount of informational masking (or vice versa). Indeed, results from Van Engen (2012) showed that adjusting listeners' speech-in-speech SNR based on their performance on the HINT test did not normalize performance across listeners on a speech-in-speech task that employed 2T babble. In other words, the ability to cope with EM was not predictive of performance in an environment involving IM (Van Engen, 2012). With respect to the characteristics of target speech materials themselves, we have shown not only that listeners utilize several intelligibility-enhancing cues to cope with difficult listening environments, but also that the benefits afforded by such cues interact with one another and with the type of masker in which they are presented. For a variety of reasons (e.g., hearing impairment, listening to a non-native language), many listeners struggle to understand speech in noisy situations. The present study suggests the need for a comprehensive approach to speech-in-noise testing, in which listeners are tested in multiple modalities with multiple types of speech targets and maskers. Such an approach would provide clinicians and researchers with valuable information about listeners' strengths and weaknesses in understanding speech in noise and utilizing the cues that can improve speech intelligibility in everyday listening situations.

## Acknowledgments

The University of Texas Longhorn Innovation Fund for Technology provided funding for this study, awarded to the third and fourth authors. We would like to thank Han-Gyol Yi, Britt Rachner, Stephanie Tchen, Kadee Bludau, and the members of the SoundBrain Lab for assistance in data collection and management.

## References

- Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research, 27*, 32–48.
- Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 5.1). Retrieved from [www.praat.org](http://www.praat.org)
- Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America, 121*, 2339–2349.
- Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112*, 272–284.
- Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking clearly for children with listening disabilities: Sentence perception in noise. *Journal of Speech, Language, and Hearing Research, 46*, 80–97.
- Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America, 110*, 2527–2538.
- Calandruccio, L., & Smiljanic, R. (2012). New sentence recognition materials developed using a basic non-native English lexicon. *Journal of Speech, Language, and Hearing Research, 55*, 1342–1355.
- Calandruccio, L., Van Engen, K. J., Dhar, S., & Bradlow, A. R. (2010). The effectiveness of clear speech as a masker. *Journal of Speech, Language, and Hearing Research, 53*, 1458–1471.
- Chandrasekaran, B., Hornickel, J. M., Skoe, E., Nicol, T., & Kraus, N. (2009). Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: Implications for developmental dyslexia. *Neuron, 64*, 311–319.
- Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects on non-native speech perception. *The Journal of the Acoustical Society of America, 123*, 414–427.
- Ferguson, S. J., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 112*, 259–271.
- Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research, 53*, 18–33.
- Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America, 109*, 2112–2122.
- Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America, 115*, 2246–2256.
- Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America, 106*, 3578–3588.
- Gagné, J.-P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology, 27*, 135–158.
- Gagné, J.-P., Querengesser, C., Folkeard, P., Munhall, K. G., & Masterson, V. (1995). Auditory, visual, and audiovisual speech intelligibility for sentence-length stimuli: An investigation of clear and conversational speech. *Volta Review, 97*, 33–51.
- Gagné, J.-P., Rochette, A.-J., & Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Communication, 37*, 213–230.
- Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America, 119*, 2445–2454.

- Gilbert, R. C., Chandrasekaran, B., & Smiljanic, R. (2014). Recognition memory in noise for speech of varying intelligibility. *The Journal of the Acoustical Society of America*, *135*, 389–399.
- Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, *104*, 2438–2450.
- Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*, 1197–1208.
- Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*, 2677–2690.
- Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *40*, 432–443.
- Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America*, *117*, 842–849.
- Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*, 1337–1351.
- Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & 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. *The Journal of the Acoustical Society of America*, *116*, 2395–2405.
- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *The Journal of Speech, Language, and Hearing Research*, *50*, 940–967.
- Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*, 329–335.
- Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*, 1085–1099.
- Nye, P. W., & Gaitenby, J. H. (1974). *The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences* (Status Report on Speech Research SR-37/38). New Haven, CT: Haskins Laboratory.
- Payton, K. L., Uchanski, R. M., & Braid, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, *95*, 1581–1592.
- Picheny, M. A., Durlach, N. I., & Braid, L. D. (1985). Speaking clearly for the hard of hearing: I. Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, *28*, 96–103.
- Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, *133*, 2431–2443.
- Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*, 1147–1153.
- Schum, D. J. (1996). Intelligibility of clear and conversational speech of young and elderly talkers. *Journal of the American Academy of Audiology*, *7*, 212–218.
- Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, *93*, B69–B78.
- Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. [Letter to the editor]. *The Journal of the Acoustical Society of America*, *118*, 2775–2778.
- Smiljanic, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass*, *3*, 236–264.
- Smiljanic, R., & Sladen, D. (2013). Acoustic and semantic enhancements for children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *56*, 1085–1096.
- Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing*, *26*, 263–275.
- Stein, B., & Meredith, M. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.
- Studebaker, G. A. (1985). A “rationalized” arcsine transform. *Journal of Speech and Hearing Research*, *28*, 455–462.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212–215.
- Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing*, *31*, 636–644.
- Uchanski, R. M., Choi, S. S., Braid, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing: IV. Further studies on the role of speaking rate. *Journal of Speech and Hearing Research*, *39*, 494–509.
- Vaden, K. I., Kuchinsky, S. E., Cude, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2013). The cingulo-opercular network provides word-recognition benefit. *Journal of Neuroscience*, *33*, 18979–18986.
- Van Engen, K. J. (2012). Speech-in-speech recognition: A training study. *Language and Cognitive Processes*, *27*, 1089–1107.
- Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, *53*, 510–540.
- Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, *121*, 519–526.
- Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PLoS One*, *7*, e43753.
- Wong, P. C. M., Uppunda, A. K., Parrish, T. B., & Dhar, S. (2008). Cortical mechanisms of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, *51*, 1026–1041.