How Competing Speech Interferes with Speech Comprehension in Everyday Listening Situations

Bruce A. Schneider*
Liang Li*
Meredyth Daneman*

Abstract

Listeners often complain that they have trouble following a conversation when the environment is noisy. The environment could be noisy because of the presence of other unrelated but meaningful conversations, or because of the presence of less meaningful sound sources such as ventilation noise. Both kinds of distracting sound sources produce interference at the auditory periphery (activate similar regions along the basilar membrane), and this kind of interference is called "energetic masking." However, in addition to energetic masking, meaningful sound sources, such as competing speech, can and do interfere with the processing of the target speech at more central levels (phonetic and/or semantic), and this kind of interference is often called informational masking. In this article we review what is known about informational masking of speech by competing speech, and the auditory and cognitive factors that determine its severity.

Sumario

Las personas que escuchan a menudo tienen problema para seguir una conversación cuando el ambiente es ruidoso. El ambiente puede ser ruidoso por la presencia de otras conversaciones no relacionadas pero significativas, o por la presencia de otras fuentes de ruido menos significativas, tales como ruidos de ventilación. Ambas fuentes de sonidos de distracción producen interferencia en la periferia auditiva (activan regiones similares a lo largo de la membrana basilar), y este tipo de interferencia se la llama enmascaramiento energético. Sin embargo, además del enmascaramiento energético, otras fuentes significativas de sonido, tales como el lenguaje en competencia, pueden interferir, y de hecho interfieren con el procesamiento de un lenguaje meta a niveles más centrales (fonético y/o semántico), y este tipo de interferencia se le llama enmascaramiento informacional. En este artículo revisamos lo que se conoce sobre enmascaramiento informacional del lenguaje por lenguaje competitivo, y los factores auditivos y cognitivos que determinan su severidad.

*Centre for Research on Biological Communication Systems, University of Toronto at Mississauga, Mississauga, Ontario

Bruce Schneider, Department of Psychology, University of Toronto at Mississauga, Mississauga, Ontario L5L 1C6; Phone: 905-828-3963; Fax: 905-569-4850; E-mail: bschneid@utm.utoronto.ca

The ability of people to follow or participate in a conversation decreases as the complexity of the auditory scene increases. For example, when there is only one person talking in a quiet non-reverberant environment, people with good hearing find listening to be easy and effortless. However, as the auditory scene increases in complexity (more and louder sound sources, greater reverberation), so does one's difficulty in following a conversation. For example, participating in a four-person conversation in a crowded, noisy, highly-reverberant restaurant is quite difficult and tiring, even for young listeners with good hearing. For older listeners, or for those with hearing impairments, communicating in such environments is often virtually impossible.

Why is it so difficult to comprehend spoken language in complex auditory environments? One obvious factor that no doubt contributes to communication difficulties in such situations is that the signal-to-noise ratio (SNR) is often so low in such environments that the energy in the competing sound sources simply overwhelms (masks) the energy in the signal (energetic masking). A second, less obvious, contributor to communication difficulties in complex listening situations is that the listener cannot easily identify, locate, and separate the different sound sources in the auditory scene. For example, it is sometimes quite difficult to attend to one person who is talking when there are two other people nearby who are also talking. What may be happening in such situations is that information from the competing talkers intrudes into the message conveyed by the target talker either because the listener cannot perceptually separate the two streams of information, or because attention switches back and forth between the target talker and one or more of the competing talkers. In other words, listeners might experience difficulties in such situations because they are unable to parse the auditory scene into its different component sources so that they may attend to one source and ignore the others. Hence a failure to perceptually segregate sound sources can contribute to the masking of speech by competing sounds.

Another factor that might be contributing to comprehension difficulties in complex scenes is that competing sound sources may initiate phonetic, semantic, and/or linguistic activity that interferes with the processing of the speech target. When the target is speech and the masker is noise, the target will elicit activity in the phonetic, semantic, and linguistic systems whereas the masker is unlikely to do so. However, when the target is speech and the masker is also speech, both are likely to initiate activity in the systems involved in language processing. Hence, the activation elicited by the competing speech could interfere with the processing of information in the target speech at a cognitive level. It should be noted that semantic and linguistic interference is more likely to occur if there are breakdowns in auditory scene analysis such that listeners find it difficult to parse the auditory scene into its component sounds sources. Finally, when the task consists of attending to more than one person, listeners might experience difficulties in switching attention from one talker to another due to (1) energetic masking, (2) scene analysis failures, or (3) semantic and linguistic interference. The effects of such non-energetic factors on spoken language comprehension are sometimes referred to as "informational masking," or "central masking," or "perceptual masking." In this paper we will discuss the sources contributing to masking of speech by competing speech, and the acoustic, perceptual, and cognitive factors that can reduce it. It is not our intention here to consider all of the different kinds of effects that fall under the rubric of "informational masking." Here, we limit our discussion to studies of masking speech by speech.

We will begin by indicating how the energetic contribution of a speech masker is typically evaluated. We will then show how speech comprehension in the presence of speech maskers is affected by one's ability to effectively parse an auditory scene into its component sound sources. This will be followed by a discussion of how the information in the speech masker can interfere with the processing of the speech signal at more central (cognitive) levels of processing, and how the degree of informational masking produced at these different processing levels can be significantly reduced. Finally, we will end by describing how informational masking is affected by age and hearing loss, and the implications of these studies for clinical practice.

CONTROLLING FOR ENERGETIC MASKING WHEN THE MASKER IS COMPETING SPEECH

Later in this article, we will be discussing factors that can lead to a release from masking when the masker is competing speech. One such factor that reduces speech on speech masking is the spatial separation of the target speech from the speech masker. This release from masking could be due to a reduction in peripheral (energetic) masking, and/or a reduction in the amount of interference produced at more central (cognitive) levels.[1] To determine how much of the release from masking is due to central interference, researchers often include a second condition in which the masker is steady-state speech-spectrum noise. Because such a masker is unlikely to initiate any competing phonetic, semantic, or linguistic activity, it should not interfere with speech comprehension at these more central levels.[1] Therefore, if the reduction in masking due to a manipulation like spatial separation is larger when the masker is speech than when the masker is speech-spectrum noise, we can infer that the manipulation is effective in reducing interference at more central levels.

Because the effects of energetic masking on speech comprehension have been widely studied (e.g., Plomp and Mimpen, 1979), we will not review that literature here. Rather, we will focus our attention on factors in the auditory scene, other than energetic masking, that affect speech comprehension in the presence of competing speech. In doing so, we will follow the convention in the speech literature of referring to the non-energetic effects of a speech masker on speech as "informational" masking effects.[2]

ROLE OF SCENE ANALYSIS IN THE MASKING OF SPEECH

Source Segregation and the Precedence Effect

When there are several simultaneous sources of sound, the information available at the ears of the listener consists of the sum of the direct waves from these sources plus all of their multitudinous reflections. Consider first a situation in which there is but a single sound source in an environment with only one sound reflecting barrier. This kind of environment could be achieved in an anechoic chamber by placing a sound source (e.g., a loudspeaker) to the left and a reflecting surface (e.g., a plane of glass) to the right of the listener. When a sound is played over the loudspeaker, the direct wavefront from the source will be followed shortly thereafter by a filtered version of the same wavefront coming from the opposite side of the head. Alternatively, one could place the listener in an anechoic environment with two sound sources, one to the left and one to the right of the listener. If the right sound source produced a filtered and time-delayed version of the waveform produced by the left sound source, the listener would first receive a wavefront from the left followed by a filtered and time-delayed version of the same wavefront from the right. In other words a sound reflection could be mimicked by having another sound source in the environment producing a filtered and time-delayed version of the original wavefront. The task facing the auditory system is to decide whether these two wavefronts represent a single sound source and a reflection, or two different sound sources. When the time delay is short, and the spectral-temporal characteristics of the delayed wave are reasonably close to that of the direct wave, the perceptual system of the listener tends to fuse the information coming from the two wavefronts, and usually locates the source of the sound as being at or near the position of the source producing the leading wavefront. This kind of capture of the reflection by the direct wave, and its fusion into a single auditory object, is often referred to as the precedence effect (e.g., Li and Yue, 2002; Litovsky et al, 1999; Zurek, 1980).

Parsing an Auditory Scene Containing Multiple Sound Sources

When there are multiple sound sources in a reverberant environment, the listener's task becomes much more complex. To parse the auditory scene correctly, the direct wave from each and every sound source has to capture its own reflections and not those of other sound sources. Failure to do so will lead to confusion. For example, consider a simplified case in which there are only two sources. If the direct wavefront from source B, or any of the sound reflections from source B are captured by the direct wave from source A, the listener might attribute some of the information coming from source B to source A, thereby producing some informational masking. How this might occur is illustrated in a recent study by Li et al (2005). These investigators presented the same 3-s burst of white noise from two loudspeakers, one situated 45^{o} to the right, and the other 45^{o} to the left of the listener with the left-speaker noise starting 2 ms before the right-speaker noise in Condition 1 (Left Leading), and with the right-speaker noise starting 2 ms before the left speaker noise in Condition 2 (Right Leading). The Left Leading Condition simulated a sound source 45^{o} to the left and a single sound reflection coming 45^{o} from the right. Under such circumstances, listeners perceived only a single sound source whose location is to the left of the listener. In the second condition (Right-Leading Condition), the opposite occurred, and listeners perceived a single sound source located on the right. The experimenters then introduced a gap in the noise emanating from the right speaker in both conditions. When the right-side noise preceded the left side noise by 2 ms, and the gap occurred in the right-side sound only, listeners reported hearing a gap in the sound on the right. This is what we would expect because the gap was in the leading sound. However, when the left-side noise led the right-side noise by 2 ms, so that the fused noise was perceived on the listener's left, and the gap was in the right-side noise only, all of the listeners reported they heard a gap in the left-side noise. Hence, an attribute of a sound emanating from the right was allocated to a sound presented from the left. In other words, the leading sound captured an attribute of a lagging sound. Interestingly, when the left and right-side noises were uncorrelated and therefore not fused (the listener perceived two noises, one on the left and the other on the right), but the left-side sound preceded the right-side sound by 2 ms, listeners sometimes reported hearing a gap in the noise on the left side when the gap was only in the right side noise (they always heard the gap in the noise from the right side). Hence, even when two sound sources are independent, and are perceived to come from two different locations, occasionally an attribute from a lagging sound can be captured by a leading sound. If this occurs when two or more people are talking simultaneously, one of the auditory streams (e.g., talker A) may capture one or more attributes of another auditory stream (e.g., talker B), which could lead to errors about what one of the talkers was saying.

The Li et al. (2005) study raises the possibility that errors on the phonemic level could occur if one speech stream captures attributes from another. Obviously, the better the listener is at segregating speech signals, the less likely it is that such captures or intrusions of one stream into another will occur. Studies of auditory

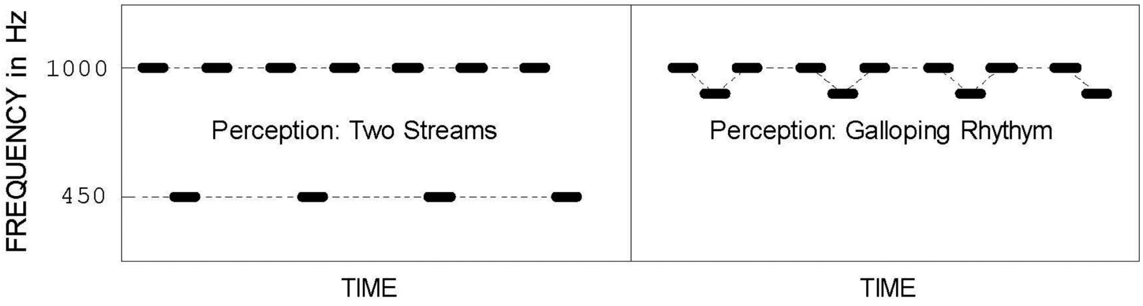

Figure 1. An experimental paradigm to evaluate how frequency separation affects auditory streaming. Both left and right panels depict sequences of tones, with all tones being of equal duration. Let H represents the higher pitched tone, L the lower pitched tone, and S a silent period whose duration is equal to that of the tone. The sequence of tones in both panels is HLHSHLHSHLHS. When there is a small frequency separation, as indicated in the right-hand panel, listeners perceive a galloping rhythm (HLHS). However, when the frequency separation is large, as shown in the left-hand panel, listeners perceive two distinct auditory streams (HSHSHSHSHS) and (LSSSLSSSLSSS). Adapted from Bregman and Ahad.

streaming (see Bregman, 1990) have identified several acoustic level factors that promote stream segregation. One of these is spectral separation. Figure 1 presents a schematic representation of a repeating sequence of two tones. When the frequency separation between the tones is minimal, one tends to perceive a galloping rhythm. As spectral separation increases, a point is reached where two auditory streams are perceived: one, a high-frequency stream, the other, a low-frequency stream. It should also be noted that the listener at some point will lose the perception of their being two separate streams as the temporal gaps between the tones is increased. Hence, we might expect a listener to segregate two different voices, either on the basis of spectral differences (either in fundamental frequency or in formant structure), or on prosodic differences between talkers. Brungart (2001) and Brungart et al (2001) reported that a listener, when presented with two competing speech messages, experienced more difficulty in segregating the content of the target phrase from that of the competing phrase when the competing phrase was spoken by the same talker than when the competing phrase was spoken by a different talker of the same gender, or by a talker of a different gender. Increasing the spectral difference between talkers could lead to a reduction in energetic masking. However, it could also be improving performance by leading to better stream segregation, thereby reducing the amount of informational masking. Furthermore, Brungart et al. (2001) also found that prior experience of the target talker's voice improved the ability of the listener to segregate the speech streams. Hence, we would expect better segregation of speech streams the more familiar the listener is with the voices of the talkers. In other words, perceptual level factors (separation in fundamental frequency) and cognitive level factors (voice familiarity), can play an important role in stream segregation and could lead to reductions in masking of speech by speech.

Clearly, to follow a conversation one must be able to parse the peripheral auditory signal into one or more auditory streams (voices). Failure to do so will make it more difficult for higher-order, more cognitive level processes to extract the linguistic and semantic information from the targeted voice. Finally, at the cognitive level, listeners must be able to focus their attention on one auditory stream (voice) in order to extract the meaning from that stream, while simultaneously inhibiting the processing of information from other auditory streams, or, if the information from the second stream is processed, prohibiting it from interfering with the processing of the targeted voice. Failure to do so will result in interference at more central levels of processing (additional evidence in support of this argument can be found in Alain et al, 2006).

COGNITIVE CONTRIBUTIONS TO MASKING OF SPEECH BY SPEECH

In order to participate in a conversation, listeners not only have to hear the individual words and phrases spoken by each person, they must also integrate this information with past input and world knowledge to extract each person's meaning and point of view. To accomplish this when the auditory scene is complex (e.g., two or more people talking simultaneously), the listener must either 1) focus attention on one stream and suppress the information coming from other sources, or 2) attempt to simultaneously process more than one stream at a time. If it becomes difficult for the listener to inhibit the processing of irrelevant information or to simultaneously process more than two information streams, the listener is likely to require a higher SNR for speech comprehension.

How does the listener go about inhibiting information from irrelevant sources? Many cognitive psychologists hypothesize that inhibiting the processing of irrelevant information is one of the functions of working memory. Working memory is considered to be a limited capacity system responsible for the processing and temporary maintenance of task-relevant information during the performance of everyday cognitive tasks such as listening comprehension (Baddeley, 1986; Baddeley and Hitch, 1974; Daneman and Carpenter, 1980; Miyake and Shah, 1999). According to Hasher and Zacks (1988), the key to successful processing is the ability to keep irrelevant information from cluttering working memory, either by excluding it from gaining access to working memory in the first place, or by deleting it from working memory when it does intrude (see also Hasher et al, 1991; Hasher et al, 1999; Stoltzfus et al, 1996). If the listener's goal is to focus on the contributions of only one of several talkers, success at comprehending the target talker will depend on the listener's ability to inhibit the processing of speech from other talkers. Clearly, the ability to inhibit the processing of irrelevant material in working memory will affect the degree of masking that is likely to occur. Of specific interest in this regard is the claim that older adults experience an inhibitory deficit in the sense that they are not as good as younger adults in either preventing irrelevant information from intruding into working memory or in deleting such information if it does intrude (Hasher and Zacks, 1988; McDowd, Oseas-Kreger, and Fillion, 1995). If this were true, we would expect a greater degree of cognitive interference and, hence, more informational masking in older adults than would be found in younger adults.

If, on the other hand, listeners attempt to simultaneously process two or more auditory streams, their ability to do so will be limited by their working memory capacity. Hence, deficits in working memory capacity could also lead to a greater degree of masking. It is interesting to note, in this regard, that older adults are often thought to have smaller working memory capacities than have younger adults (Brébion, 2003; Kirasic et al, 1996; Stine and Wingfield, 1990). If this were indeed true, we would also expect more interference of competing speech on the targeted speech in older than in younger adults.

It is also worth noting that a number of studies have indicated that there are large individual differences in working memory capacity (Daneman and Carpenter, 1980; Daneman and Merikle, 1996; Miyake and Shah, 1999). If the ability to inhibit irrelevant information or to process multiple information streams are working memory functions, then we would expect to find individual differences in working memory to be correlated with individual differences in susceptibility to informational masking. Interestingly, intersubject variability is much larger in informational masking than in energetic masking. Hence, it is possible that intersubject differences in working memory capacity (which can be substantial) could account for some of the intersubject variability in informational masking.

It is important to note that the amount of cognitive level interference should also be modulated by perceptual level effects such as stream segregation. If a person is unable to perceptually segregate one talker from another, there are likely to be more intrusions of irrelevant material into working memory, and, correspondingly, greater difficulty in deleting such intrusions. Indeed, much of the work on informational masking of speech has been concerned with the factors that are likely to lead to a release from informational masking.

FACTORS LEADING TO RELEASE FROM INFORMATIONAL MASKING

Spatial Separation

It has long been known that spatially separating the target speech from the masker improves target detection and recognition (e.g., Freyman et al, 1999). In other words, spatial separation releases the target from masking. However, some of this release from masking is likely due to release from energetic masking. Compare a situation in which the target and masker are coming from the same loudspeaker located to the listener's right, to one in which the target is coming from the right loudspeaker and the masker from a loudspeaker located to the listener's left. It is easy to see that the SNR at the right ear of the listener will be much higher when the target and masker are spatially separated than when they are coming from the same source because of the shadow cast by the listener's head. Increasing the SNR to the right ear will, of course, reduce the amount of energetic masking at the right ear. Hence, we would expect an improvement in detection and/or recognition due to a release from peripheral or energetic masking.

Spatially separating the sound sources should also improve auditory stream segregation. Accordingly, we might expect spatial separation to lead to a reduction in informational masking because it would make it easier for the listener to focus in on the target and to ignore the informational content of the masker. How, then, do we go about measuring the degree to which spatial separation leads to a release from informational masking? One way of doing so was developed by Freyman et al. (1999), who used the precedence effect to produce a perceived spatial separation. The advantage of using precedence to achieve perceived spatial separation is that shifting the perceived location of the masker using precedence does not improve the SNR in either ear. To see this, consider the following condition reported in Freyman et al (2001). In their experiment, the target, nonsense sentences spoken by a female voice (e.g., "A *shop* could *frame* a *dog*"), were always presented over a loudspeaker located directly in front of the listener. Participants were asked to repeat the target sentence after it was presented, and the number of key words (those in italics) that were correctly identified was recorded. In all conditions the target sentences were presented along with a masker. The masker was either one or two other people speaking other nonsense sentences. There were two masking conditions. In the baseline condition, both the masker and target were presented from the front speaker (Condition F-F, where the first F indicates that the target location was frontal, and the second F that the masker location was also frontal). In the second condition, the target again was presented from the frontal speaker, with the masker presented from both the front and right speakers, with the right speaker leading the frontal speaker by 4 ms (Condition F-RF). Note that the masker in Condition F-RF will be perceived on the right because of the precedence effect. Note also that Condition F-F is the same as Condition F-RF except that the masker is played over the right speaker as well as the frontal speaker. This means that although the signal at the right and left ears remains the same in both conditions, the energy in the masker reaching each ear from the frontal loudspeaker in Condition F-RF is augmented by energy in the masker reaching each ear from the right speaker. Hence, the energy in the masker in Condition F-RF should be higher in both ears than the energy in the masker in Condition F-F, and the SNR correspondingly lower.[3] In a previous experiment, Freyman et al. (1999) found that when the masker was speech spectrum noise, changing the perceived location of the masker from frontal (F-F) to right (F-RF) using precedence led to a small decrease in performance. This is what we would expect from a non-informational masker (steady state noise) given that the SNR was lower in condition F-RF than in condition F-F. Yet as Figure 2 shows, there was a large reduction in masking when the perceived spatial position of a speech masker (either single talker or two talker maskers) was shifted away from that of the target using the precedence effect. Because this reduction could not be attributed to reduction in energetic masking, we can conclude that separating the perceived spatial location of the masker from that of the target can result in a rather large reduction in informational masking when the masker is speech (on the order of 4-9 dB).

Figure 2 also shows that there was a greater degree of release from informational masking for two talkers than for one talker, suggesting that the degree of informational masking changes with the number of talkers. Freyman et al (2004) systematically investigated the effect of the number of talkers using the same paradigm and the same two conditions (Condition F-F, Condition F-RF) as described above. They found that the amount of release from informational masking decreased as the number of talkers increased from two to ten. This is what we might expect because, as the number of talkers becomes large, it becomes more and more difficult to hear individual words. A multi-talker condition is not as likely as a one- or two-talker condition to lead to competing activity in the semantic and linguistic systems. Hence, there will be less interference at semantic and linguistic levels, and a smaller release from masking due to perceived spatial separation.

Image Compactness

It is worth noting that in the studies by Freyman and his colleagues described in Figure 2, the baseline condition consisted of the presentation of the signal and masker over a single loudspeaker. When the position of the masker is shifted using precedence, in addition to the shift in spatial position, other characteristics of the masker are changed. Specifically, the addition of the sound from the delayed loudspeaker also affects the timbre of the masker, and produces a sound image that is much more spacious (less compact) than that produced when the sound is played over only a single loudspeaker (Blauert, 1983). These changes, by themselves, could make it easier to segregate the speech target (which has a compact spatial image) from that of the masker (which has a more diffuse spatial image), leading to a reduction in informational masking. That a change in image size and timbre alone can also produce a release from informational masking was shown in Freyman et al. (1999). In this study they compared condition F-F (where both masker and target were only presented over the frontal loudspeaker), to a condition in which the target was presented over the frontal loudspeaker, but the masker was presented over both loudspeakers with the frontal loudspeaker leading the right loudspeaker by 4 ms (Condition F-FR). Note that in both conditions, both target and masker were perceived to be located frontally. However, even though the location of all images in both conditions remained at the frontal position, the image of the target in Condition F-FR was more compact than that of the masker, whereas the masker and target had the same degree of compactness in the baseline condition (Condition F-F). Freyman et al. (1999) found that a change from Condition F-F to Condition F-FR produced a large release from masking. This comparison suggests that differences in the compactness of target and masker can lead to a release from masking, presumably because it enables the listener to more accurately parse the auditory scene into two different sound sources.

Finally, it should be noted that when the compactness of the target and masker remain the same, a shift in the perceived location of the masker is sufficient to produce a release from informational masking. This was shown by Freyman et al. (1999) when they tested a condition in which both the masker and target were located frontally using precedence (Condition FR-FR, in which both the masker and the signal were played over both loudspeakers, with the frontal loudspeaker leading the right loudspeaker by 4 ms) to one in which the target was perceived frontally, and the masker was perceived to be on the right (Condition FR-RF). Because all images were presented

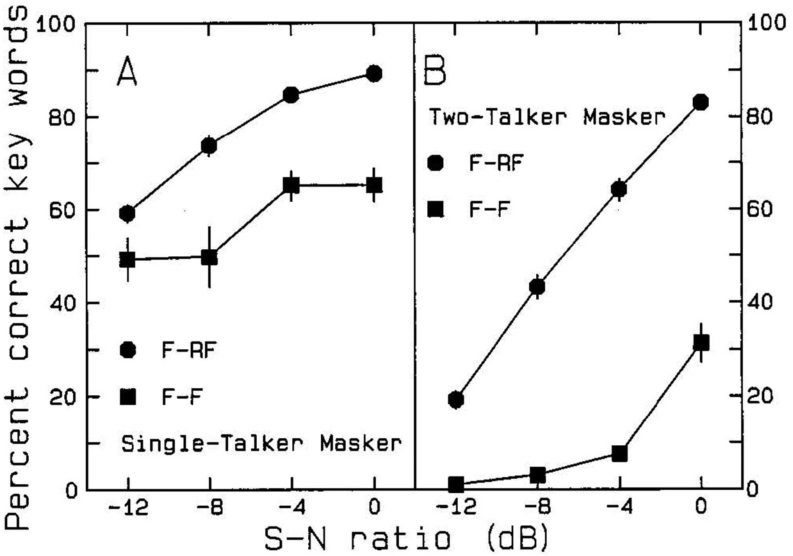

Figure 2. Percent correct identification of key words as a function of SNR in the presence of a single-talker masker (A) and a two-talker masker (B). Because the target speech was always presented over the front loudspeaker, the listener perceived the target talker to be located directly ahead. The masker was either perceived to be co-located with the target (Condition F-F), or to the target's right (Condition F-RF). Performance was substantially better when the listener perceived the target and masker as spatially separate. From Freyman et al. (2001).

over both loudspeakers, they were approximately equally diffuse in all conditions. Yet, a shift in the perceived spatial location of the masker resulted in a reduction in masking, indicating that spatial position, in and of itself, can produce a substantial release from informational masking.

So far, we have seen that a change in perceived spatial position, or a change in perceived timbre or image compactness can lead to reductions in informational masking. Previous studies have also shown (e.g., Brungart, 2001) that a change in vocal qualities can also produce a reduction in informational masking. One could argue that all three factors lead to a release from masking of speech by speech because they make it easier for the listener to perceptually segregate the speech target from the speech masker. This, in turn, could make it easier to inhibit the processing of irrelevant information in the phonetic, semantic and linguistic systems, leading to a reduction in the amount of informational masking. We might also expect that any other acoustic factors that would help to parse the auditory scene would lead to reductions in informational masking. Correspondingly, any acoustic factors, such as excessive reverberation, that might make it more difficult to isolate sound sources, would increase the amount of informational masking. In this regard, it is interesting to note that both hearing-impaired, and older adults with clinically normal audiometric thresholds in the speech range, find it difficult to comprehend speech in highly reverberant conditions (e.g., Helfer, 1992).

Familiarity with the Content of the Message

Listening to a conversation in a challenging auditory environment is much easier when the listener knows, and is familiar with, the topic of conversation. The most likely reason for this is that a priori knowledge of the topic facilitates language processing. In particular, if listeners miss some of the words or phrases because the listening situation is difficult, they may be able to recover the lost information from the context provided by the parts of the conversation they have heard, and their knowledge of the topic in question. However, there is also another possible reason why it is easier to follow a conversation when one has knowledge of the topic under discussion. Suppose that this conversation takes place in a background of other conversations, as is likely to happen at many social events. It is possible that knowledge of the topic helps the listener to focus attention on the relevant voice. For example, if the topic is about cochlear implants in children, and the listener perceives the following sentence fragment "patient 3's acquisition of language," it is quite likely that this voice is a relevant part of the conversation. Hence, it would make sense for listeners to focus their attention on this auditory stream.

Recently, Freyman et al. (2004) showed that listeners are capable of using their knowledge of part of a phrase to recognize the end word of the phrase when that phrase is embedded in a speech masker. Specifically, participants listened to a target sentence in a background masker and then repeated it. Both target and masker were presented over a single loudspeaker located in front of the listener. Two types of maskers were employed: a speech-spectrum noise masker; and a two-talker masker. In Condition 1 (no priming), the listener pressed a button which presented the sentence target and the masker. However, in Condition 2, when listeners pressed the button, they first heard all but the last word in the nonsense sentence in quiet, followed by that same sentence presented in either the noise or speech masker. For example, in Condition 2, if the target sentence was "A corn took their wire," they first hear the first four words ("A corn took the") followed by a noise burst. This priming sentence was presented in quiet. They then heard the full sentence in the masker. Note that because the sentences were nonsense sentences, knowledge of the first four words could not be used to predict the final word. Yet, when the full sentence was presented in the speech masker, they could correctly identify it at a much lower SNR than when full sentences were presented without a prime. In fact, the presentation of the prime improved the threshold for final word recognition by 4 dB when the full sentence was presented in a speech masker. By way of contrast, when the same conditions were run in a noise masker, the prime only improved the threshold by 1.3 dB. Hence, partial knowledge of the sentence that was to follow reduced the amount of informational masking that occurred when the masker was two-talker speech.

It is interesting to note that the same reduction in informational masking was obtained when the prime was spoken by a different voice than that of the target sentence (prime voice was male, target voice was female), and when the participant read the prime instead of listening to it. Clearly, knowledge of the words alone is sufficient to lead to a reduction in informational masking. Because the last word cannot be predicted from the preceding four words, knowing part of the sentence must help the listener to identify and focus in on the target stream. Hence, it is reasonable to hypothesize that the listener in a complex acoustic environment is capable of using knowledge about the nature of conversation to identify and focus in on the talkers participating in a discussion.

A Priori Knowledge of Spatial Location

If listeners can use knowledge about the content of a conversation to focus in on the target talker, it is not unreasonable to expect that they might be able to use a priori knowledge of other characteristics of the auditory scene, such as familiarity with a speaker's voice, or even knowledge of a speaker's location. Recently, Kidd et al (2005) found that listeners can indeed use a priori knowledge about a target's spatial position in order to focus attention on a target. In their task, listeners heard three different sentences spoken simultaneously over three spatially separated loudspeakers. The sentences were from the Coordinate Response Measure (CRM) corpus (Bolia et al, 2000), and were of the type "Ready [call sign] go to [color] [number] now." For each talker there were eight possible call signs, four possible colors, and eight numbers. The listener was instructed to report the color and number associated with a particular call sign. For instance, if the three utterances were "Ready Baron, go to green 4 now," "Ready Fox, go to red 3 now," and "Ready Alpha, go to blue 8 now," and the listener was given the call sign "Baron," the correct response was "green 4." In one part of the experiment, the call sign was announced to the participant after the three sentences were played. When the target was randomly assigned to one of the three loudspeakers, and the call sign was given after the three sentences were presented, listeners were able to correctly identify the color and number associated with the call sign approximately 1/3 of the time. However, when the a priori probability that a call sign would be presented from a particular loudspeaker was 1.0 (participants were given this information prior to the experimental block), participants were able to correctly report the color and number over 90% of the time. In other words, a priori knowledge of where the target would appear allowed the participant to attend to that location and report on that auditory stream. Hence, a priori knowledge of content, voice, and/or spatial location of the target speaker can reduce the amount of masking.

Visual Speech Cues

It has long been known that visual speech cues (the sight of the person talking) can increase speech recognition by providing supplemental information as to the identity of phonemes (e.g., Sumby and Pollack, 1954). Helfer and Freyman (2005) have recently shown that visual speech cues can also lead to a reduction in informational masking. These investigators compared energetic and informational masking of speech under auditory only and auditory-visual conditions. When the speech target and a noise were presented over a single loudspeaker located frontally, the addition of a visual component reduced the amount of masking by about 3 dB. However, under the same conditions, when the masker was speech, the addition of a visual component resulted in a 9 dB reduction in masking. Presumably, the addition of the visual component when the masker was speech facilitated the segregation of the speech target from the speech maskers, thereby reducing the amount of informational masking. Hence, there appear to be a number of perceptual and cognitive factors that either contribute to or alleviate the effects of informational masking.

INFORMATIONAL MASKING IN HEARING IMPAIRED AND IN OLDER ADULTS

To our knowledge, there are only a handful of studies that have examined informational masking of speech by speech in the hearing impaired, and in good hearing older adults. Arbogast et al (2005) found that the amount of release from informational masking due to spatial separation was less for the hearing-impaired group than for the normal-hearing group. However, it was not possible in that study to determine whether this difference between normal and hearing-impaired participants was due to sensorineural hearing losses of cochlear origin, or to more central auditory deficits, such as a diminution in the ability to benefit from spatial separation. Summers and Molis (2004) investigated the effect of masker level on sentence recognition in hearing-impaired listeners when the masker was a single sentence, a reversed sentence, and a steady state speech spectrum noise (stimulus presentation was monaural). Of interest was whether increasing the level of target and masker (simple amplification) would improve performance for the hearing-impaired listeners when the masker was informational (single sentence). Summers and Molis found that increases in the level of the sentence improved performance in 2 of the 6 hearing-impaired listeners but worsened performance in 2 other hearing-impaired listeners. Hence, the benefit of overall amplification in hearing impaired listeners when the masker is informational may vary from individual to individual. Hornsby et al (2006), in a sound-field study in which speech was masked by speech, found that, for hearing-impaired listeners, there was very little difference in performance between aided and unaided conditions. These results suggest that hearing-impaired individuals may benefit less than normal-hearing individuals from at least one of the factors (spatial position) that can provide release from informational masking, that the benefits of simple amplification are uncertain, and that aided hearing in the sound field may not improve performance when other talkers are present.

The results from the Li et al. (2004) study of good-hearing older adults are interesting because aging is associated with both hearing loss and cognitive decline. Hence, it would be interesting to see how these two factors (sensory and cognitive) interact in a complex listening situation in older listeners. In Li et al., the target sentences (which were the same as those in Freyman et al., 2001) were presented over two loudspeakers (one to the left, the other to the right of the listener) with the right speaker leading the left by 3 ms so that the target sentence was always perceived as coming from the right. In one condition, the perceived location of the masker was set to be the same as that of the target. In two other conditions, the lag for the masker, but not for the target, was changed so that the masker was perceived as originating from the center (no lag between left and right) or from the right (right leading the left by 3 ms). Two types of maskers were employed: a speech-spectrum masker, and two-talker masker (the same as in Freyman et al. 2001). Because changing the perceived location using precedence does not change the SNR at either ear in any significant way (see Freyman et al., 1999, and the Appendix in Li et al., 2004, for a discussion of this), any age difference in informational masking found in this experiment cannot be attributed to differences in peripheral or energetic masking between younger and older adults.

According to one cognitive theory, normal aging is associated with reduced inhibitory mechanisms for suppressing the activation of “goal-irrelevant” information (Hasher and Zacks, 1988; Hasher et al., 1999), so that interfering signals will intrude into working memory (Daneman and Carpenter, 1980). A prominent feature of this theory is that it becomes more difficult for older adults to inhibit the processing of irrelevant stimuli as the similarity between target and distractors increases. Clearly a speech distractor is more similar to a speech target than is a noise distractor, and spatial separation between target and distractor should aid in the inhibition of the processing of irrelevant information. Thus, this theory predicts that older adults should demonstrate more interference from informational masking than do

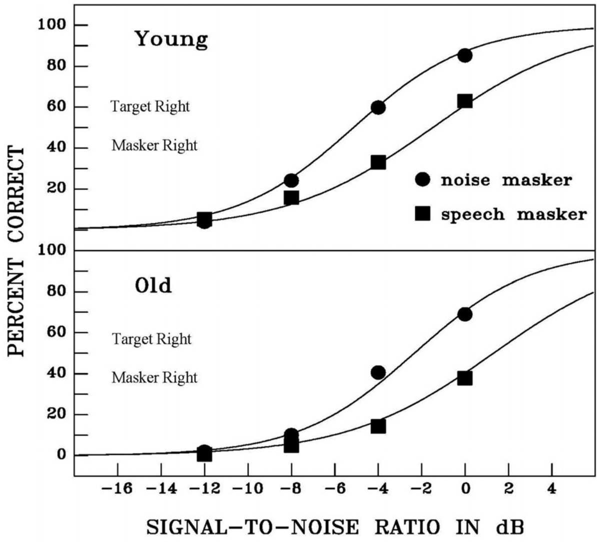

Figure 3. Mean percentage correct as a function of SNR when the perceived location of the masker was the same as that of the target speech for both young (top panel) and old (bottom panel) participants. The same psychometric functions that fit the data of the younger participants, when shifted to the right by 2.8 dB, also fit the data of the older participants. From Li et al. (2004).

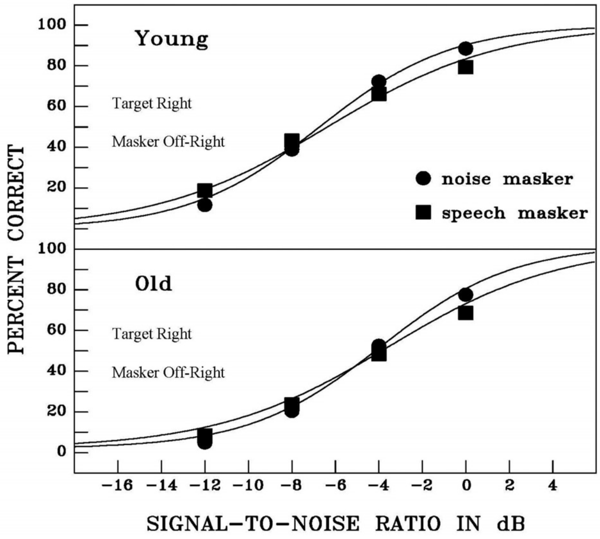

Figure 4. Mean percentage correct as a function of SNR when the perceived location of the masker differed from that of the target speech for both young (top panel) and old (bottom panel) participants. The masker off-right condition is an average of the two off-right (masker-left, masker-center), which did not differ from each other. The same psychometric functions that fit the data of the younger participants, when shifted to the right by 2.8 dB, also fit the data of the older participants. From Li et al. (2004).

younger adults especially when there is no perceived spatial separation between target and masker.

Consider first the conditions in which both masker and target are perceived to be emanating from the same location. Figure 3 shows that under these conditions, a switch from a speech masker (squares) to a noise masker (circles) significantly improves speech recognition. Note, however, the degree of improvement is the same for younger and older adults. In fact, the only difference between young and old is that the older adults require a higher SNR (about 2.8 dB) for 50% detection than do the younger adults. Otherwise there are no differences between the psychometric functions of young and old. Now consider what happens when masker and target are perceived to be originating from two different locations in space. Spatially separating masker and target should attenuate the degree to which information in the masker interferes with target recognition. Figure 4 shows that, under these conditions (spatial separation of masker and target), switching from speech to noise masking has very little effect (if any) on performance. In other words, perceived spatial separation of target and masker appears to virtually eliminate the additional interference imposed by having an informational masker. Again, the only difference between young and old listeners is that the older listeners required approximately a 2.8 dB higher SNR than did younger adults in all conditions.

Because the older adults in this experiment were in the early stages of presbycusis, it is not too surprising that they required a higher SNR (2.8 dB) for speech recognition in noise. What is surprising from a cognitive point of view is that the age-related differences did not increase when participants listened to the nonsense sentences in a two-talker masker, and that older adults benefitted as much as did younger adults from perceived spatial separation. If older adults were less able to inhibit the processing of irrelevant information than younger adults, the old-young difference in SNR should be greater when an informational masker is used than it is when a purely energetic masker is used. Moreover, if it were the case that older adults experienced more difficultly than younger adults in using perceived spatial separation to perceptually segregate the target talker from the two-talker masker, we would expect that they would have a smaller degree of reduction in informational masking than do younger adults. Contrary to this expectation, the degree of informational masking was the same for both age groups. It appears, then, that healthy older adults can benefit as much as younger adults from perceived spatial separation.

IMPLICATIONS FOR CLINICAL PRACTICE

The primary reason why people seek the help of an audiologist is that they want to be able to communicate better in everyday situations. To hear in complex listening situations, a person needs to overcome peripheral (energetic) masking, parse the auditory scene appropriately, focus attention on the target talker, suppress the processing of irrelevant information, and, when appropriate, switch attention from one talker to another. Clearly, a person's ability to function well in complex auditory environments will depend on the status of that person's auditory and cognitive systems. Cochlear pathology can result in a greater susceptibility to energetic masking, central auditory deficits (e.g., loss of binaural hearing, loss of neural synchrony) will interfere with scene analysis, and cognitive declines (such as a loss in working memory capacity) can make it more difficult to a) suppress irrelevant information, b) handle multiple streams of information, and c) rapidly switch attention from one talker to another. At present, audiologists can assess various cochlear and retro-cochlear functions, and determine a person's ability (either aided or unaided) to overcome energetic masking using one or more speech-in-noise tests. However, at present there are no tools in the audiologist's toolbox to assess a person's ability to use the available auditory cues to parse the auditory scene and suppress the processing of irrelevant information, even though the ability to do so can reduce, in some situations, the SNR needed for speech recognition by 4 to 9 dB! Hence, it is worth considering whether it would be useful to develop a clinical test of informational masking based on the paradigm developed by Freyman and his colleagues. Finally, because cognitive factors also play a significant role in communication situations, one could argue that audiologists might wish to take into account the cognitive status of the client (see also Humes, and Lunner and Sundewall-Thorén, this issue).

NOTES

1. It is important to note that using a modulated noise rather than a steady-state noise could lead to some degree of interference at more central levels. For example, Shannon et al (1995) have shown that speech recognition is possible with as few as two amplitude-modulated noise bands. Clearly, amplitude modulation of noise must be capable of eliciting phonetic, semantic, or linguistic activity if it can lead to speech recognition under some circumstances.

2. There is no general agreement as to what non-energetic masking effects should be considered informational or even how informational masking itself should be defined (see Durlach et al, 2003, for a discussion of this issue). Kidd et al refer to informational masking as interference with signal recognition that "is not due to signal-to-noise constraints, but rather is due to our inability to correctly determine that the signal is a distinct entity or object separate from the masker or to recognize the pattern of acoustic information it carries" (1994, p. 3475). We will use this definition to distinguish between the informational and energetic effects of a speech masker of speech.

3. Adding a delayed signal to itself results in a filtered version of the original signal at the receiver, in which the spectral profile of the filter resembles a comb. Hence the spectral power at some frequencies is increased over the original signal while it is decreased at other frequencies when the masker is played over both frontal and right-side speakers with a 4 ms delay between them. For a discussion of comb filtering effects related to the precedence effect, see the appendix in Li et al (2004).

REFERENCES

Arbogast TL, Mason CR, Kidd Jr G. (2005) The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 117:2169–2180.

Alain C, Dyson BJ, Snyder JS. (2006) Aging and the perceptual organization of sounds : a change of scene? In: Conn M, ed. *Handbook of Models for the Study of Human Aging*. Academic Press, 759–769.

Baddeley AD. (1986) *Working Memory*. Oxford: Oxford University Press.

Baddeley AD, Hitch G. (1974) Working memory. In: Bower GH, ed. *The Psychology of Learning and Motivation*. Vol. 8. New York: Academic Press, 47–89.

Blauert J. (1983) *Spatial Hearing*. Cambridge, MA: MIT Press.

Bolia RS, Nelson WT, Ericson MA, Simpson BD. (2000) A speech corpus for multitalker communications research. *J Acoust Soc Am* 107:1065–1066.

Brébion G. (2003) Working memory, language comprehension, and aging: four experiments to understand the deficit. *Exp Aging Res* 29:269–301.

Bregman AS. (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Bregman AS, Ahad P. Demonstrations of Auditory Scene Analysis: The Perceptual Organization of Sound [Audio CD and booklet].

Brungart DS. (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* 109:1101–1109.

Brungart DS, Simpson BD, Ericson MA, Scott KR. (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J Acoust Soc Am* 110:2527–2538.

Daneman M, Carpenter PA. (1980) Individual differences in working memory and reading. *J Verbal Learn Verbal Behav* 19:450–466.

Daneman M, Merikle PM. (1996) Working memory and language comprehension: a meta-analysis. *Psychonomic Bull Rev* 3:422–433.

Durlach NI, Mason CR, Kidd Jr G, Arbogast TL, Colburn HS, Shinn-Cunningham B. (2003) Note on informational masking. *J Acoust Soc Am* 113:2984–2987.

Freyman RL, Balakrishnan U, Helfer KS. (2001) Spatial release from informational masking in speech recognition. *J Acoust Soc Am* 109:2112–2122.

Freyman RL, Balakrishnan U, Helfer KS. (2004) Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J Acoust Soc Am* 115:2246–2256.

Freyman RL, Helfer KS, McCall DD, Clifton RK. (1999) The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am* 106:3578–3588.

Hasher L, Stoltzfus ER, Zacks RT, Rypma BA. (1991) Age and inhibition. *J Exp Psych Learn Mem Cogn* 17:163–169.

Hasher L, Zacks RT. (1988) Working memory, comprehension, and aging: a review and a new view. In: Bower GH, ed. *The Psychology of Learning and Motivation*. Vol. 22. San Diego, CA: Academic Press, 193–225.

Hasher L, Zacks RT, May CP. (1999) Inhibitory control, circadian arousal, and age. In: Gopher D, Koriat A, eds. *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*. Cambridge, MA: MIT Press, 653–675.

Helfer KS. (1992) Aging and the binaural advantage in reverberation and noise. *J Speech Hear Res* 35:1394–1401.

Helfer KS, Freyman RL. (2005) The role of visual speech cues in reducing energetic and informational masking. *J Acoust Soc Am* 117:842–849.

Hornsby BW, Ricketts TA, Johnson EE. (2006) The effects of speech and speechlike maskers on unaided and aided speech recognition in persons with hearing loss. *J Am Acad Audiol* 17:432–437.

Kidd Jr G, Arbogast TL, Mason CR, Gallun FJ. (2005) The advantage of knowing where to listen. *J Acoust Soc Am* 118:3804–3815.

Kidd Jr G, Mason CR, Deliwala PS, Woods WS, Colburn HS. (1994) Reducing informational masking by sound segregation. *J Acoust Soc Am* 95:3475–3480.

Kirasic KC, Allen GL, Dobson SH, Binder KS. (1996) Aging, cognitive resources, and declarative learning. *Psych Aging* 11:658–670.

Li L, Daneman M, Qi J, Schneider BA. (2004) Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults? *J Exp Psych Human Percep Perf* 30:1077–1091.

Li L, Yue Q. (2002) Auditory gaiting processes and binaural inhibition in the inferior colliculus. *Hear Res* 168:113–124.

Li L, Qi JG, Yu H, Alain C, Schneider BA. (2005) Attribute capture in the precedence effect for long-duration noise sounds. *Hear Res* 202:235–247.

Litovsky RY, Colburn HS, Yost WA, Guzman SJ. (1999) The precedence effect. *J Acoust Soc Am* 106:1633–1654.

McDowd JM, Oseas-Kreger DM, Fillion DL. (1995) Inhibitory processes in cognition and aging. In: Dempster FN, Brainerd CJ, eds. *Interference and Inhibition in Cognition*. San Diego, CA: Academic Press, 363–400.

Miyake A, Shah P. (1999) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. New York: Cambridge University Press.

Plomp R, Mimpen AM. (1979) Speech-reception threshold for sentences as a function of age and noise level. *J Acoust Soc Am* 66:1333–1342.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. (1995) Speech recognition with primarily temporal cues. *Science* 270:303–304.

Stine EAL, Wingfield A. (1990) How much do working memory deficits contribute to age differences in discourse memory? *Eur J Cog Psych* 2:289–304.

Stoltzfus ER, Hasher L, Zacks RT. (1996) Working memory and aging: current status of the inhibitory view. In: Richardson JTE, Engle RW, Hasher L, Logie RH, Stoltzfus ER, Zacks RT, eds. *Working Memory and Cognition*. Oxford: Oxford University Press, 66–88.

Sumby WH, Pollack I. (1954) Visual contributions to speech intelligibility in noise. *J Acoust Soc Am* 26:212–215.

Summers V, Molis MR. (2004) Speech recognition in fluctuating and continuous maskers: effects of hearing loss and presentation level. *J Speech Lang Hear Res* 47:245–256.

Zurek PM. (1980) The precedence effect and its possible role in the avoidance of interaural ambiguities. *J Acoust Soc Am* 67:952–964.