Attention, Perception, & Psychophysics (2020) 82: 1443–1458
https://doi.org/10.3758/s13414-019-01808-2

Can the diffuseness of sound sources in an auditory scene alter speech perception?

Meital Avivi-Reich[1,2] · Brendan Fifield[1] · Bruce A. Schneider[1]

Published online: 13 August 2019

Abstract

When amplification is used, sound sources are often presented over multiple loudspeakers, which can alter their timbre, and introduce comb-filtering effects. Increasing the diffuseness of a sound by presenting it over spatially separated loudspeakers might affect the listeners' ability to form a coherent auditory image of it, alter its perceived spatial position, and may even affect the extent to which it competes for the listener's attention. In addition, it can lead to comb-filtering effects that can alter the spectral profiles of sounds arriving at the ears. It is important to understand how these changes affect speech perception. In this study, young adults were asked to repeat nonsense sentences presented in either noise, babble, or speech. Participants were divided into two groups: (1) A Compact-Target Timbre group where the target sentences were presented over a single loudspeaker (compact target), while the masker was either presented over three loudspeakers (diffuse) or over a single loudspeaker (compact); (2) A Diffuse-Target Timbre group, where the target sentences were diffuse while the masker was either compact or diffuse. Timbre had no significant effect in the absence of a timbre contrast between target and masker. However, when there was a timbre contrast, the signal-to-noise ratios needed for 50% correct recognition of the target speech were higher (worse) when the masker was compact, and lower (better) when the target was compact. These results were consistent with the expected effects from comb filtering, and could also reflect a tendency for attention to be drawn towards compact sound sources.

Keywords Speech perception · Hearing · Scene perception

Abbreviations

T_C	Compact target sound source
T_D	Diffuse target sound source
M_C	Compact masker sound source
M_D	Diffuse masker sound source

Introduction

The variety and nature of the auditory scenes in which daily communication takes place have changed significantly over the years due to the growing use of electronic amplification and surround-sound systems. As amplification becomes more common in both public spaces and in private homes, it is important to understand how the changes it creates in the auditory scene may affect one's ability to communicate efficiently in it.

When amplification is used, sound sources often are presented over more than a single loudspeaker (e.g., surround-sound systems), creating a broader and more diffuse auditory image of the original sound source. In addition, when the same sound is played from two or more spatially separated loudspeakers, the direct sounds from the different loudspeakers are likely to arrive at the ear of a listener at slightly different times. This can significantly alter the spectrum of the sound received by the ear due to comb filtering. Figure 1 plots (in red) the two-sided spectrum of a band-limited white noise (0–6 kHz) whose RMS amplitude is 1. This figure also plots, in blue, the spectrum of the sum of that noise plus a time-delayed version of it where the delay is .00031 s, after the sum of the two noises has been adjusted to have a RMS amplitude of 1 (the same as the RMS of the original noise). The addition of these two sounds at this delay changes the spectrum from flat to one consisting of peaks and troughs. Hence, a target sound source with spectral energy in the regions where there are troughs in the masker will be partially unmasked in those regions.

✉ Bruce A. Schneider
bruce.Schneider@utoronto.ca

[1] Department of Psycholoygy, University of Toronto at Mississauga, 3359 Mississauga Road N., Mississauga, Ontario L5L 1C6, Canada

[2] Communication Arts, Sciences and Disorders, Brooklyn College, City University of New York, Brooklyn, NY, USA

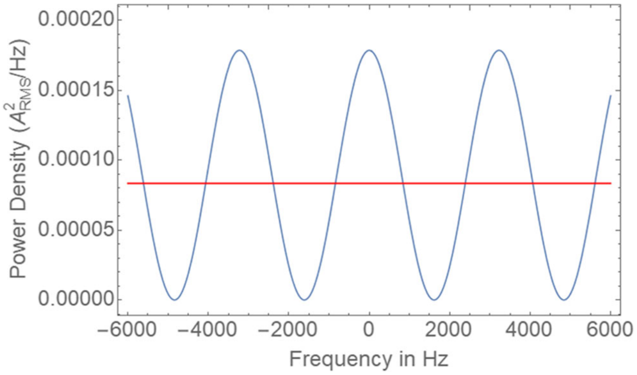

Fig. 1 The two-sided power density function of a band-limited (0–6 kHz) white noise (n[t]), the RMS amplitude of which is 1.0, is shown in red. Shown in blue is the two-sided power density function of n[t] + n[t – 0.00031 s], the RMS amplitude of which has been adjusted so that it too has an RMS value of 1.0

In most natural settings (no amplification), sound sources typically have a compact image emanating from the actual source location. In such cases, and in the absence of excessive reverberation, when the sound has a single source, it should be relatively easy for the listener to fuse the sound coming directly from the source and any secondary streams (e.g., reflections) into one auditory image. Moreover, the greater the relative amount of direct sound energy from the original source compared to the energy from additional secondary streams, the closer the source will be perceived (Mershon & King, 1975), and the magnitude of any comb-filtering effects due to the summation of the direct and reflected waveforms will be reduced.

Hence, a compact sound source is likely to be perceived as being closer to the listener than a diffuse sound source, with a spectrum level at the ear of the listener that is closer to that of the original source. Because sound sources close to the listener are likely to have a higher ecological salience than sources that are further away (in vision, this is referred to as the behavior urgency hypothesis; Franconeri & Simons, 2003), it could be that compact sound sources are more likely to capture the attention of the listener than those that are further afield.

The auditory scene, however, is substantially changed when amplification is introduced, and sound is presented over multiple loudspeakers. For example, imagine watching a play in which the actors are conversing in a marketplace. When no amplification is used, the voices that emanate from the actors on the stage in front of you will have a relatively compact image (the direct wave from a voice along with its reflections will fuse into a single auditory object with a precise location in space, provided the amount of reverberation is not excessive). The voice of an actor on the left side of the stage will appear to be located to the listener's left, whereas the voice of another actor on the right side of the stage will appear to be located to the listener's right. Moreover, based on the auditory information alone, the listener will be able to locate the sound source with some degree of precision.

However, the director, in order to increase the degree of realism, may have recorded activity in an actual marketplace, and play this recording over loudspeakers placed at various locations in the theatre. In this situation, the actors' voices are likely to have compact and more easily localized images, while the marketplace noise will have a diffuse image that appears to fill a large volume of space. One could imagine that the contrast in the timbre of the voices versus that of the marketplace noise could facilitate stream segregation of the actor's voices from the crowd noise. In addition, the comb filtering that occurs when the marketplace noise is played over spatially separated loudspeakers could produce troughs in the spectrum of the marketplace sounds reaching the listener's ears, thereby partially unmasking the actor's voices within those spectral regions where the troughs occur.

Now imagine watching the play with the actors' voices amplified and played along with the marketplace noise over the same loudspeakers, creating a diffused image of both the voices and the marketplace activity. In this situation, the voices of the actors will be perceived to be located in front of the listener, but, because of the loss of compactness in the image, will not be as precisely located. Moreover, listeners will not have the timbre contrast between the actors' voices and the marketplace noise to facilitate streaming, thereby reducing their ability to focus attention on the actors' voices. In addition, due to comb-filtering effects, troughs in the spectra of the actors' voices will occur in the same spectral regions as troughs in the spectra of the marketplace noise. Hence, presenting both the actors' voices and the marketplace noise over the same loudspeakers will not result in the unmasking of the actors' voices due to comb-filtering effects because the comb-filtering effects will be the same for both the actors' voices and the marketplace noise. Thus, listeners might find it more difficult to follow the play than when the actors' voices are not amplified, and there is a timbre contrast between the actors' voices, and the marketplace scene.

In this study, we attempted to determine in situations such as those described above, whether it would be easier to understand the actors when their voices are compact (coming from a single loudspeaker), and the background is diffuse (coming from three different, spatially separated loudspeakers), than when both the voices and the background sounds are compact. In addition, we wanted to determine the extent to which a nearby conversation (compact sound sources) would interfere when listening to diffuse target voices presented over a surround-sound system. Will compact sound sources located several seats away distract attention away from the diffuse voices of the play more than the same competing voices would if they were also introduced into the diffuse sound broadcast? In addition, the voices of the actors are subjected to comb filtering because they are played over several loudspeakers, whereas that of the nearby talkers are not. What effect might this have on the listener's ability to follow the play? The experiments reported here were motivated by a consideration of such situations, and how the configuration of sound sources in the auditory scene may affect speech perception. In addition, to better identify which levels of auditory processing are affected by timbre contrasts between masker and target, three different types of maskers were used: speech-spectrum noise, 12-talker babble, and two-taker competing speech.

The current study addresses three questions. First, will the contrast between the diffuseness levels of the target and competing masker provide listeners with acoustic information that would help them to better analyze the acoustic scene and segregate the incoming auditory streams into different sound sources? Second, when there is a contrast in timbre, is there a difference in speech perception between situations in which the target is compact and the masker diffuse (T_CM_D), versus when the target is diffuse and the masker compact (T_DM_C)? Third, does the effect of a timbre contrast differ among the three types of maskers (steady-state noise, a babble of voices, or two-talker speech)? We believe this to be the first systematic attempt to investigate how a difference in diffuseness of target voices relative to background sounds affects speech recognition.

Auditory streaming: Energetic and informational masking

To ascertain where in the auditory processing stream timbre differences between masker and target could be affecting listeners' ability to process the target speech requires a consideration of the different levels of processing involved in perceiving speech. Everyday speech perception can be a demanding task both at peripheral and more central processing levels, as most verbal communication takes place in the presence of other sound sources. A listener must first analyze the auditory scene into its components, identify the target stream, and extract it from the mixture of the competing sounds, before allocating attention to the target sound source. Any competing source that temporally and spectrally overlaps the target speech signal can interfere with the processing of the target speech at the auditory periphery by creating overlapping excitation patterns in the cochlea or the auditory nerve. This competition between target and masker at the periphery of the auditory system is often referred to as *energetic masking* or *peripheral masking* (Durlach, Mason, Kidd, Arbogast, Colburn, & Shinn-Cunningham, 2003). However, additional masking can occur at higher levels of auditory processing when the masker consists of meaningful speech that could interfere with the linguistic and semantic processing of the target speech. When listeners fail to successfully segregate the elements of the target signal from other similar sounds, this failure may allow the content of irrelevant streams to intrude into working memory and interfere with the processing of the target message. This type of interference, which is often referred to as *informational masking*, can occur independently of energetic masking (Durlach et al., 2003; Freyman, Helfer, McCall, & Clifton, 1999; Kidd, Mason, Richards, Gallun, & Durlach, 2008; Schneider, Pichora-Fuller, & Daneman, 2010; Schneider, Li, & Daneman, 2007). Based on the results of previous studies that have examined the differences between energetic and informational masking, we would expect the benefit obtained from a contrast in timbre to be larger when the masker causes substantial informational masking rather than when the masker is primarily energetic (e.g., Arbogast, Mason, & Kidd, 2002; Avivi-Reich, Puka, & Schneider, 2018; Ezattian, Avivi, & Schneider, 2010; Freyman, Balakrishnan, & Helfer, 2004). With respect to the current study, we might expect a timbre contrast to produce a greater release from masking when the masker is babble or competing speech than when the masker is steady-state noise.

Stream segregation and release from masking

The listener's ability to successfully segregate competing streams largely depends on the perceptual similarities between the target signal and other irrelevant sound sources present in the auditory scene. Stream segregation is especially challenging when the target and the competing streams share similar acoustical characteristics. Any dissimilarity between them may serve as an assisting cue that could help the listener to perceptually segregate the streams, and enhance release from masking (Bregman, 1990). For example, competing same-gender voices are more likely to informationally mask the target voice than, say, different-gender voices (e.g., Brungart, Simpson, Ericson, & Scott, 2001; Humes, Lee, & Coughlin, 2006; Vongpaisal & Pichora-Fuller, 2007). In such cases, the listener may be unable to parse the auditory scene into its components and keep the streams separated as the target speech unfolds.

A large number of acoustic cues that could assist stream segregation have been investigated in order to assess their potential to release the target signal from masking. One such cue, which has been extensively studied, and was found to provide a substantial release from masking, is the presence of spatial separation between the target signal and the other sound sources (e.g., Arbogast et al., 2002; Brungart & Simpson, 2002; Ezzatian et al., 2010). Several studies that have investigated the benefit of spatial separation did so by creating a perceived spatial separation using the precedence effect to change the virtual location of the sound sources (e.g., Avivi-Reich, Daneman, & Schneider, 2014). This effect can be achieved by presenting the same sound over two loudspeakers located to the right and left of the listener, with the sound coming from one of the speakers lagging the other one by a couple of milliseconds. In such a scenario, the listener perceives the sound as emanating from the leading loudspeaker (e.g., Rakerd, Aaronson, & Hartmann, 2006). Using the precedence effect allowed the investigators to study the impact of spatial separation without altering the signal-to-noise ratio (SNR) at each ear (Freyman et al., 1999). However, use of the precedence effect also alters the overall image of the sound source, creating a more diffused image that cannot be as precisely located in space as a sound source that is presented over a single loudspeaker (Avivi-Reich et al., 2014). Moreover, presenting the same signal from more than a single source is likely to create comb-filtering effects. Although the possible role that timbre differences might play in stream segregation when the precedence effect is used to spatially separate sounds sources has been briefly discussed in previous studies (e.g., Freyman et al., 1999), as far as we know, there have been no systematic investigations of how a timbre difference might affect stream segregation in the absence of real or perceived spatial separation.

Recent results, however, suggest that timbre differences among various sound sources might affect a listener's ability to comprehend what is being said. Avivi-Reich et al. (2014) and Avivi-Reich, Jakubczyk, Daneman, and Schneider (2015) tested speech recognition using the R-SPIN sentences (Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984) in two different spatial conditions. In the first study, Avivi-Reich et al. (2014) used a real no-separation condition, in which both the target voice and the babble masker were presented over the central loudspeaker only. In the second study (Avivi-Reich et al., 2015), the target voice was presented over the central loudspeaker only, while the babble masker was playing from three loudspeakers placed symmetrically in front of the listener. Even though there was no perceived spatial separation between the target and masker in both studies, the SNRs yielding 50% correct repetition of the target words were somewhat higher in the first study than those obtained in the second study (1.86 vs. -1.51dB SNR for a 3.37 dB difference). These results could imply: (1) under certain conditions, a timbre difference could be used as an acoustic differentiating cue to improve stream segregation and ease the listening difficulty experienced by the listener; (2) listeners may find it easier to form auditory objects when the sound source is compact rather than diffuse. In addition, the fact that compact sources have a more precise location in space than diffuse sources may attract listeners' attention to the compact sources. If so, we might expect to find better speech recognition when the target is compact and the masker diffuse (T_CM_D) than when the target is diffuse and the masker is compact (T_DM_C). Moreover, when the target is compact and the masker is diffuse (T_CM_D), the presence of troughs in the spectrum of the masker produced by comb filtering could facilitate recognition of the target, because the presence of troughs in the masker would partially unmask the energy in the speech signal falling into those troughs. This would have the effect of increasing the intelligibility of the speech signal. By way of contrast, when the target is diffuse and the masker is compact (T_DM_C), the advantage of a difference in timbre between the masker and target could be offset by the fact that comb filtering introduces troughs in the spectrum of the diffuse target that are not present in the masker. This could increase the degree to which the target is masked.

Method

Participants

The participants were 24 younger normal-hearing listeners whose first language was English. The participants were divided into two experimental groups: 12 young adults (mean age: 21.93 years; *SD*: 2.02) were tested using a compact target speech source (T_C); and a different group of 12 young adults were tested when the target speech source was diffuse (T_D; mean age: 20.14; *SD*: 1.76). Listeners were all born and raised in a country in which the primary language was English and were not fluent in any other language at the time of participation. Participants were students recruited from the University of Toronto. All participants were asked to complete a questionnaire regarding their general health, hearing, vision, and cognitive status. Only participants who reported that they were in good health and had no history of serious pathology (e.g., head injury, neurological disease, seizures, and the like) were included. None of the participants had any history of hearing disorders, and none used hearing aids. The study reported here was approved by the Ethics Review Board of the University of Toronto.

Materials, apparatus, and procedure

Audiometric thresholds, Nelson-Denny reading comprehension skill (Brown, Bennett, & Hanna, 1981), and Mill Hill vocabulary knowledge (Raven, 1965) were measured during each participant's first session. The speech recognition task was administered during a second experimental session. Each of the two sessions was typically 1–1.5 h in duration. All participants gave their written informed consent to participate in the experiments and were paid a modest stipend ($10/h) for their participation.

Hearing measures

Audiometric testing Pure-tone air-conduction thresholds were measured at nine frequencies (0.25–8 kHz) for both ears using an Interacoustics Model AC5 audiometer (Interacoustic, Assens, Denmark). All participants were required to have a pure-tone threshold 20 dB HL or lower from 0.25–8 kHz. In addition, participants who demonstrated unbalanced hearing (more than a 15-dB difference between ears at any of the nine tested frequencies) were excluded from participation. The average audiograms for the two groups of participants are shown for the right and the left ears in Fig. 2. The two groups of younger adults had similar hearing levels at all frequencies.

Language proficiency measures

Vocabulary knowledge Participants were asked to complete the Mill Hill vocabulary test (Raven, 1965), which is a 20-item synonym test. In this test, participants were required to match each test item with its closest synonym from six listed alternatives. No time restrains were applied.

Reading comprehension skill The Nelson-Denny test (Brown et al., 1981) was used to assess the reading comprehension skills of each participant. In this test, the participants had to read through a series of eight independent passages and answer multiple-choice questions regarding the content of the passages. This test includes a total of 36 questions and was limited to 20 min. Participants were instructed to answer as many questions as possible within the time given.

Semantically anomalous sentences-recognition task

During the experimental recognition task, the listener was seated in a chair located in the center of an Industrial Acoustic Company (IAC) sound-attenuated chamber, the internal dimensions of which were 283 cm in length, 274 cm in width, and 197 cm in height. Two loudspeakers were placed symmetrically in the frontal azimuthal plane at 45^{o} angles to the left and right of the listener, and a third loudspeaker was placed directly in front of the listener. The distance between the center of the listener's head and each one of the speakers was approximately 170 cm. The height of the loudspeakers was adjusted to match the ear level of a seated listener of average body height. All the acoustic stimuli used for the current study were digitized at 20 kHz sampling rate using a 16-bit Tucker Davis Technologies (TDT, Gainesville, FL, USA) System II and custom software. The digital signals were converted to analog form using Tucker-Davis Technologies digital-to-analog converters under the control of a Dell computer with a Pentium 4 processor. The analog outputs were low-passed at 10 kHz, attenuated by two programmable attenuators, and then presented to the participant either through the central loudspeaker (when presenting compact target speech and/or compact maskers) or from all three speakers (when presenting diffuse target speech and/or diffuse maskers).

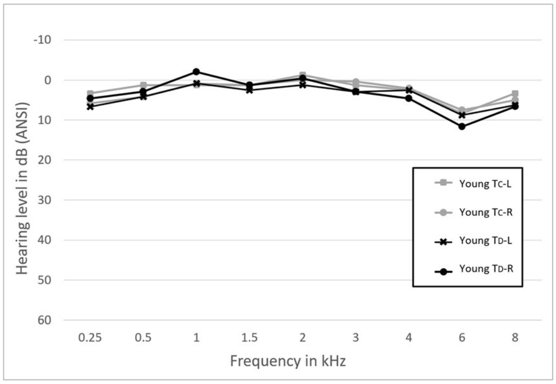

Fig. 2 Average audiograms for the two groups of participants are shown for the right and the left ears

Presenting the sound from three different loudspeakers rather than a single loudspeaker alters the timbre of the sound due to comb filtering, because the sound waves from the three different loudspeakers arrive at an ear at slightly different times. This produces peaks and troughs in the spectrum of the sound which changes its timbre. In addition, if: (1) the center of the person's head is not fixed precisely at the same distance from each loudspeaker, and/or (2) the head is not a perfect sphere; and/or (3) there is any asymmetry with respect to reverberation in the chamber, there will be interaural differences in the signals arriving at the ears. Such interaural differences could lead to the stimulus being perceived as diffuse (Lavandier & Culling, 2008). Because the participant's head was not held in position by a bite bar, it could not be precisely centered with respect to the three loudspeakers. Hence, in addition to timbral differences due to comb filtering in the three-loudspeaker situation, interaural differences in the signals arriving at the two ears could lead to a three-loudspeaker sound being perceived as more diffuse than the sound emanating from only a single loudspeaker.

In order to confirm that the three-loudspeaker condition (L3) was perceived as producing a more diffuse sound than the one loudspeaker condition (L1), we asked eight different young adults to rate the perceived diffuseness of the L1 and L3 conditions. The female target talker and each of the three types of masker stimuli used in the study (12-talker babble, two competing female talkers, speech spectrum noise) were presented either from the center loudspeaker only (L1) or over all three loudspeakers (L3), which resulted in a total of eight different conditions tested. Loudspeakers were positioned at the exact location as in the study and the same settings and intensity were used as in the study. Eight young adults (18–24 years old), who were undergraduates at the University of Toronto Mississauga, were tested individually in the same double-walled sound-attenuated booth. Each participant listened to each of the eight conditions (a segment that was equal in time to seven target sentences). The order of the different types of stimuli was counterbalanced between participants, as well the order of the L1 and L3 conditions within each type. Participants were asked after each stimulus presentation to indicate: "On a scale of 1 to 10, how much do you think the sound filled the room?" In addition, after the presentation of each L1 and L3 pair, participants were asked to determine whether they felt that the L1 or L3 condition had a more spreadout sound.

A repeated-measures ANOVA of the degree to which the sound appeared to fill the room, with the four types of stimuli and the two loudspeaker conditions as within-participant variables, found that participants rated the sound played from all three loudspeakers (L3) as filling the room more than the same sounds played from a single loudspeaker (*F*[1,7] = 28.985, p = 0.001). Neither the type of stimuli nor any of the interactions between stimulus type and the L1–L3 factor reached statistical significance. On the comparison question, participants identified the L3 sounds to be more spread-out than the L1 sounds in 96.875% of the comparisons. Hence the L3 sounds were perceived to be more diffuse than the L1 sounds.

Target sentences consisted of 312 syntactically-correct-but-semantically-anomalous sentences spoken by a female talker, which were developed by Helfer (1997) and previously used in experiments by Freyman et al. (1999), Li, Daneman, Qi, and Schneider (2004), and Ezzatian et al. (2010). Each of these sentences contained three target words in sentence frames such as "A *spider* will *drain* a *fork*," or "A *shop* can *frame* a *dog*" (target words italicized). The sentences were divided into 24 lists containing 13 sentences each. In the Compact-Target group the target sentences were presented over the front loudspeaker while the masker was either presented over all three loudspeakers to create a diffused image or over the central loudspeaker only to create a compact image of the masker. In the Diffuse-Target group the target sentences were presented over all three loudspeakers to create a diffused target image while the masker was either presented from all three loudspeakers to create a diffused image, or over the central loudspeaker only to create a compact image of the masking sound source.

Target sentences were presented with either one of three types of masking stimuli: noise, babble, or speech. The noise masker was a steady-state speech-spectrum noise recorded from an audiometer (Interacoustic [Assens, Denmark] model AC5), the babble was a 12-talker babble taken from the modified Speech Perception In Noise (SPIN) test (Bilger et al., 1984), and the speech masker was a 315-s long track created using an additional set of semantically anomalous sentences uttered by two female talkers and repeated in a continuous loop. The target sentences were presented at an average sound pressure of 55 dBA at the estimated center of a listener's head, whether a single loudspeaker was playing the sentences (T_C; compact target) or all three (T_D; diffused target). The sound pressure was measured using a Brüel and Kjær (Copenhagen, Denmark) KEMAR dummy-head. Masker intensity was measured separately for the conditions in which the masking sounds were played only over central loudspeaker (compact masker), and when they were simultaneously played over all three loudspeakers (diffused masker). The voltages of the sounds presented in the three loudspeaker conditions were adjusted so that the sound pressure produced at the KEMAR head in the three-loudspeaker conditions matched the sound pressure produced at the KEMAR head in the single-loudspeaker conditions. Hence, the voltage at each of the three loudspeakers that produced a specified dB SPL level at the ear of the dummy head when all three loudspeakers were in use was lower than the voltage level of the signal when it was presented over the central loudspeaker only.

We also checked the correctness of our sound level calibrations by placing a Bruel and Kjaer sound level meter (Model 2260) at the location corresponding to the approximate center of a participant's head. The readings from this sound-level meter were between .5 and 1.5 dB higher than those found using the dummy head in all four of the conditions in this experiment. The slightly higher levels found using the sound level meter in the free field are expected because they do not include the head-related transfer functions.

While the target's sound pressure level was kept constant at 55 dBA throughout the experiment, the sound pressure level of the masker was adjusted in order to produce four different SNRs depending on the Masker Type and the Timbre Condition tested. The different SNRs used were initially chosen based on previous studies that used similar stimuli in noise (e.g., Ezzatian et al., 2010) and then altered according to the results of preliminary pilot testing done under the present listening conditions. The SNRs used in the current study are presented in Table 1. A single list of 13 sentences was used for each of the SNR values that appear in the table.

The sentences in each of the 24 target lists were presented at a constant SNR in all of the four Timbre Conditions: (1) target compact, masker compact (T_CM_C); (2) target compact, masker diffuse (T_CM_D,); (3) target diffuse, masker diffuse (T_DM_D); and (4) target diffuse, masker compact (T_DM_C). Sentence lists and SNRs were counterbalanced across participants such that each list was presented at each of the four

Table 1 The values of the four SNRs used under each condition (compact target and maskers (T_CM_C), compact target and diffuse maskers (T_CM_D), diffused target and maskers (T_DM_D), diffused target and compact maskers (T_DM_C)), for each of the three masker types, presented separately for each of the two experimental groups

T_CM_C			T_CM_D		
S	N	B	S	N	B
3	2	-6	2	1	-11
-3	-3	-12	-4	-4	-17
-9	-8	-18	-10	-9	-23
-15	-13	-24	-16	-14	-29
T_DM_C			T_DM_D		
S	N	B	S	N	B
6	6	-4	6	3	-5
0	1	-10	0	-2	-11
-6	-4	-16	-6	-7	-17
-12	-9	-22	-12	-12	-23

different SNRs an equal number of times in each group. Additionally, each sentence list was presented in each of the Timbre Conditions (T_CM_C, T_CM_D, T_DM_D, T_DM_C) and Masker (speech, babble, noise) combinations an equal number of times. In each experimental group (T_C, T_D), six participants were first tested with a diffused masker (M_D) for the first 12 lists, and with a compact masker (M_C) for the remaining 12. The other six participants were tested in the reverse order. Before beginning the experimental session, an explanation was given to familiarize the participant with the task. Participants were asked to repeat the target semantically anomalous sentence after each presentation and were scored for any keyword that was repeated correctly. Performance was assessed both online while the session was taking place and later by a second research assistant who listened to the participant's recorded responses. After the participant had responded, the researcher initiated the presentation of the next trial. Each trial started with the masker sound, which was followed 1 s later by a target sentence. The masker remained on during the sentence, then the masker was gated off when the target sentence was turned off. After completing 12 lists, a short break was offered to the participants.

Results

Table 2 presents the gender breakdown, mean age, Mill Hill test of vocabulary knowledge, and Nelson-Denny test of reading comprehension results for each of the two groups. The vocabulary scores and reading comprehension scores were similar in the two groups. There was a slight age difference (1.79 years) between the two groups ($t(22)=2.30$, $p=0.031$).

Figure 3 shows the percentage of correctly identified keywords, averaged over the 12 participants in each group, as a function of SNR, when the masker was speech spectrum noise (left panels), two-talker speech (middle panels), or 12-talker babble (right panels). The top panels present the psychometric functions when there is no contrast in timbre between the target and masker (T_CM_C and T_DM_D). The bottom panels present the corresponding data when there is a contrast between the target and masker (T_CM_D and T_DM_C). Circles represent the data for compact targets (T_C) with squares representing the data for diffuse targets (T_D). Logistic psychometric functions of the form $y=\frac{1}{1+e^{-\sigma(x-\mu)}}$ were fit to these data points. The parameter μ denotes the 50% point on the psychometric function (the threshold), and σ controls the slope of the function (for a description of the fitting procedure see Yang, Chen, Huang, Wu, Wu, & Schneider, 2007). The estimated 50% points are indicated by the dashed vertical lines when the target speech was compact (T_C), and solid vertical lines for when the target speech was diffuse (T_D).

Table 2 Demographic information (mean age, gender distribution, mean vocabulary, and reading comprehension scores) for the participants divided into the two experimental groups tested

	Group A (T_C)		Group B (T_D)	
	M	SD	M	SD
Age	21.93	2.02	20.14	1.764
Gender	8 F + 4M		11F + 1 M	
Vocabulary (max=20)	14.5	1.24	13.08	2.71
Reading comprehension (max=36)	28.58	3.92	25.75	6.30

An examination of this figure suggests that when there is no contrast in timbre between the target and the masker (T_CM_C or T_DM_D), speech recognition seems to be independent of whether the target sound source is diffuse or compact. However, when one sound source is compact and the other is diffuse (T_CM_D or T_DM_C), performance seems to be significantly better when the target speech is the compact sound source. The estimated slopes of the psychometric functions when the masker is noise appear to be steeper than those estimated when the masker is babble or speech.

These visual impressions were mostly confirmed by statistical analyses performed on the parameters of the individual psychometric functions. Specifically, psychometric functions were fit to all individuals in order to obtain individual estimates of the threshold, μ, and the slope, σ. To confirm these visual patterns, we conducted a 2 Target Timbre (T_D vs. T_C) × 3 Masker Types (Noise, Babble, Speech) × 2 Masker Timbre conditions (M_D vs. M_C) ANOVA with Target Timbre as between-subjects factor and Masker Type and Masker Timbre as within-subject factors.

Thresholds

The ANOVA for thresholds (μ) revealed that all three factors have a significant main effect on thresholds (Target Timbre: F [1, 22] = 71.218, p <0.001; Masker Timbre: F [1, 22] = 35.51, p <0.001; Masker Type: F [2, 44]= 1594.75, p <0.001). In addition, a significant two-way interaction was found between Masker Type and Target Timbre (F [2, 44]=7.98, p =0.001), as well as a significant three-way interaction between Masker Type, Masker Timbre, and Target Timbre (F [2, 44]=4.41, p =0.018).

The nature of the three-way interaction is illustrated in Fig. 4, which plots the SNR corresponding to 50% correct recognition for the two no-timbre-contrast conditions (T_CM_C, T_DM_D) on the left panel, and for the timbre-contrast conditions (T_CM_D, T_DM_C) on the right, for each of the three Masker Types separately. The left panel shows that when there is no timbre contrast between the target and the masker (T_CM_C and T_DM_D), speech recognition performance is similar in both conditions, which implies that the Target Timbre has no significant effect when there is no timbre contrast. However, when looking at the right panel, which presents the performance under the two timbre-contrast conditions (T_CM_D and

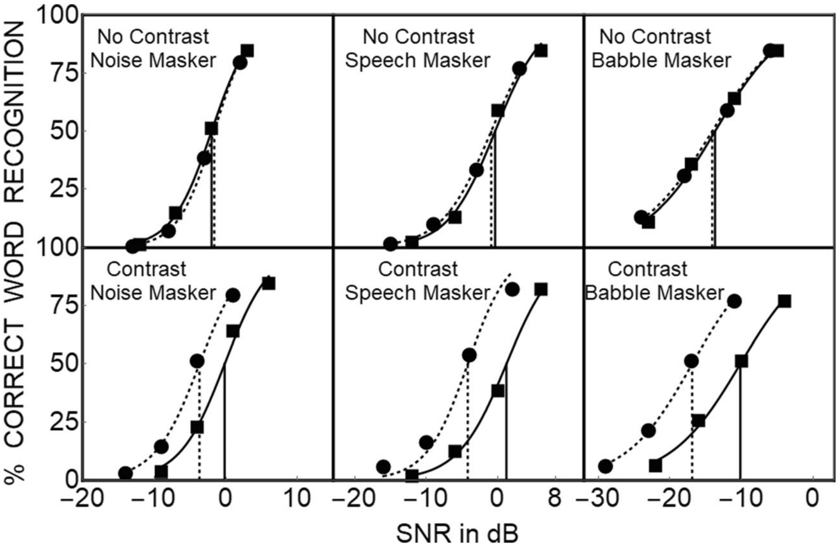

Fig. 3 Circles represent the average data when the target was compact (T_C), squares represent the data for diffuse targets (T_D). **Top panels**: Average percent correct word identification as a function of signal-to-noise ratio (SNR) in dB when there is no contrast in timbre conditions (T_CM_C and T_DM_D) for the three types of maskers (Noise, Speech, and Babble). **Bottom panels**: Average percent correct word recognition as a function of SNR when there was a timbre contrast between target and masker (T_CM_D and T_DM_C). Thresholds (SNRs corresponding to 50% correct on the psychometric functions) are indicated by solid vertical lines when the target speech was a diffuse sound source (T_D) and dashed vertical lines when the target speech was a compact sound source (T_C)

T_DM_C), there is a difference between the two conditions. Overall, the SNRs corresponding to 50% correct repetition are lower (better) when the target is the compact sound source than when the masker is the compact sound source. In addition, the right panel in the figure suggest that the differences found between the two timbre-contrast conditions are dependent on the type of masker condition. In both panels, thresholds are much lower when the masker is babble than when the masker is either speech-spectrum noise or competing speech.

To better understand the nature of these interactions, the ANOVA was repeated separately for the conditions in which there was a contrast between the Target Timbre and the Masker Timbre, and then again for the conditions in which there was no such contrast in timbre. The results showed that when there was no timbre contrast (Either T_DM_D or T_CM_C), the only main effect that was found to be statistically significant was Masker Type (F [2, 44] = 1031.5, p < *0.001*). There was no evidence of a difference due to Target Timbre (F [1, 22] > 1, p = *0.625*), or any interaction between Target Timbre and Masker Type (F [2, 44] = 1.445, p = *0.247*). However, when there was a timbre contrast (either T_CM_D or T_DM_C), both the main effect of Masker Type as well as Target Timbre were found to be statistically significant (F [2, 44] = 760.88, p < *0.001*, F [1, 22] = 85.42, p < *0.001*, respectively), as well as the interaction between the two (F [2, 44] = 9.372, p < *0.001*). Hence, the two-way interaction between Masker Type and Target Timbre only appears when there is a timbre contrast between Target Type and Masker Type.

To get a better picture of the nature of the interaction between Target Timbre and Masker Type when there is a timbre

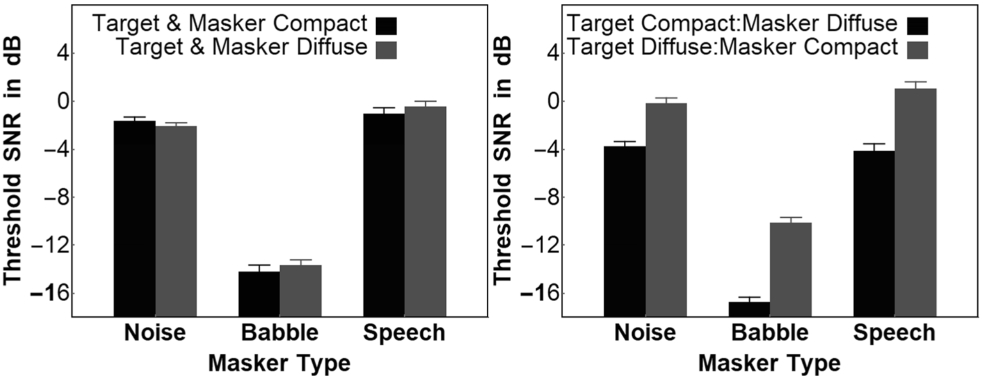

Fig. 4 The signal-to-noise ratios (SNRs) corresponding to 50% correct recognition for the two no-timbre-contrast conditions (T_CM_C, T_DM_D) are presented on the left panel, and those for the timbre contrast conditions (T_CM_D, T_DM_C) are presented on the right, for each of the three Masker Types separately

contrast between the target and masker (T_CM_D, T_DM_C), we first computed the average threshold for each of the Masker Types.

$$\overline{\mu}_{Noise} = \frac{\overline{u}_{Noise,T_CM_C} + \overline{u}_{Noise,T_DM_D} + \overline{u}_{Noise,T_CM_D} + \overline{u}_{Noise,T_DM_C}}{4}$$

$$\overline{\mu}_{Babble} = \frac{\overline{u}_{Babble,T_CM_C} + \overline{u}_{Babble,T_DM_D+}\overline{u}_{Babble,T_CM_D} + \overline{u}_{Babble,T_DM_C}}{4}$$

$$\overline{\mu}_{Speech} = \frac{\overline{u}_{Speech,T_CM_C} + \overline{u}_{Speech,T_DM_D} + \overline{u}_{Speech,T_CM_D} + \overline{u}_{Speech,T_DM_C}}{4}$$

Second, we subtracted the average threshold for a Masker Type from the thresholds for the timbre-contrast conditions (T_CM_D, T_DM_C) for that Masker Type. Figure 5 plots these adjusted thresholds for the three maskers. As can be seen clearly in Fig. 5, when the target is compact and the masker is diffuse (T_CM_D), the difference is negative, indicating that the contrast in timbre facilitated speech recognition for all three maskers. But when the target is diffuse and the masker is compact (T_DM_C), the difference is positive, indicating that the timbre contrast has a detrimental effect on speech recognition.

Interestingly, the size of the difference in thresholds between the two types of timbre contrast appears to be larger for Babble and Speech maskers than it is for the Noise masker. A t-test of whether the size of the difference in thresholds was the same for Babble and Speech maskers did not reach significance (t[11] = -1.9051, p = .07). Hence, we averaged the thresholds across the Babble and Speech maskers, and compared these average thresholds to those for the Noise masker. The difference in thresholds between the T_CM_D and T_DM_C conditions for the average of the Babble and Speech maskers was significantly greater than the comparable difference in thresholds for the Noise masker (t[11] = 3.9955, p < .001). Hence, the difference in thresholds between T_CM_D and T_DM_C is significantly larger for the maskers (Babble and Speech) that are informationally more complex than a Noise masker.

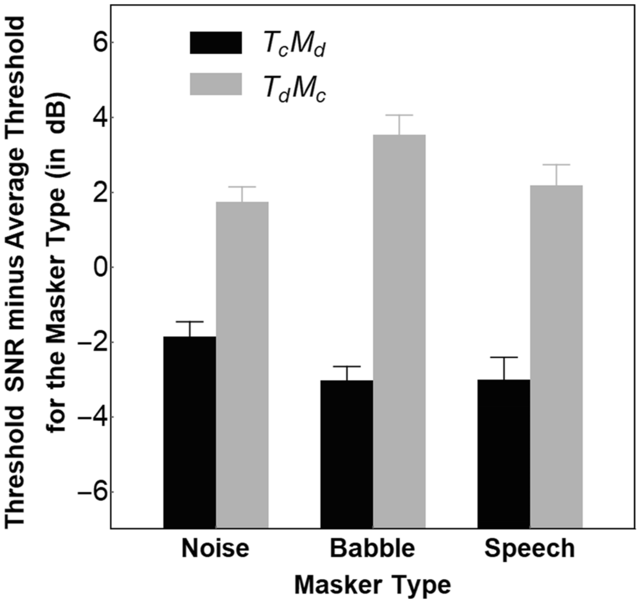

Fig. 5 The average signal-to-noise rations (SNRs) corresponding to 50% correct recognition under each Timbre Condition and Masker Type minus the average SNR threshold measured across all four conditions (T_CM_C, T_DM_D, T_CM_D, T_DM_C) calculated for the same Masker Types

The contribution of vocabulary knowledge and reading comprehension to thresholds

To determine whether individual differences in linguistic competence (vocabulary and reading comprehension skills) could account for a significant portion of the variance in the speech recognition task, the Mill Hill and Nelson-Denny scores were centered within each experimental group (target compact and target diffuse), and an ANCOVA analysis was then conducted with Mill Hill vocabulary scores, and Nelson-Denny reading comprehension scores as covariate measures following the procedure recommended by Schneider, Avivi-Reich, and Mozuraitis (2015). The ANCOVA results showed a significant interaction between vocabulary Mill-Hill scores and the performance under the three different Masker Types (*F*(2,40)=4.545, *p=0.017*). However, no similar interaction between Masker Type and Nelson-Denny scores was found (*F*(2,40)=1.19, *p=0.315*). Figure 6 presents the correlation found between Mill-Hill vocabulary scores and the 50% correct SNR thresholds under each of the Masker Type levels (Noise, Babble, Speech). The correlation was found to be significant only when the Masker Type was Speech (p=0.05). When the background masker was competing speech, the 50% correct SNR dB thresholds tended to be lower (better) when the vocabulary scores were higher.

Slopes of the psychometric functions

Figure 7 presents the average slopes of the psychometric functions. The left panel presents the slopes when there was no contrast in timbre between the target speech and masker (T_CM_C and T_DM_D), while the right panel shows the slopes when there was a contrast (T_DM_C and T_CM_D). Slopes appear to be steeper for Noise than for Speech, and steeper for Speech than for Babble. In addition, the slopes in the absence of a timbre contrast appear to be steeper than in the presence of a timbre contrast. Finally, the slope difference between when the target was compact and the masker was diffuse (T_CM_D) versus when the target was diffuse and the masker was compact (T_DM_C), appears to be larger when the masker was Speech than when it was either Noise or Babble. To confirm these observations, the slopes (σ) of the individual psychometric functions were also analyzed using a 2 Target Timbre Condition × 3 Masker Type × 2 Masker Timbre condition ANOVA. This analysis revealed a significant main effect of Masker Type on slopes (*F* [2, 44] = 54.07, *p <0.001*), as well as a significant two-way interaction between Masker Timbre

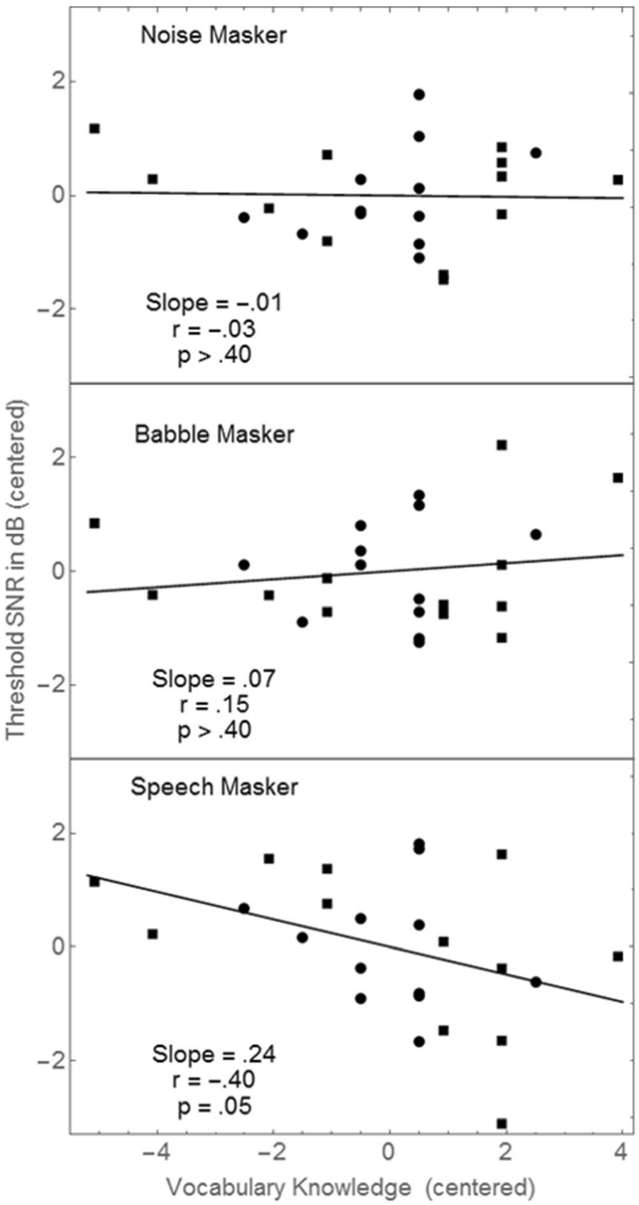

Fig. 6 Centered 50% correct recognition SNR thresholds plotted against the vocabulary Mill-Hill scores under the three different Masker Types. Slopes, p- and r-values are reported for each Masker Type level

and Target Timbre Condition (*F* [1, 22]=4.46, *p=0.046*) and a three-way interaction between Masker Type, Masker Timbre, and Target Timbre (*F* [2, 44]=5.01, *p=0.011*).

To pinpoint the source of the three-way interaction, the slopes were analyzed separately for when there was a timbre contrast between target and masker (T_CM_D and T_DM_C), and when there was no such contrast (T_CM_C and T_DM_D). The results showed that in both types of conditions, with and without timbre contrast, the main effect of Masker Type is statistically significant (*F*[2,44]=21.21, *p<0.001*), *F* [2, 44]=47.6, *p<0.001*, for timbre contrast and no timbre contrast, respectively). For the no timbre-contrast conditions, neither the main effect of Target Type (F[1,22] < 1) nor the interaction between Masker Type and Target Type (F[2,44] < 1) were significant. The main effect of Target Type also was not significant under timbre-contrast conditions (F[1,22] = 1.712, p = .204). However, under the timbre-contrast conditions the two-way interaction between Masker Type and Target Timbre was found to be statistically significant (*F*[2, 44]=4.388, *p=0.018*). *Post hoc*, univariant ANOVAs were conducted for each of the maskers separately when there was a contrast between target and masker (right panel of Fig. 7). Only when the masker was Speech was a significant difference found between Compact and Diffuse Target Timbre (*F*[1,22] = 6.082, *p* =*0.022*). Hence, this difference in slopes for the Speech masker condition when there is a timbre contrast between target and masker (right panel of Fig. 7) is responsible for the three-way interaction, and also for the two-way interaction between Masker Type and Target Timbre mentioned before.

Discussion

Signal-to-noise (SNR) thresholds and performance

The primary result of the current study is captured by the three-way interaction found between Target Timbre, Masker

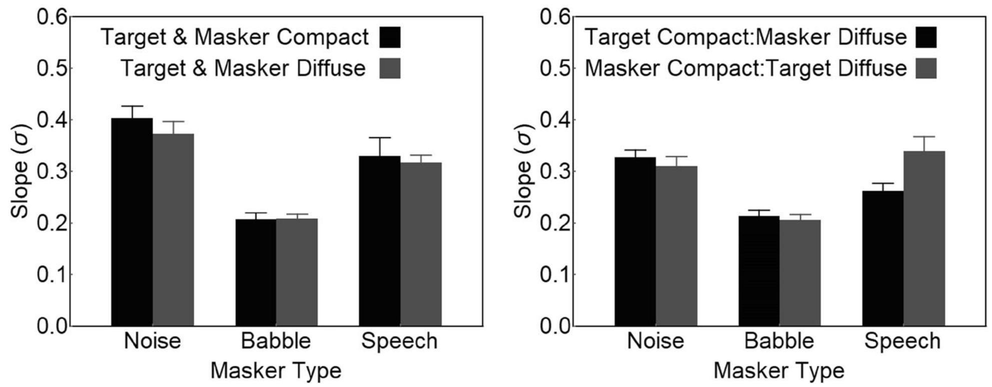

Fig. 7 The average slopes (σ) of the psychometric functions for the two no-Timbre-contrast conditions (T_CM_C, T_DM_D) are presented on the left panel, and those for the Timbre contrast conditions (T_CM_D, T_DM_C) are presented on the right, for each of the three Masker Types separately

Timbre, and the Masker Type. This interaction reveals that the Target Timbre has no significant effect when there is no timbre contrast (T_CM_C and T_DM_D). However, when comparing the SNR thresholds when such timbre contrast exists to the SNR thresholds found when there is no timbre contrast, the SNR thresholds are lower (better) when the target is the compact sound-source (T_CM_D) and higher (worst) when the masker is the compact sound source (T_DM_C). In addition, the right panel in Fig. 4 suggests that the differences found between the Timbre Conditions are dependent on the Masker Type. Specifically, the effect of timbre contrast appears to be larger when the masker has some informational content (babble and two-talker speech) as opposed to when it is primarily energetic.

These results are consistent with the hypothesis that compact sources with a precise location may attract the attention of the listener. When the target is compact and the masker is diffuse (T_CM_D), drawing attention to the target has a beneficial effect on word recognition. On the other hand, when the masker is compact and the target is diffuse (T_DM_C), the fact that attention is drawn toward the compact sound source interferes with recognizing the words in the diffuse target sentence.

The results are also consistent with the advantages and disadvantages in the SNRs that might be found in the different combinations of Timbre Conditions. Figure 8 presents a diagram of the loudspeaker arrangement in this experiment. The distances between the loudspeakers and the position of the center of the listener's head was 1.7 m. The central loudspeaker was located at 0° azimuth, with the left and right loudspeakers offset by 45°. Assuming that each of the listener's ears is approximately 3 in. from the center of the head, we can calculate the distance from each loudspeaker to the position of each ear opening, and thereby the time it takes for signals simultaneously presented over each loudspeaker to reach each ear. Calculations for the left ear are presented in the Appendix.

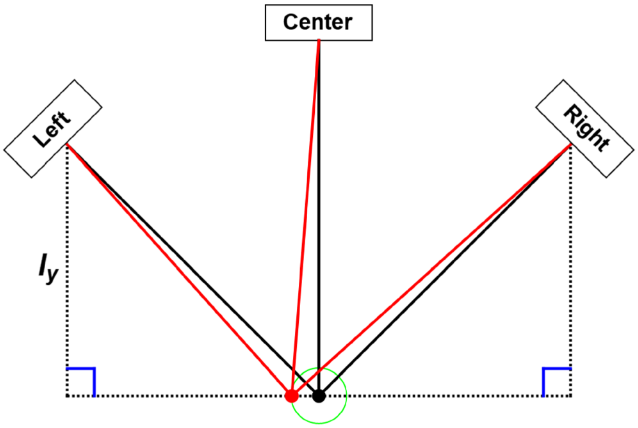

Fig. 8 The configuration of the loudspeakers in this experiment. The green circle represents the position of the listener's head. The center of each loudspeaker was located 1.7 m from the center of the head (black lines). The central loudspeaker was positioned directly ahead with the other two loudspeakers positioned 45° to the left and right of center. The red lines represent the distance to the opening to the left ear, which was set to 3 in. to the left of the center of the head

The red line in Fig. 9 plots the positive half of the long-term spectrum of a band-limited white noise (0–6 kHz), presented only over the central loudspeaker, as measured at the opening of the left ear of the hypothetical listener shown in Fig. 8. The calculation of this spectrum did not take into account head-related transfer functions and assumed an anechoic environment (see Appendix). The green line represents the long-term spectrum of the noise arriving at the left ear when the same noise is presented simultaneously over all three loudspeakers. For this case, the amplitude of the noise presented over each of the three loudspeakers was reduced relative to the noise present over the single central loudspeaker so that the overall RMS amplitude of the three-loudspeaker noise was the same as the RMS of the single loudspeaker noise. Note that at the location occupied by the ear, the spectrum of the single noise is flat whereas the spectrum of the three-loudspeaker noise is notched due to comb filtering.

A comparison of the condition where both target and masker are compact (T_CM_C) to the condition where the target is compact and the masker is diffuse (T_CM_D) shows that, in most spectral regions, the SNR is higher (more favorable) when the target is compact and the masker diffuse (T_CM_D) than it is when both target and masker are compact (T_CM_C). Here, a timbre contrast improves the SNR. On the other hand, the SNR appears to lower in most spectral regions when the target is diffuse and the masker compact (T_DM_C) than it is when both the target and the masker are diffuse (T_DM_D). Here, a change in timbre is detrimental with respect to SNR rather than beneficial.

When we compare the condition where both target and masker are compact (T_CM_C) to the condition where both are diffuse (T_DM_D), we see that the SNRs are comparable in both conditions. This is consistent with the results (see Fig. 4) that show that thresholds for sentence recognition are comparable for these two conditions across all three Masker Types. However, when there is a timbre contrast, Fig. 9 indicates that performance should worsen when the target is diffuse and masker is compact (T_DM_C) and improve when the target is compact and the masker is diffuse (T_CM_D), a result that is consistent with the data in Fig. 5. Hence, the pattern of results found here is consistent with what we would expect from the comb-filtering effects that occur when the same sound is played over multiple loudspeakers versus when it is played over a single loudspeaker only.

The degree to which comb filtering might affect speech perception will, of course, depend on the spatial locations of the loudspeakers, and time delays in the signals played over them, and the position and distance of listeners with respect to the locations of the loudspeakers. It will also depend on the head-related transfer function of the listener, as well as the

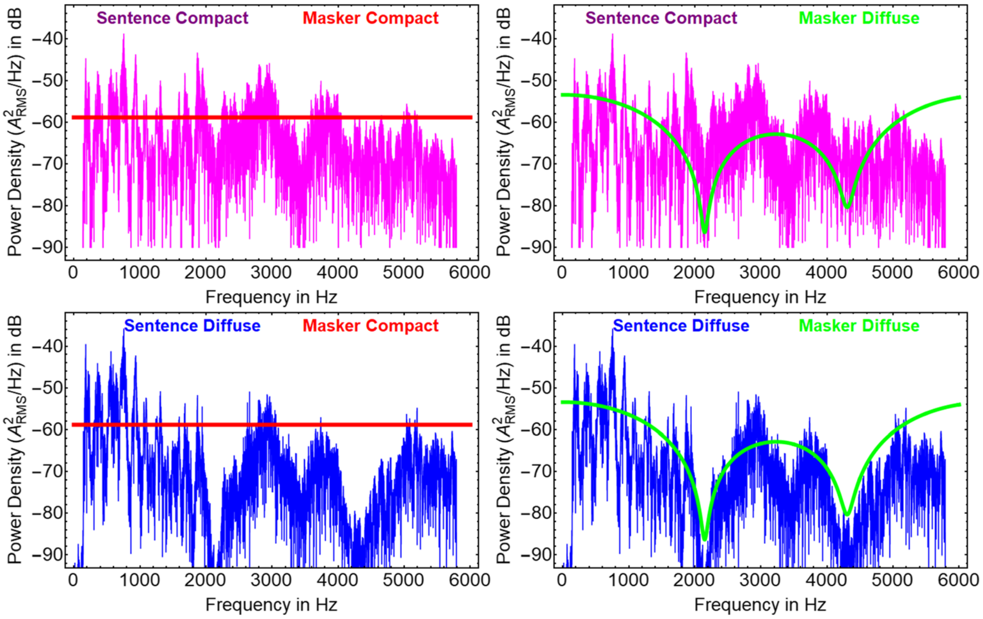

Fig. 9 Average spectra for two stimuli at the left ear of the hypothetical listener in Fig. 8: (1) a band-limited white noise (0–6 kHz) and (2) a sentence ("That ocean could shadow our peak"). These two stimuli could be presented either over a single loudspeaker or over all three loudspeakers simultaneously, and all four combinations were considered. All four stimuli have been equated with respect to their RMS amplitudes. The spectrum of the noise masker, when presented over a single loudspeaker (masker compact), is shown in red. The spectrum of the noise masker when played over all three loudspeakers simultaneously (masker diffuse) is shown in green. The spectrum of the sentence when played over a single loudspeaker (target compact) is shown in purple. The spectrum of the sentence when played over all three loudspeakers simultaneously (target diffuse) is shown in blue

orientation of the listener's head with respect to the loudspeakers. Finally, it will depend on the sound-attenuating characteristics and the distances of all sound-reflecting surfaces in the sound field. Hence, any additional comb-filtering due to the presentation of the same sound over multiple loudspeakers that might occur in a reverberant environment would be hard to predict. Nevertheless, Fig. 9 indicates that it is quite possible that the presentation of a diffuse target when the masker is compact (T_DM_C) may lead to an increase in the SNR required for speech recognition when compared to situations where the timbres of the target and masker are the same (T_CM_C, T_DM_D).

However, when the target is compact and the masker is diffuse (T_CM_D), it is highly likely that comb filtering will reduce the SNR needed for speech recognition when compared to situations where the timbres of the target and masker are identical (T_CM_C, T_DM_D). The reason for this is that the comb-filtering that results when the masker is played over multiple loudspeakers will lead to troughs in the spectrum of the masker. The SNR will be improved in those regions where there is a trough. This should help to unmask the target speech. The degree of unmasking that will occur will depend on locations of the troughs in the masker relative to spectrum of the energy in the speech target. Because the location of these troughs will depend on the configuration of the array of loudspeakers in a surround-sound system, and the position of the listener's head with respect to them, the degree of unmasking in this situation is hard to predict. Nevertheless, we would expect some degree of unmasking when the target is compact and the masker is made to be diffuse by presenting it over multiple loudspeakers.

Overall, speech recognition thresholds were lowest for the Babble masker, next lowest for Noise, and highest for Speech. The overall lower thresholds found for the Babble masker most likely is due to spectral difference between the target sentences and the Babble masker. The 12-talker babble used in the current study was taken from the R-SPIN test. As such, its spectral composition matched that of the male target voice used in the SPIN test. The target voice in this study was female. When both the target voice and the Babble masker were matched with respect to overall RMS, the spectral composition of the target voice had less energy in the low-frequency region and much more energy in the high frequency region (see Ben-David, Tse, & Schneider, 2012, for an example of how the spectral composition of a female voice differs from that of the babble masker). The fact that the target sentences contained a considerable degree of energy in the high-frequency region (because they were spoken by a female) most likely is responsible for the lower thresholds in babble compared with noise or competing speech.

The speech masker used in this study consisted of two female (same gender) talkers who were speaking at a similar rate. Their utterances were short semantically anomalous sentences that were recognizable, and as such they most likely create a substantial amount of informational masking. However, any differences found between speech recognition performances under the noise conditions versus the speech conditions cannot be attributed solely to a difference in the degree of informational masking between competing voices and noise. As mentioned previously, speech signals contain amplitude fluctuations that allow the listeners to take advantage of troughs in the amplitude envelope (Cooke, 2006). Therefore, these differences are likely to reflect a combination of greater informational masking as well as the ability to focus attention on the target speech in the troughs in the envelope of the Speech masker.

The result that the effects of a timbre contrast (T_DM_C – T_CM_D) are larger for informational maskers (Babble and Speech) than they are for a Noise masker is consistent with the general result found for stimulus conditions that produce a release from masking (Avivi-Reich et al., 2018; Ezzatian, et al., 2010; Freyman et al., 2004). The difference here is that the direction of the contrast effect depends on whether the target is compact versus when it diffuse.

Slope differences and interaction patterns

In general, examining the slopes provides valuable information regarding how increases in SNR are translated into increases in speech recognition performance under the different conditions.

A noise masker is unlikely to elicit any activation in the semantic or linguistic processes. As such, the interference it causes is essentially energetic. Energetic masking is considered to be less subject to listener control compared with informational masking (Mattys, Davis, Bradlow, & Scott, 2013). When the masker is noise it is reasonable to assume that a greater weight will be assigned to basic auditory processes, rather than to high-order processes, in order to minimize the impact of the energetic masking. Therefore, it is not surprising that the slopes calculated for the Noise condition are steeper than those found for Babble and Speech.

In the current study, participants were also asked to complete two tests that are measures of language competence. The Mill Hill provides an estimate of the individuals' vocabulary knowledge, while the Nelson-Denny reflects the processes and skills involved in reading and comprehending written prose. The individual scores were centered within each group and the individual differences were correlated with the speech recognition results in order to examine whether the processes and skills that these two cognitive tests measure could account for individual differences in speech recognition performance. The results showed that the vocabulary knowledge of the listeners interacted with Masker Type in this study. When the masker was Speech, higher vocabulary knowledge was significantly correlated with lower (better) SNR recognition thresholds. However, there was no indication that reading comprehension skills were related to individual differences in speech recognition. These results imply that the speech recognition task employed here did not require a substantial engagement of the types of cognitive and linguistic processes tapped by the reading comprehension test. The young native-English listeners who participated in the current study do not seem to feel the need to engage higher-order processes to complete the recognition task given here.

The results of the current study may have important practical implications as they call for a reassessment of how surround-sound systems should be designed and soundtracks should be mixed and assigned to channels in order to enhance speech recognition. For example, when amplification is used, theatres could assist their audience, especially those experiencing difficulties, by presenting the voice of an actor or actress using a single loudspeaker to maintain compact images for the voices while presenting the background sounds using loudspeakers placed all around the audience to create a contrast between the compactness of the target voices and the diffuseness of the background, and reduce the SNR needed for speech recognition because of the troughs in the masker's spectrum created by comb filtering. On a similar note, television and movie sound technicians may want to mix the target voices into a limited number of channels so that they have a more compact location in space and are less subject to comb-filtering effects.

Future studies should further investigate the effects of different amplification compositions on the ability of listeners of different ages and hearing statuses, to analyse the auditory scene and successfully perceive the target speech. In addition, with the current results in mind, it would be of value to design a future study that would differentiate the effects due to comb filtering from those due to timbre differences.

Authors' Note This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN 9952-13). The authors thank Jane Carey for assistance in recruiting participants and collecting data. The authors also would like to thank James Qi for his work in programming the experiments.

Open Practice Statement The data and materials for all experiments reported here are available upon request from the first author. The experiment was not preregistered.

Appendix: Comb-filtering Effects

The arrangement of loudspeakers in the sound attenuating chamber is illustrated in Fig. 8. Each of three loudspeakers was positioned at approximately 1.7 m from the center of the listener's head (green circle). Two of the loudspeakers were located 45° to the left and right of the listener. The remaining loudspeaker was located directly ahead. It is assumed that the distance from the listener's left ear to the center of her or his head is 3 in. (.0762 m). The calculations below assume an anechoic room, with the position of the center of the listener's head being fixed at precisely 1.7 m from each of the loudspeakers. In reality, the listener, although instructed to orient to the central loudspeaker and not move her or his head, was not constrained to do so. In addition, the room was not anechoic, so that sound from each of the loudspeakers reflected off the walls, floor, and ceiling also reached the left ear. These factors were ignored in specifying the nature of the summed waveform reaching the left ear from each of the loudspeakers.

If we set the coordinates in the diagram such that the center of the listener's head was position at {x = 0,y = 0}, the coordinates of the central loudspeaker become {0,1.7}, and the coordinates of the left and right loudspeakers can be obtained from Pythagarus' theorem. Consider the left loudspeaker. The vertical dashed line from the center of the left loudspeaker to the x-axis, and from this point to the center of the head form a right angle. Hence the length of the vertical line (l_y) is obtained by solving the equation

$$1.7^2 = 2*l_y^2, l_y = \sqrt{1.7^2/2}$$

Note that this is also the length of the horizontal dashed line from the center of the head to the point where it intersects the vertical line l_y to the left of the head at coordinates {0,- l_y}. It follows that the length of the horizontal dashed line from the left ear to the point {0,- l_y} is $\sqrt{1.7^2/2}–.0762$ m, and the distance from the left ear to the point where the vertical line from the right loudspeaker intersects the horizontal axis is $\sqrt{1.7^2/2} + .0762$ m. Hence the length of the red line connecting the left loudspeaker to the left ear is

$$\sqrt{1.7^2/2 + \left(\sqrt{1.7^2/2}–.0762\right)^2} = 1.647\ m$$

The length of the red line connecting the right loudspeaker to the right ear (ignoring the distance required to travel around the head) is

$$\sqrt{1.7^2/2 + \left(\sqrt{1.7^2/2} + .0762\right)^2} = 1.75471\ m$$

Finally, the length of the red line connected the central loudspeaker to the left ear is

$$\sqrt{1.7^2 + (.0762)^2} = 1.70171\ m$$

When the same signal is played simultaneously from all three loudspeakers, the time it takes for the signal from the left loudspeaker to reach the left ear is 1.647/343 = 0.00480174 s (assuming that the speed of sound is 343 m/s). The corresponding times it takes the signal from the right and central loudspeakers to reach the left ear are 1.75471/343 = .00511576 s, and 1.70171/343 = .00496124 s, respectively. This means that the signal from the central loudspeaker reaches the left ear .00511576 - .00496124 = .00015452 or approximately .00015 s earlier than the signal from the right loudspeaker, and the signal from the left loudspeaker reaches the left ear .00511576 -.00480174 = .00031402 or approximately .00031 s earlier than the signal from the right loudspeakers. If g[t] is the signal from the right loudspeaker reaching the left ear at time t s, g[t-.00015] is the same signal reaching the left ear from the central loudspeaker at time t, and g[t-.00031] is the signal reaching the left ear from the left loudspeaker.

Now let h[t] be a steady-state bandpass white noise (0 – 50 kHz) whose RMS amplitude is 1. The combined signal from the three loudspeakers arriving at the left ear is

$$h[t] + h[t–.00015] + h[t–.00031]$$

Now let's assume that we take a 1-s sample of $h[t]$ (sampling rate = 100,000 samples per second). The Fourier expansion of this 1-s segment will have a fundamental frequency of 1 Hz, and contain 50,000 multiples of this fundamental frequency. The power at each of these frequencies is $A_k^2/2$ where k specifies the kth frequency in the Fourier expansion.[1] The expected power at each of these frequencies is 1/50,000 for this band-limited white noise. Hence the value of A that will yield a power of 1/50,000 is $\frac{1}{50\sqrt{10}}$. Now if we add the left and center loudspeaker noises to the one coming from the right loudspeaker, the expected value of the sum of the three waveforms at frequency f_k is $\frac{1}{50\sqrt{10}}(Cos[2\,Pi\,f_k\,t + \theta_k] + Cos[2\,Pi\,f_k\,(t–t_C) + \theta_k\,] + Cos[2\,Pi\,f_k\,(t–t_L) + \theta_k\,])$. The expected power at frequency f_k is

$$f_k\int_0^{1/f_k}\left(\frac{1}{50\sqrt{10}}(Cos[2\,Pi\,f_k\,t + \theta_k] + Cos[2\,Pi\,f_k\,(t–t_C) + \theta_k\,] + Cos[2\,Pi\,f_k\,(t–t_L) + \theta_k\,])\right)^2 dt$$

$$= \frac{3}{50,000} + \frac{Cos[2\,\pi\,f_k\,t_C]}{25,000} + \frac{Cos[2\,\pi\,f_k(\,t_C–t_L)]}{25,000} + \frac{Cos[2\,\pi\,f_k\,t_L]}{25,000}$$

If we split the power between positive and negative frequencies, the expected power in the two-sided power density function at frequency f_k

$$W[f_k] = \frac{3}{100,000} + \frac{Cos[2\,\pi\,f_k\,t_C]}{50,000} + \frac{Cos[2\,\pi\,f_k(\,t_C–t_L)]}{50,000} + \frac{Cos[2\,\pi\,f_k\,t_L]}{50,000}$$

where $W[f_k]$ is the power at frequency f_k in the two-sided Fourier expansion of

$$h[t] + h[t–t_C] + h[t–t_L]$$

[1] The power at frequency k in the Fourier expansion is given by $\frac{\int_o^{1/f_k}(A*Cos[2\ Pi\ f_k\ t+\theta_k)^2 dt}{1/f_k} = \frac{A_k^2}{2}$.

Therefore, the power in $h[t]+h[t-t_C]+h[t-t_L]$ is $2\sum_{k=1}^{50,000} W[k]$.

Since the expected total power in $h[t]$ is 1.0, we have to divide the total power in $h[t]+h[t-t_C]+h[t-t_L]$ by $2\sum_{k=1}^{50,000} W[k]$ so that the total power in $h[t]+h[t-t_C]+h[t-t_L]$ is the same as the total power in $h[t]$. Once both $h[t]$ and $h[t]+h[t-t_C]+h[t-t_L]$ were equated, they were low-pass filtered at 6 kHz, and rescaled so that both noises had an RMS amplitude of .125. The reason why we did this is that sound file of the sentence we used had an RMS amplitude of .125. Hence, all three files had the same RMS amplitude.

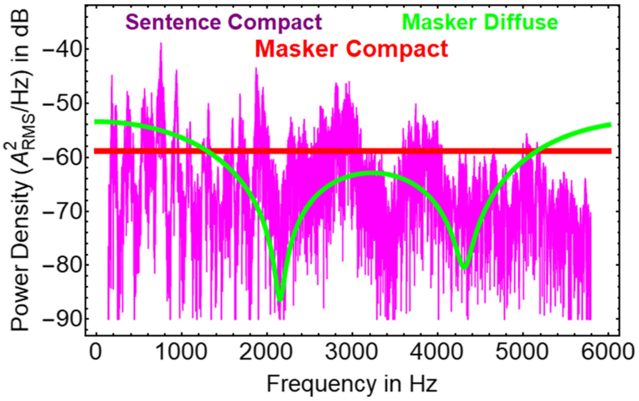

Fig. 10. Power density functions (in dB) for three different stimuli at the point represented by the opening to the left ear of the hypothetical observer shown in Fig. 8. Anechoic conditions were assumed and only the positive frequency components of the two-sided power density function are plotted. The RMS amplitudes of all three stimuli at that point of entrance to the left ear were set to .125. The red line specifies the power density function of a band-limited white noise (0 – 6 kHz) played over a single loudspeaker at the left ear of the listener. The green specifies the power density function of the signal arriving at the left ear when the band-limited white noise was played over all three loudspeakers simultaneously. The purple function is the power spectral density function at the left ear of the listener when the sentence "That ocean could shadow our peak" was played over the center loudspeaker only. Note that the SNR of the sentence to the noise over the region from approximately 1300 Hz to 4800 Hz is greater when the noise masker is diffuse (3 loudspeaker case) than when it is compact (single-loudspeaker case)

Figure 10 plots the power density function for $h[t]$ (in red), along with the power density function for $h[t]+h[t-t_C]+h[t-t_L]$ (in green) for the case where both stimuli had RMS amplitudes = .125. The values used for the two delays were $t_C = .00015$ s and $t_L = .00031$ s. Also shown (in purple) is the power density function for the sentence "That ocean could shadow our peak" that we would expect at the left ear when the sentence is played from the center loudspeaker only. This figure illustrates that the power in the sentence in the frequency region between 1,500 and 4,500 Hz is substantially higher than the power in the noise when the noise is played over all three loudspeakers than when the noise is played over a single loudspeaker. Hence, we would expect the sentence to be more intelligible when the masker is played over all three loudspeakers than it would be when the masker is coming from a single central loudspeaker (see Fig. 9).

We also modified the sentence using the filter profile arising from the comb filtering produced by playing a stimulus simultaneously over the three loudspeakers (the blue spectrum in Fig. 9). This enabled us to plot power spectra for the four conditions in this experiment for a single target sentence. When the masker is diffuse, and the target sentence is compact (T_CM_D), the spectral profile of the target exceeds that of the masker much more frequently than when the target is compact and the masker is compact (T_CM_C). Note that when both target and masker are compact (T_CM_C), or when both target and masker are diffuse (T_DM_D), the spectrum of the target sentence exceeds that of the masker by approximately the same amount. Hence, we would expect performances in these two conditions to be equivalent.

Figure 9 also shows that when the target sentence is diffuse, its profile protrudes above that of the diffuse masker (T_DM_D), more than the when the target is diffuse and the masker is compact (T_DM_C). Hence, the spectral profiles for these four conditions are consistent with the notion that speech recognition depends substantially on the power spectral profiles of the target sentences relative to that of the masker.

References

Arbogast, T. L., Mason, C. R., Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America, 112(5),* 2086-2098.

Avivi-Reich, M., Daneman, M., and Schneider, B. A. (2014). How age and linguistic competence alter the interplay of perceptual and cognitive factors when listening to conversations in a noisy environment. *Frontiers in Systems Neuroscience, 8 21.* https://doi.org/10.3389/fnsys.2014.00021

Avivi-Reich, M., Jakubczyk, A., Daneman, M., Schneider, B.A. (2015). How age, linguistic status, and the nature of the auditory scene alter the manner in which listening comprehension is achieved in multitalker conversations. *Journal of Speech, Language, and Hearing Research, 58(5)*, 1570-1591. https://doi.org/10.1044/2015_JSLHR-H-14-0177.

Avivi-Reich, M., Puka, K., Schneider, B.A. (2018). Do age and linguistic background alter the audiovisual advantage when listening to speech in the presence of energetic and informational masking? *Attention, Perception and Psychophysics, 80 (1),* 242-261.

Ben-David, B. M., Tse, V. Y. 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,* 55-63. https://doi.org/10.1016/j.heares.2012.04.022

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*, 32-38.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, Mass: MIT Press.

Brown, J. I., Bennett, J. M., Hanna, G. (1981). *The Nelson-Denny reading test.* Chicago: Riverside.

Brungart, D. S., Simpson, B. D. (2002). The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *Journal of the Acoustical Society of America, 112(2),* 664-676.

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. *Journal of the Acoustical Society of America, 110(5),* 2527-2538.

Cooke, M. P. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America, 119(3),* 562-1573.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., Shinn-Cunningham, B.G. (2003). Note on informational masking. *Journal of the Acoustical Society of America, 113(6),* 2984-2987. https://doi.org/10.1121/1.1570435

Ezzatian, P., Avivi, M., Schneider, B. A. (2010). Do nonnative listeners benefit as much as native listeners from spatial cues that release from speech masking? *Speech Communication, 5,* 919-929.

Franconeri, S. L., Simons, D. J. (2003). Moving and looming stimuli capture attention. *Perception & Psychophysics, 65,* 999-1010.

Freyman, R.L., Balakrishnan, U., Helfer, K.S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *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. *Journal of the Acoustical Society of America, 106(6),* 3578-3588.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 40(2),* 432-443.

Humes, L. E., Lee, J. H., Coughlin, M. P. (2006). Auditory measures of selective and divided attention in young and older adults using single-talker competition. *Journal of the Acoustical Society of America, 120(5),* 2926-2937.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., Durlach, N. I. (2008). Informational masking. In: Yost, W. A., Popper, A. N., Fay, R. R. (eds) Auditory perception of sound sources, New York, NY: *Springer Handbook of Auditory Research*, pp. 143–190.

Lavandier, M., & Culling, J. F. (2008). Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer. *Journal of the Acoustical Society of America, 123,* 2237-2248.

Li, L., Daneman, M., Qi, J. G., Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults? *Journal of Experimental Psychology: Human Perception and Performance, 30,* 1077-1091.

Mattys, S., Davis, M. H., Bradlow, A. R., Scott, S. (2013). *Speech Recognition in Adverse Conditions: Explorations in Behaviour and Neuroscience.* New York: Psychology Press.

Mershon, D.H., King, L.E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics, 18 (6),* 409–415.

Rakerd, B., Aaronson, N. L., Hartmann, W. M. (2006). Release from speech-on-speech masking by adding a delayed masker at a different location. *Journal of the Acoustical Society of America, 119(3),* 1597-605. https://doi.org/10.1121/1.2161438

Raven, J. C. (1965). *The Mill Hill Vocabulary Scale*. London: H.K. Lewis.

Schneider, B. A., Avivi-Reich, M., Mozuraitis, M. (2015). A cautionary note on the use of the Analysis of Covariance (ANCOVA) in classification designs with and without within-subject factors. *Frontiers in Psychology, 6,* 474. https://doi.org/10.3389/fpsyg.2015.00474

Schneider, B. A., Li, L. Daneman, M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *Journal of the American Academy of Audiology, 18,* 578-591. https://doi.org/10.3766/jaaa.18.7.4.

Schneider, B. A., Pichora-Fuller, M. K., Daneman, M. (2010). The effects of senescent changes in audition and cognition on spoken language comprehension. In S. Gordon-Salant, R. D. Frisina, A. N. Popper, & R. R. Fay (Eds.), *Springer Handbook of Auditory Research: The Aging Auditory System: Perceptual Characterization and Neural Bases of Presbycusis* (167-210). New York: Springer.

Vongpaisal, T., Pichora-Fuller, M. K. (2007). Effect of age on use of F0 to segregate concurrent vowels. *Journal of speech, hearing and language research, 50,* 1139-1156.

Yang, Z. G., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. *Speech Communication, 49,* 892-904.