Chapter 4
Informational Masking in Speech Recognition

Gerald Kidd Jr. and H. Steven Colburn

Abstract Solving the "cocktail party problem" depends on segregating, selecting, and comprehending the message of one specific talker among competing talkers. This chapter reviews the history of study of speech-on-speech (SOS) masking, highlighting the major ideas influencing the development of theories that have been proposed to account for SOS masking. Much of the early work focused on the role of spectrotemporal overlap of sounds, and the concomitant competition for representation in the auditory nervous system, as the primary cause of masking (termed energetic masking). However, there were some early indications—confirmed and extended in later studies—of the critical role played by central factors such as attention, memory, and linguistic processing. The difficulties related to these factors are grouped together and referred to as informational masking. The influence of methodological issues—in particular the need for a means of designating the target source in SOS masking experiments—is emphasized as contributing to the discrepancies in the findings and conclusions that frequent the history of study of this topic. Although the modeling of informational masking for the case of SOS masking has yet to be developed to any great extent, a long history of modeling binaural release from energetic masking has led to the application/adaptation of binaural models to the cocktail party problem. These models can predict some, but not all, of the factors that contribute to solving this problem. Some of these models, and their inherent limitations, are reviewed briefly here.

G. Kidd Jr. (✉)
Department of Speech, Language and Hearing Sciences, Hearing Research Center,
Boston University, 635 Commonwealth Avenue, Boston, MA 02215, USA
e-mail: gkidd@bu.edu

H.S. Colburn
Department of Biomedical Engineering, Hearing Research Center, Boston University,
44 Cummington Street, Boston, MA 02215, USA
e-mail: colburn@bu.edu

J.C. Middlebrooks et al. (eds.), *The Auditory System
at the Cocktail Party*, Springer Handbook of Auditory Research 60,
DOI 10.1007/978-3-319-51662-2_4

Keywords Adverse listening conditions · Auditory masking · Auditory scene analysis · Binaural models · Cocktail party problem · Energetic masking · Informational masking · Speech comprehension · Speech in noise · Speech perception

4.1 Introduction

Of all of the important uses for the sense of hearing, human listeners are perhaps most dependent in their everyday lives on selectively attending to one talker among concurrent talkers and following the flow of communication between participants in conversation. This ability is fundamental to a wide range of typical social interactions and, for listeners with normal hearing at least, usually is accomplished successfully and fairly effortlessly (see Mattys et al. 2012; Carlile 2014; and Bronkhorst 2015 for recent reviews). It has long been recognized, though, that these are highly complex tasks that must be solved by the concerted actions of the ears and the brain (and, in many cases, the eyes as well). Extracting a stream of speech from one talker among a mixture of talkers or other sounds depends on perceptually segregating the different sound sources, selecting one to focus attention on, and then recognizing and comprehending the flow of information emanating from the chosen source. These tasks usually are performed while the listener remains attuned—to some degree—to sources outside of the primary focus of attention in the event that attention needs to be redirected. The sounds a listener may wish to receive ("targets") often overlap in time and frequency with competing sounds ("maskers"), resulting in what is known as "energetic masking" (EM). Even in the absence of spectral or temporal overlap, however, a variety of other factors may act to limit target speech recognition. These factors are broadly categorized as "informational masking" (IM).

The present chapter compares and contrasts EM and IM for the case of speech-on-speech (SOS) masking. The chapter is divided into three sections. First, the early work on the masking of speech by speech and other sounds is reviewed in an attempt to explain how the major ideas developed and the evidence on which they were based. Second, the issues involved in measuring SOS masking are discussed, focusing on how the distinction between EM and IM is made. Finally, some models of masking are considered—in particular those binaural models addressing the benefit of spatial separation of sources—with respect to how they may be applied to the masking of speech by other speech.

4.2 The History of Study of the Special Case of SOS Masking

In his seminal article describing the masking of speech, George A. Miller (1947) writes, “It is said that the best place to hide a leaf is in the forest, and presumably the best place to hide a voice is among other voices” (p. 118). Although he concluded in that article that the masking of speech by other speech was largely a consequence of overlapping energy in time and frequency, this analogy serves to illustrate a fundamental problem in the design of speech masking experiments: when the sound field contains many distinct but similar sources, how do we ask the question of whether one specific source is present or what information is being conveyed by that particular source? In a typical communication situation comprising multiple concurrent talkers, a listener normally may use a variety of cues—often relying heavily on context—to segregate the sounds and determine which source should be the focus of attention.

Cherry (1953) suggested several factors facilitating the process of separating one talker from others, including differences in source direction, lip-reading and gestures, differences in vocal properties and accents between talkers, and various transition probabilities. In designing experiments in the laboratory to measure aspects of this formidable ability, such as determining the strength of source segregation cues or measuring the ability to shift attention from one source to another, the means by which one source is designated as the target and so distinguished from those sources that are maskers may exert a profound influence on the outcome of the experiment. Thus, assessing the potential benefit that might result from another variable under test is strongly influenced by the way that the target is designated as the target, and a different answer about the role of such factors may be obtained with a different means for designating the source. This issue pervades the literature on SOS masking and has become increasingly relevant as a finer distinction is drawn between the sources of interference from competing talkers (i.e., whether they produce primarily EM or IM).

The issue of source designation in SOS masking was raised early on by Broadbent (1952a), who demonstrated that the manner of target designation could affect the amount of masking produced by a concurrent talker. In summarizing a study of the factors that underlie the recognition of the speech of one talker in competition with another, he observes: “From the practical point of view, these experiments show that there is a possibility, when two messages arrive simultaneously, of identification of the message to be answered becoming a more serious problem than the understanding of it once identified” (p. 126). However, because the majority of early work on the topic of masking relied on noise maskers—regardless of whether the target was speech or other sounds such as pure tones—the issues of source designation and listener uncertainty (e.g., possibility for source confusions) were not given extensive consideration (a notable exception is the topic of signal frequency uncertainty; cf. Kidd et al. 2008a). Likewise, in Cherry’s (1953) study, designating one ear as containing the target with the other ear containing the masker provided a simple, unambiguous means of source designation.

The findings from much of the early work on SOS masking were, in fact, largely consistent with the more general view of masking that was prevalent at the time: that is, that one sound interferes with the reception and processing of another sound primarily by obscuring or covering up the energy of the target sound within the frequency channels ("critical bands"; Fletcher 1940) containing the target. This perspective, which is based on EM, led to the original methods proposed for predicting speech recognition in noise (e.g., Egan and Weiner 1946; French and Steinberg 1947) as well as later refinements of those methods such as the speech intelligibility index (SII; cf. ANSI 1997). The connection between detecting a tone in noise and understanding speech in noise seemed obvious. For example, Beranek (1947) states, "Of great importance in understanding the ability of the ear to interpret transmitted speech is the way in which various noises mask desired sounds. Extensive tests have shown that for noises with a continuous spectrum, it is the noise in the immediate frequency region of the masked tone which contributes to the masking…. The bandwidth at which the masking just reaches its stable value is known as a "critical band"… Bands of speech appear to be masked by continuous-spectra noises in much the same way as pure tones are masked by them. For this reason, it is possible to divide the speech spectrum into narrow bands and study each band independently of the others" (p. 882).

Using noise as a masker has many advantages: it is easy to specify based on its underlying statistical properties, and it produces masking that tends to be more repeatable across trials and subjects than that produced by speech maskers (e.g., Freyman et al. 1999; Brungart 2001; Arbogast et al. 2002). Also, importantly, one need not worry about the listener confusing the target with the masker so that attention is unlikely to be misdirected, nor does noise typically carry any special information that commands our interest (however, the effect of Gaussian noise is not confined to EM although it often is used as a high-EM control condition for comparison; cf. Culling and Stone, Chap. 3; Schubotz et al., 2016).

Some of the early findings that supported EM as the basis for SOS masking include Miller's (1947) report that the masking produced by unintelligible speech from a language other than that of the listener was about the same as for intelligible speech in the primary language. Similarly, Miller noted that uncertainty about the content or production of speech also had little effect on masking: "The content of the masking speech is a more difficult factor to evaluate [than masking by noise or other non-speech sounds]. Conversational voices were compared with loud, excited voices liberally interspersed with laughter, cheering and improbable vocal effects. The two sounds could be likened to the chatter at a friendly dinner-party versus the din of a particularly riotous New Year's Eve celebration" (p. 119). These findings led Miller to state: "Once again, it is necessary to conclude that the crucial factor is the masking spectrum. The particular way in which the spectrum is produced is of secondary importance" (p. 120). Although this work was limited by the methods available at the time, and later work produced findings inconsistent with this broad conclusion, Miller's comments presaged both the "cocktail party problem" and, importantly, the role that uncertainty could play in SOS masking.[1]

The proposition that central factors—and not just peripheral overlap—may contribute to speech signals masking other speech signals was given strong empirical support by Broadbent (1952b). In a clever paradigm, he interleaved target and masker words in sequence finding that, despite the fact that the words had no spectrotemporal overlap and therefore ostensibly no EM, performance in target speech recognition nonetheless was degraded by the presence of the intervening masker words. Furthermore, certain nonacoustic aspects of the stimuli (e.g., familiar target voice; cf. Johnsrude et al. 2013; Samson and Johnsrude 2016) also influenced performance. Broadbent considered that his results revealed a "failure of attention in selective listening" because a perfect selection mechanism could simply gate "on" only the target words and gate "off" the masker words so that they would have no masking effect. Later, Broadbent (Broadbent 1958; pp. 11–29) concluded that these findings provided strong evidence for "central factors" in masking.

In an article that identified and evaluated several factors contributing to SOS masking that involved both peripheral and central mechanisms, Schubert and Schultz (1962) measured the benefit of imposing differences in interaural timing between the target talker and masker talkers. This study exemplified some of the difficulties inherent to the study of SOS masking because multiple variables influenced the results, but it also identified several ways that SOS masking could be released by central factors. The binaural differences they imposed were phase inversion (i.e., the target was π radians out of phase at the two ears while the masker was in-phase at the two ears; $S_{\pi}M_{0}$) or broadband time delays. Those manipulations were logical extensions of earlier work demonstrating masking level differences (MLDs) for detecting tones in noise (e.g., Hirsh 1948) and intelligibility gains for speech in noise (Licklider 1948), and therefore aimed to reduce EM (see Sect. 4.4). Other manipulations tried by Schubert and Schultz (1962), however, appear to have stemmed from intuitions about the perceptual basis upon which sources are segregated. This is apparent in their Table 1, in which they proposed a hierarchical arrangement of the effects of the masking stimuli according to a rough, qualitative estimate of similarity to the target. In that hierarchy, the most similar masker was the target talker's own voice, followed by single same-sex talker, single different-sex talker, multiple talkers, and ultimately multiple talkers reversed in time. It is clear from that hierarchy that their choice of masking stimuli reflected an expectation about an interaction between the binaural manipulations and these similarity-based masker properties.

In a study that has been widely cited because it identified both the masking of speech that could not be attributed to peripheral processes and the release from

[1] Irwin Pollack (2002; personal communication) attributed his use of the term "informational masking" to influential comments by George A. Miller at a seminar presented by Pollack describing the masking of speech by bands of filtered noise. According to Pollack, Miller objected to (Pollack's) use of noise as a masker considering its effects to be "secondary" to the "informational content of the messages" contained in speech maskers.

masking of speech beyond that predicted by traditional models of binaural unmasking, Carhart et al. (1969a) reported several instances of "excess masking." As with the Schubert and Schultz (1962) study, Carhart et al. (1969a) were interested primarily in understanding binaural release from masking for speech. However, that interest inevitably led to consideration of the cause of masking to begin with. It became clear that explanations were required for this excess masking effect—which they termed "perceptual masking"—that extended beyond traditional EM-based theories and models (see also Carhart et al. 1969b).

4.3 Determining Energetic and Informational Masking in SOS Masking

Although there are several methods that researchers have employed in an attempt to separate energetic and informational factors in masking experiments, the two most common are—broadly speaking—to vary the degree of target and/or masker uncertainty in the task and to control the amount of spectrotemporal overlap that is present between target and masker. In the former case, this is usually accomplished by manipulating the variability in the stimulus or the manner in which it is presented to the listener. In the latter case, an attempt is made to hold EM constant (or is taken into account by modeling) while factors that do not influence EM (e.g., linguistic aspects of speech) are varied, with the rationale being that any observed changes in performance may then be attributed to the influences of IM.

4.3.1 Uncertainty

Manipulating observer uncertainty by imposing stimulus variability is an empirical approach that was commonly employed in the early studies of IM using nonspeech stimuli (see Kidd et al. 2008a for a review). For example, in the series of studies by Watson and colleagues (summarized in Watson 2005), the task often was to detect an alteration in the frequency or intensity of a tone pulse embedded in a sequence of similar pulses or "context tones." The way that the context tones were presented—specifically, whether they varied in composition from trial to trial within a block of trials or were held constant across trials within a block—was used to manipulate listener uncertainty and often produced large differences in performance. Although less common in the SOS masking literature, analogous manipulations are possible. Brungart and Simpson (2004) explicitly varied the degree of uncertainty in a SOS masking paradigm. They used a closed-set, forced-choice, speech identification task (the "Coordinate Response Measure," CRM, test) in which the target voice is followed throughout the sentence after a specified "callsign" occurs until two test words—a color and a number—are presented (cf. Brungart 2001; Iyer et al. 2010).

Both the masker talkers and/or the semantic content could be fixed or randomized across trials. Somewhat surprisingly based on a logical extrapolation of the findings from the nonspeech IM literature, increasing masker uncertainty caused little decrement in performance, with variability in semantic content producing the only statistically significant difference. Similarly, Freyman et al. (2007) tested a condition in which masker sentences were held constant across trials or varied randomly across trials. Consistent with the small effects of masker uncertainty reported by Brungart and Simpson (2004), no significant effect on performance was found due to masker uncertainty for variation in talker, content, or target-to-masker ratio (T/M). The open-set target speech materials used by Freyman and colleagues were nonsense sentences while the maskers were similar nonsense sentences from a different corpus. It is possible that the time available to focus on these relatively long stimuli allowed the listener to overcome any initial uncertainty about the characteristics of the target source. With a clear cue to source designation (e.g., the callsign for the CRM test), the ability to select the target source was sufficient to overcome the relatively minor uncertainty caused by the stimulus variation that was present.

Uncertainty about some aspects of the stimulus or its presentation *can* affect the amount of IM in SOS masking. For example, Kidd et al. (2005) demonstrated that uncertainty about the spatial location of a target talker influenced speech identification performance in a multiple-talker sound field. By manipulating the a priori probability of target presentation (one of three concurrent talkers) from one of three locations separated in azimuth, Kidd and colleagues found large differences in performance depending on whether the listener was provided with the cue designating the target sentence (the "callsign") before or after the stimulus. When the listener had no a priori knowledge about target location and did not receive the callsign designating the target until after the stimulus, performance was relatively poor—near the value expected simply from choosing to focus attention on only one of the three locations. When the target sentence was cued/designated before the trial, but location was uncertain, performance improved significantly relative to the uncued case. When the probabilities about source location were provided before the stimulus, performance improved significantly for both cued and uncued conditions. If the location of the target was certain, proportion correct identification performance was higher than 0.9 independent of whether the target was cued beforehand. These findings are shown in Fig. 4.1A. Similar effects of location uncertainty have been reported by Best and colleagues (2007) and by Kidd and colleagues (2014) using different paradigms. In those studies, as in the Kidd et al. (2005) study just described, the conclusion was that a priori knowledge about target source location can improve speech recognition under multiple-talker competition..

An example of the type of error analysis that reveals confusions among sources is found in Fig. 4.1B, reproduced from Kidd et al. (2005). This panel shows a breakdown of error types for each condition. For the condition with the greatest certainty about location, the most frequent error was to mix one target word (color or number) with one masker word. For the uncertain cases, the most common error was to report both color and number words from one of the two masker sources.

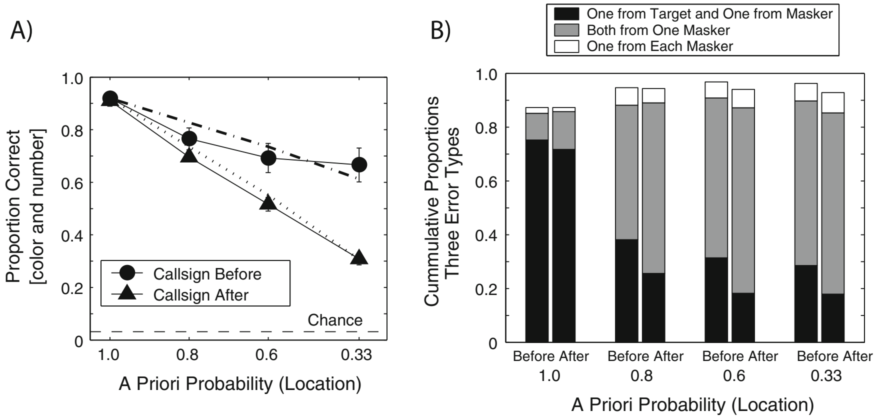

Fig. 4.1 **(A)** Proportion correct speech identification scores as a function of the a priori probability of occurrence at one of three locations. The data points are group means with standard errors. The straight lines are predictions of a simple probability-based model. The circles show performance when the callsign designating the target sentence was provided before the stimulus while the triangles show performance when the callsign was provided after the stimulus. Chance performance is indicated by the dashed line at the bottom. **(B)** The error analysis associated with the results shown in **A**. The bars are composite histograms indicating the proportions of error types that occurred. (**A** and **B** from Kidd et al. 2005, *The Journal of the Acoustical Society of America*, with permission.)

The difference between the height of each composite bar and 1.0 indicates the proportion of errors not attributable to confusions that could be due to EM. The authors concluded that in nearly all cases the three talkers likely were each audible but that errors occurred because of source confusions/misdirected attention.

It is clear from the preceding discussion that the structure of the SOS masking task can affect the outcome of the experiment. This observation may seem obvious but what is (or historically has been) less obvious is that it applies much more strongly for speech masked by other speech than for speech masked by noise and is at the heart of the IM–EM distinction. The conditions that produce the highest IM tend to be those in which confusions are possible such as happens when both target and masker share similar low-level features (e.g., same-sex talkers or even same talker as masker) and the masker words are allowable response alternatives in closed-set paradigms (see Webster 1983 for a review of early work on closed-set speech tests). Using very different types of materials for target and masker(s) can greatly reduce uncertainty and therefore reduce IM. Natural communication situations may of course vary widely in the degree to which source or message uncertainty is present and expectation based on context and a priori knowledge often determines success.

4.3.2 Controlling/Estimating Energetic Masking

When two or more independent talkers are speaking concurrently, the acoustic overlap between the sounds varies considerably from moment to moment. The spectrotemporal overlap of the speech from different sources depends on a variety of factors including inherent differences in source characteristics (e.g., size and shape of the vocal apparatus, acquired speaking patterns, etc.), the speech materials that are being uttered by the various sources, and the acoustic environment (e.g., reverberation), among others. Moreover, speech sources in real sound fields typically originate from different locations meaning that the waveforms arrive at the listener's ears with differing interaural time and intensity values. For this reason, perhaps, much of the work on the "cocktail party problem" has addressed multiple source segregation and selection cues that occur concurrently and include such explicit factors as binaural difference cues and fundamental frequency/formant resonance differences, etc., in addition to the source designation methods discussed in Sect. 4.2. Ultimately, determining the precise way that the sounds overlap in their representations in the auditory system can be a very complex problem involving models of how the ear codes the relevant sound parameters dynamically and the interaural differences in the sound inputs.

Because the early stages of the peripheral auditory system are tonotopically organized, one nearly universal way of thinking about EM is to divide the stimulus into physiologically inspired frequency channels and to consider how the representations of the competing speech sounds are preserved within these channels over time. To test hypotheses about how these representations interact under different assumptions, a variety of experimental approaches have been devised that reduce the acoustic stimulus to limited frequency regions so as to manipulate the overlap that occurs within auditory channels.

Among the first studies to attempt to separate EM from IM in SOS masking by taking advantage of the tonotopic organization of sounds in the auditory system was Arbogast et al. (2002). They used a tone-vocoding procedure to process two independent speech sources into acoustically mutually exclusive frequency channels (within the limits of the procedure). This is illustrated in Fig. 4.2.

The upper panels show the magnitude spectra of the processed target plus masker while the lower panels show the waveforms. The two types of masker shown are "different-band speech" (DBS), which consists of intelligible speech in narrow frequency bands that do not contain target speech and "different-band noise" (DBN), which consists of equally narrow (unintelligible) bands of noise in the bands that do not contain target speech. Pilot tests showed that sufficient speech information was present in the envelopes of the small number of spectral bands for the target and masker speech sources each to be intelligible separately. To solve the task the listener had to distinguish the target speech from another similar CRM sentence (DBS condition) spoken by a different talker. The key to determining the amount of IM present was to compare performance obtained using the speech masker (DBS) with the performance obtained using the noise masker (DBN).

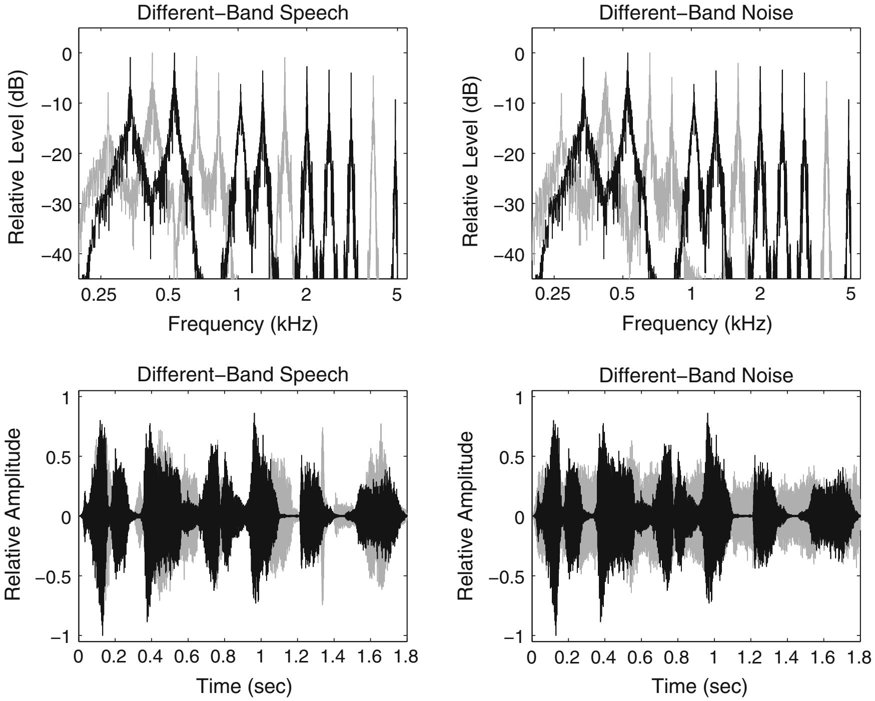

Fig. 4.2 The *upper two panels* show the magnitude spectra for the "different-band speech" and "different-band noise" maskers (light gray) plus target (dark gray) while the *lower two panels* show the associated waveforms (same shading). As may be seen from the upper panels, the target and maskers are processed into mutually exclusive frequency channels that are chosen randomly on every presentation. (Adapted from Arbogast et al. 2002, *The Journal of the Acoustical Society of America*, with permission.)

Because the amount of EM for the DBS and DBN maskers was expected to be about the same, the greater masking caused by the speech (about 18 dB) was attributed to IM. The large amount of IM found in this experiment depended in part on the specific way that the stimuli were processed which was designed to minimize EM while preserving enough of the speech for high intelligibility. The important finding from Arbogast et al. (2002) for the current discussion is that maskers that were equated for EM were shown to produce significantly different amounts of IM depending on whether the masker was intelligible.

Brungart et al. (2006) proposed a method of processing speech into highly quantized elements so that the EM and IM present in SOS masking could be estimated/controlled. Not only did they analyze the speech stimulus into narrow frequency channels but they also then subdivided each channel into brief time intervals. Essentially, the result was a matrix of values representing energy contained in fine time–frequency (T–F) units. Based on a priori knowledge of the stimuli, the T/M in each bin was computed and a criterion for sorting the bins based

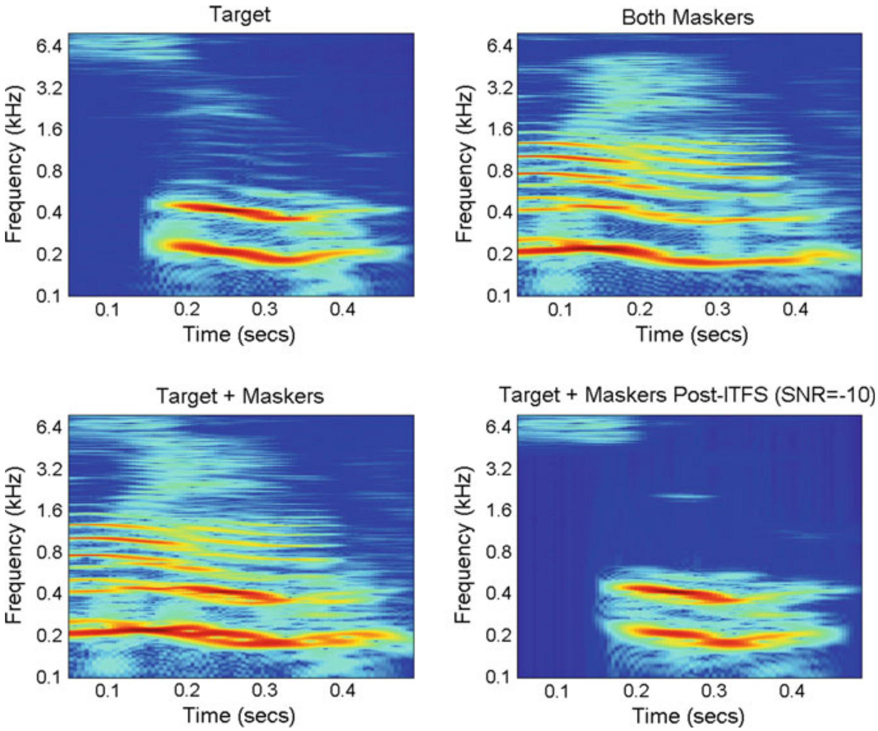

Fig. 4.3 Results of the processing of target and masker stimuli into time–frequency bins following the procedure used by Brungart et al. (2006). The abscissa is time while the ordinate is frequency on a log scale. Red/blue shading represents high/low intensity. The *upper left panel* shows the spectrogram of the target; the *upper right panel* shows the spectrogram of the two-talker masker; the *lower left panel* shows the combination of the target and maskers; and the *lower right panel* shows the T–F units of the combined stimulus for which T > M (stimuli and analysis)

on T/M was applied. The criterion could be used to exclude bins based on T/M—discarding the bins below the criterion—with the remaining bins reassembled into a speech stimulus. The results of this procedure applied to multiple speech sources are shown in Fig. 4.3.

The top left panel is a spectrogram of the original target speech; the top right panel shows the masker speech (two independent maskers); the lower left panel shows the mixture of target and masker signals, while the lower right panel shows only the T–F units that remain after discarding those in which the masker energy is greater than the target energy (an "ideal binary mask"). In the procedure used by Brungart et al. (2006) the difference in intelligibility between the two sets of stimuli shown in the lower panels is taken as an estimate of IM. The finding of a significant improvement in speech identification performance by removal of the low T/M bins argued for a strong role of IM. This is a crucial finding on a theoretical level because the usual assumption about combining the information from different T–F units based on speech in noise tasks is that each unit containing target energy contributes some increment—even if infinitesimal—to overall intelligibility. The worst a T–F unit could do is to produce no appreciable gain. However, the presence of IM means that the presence of units with little or no target information *reduces* overall intelligibility. In fact, not only will including these units "garble" the target, they also may yield an alternate, intelligible source that is confused with the target source. In contrast, a parallel manipulation using noise as a masker revealed minor detrimental effects of presentation of the unprocessed versus processed stimulus thereby eliminating differences in EM as the cause of the effect. The findings of Brungart et al. (2006) were significant not only because they provided a quantitative means for separating EM from IM in SOS mixtures but also because their results revealed a dominant role of IM in SOS masking for the stimuli and conditions tested. In a later study using the procedure described above, Brungart et al. (2009) found that increasing the number of independent masker talkers to the point where the individual voices are lost in an incomprehensible—but obviously speech—babble increased EM while decreasing IM. The idea that increasing the number of similar individual elements in the sound field (like increasing the number of leaves in the forest envisioned by Miller 1947), increases EM while it (ultimately) decreases IM, is a common theme in contemporary auditory masking studies (cf. Kidd et al. 2008a). The use of unintelligible babble as a speech masker, coupled with strong target segregation/designation cues, likely contributed to the conclusion from some early studies that SOS masking was predictable solely on the basis of spectrotemporal overlap of the competing sources.

4.3.3 Linguistic Variables

A persistent question in the SOS literature is whether the degree of meaningfulness of competing maskers affects the masking that is observed. For example, randomly selected words with no syntactic structure and little semantic value are less meaningful than coherent discourse, but do they mask target speech any less? If so, does this imply that the greater the meaning, or perceived potential to carry meaning, a masker possesses the more it invokes some degree of obligatory processing? If linguistic variables affect SOS masking, then an explanation based purely on peripheral overlap of excitation falls short of providing a satisfactory account of the underlying processes governing performance. Although this point has been recognized for decades, the evidence often has been inconclusive and sometimes contradictory—partly for reasons discussed in Sect. 4.2 concerning differences in methodology. Here we review work intended to determine the role that linguistic variables play in masking target speech.

4.3.3.1 Time Reversal

Among the more obvious ways of evaluating the influence of lexical factors in SOS masking is to degrade the meaning of speech by reversing it in time. Historically, reversed speech has been an intriguing stimulus because it largely maintains its frequency and envelope spectra while losing intelligibility (cf. Kellogg 1939; Cherry 1953; Schubert and Schultz 1962). Differences in the amount of masking produced by time-forward speech and the same speech time-reversed therefore could be due to the difference in "meaningfulness." Based on the premise that "speech perception cannot be explained by principles that apply to perception of sounds in general" (p. 208), Hygge et al. (1992) reasoned that "…it can be expected that a normal background speech condition should interfere more with a speech comprehension task than a noise control that does not carry any phonological information (and)…normal (i.e., forward) speech should interfere more than the same speech played in reverse…" With respect to early work examining this issue, an article by Dirks and Bower (1969) was particularly influential. In their careful and systematic study, short "synthetic" sentences (Speaks and Jerger 1965) spoken by a male talker were masked by unrelated, continuous discourse spoken by the same talker played forward or backward. The observed performance-level functions indicated nearly identical results in all cases. Likewise, in the Hygge et al. (1992) study, in which the target talker was female and the masker was a single male talker, no significant difference in the amount of masking (using a subjective "just understandable" criterion and method of adjustment) was found when the masker talker was presented normally versus time reversed. In this case the speech materials (both target and maskers) were relatively long (3 min) passages of connected speech. The conclusion drawn from these studies, supported by the original findings from Miller (1947) noted in Sect. 4.2, was that the main determinant of SOS masking is the spectrotemporal overlap of the sounds and that linguistic factors per se were of little import. These studies suggest that the outcomes of SOS masking experiments are very sensitive to the specific methods that are used. When the masker differs in fundamental ways from the target—on a semantic level, as is the case with very different types of speech materials, or on a more basic acoustic level as with the differences in source characteristics for male versus female talkers—uncertainty may be minimal and subsequent manipulations intended to examine other factors (e.g., masker time reversal) may produce negligible effects.

In a pivotal article in the IM literature concerning speech, Freyman et al. (2001) reported a large difference (4–8 dB) between the masking effectiveness of forward and time-reversed masker speech. The speech corpus for both target and masker consisted of simple sentences spoken by female talkers that were semantically implausible but syntactically correct. Importantly for the discussion that follows regarding spatial release from IM, the additional release from IM (beyond that obtained by time reversal) due to a perceived difference in location between target and masker was relatively small when compared to the same perceived location difference for forward speech. These findings suggested that the high IM produced by the SOS masking conditions tested could be released by *either* time reversing the masker—causing it to be unintelligible—*or* by perceptually segregating the apparent locations of the sources.

The large benefit due to time-reversing the masker obtainable in some SOS conditions subsequently has been confirmed in several other studies. Marrone et al. (2008; see also Best et al. 2012) used the closed-set CRM test spoken by a female target talker masked by two female masker talkers with the specific voices randomized from trial to trial. Marrone and colleagues varied the locations from which the maskers were presented using the co-located case as a reference for determining spatial benefit. When target and masker talkers were co-located, time-reversing the maskers yielded a large advantage over natural presentation with the T/Ms at threshold lower by about 12 dB—nearly the same release from masking as was obtained from spatial separation of sources. Even larger reductions in T/M due to masker time reversal—about 17 dB, on average—in co-located conditions have been reported by Kidd et al. (2010). They used a different closed-set speech identification test with female target and two female masker talkers uttering five-word sentences with all of the syntactically correct sentences drawn from the same corpus. As with Marrone et al. (2008), the specific talkers were selected randomly from a small closed set of talkers on every trial. The large "reversed masking release" (RMR) reported by Marrone et al. and Kidd et al. in the co-located condition likely reflects a reduction in IM based on the assumption that the amount of EM remains the same when the masker is time reversed. However, the extent to which time reversal preserves the EM of a speech masker is a matter of some conjecture. It is possible, for example, that time reversal affects the temporal masking that one phoneme can exert on another. Moreover, closed-set tests that use the same syntactic structure for target and masker speech, with some degree of synchrony, could result in more EM if the envelopes were highly correlated reducing "glimpses" of the target in masker envelope minima.

Rhebergen et al. (2005) proposed that time reversal of masking speech may not produce EM that is equivalent to natural speech. They noted that the envelopes of speech produced naturally often tend to exhibit an asymmetric shape with quick onsets (attributed to plosive sounds) followed by slower decays. Time reversal alters this shape so that the rise is more gradual and the offset more abrupt. The consequence of this reversal is that some soft sounds would be masked (via forward masking) in one case but not in the other so that EM could effectively differ. In the key finding from their study, the masking produced by a different-sex masking talker uttering a language that was not known to the listeners was greater when the masker was time reversed than when it was played forward. The greater amount of masking from the reversed speech was small, about 2 dB, but was judged to be significant. The greater EM for reversed speech means that release from IM due to time reversal may be *under*estimated by an amount that depends on the increase in EM due to greater forward masking from the reversed envelope.

Concerns about potential differences in EM due to time reversal, and the possibility that these differences are exacerbated when the target and masker sentences are similar in structure and spoken nearly in cadence, led Marrone et al. (2008) to test a "control" condition explicitly examining whether time-reversed speech generated greater EM than the same speech played forwards. In their experiment, the target speech was masked by two independent speech-spectrum–shaped speech-envelope–modulated noises that were co-located with the target. The speech envelopes that modulated the maskers were presented time-forward versus time-reversed. No significant difference was observed in threshold T/Ms between these two noise masker conditions, suggesting that EM was the same for both because the small amount of IM expected from modulated noise maskers would be the same as well. They concluded that the large reduction in masking found in the actual SOS conditions (about 12 dB) therefore was due to a release from IM and not to a difference in EM. Recent work from Kidd et al. (2016) using the ideal T–F segregation technique (e.g., Fig. 4.3) applied to time-forward and time-reversed speech supports the conclusion by Marrone and colleagues that the amount of EM for the two cases is the same. It should be noted that both Marrone and colleagues and Kidd and colleagues used (different) closed-set speech tests that have been shown to produce high IM. It is not yet clear whether the conclusion above generalizes to other types of speech materials and testing procedures and perhaps accounts for the small difference with the findings by Rhebergen et al. (2005) noted earlier in this section.

Further evidence that the meaningfulness of the masker may exert a strong effect in SOS masking comes from Kidd et al. (2008b; see also Best et al. 2011), who employed a variation on the “every other word” paradigm devised by Broadbent (1952b). In that paradigm, as implemented by Kidd and colleagues, five-word sentences from a closed-set corpus consisting of one random selection from each of five word categories (name, verb, number, adjective, object) were used to generate syntactically correct sentences (e.g., “Sue bought four old toys”). On any given trial, the target words formed the odd-numbered elements in a sequence with the even-numbered elements being masker words, time-reversed masker words, or noise bursts. When the masker was bursts of noise, performance was the same as when no masker was present. A small decrement in performance was found for the time-reversed speech masker but much less than was found for the meaningful time-forward speech (however, as noted in Sect. 4.3.3, masker syntax did not affect performance). This is a clear case in which speech caused significant IM with little or no EM. It should be pointed out that the small difference between the effect of the noise masker and the time-reversed speech masker is consistent with the view that even unintelligible speech—or masking stimuli that mimic the properties of speech such as speech-shaped speech-envelope–modulated noise—produces some amount of IM.

Swaminathan et al. (2015) reported large reductions (16–18.5 dB) in T/M at threshold for a target talker masked by two independent, same-sex masker talkers when the masker talkers were time reversed relative to when they were presented naturally. These large threshold reductions were obtained using the same closed-set speech materials employed by Kidd et al. (2010) in the study noted earlier in this section. Swaminathan and colleagues examined one factor potentially related to the individual differences observed between subjects: musical training. The results of this study are shown in Fig. 4.4A.

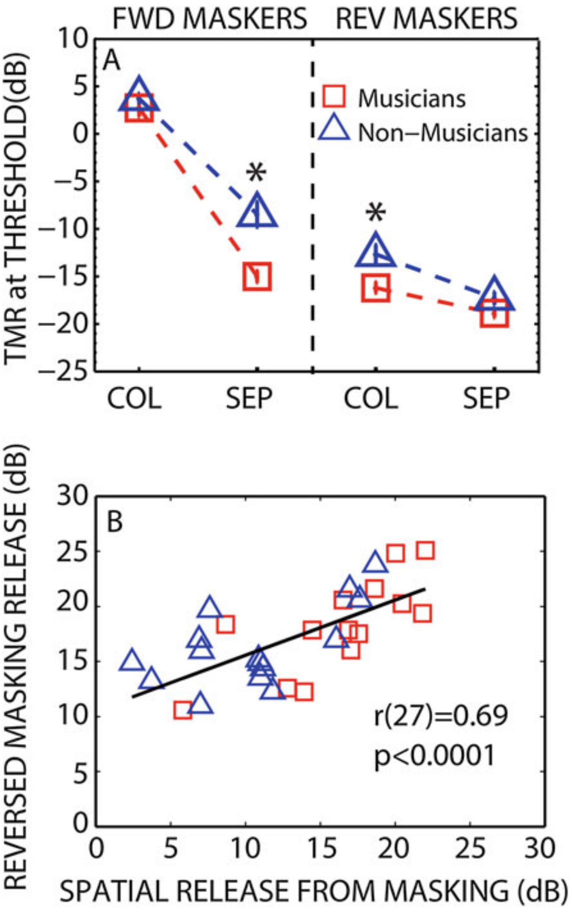

Fig. 4.4 **(A)** Group mean thresholds (target-to-masker ratio, TMR, in decibels) and standard errors for co-located (COL) and spatially separated (SEP) conditions for natural (FWD) and time-reversed (REV) speech maskers. The squares show the results from the musician group while triangles are for nonmusicians. The asterisks indicate statistically significant group differences. (From Swaminathan et al. 2015, *Scientific Reports*, with permission.) **(B)** Results from individual listeners plotted as reversed masking release (RMR) as a function of spatial masking release (SRM)

Group mean T/Ms at threshold are plotted for musicians and nonmusicians for time-forward and -reversed speech maskers presented in co-located and spatially separated configurations. Thresholds in the co-located condition for the forward speech maskers were about the same for the different subject groups, with relatively small differences observed across subjects. Either spatial separation or time reversal produced large reductions in T/Ms at threshold. Musicians as a group showed greater masking release for both variables than did their nonmusician counterparts. Large individual differences were observed for both subject groups. This is illustrated in Fig. 4.4B, in which the spatial release from masking (SRM) is plotted against the reduction in threshold that occurred due to masker time reversal (RMR) for individual subjects. The two subject groups are indicated by different symbols. The significant correlation between these variables suggests that subjects tended to exhibit a similar proficiency in using either variable to overcome IM (see also Kidd et al. 2016). It also is clear that, despite the overlap in the distributions, most individual musically trained listeners exhibited greater masking release than the nonmusicians. Supporting evidence for this finding was reported by Clayton et al. (2016; see also Başkent and Gaudrain 2016) who found that the best predictors of individual differences in SRM were musicianship and performance on a visual selective attention task. Swaminathan and colleagues argued that the differences between groups were more likely due to central factors related to training and/or innate ability than to differences in peripheral auditory mechanisms. They employed a physiologically inspired model of the responses of the auditory nerve (AN) to determine whether the large RMRs found experimentally could be accounted for by a decrease in EM. The performance predicted by the AN model, however, was roughly equivalent for the time-forward and -reversed conditions. Swaminathan and colleagues concluded that the large RMRs found in their study were not due to differences in EM but rather to differences in IM.

4.3.3.2 Familiar Versus Unfamiliar Languages as Maskers

As noted in Sect. 4.2, the attempt to determine whether the masking produced by a familiar language was greater than that produced by an unfamiliar language dates at least to the report by Miller (1947). Although early work did not find much evidence that SOS masking varied depending on whether the masker was understandable or not, more recent work clearly has shown that this can be the case. Freyman et al. (2001) reported small differences in masking between Dutch and English sentence-length maskers on the intelligibility of English target speech by native English listeners who did not understand Dutch. The differences they reported were as large as 10 percentage points at low T/Ms in a reference condition in which the target and masker were co-located (the study focused on the benefit of perceptual segregation of sources based on apparent location differences). In the Rhebergen et al. (2005) study discussed in Sect. 4.3.3.1, only a 2-dB difference in masked speech reception thresholds (SRTs) was found for maskers in familiar (Dutch) versus unfamiliar (Swedish) languages.

In an important study specifically designed to determine whether knowledge of the language spoken by the masker talker affects the amount of SOS masking, Van Engen and Bradlow (2007) tested the recognition of simple meaningful English sentences masked by speech in either a known (English) or unknown (Mandarin) language. The maskers were two or six concurrent talkers uttering semantically anomalous (implausible) sentences. The target speech was distinguished from the masking speech by the nature of the materials and by a temporal offset between the masker and the target. Van Engen and Bradlow found that speech recognition performance was poorer when the masker was English, particularly at low T/Ms, and comprised two masker talkers rather than six. The broad conclusion was that greater masking occurs when the masker is intelligible to the listener. Thus, English is a more effective masker than Mandarin for English-speaking listeners, especially when the maskers comprise distinctly individual, salient sources as opposed to multitalker babble.

A number of other studies have provided evidence that the amount of masking obtained in SOS masking experiments is greater when the masker language is familiar to the listener than when it is not, even after accounting for language-specific acoustic differences (e.g., Calandruccio et al. 2010, 2013). When the masker language is unfamiliar to the listener, there is little reason to expect that the masking it produces is substantially different from that produced by a familiar language that is unintelligible due to time reversal. The relatively small effects of maskers in familiar versus unfamiliar languages reported to date thus seems inconsistent with the large—and in some cases very large—masking release found for masker time reversal noted in Sect. 4.3.3 (e.g., 15–19 dB by Kidd et al. 2010 and Swaminathan et al. 2015). The reason for this discrepancy is not clear at present but may be due in part to differences in the procedures that have been used to study these issues.

The semantic content of speech may influence its effectiveness as a masker when the language in which it is spoken is native or otherwise well known to the listener. However, a much more complicated case arises when the target speech or the masker speech, or both, are spoken in a language known to the listener but are not the native or primary language (e.g., Cooke et al. 2008; Brouwer et al. 2012; Calandruccio et al. 2013). There are several possible combinations of talker–listener languages that may occur, and there are the further complications of the linguistic similarity between the target and masker speech together with the possibility that the unfamiliar language is actually partially comprehensible by the listener. If the target speech is in a language that is not well known/native to the listener, so that it requires greater effort and/or time for the listener to decipher, then it may be more susceptible to interference from other speech, especially if that speech is in the primary language. Conversely, if the target is in the primary language but the masker speech is not, the masker speech likely may be less distracting than if it is easily recognized (in the limit, as above, a completely unfamiliar language would cause relatively little IM). A general principle that appears to summarize many of the observations about primary and secondary language SOS masking, as well as other higher-level effects, was proposed by Brouwer and colleagues (2012) and is referred to as the "linguistic similarity" hypothesis.

In a study that emphasized the importance of linguistic factors, Ezzatian et al. (2010) measured SOS performance when the target and masker speech were in an acquired secondary language (English) as a function of the age of acquisition by the listener and compared performance to that of native English language listeners. They measured performance at different T/Ms for two spatial conditions, one where the target and masker were co-located and a second where target and masker were perceived at different spatial locations using the method of Freyman et al. (1999). The key findings are depicted in Fig. 4.5; open and filled symbols represent co-located and spatially/perceptually separated conditions, respectively.

The left column shows the results from a noise masker used as a high-EM control while the right panel shows results from a two-talker same-sex masker.

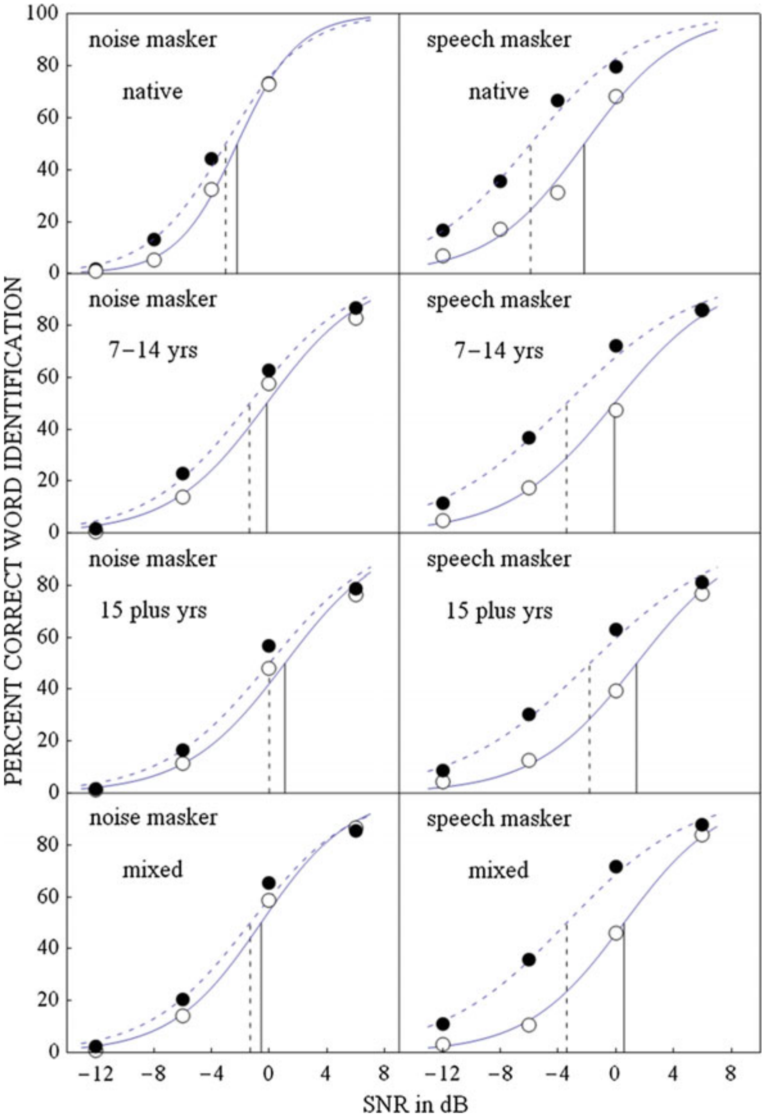

Fig. 4.5 Word identification performance as a function of signal-to-noise ratio (SNR) in decibels for four groups based on age of acquisition of English (native listeners; 7–14 years; 15-plus years; mixed: those who were raised in a non-English environment but learned to speak English at an early age). The left column is for a noise masker while the right column is for a speech masker. The open circles/solid lines represent spatially co-located target and masker. Solid circles/dashed lines indicate target and masker perceived from different locations. Thresholds (50% points on the psychometric functions) are indicated by the solid vertical lines for the co-located conditions and by the dashed vertical lines for the separated conditions. (From Ezzatian et al. 2010, *Speech Communication*, with permission.)

The rows are for different groups divided according to the age at which English was acquired. The important findings for this discussion are that performance was generally better (masking was less) when English was native (top row) or acquired early (7–14 years of age) as opposed to later (15 years or older) or in a "mixed" language environment where there was exposure to English from childhood but not as the primary language. Age of acquisition was less of a factor for the noise masker. A related finding concerning age of language acquisition was reported by Newman (2009). She tested infants' ability to recognize their own names (respond preferentially re other names) against different types of backgrounds including the speech of a single talker presented naturally or reversed in time. She concluded that linguistic influences on IM develop as language is acquired and that infants have not yet acquired language to the point where meaningful speech interferes more than similar nonmeaningful speech. In a recent study, Newman et al. (2015) found that the greater masking effectiveness for meaningful speech, compared to the same speech rendered unintelligible by time reversal, was apparent for children by the age of 4–6 years. These findings suggest that susceptibility to IM in SOS masking is influenced by the degree of linguistic competence in the target language, at least as indicated by age/length of time of acquisition (see also Buss et al., 2016, and Calandruccio et al., 2016).

4.3.3.3 Syntactic and Semantic Content: Predictability and Obligatory Processing

Cherry's (1953) seminal article exploring the factors governing communication performance in a "cocktail party" environment continues to be cited frequently for highlighting the importance of binaural processing of sounds and, less frequently, for identifying other relevant factors for separating competing talkers such as vocal characteristics and speech reading. However, what is often overlooked is that Cherry also emphasized the important role of predictability in natural communication and, indeed, the first experiment in his 1953 article was devoted to determining the effect of varying the predictability of speech by manipulating speaker transition probabilities. He states, "The logical principles involved in the recognition of speech seem to require that the brain have a vast "store" of probabilities, or at least of probability rankings. Such a store enables prediction to be made, noise or disturbances to be combatted, and maximum-likelihood estimates to be made" (p. 976). A number of speech corpora and tests have been developed subsequently that explicitly varied target speech predictability (e.g., Speaks and Jerger 1965; Kalikow et al. 1977; Uslar et al. 2013; Helfer and Jesse 2015).

Recently, Kidd et al. (2014) provided evidence suggesting that the predictability of sequences of words, as reflected by the conformance to a known syntax, can be beneficial in selectively attending to one of three spatially distributed speech sources. In their experiment, the intelligibility of target speech that comprised randomly selected words was compared to similar target speech arranged into brief,

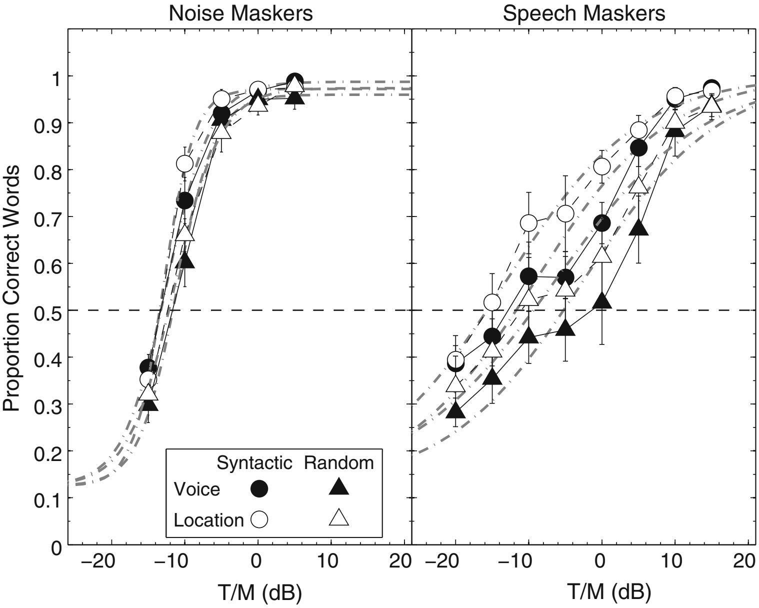

Fig. 4.6 Speech identification performance as a function of target-to-masker ratio (T/M) in decibels. The *left panel* contains the results for noise maskers while the *right panel* contains the results for speech maskers. The data points are group mean proportion correct scores and standard errors of the means. The fits are logistic functions (dashed-dotted lines) from which thresholds were obtained at the 0.5 proportion correct point (horizontal dashed line). The filled symbols are for conditions in which the target was indicated by constant voice while the open symbols are for conditions in which the target was indicated by constant location. Circles indicate that the target sentence was syntactically correct (*syntactic*) while triangles are for syntactically incorrect (*random*) target sentences. (From Kidd et al. 2014, *The Journal of the Acoustical Society of America*, with permission.)

syntactically correct, simple sentences masked by two competing talkers or by noise. Group mean results from that study are depicted in Fig. 4.6.

The left panel shows the intelligibility results obtained under two competing noise maskers while the right panel shows the results for two competing speech maskers. The primary cues to the target were constant talker voice or location, which were paired with correct or random target sentence syntax. In all cases, performance was better when the target conformed to correct syntax, but the differences—expressed as a reduction in T/M—were much larger when the maskers were speech. The authors concluded that the predictability of the target words conforming to a known syntax was particularly beneficial under conditions that were high in IM.

An earlier study by Freyman et al. (2004) demonstrated that priming a target sentence could improve performance under speech masking (but not noise masking) conditions relative to unprimed sentence presentation. They provided a prime by presenting a fragment of the target sentence spoken by the same talker that

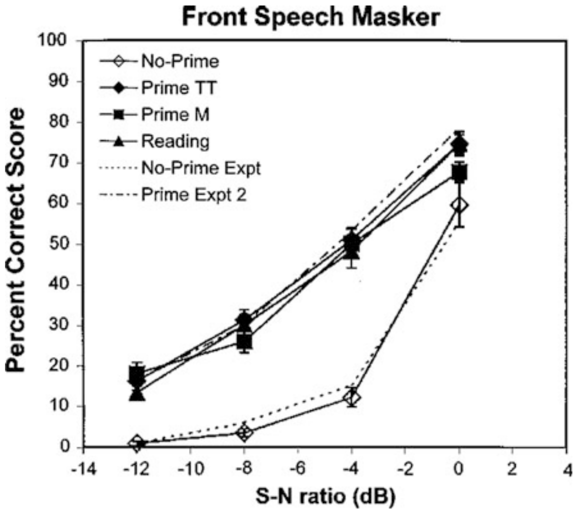

Fig. 4.7 Comparison of group mean percent correct scores and standard errors as a function of signal-to-noise ratio (S-N) for different priming conditions with target and masker co-located in the front. The control was the "no-prime" condition (open diamonds). "Prime TT" (filled diamonds) refers to the condition in which the target talker produced the priming utterance. "Prime M" (filled squares) is the condition in which the priming utterance was produced by a male (nontarget) talker. "Reading" (filled triangles) refers to the prime presented in print. Dashed/dotted lines without symbols show the primed and unprimed percent correct scores obtained in a separate experiment. (From Freyman et al. 2004, *The Journal of the Acoustical Society of America*, with permission.)

subsequently repeated the entire sentence as well as primes that were a different same-sex talker uttering the sentence fragment prime or the sentence fragment presented in written, rather than spoken, form. The results of this experiment are shown in Fig. 4.7. Rather remarkably, these three primes were equally effective in enhancing speech recognition performance. These effects were obtained using syntactically correct nonsense sentences masked by similar sentences from a different corpus for two co-located same-sex (as the target) masker talkers. Freyman and colleagues concluded that the benefit of the prime was to partially release IM by reducing the attentional resources devoted to the maskers.

Brouwer et al. (2012) proposed that the greater the degree of linguistic similarity between target and masker speech sources, the greater the IM that results. To test this "linguistic similarity hypothesis," they varied the language of the target and masker talkers (i.e., target in one language, masker in the same or different language), measuring performance when the languages were primary, secondary, or the masker was not understood by the listener. They also varied the semantic value of the target and masker speech. For both manipulations—language and semantic content—the observed amount of masking increased when the target and masker speech were similar, as compared to dissimilar according to their criteria. Some of their findings are shown in Fig. 4.8.

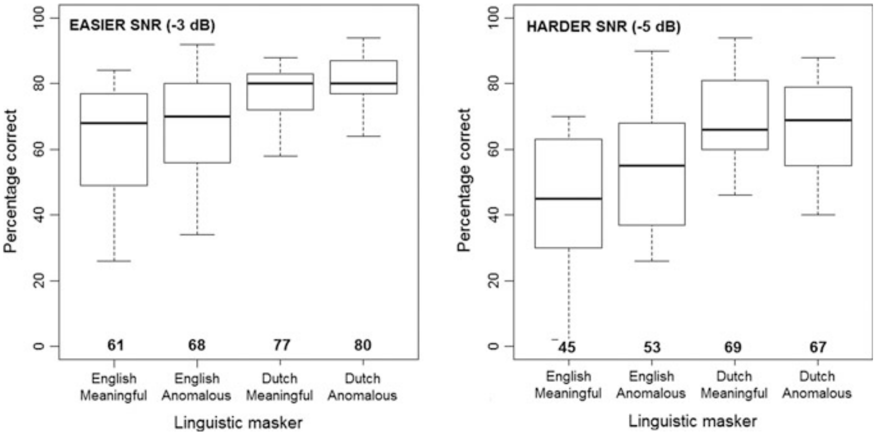

Fig. 4.8 Boxplots showing the interquartile ranges of intelligibility scores (in % correct) for English listeners on English target sentence recognition. The two panels show results at different signal-to-noise ratios (SNRs). The abscissa indicates masker type ordered according to decreasing linguistic similarity to the target. The mean percent correct score is given at the bottom of each plot. (From Brouwer et al. 2012, *The Journal of the Acoustical Society of America*, with permission.)

In general, the patterns of results were interpreted as being consistent with the linguistic similarity hypothesis. For these English-speaking listeners and meaningful English targets, performance was poorer when the masking speech was also in English than when the masking speech was in Dutch, a language that was not intelligible to these listeners. The differences due to language were more pronounced at the lower T/M. Furthermore, the “meaningful” English masker sentences produced more masking than did the semantically “anomalous” English sentences. These differences in performance due to linguistic factors occurred even in the absence of reliable differences in “general auditory distance” (low-level segregation cues) between stimuli. This idea of IM increasing in proportion to linguistic similarity was further supported by Calandruccio et al. (2013), who measured the masking of English target/masker speech for English-speaking listeners and compared it to that found for two maskers in languages unfamiliar to the subjects: Dutch and Mandarin. Furthermore, they attempted to control acoustically for differences in EM across languages so that the changes in performance that were found could then be attributed to IM. Their results indicated that comprehensible English was the most effective masker of English while Dutch maskers, which were judged to be more similar linguistically to English than were Mandarin maskers, produced more masking than Mandarin even though both the Dutch and Mandarin maskers were unintelligible. All three languages produced more masking than did a speech-spectrum-shaped noise-masker control.

Although qualitative differences between conditions can readily be specified, quantifying the degree of linguistic similarity may prove to be as challenging as quantifying the degree of IM in general. Furthermore, not all of the SOS masking results support the linguistic similarity hypothesis. The Kidd et al. (2008b) study mentioned in Sect. 4.3.3 that used an adaptation of Broadbent's (1952b) "every other word" paradigm found no significant difference between masker speech that was syntactically correct versus the same speech that was not syntactically correct (presented in random word order). The target speech was also syntactically correct short sentences. A logical extrapolation of the linguistic similarity hypothesis discussed earlier in the preceding paragraph would seem to predict greater masking for the more similar masker; that is, the syntactically correct masker. However, the target sentences used by Kidd and colleagues, while syntactically correct, were low in semantic value and perhaps for that reason differences due to masker syntax were not apparent. Furthermore, although this method eliminates EM as a factor, the linguistic structure—as noted by Broadbent (1958)—may be so different than normal communication that it invokes a different form of processing than occurs in natural speech, perhaps reducing the effects of linguistic similarity that would otherwise occur.

To summarize, the available evidence suggests that predictability and linguistic similarity may exert a strong influence on the outcome of SOS masking experiments. However, disentangling linguistic effects from other factors, in particular low-level segregation cues or high-level selective attention, may be challenging and depends on the interactions of many variables such as the means of target source designation, the speech corpora used, and the specific methods that are employed. The extent to which linguistic factors govern performance in natural listening environments remains an intriguing question, with the answer likely to depend on obtaining a better understanding of the role of context and predictability in realistic sound fields.

4.4 Models of Binaural Analysis Applied to SOS Masking

As noted in Sect. 4.2, Cherry (1953) identified several factors that could affect human performance in solving the cocktail party problem. Of those factors, the spatial separation of sound sources subsequently received the greatest attention in the literature, and this attention helped to inspire the development and testing of models of the processes underlying spatial release from masking. The efforts to model binaural factors in SOS masking largely have been limited to extensions of the binaural models that have been developed to explain tone-in-noise and speech-in-noise stimulus configurations. Thus, although the speech masker produces a complex pattern of spectrotemporal overlap with a speech target, the underlying mechanism limiting performance is assumed to be energetic masking. The lack of explicit modeling applied to the issues specific to SOS masking (e.g., linguistic and cognitive factors influencing IM) likely is due, at least in part, to the multiplicity and complexity of the factors involved. Although it may be possible to construct experiments to isolate and control some of these factors, incorporating all of these influences—and their interactions—into a comprehensive model of binaural analysis is a daunting task.

In the following paragraphs, the work to date is summarized, starting with the traditional waveform-based models of EM as developed originally for detecting tones in noise, followed by a discussion of the specializations that were incorporated to extend these models to predicting speech intelligibility in noise. A brief presentation of recent work is then provided that considers the importance of the significant spectrotemporal fluctuations found in speech masked by speech. None of these existing models explicitly account for the role of IM, but by comparing predictions of models that include as many of the factors as currently may be described, it is then possible to estimate the masking that is unaccounted for and to begin to develop new models that may be more comprehensive.

The earliest binaural models were based solely on differences in the interaural values of target and masker waveforms. Stimulated by the postulate from Jeffress (1948) of a network of coincidence detectors that were sensitive to interaural time delay/difference (ITD) in the binaural stimulus, Webster (1951) suggested that ITD might be the basis for binaural advantages in detection of tones in noise (i.e., MLDs). This concept received notable support from the findings of Jeffress and colleagues (1956), and it remains a viable hypothesis about the mechanism underlying binaural advantages for detection. Another early model devised to account for binaural detection advantages was proposed by Durlach (1963) and was termed the "equalization-cancellation (EC) model." Put simply, the EC model postulated a binaural equalization of the masker using interaural time and level compensations followed by a (partial) cancellation of masker energy resulting in an improved target-to-masker ratio in the internal representation of the stimulus. Even today, these two models, or variations of these two models, form the bases for most explanations of binaural masking release and there continues to be active discussion and debate about the possible physiological mechanisms that might implement their processing.

These two models, and the modifications proposed to accommodate variations in model parameters, have evolved over the decades. Initially, work focused on tone-in-noise masking experiments with the goal of accounting for variations in parameters such as frequency and duration, and eventually the interaural parameters, of the target tone. Similar studies of how detection thresholds depended on the parameters of the Gaussian masking noise, including level, center frequency and bandwidth, and the interaural difference parameters (e.g., time delay, phase, level, and their interactions) contributed to the refinement of these models. A summary of much of this early work may be found in Colburn and Durlach (1978).

As was the case with SOS masking in general, the early models that attempted to account for the release from masking of speech resulting from interaural differences in target and masker focused on the case of speech masked by noise and assumed that the masking that occurred was predominantly EM. This view of binaural masking release for speech found considerable support from the work of Levitt and Rabiner (1967a, b), who combined the known frequency dependence of the MLD with Articulation Index (AI) theory (French and Steinberg 1947) to successfully predict both the improvements in masked speech detection and recognition scores for different interaural parameters for target speech masked by noise. The empirical manipulations tested by Levitt and Rabiner involved reversing the interaural phase or delaying the waveform of target speech relative to the masking noise and doing so for various frequency regions. The binaural gain in intelligibility for the independently contributing bands of the AI was assumed to follow directly from the magnitude of the MLD for that frequency band. The success of this approach also was extended to the case of speech masked by a speech spectrum–shaped noise in a free-field environment by Zurek (1993), who accounted for the effects of head shadow in addition to the binaural analysis underlying the MLD measured under earphones. The maximum benefit predicted by Zurek's model was 8–10 dB, divided roughly equally between interaural differences in timing (MLD) and level (head shadow). Zurek's work provided a very good description of the spatial dependence of thresholds on the angle of the target speech and the angle of the masking noise. Performance with monaural listening alone was also considered. Overall, this work gave excellent support to the idea that, for these noise-masker cases, frequency bands were processed independently and combined to exploit the signal-to-noise advantages that were available in each band. In Levitt and Rabiner (1967a, b) and Zurek (1993) the underlying mechanism responsible for the binaural advantages found empirically was not specified but was assumed to be the same as that producing the MLDs on which the model predictions were based.

It is notable that all of the modeling discussed to this point was based on interference in speech reception caused by noise, which differs from the interference caused by speech in multiple ways. In terms of acoustic differences, speech-masker envelopes have greater fluctuations than steady-state noise maskers (even narrowband filtered maskers), and there are times when the level of a speech masker may be negligible within one or more frequency bands (e.g., during gaps between the words comprising sentences or in lower-level phonemes such as voiceless consonants). One way to address this opportunity to "listen in the dips" of the masker envelope is to analyze binaural performance using a weighted combination of signal-to-noise ratios within individual time–frequency slices (T–F units; cf. Brungart and Iyer 2012; Best et al. 2015). This approach was used to model monaural speech intelligibility by Rhebergen and colleagues (2006) for both amplitude-modulated noise and speech maskers.

This time-dependent processing approach was extended to binaural models in a series of studies by a variety of investigators. Beutelmann et al. (2010) extended their binaural modeling of speech in wideband noise (see also Beutelmann et al. 2009) by allowing processing parameters to vary across time. Basically, they considered processing in separate T–F slices so that they could use appropriate parameters for maskers that were modulated in time. Their modeling was quite successful in comparing the different types of maskers. They concluded that listeners were able to process the stimulus binaurally according to separate T–F units, which supported the proposition that binaural model parameters could vary accordingly. This time-dependent EC processing was also suggested and used by Wan et al. (2010, 2014) to model the case of multiple speech maskers. They

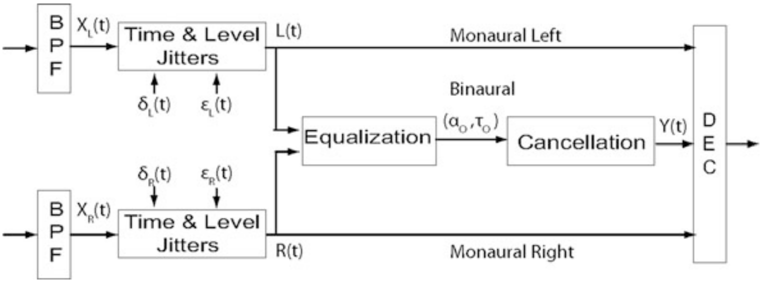

Fig. 4.9 Equalization–cancellation model of Durlach (1963) modified to include time-varying jitter. The *leftmost boxes* indicate the bandpass filtering stage (BPF) and the added time and level "jitter" for the left and right monaural channels. The binaural processing stages of equalization and cancellation are shown in *center boxes* followed by a decision mechanism (DEC). In the short-time EC (STEC) model used here, the equalization parameters α_O and T_O are adjusted to optimize cancellation within each time window. (From Wan et al. 2014, *The Journal of the Acoustical Society of America*, with permission.)

reasoned that independent speech-masker sources dominate in different time and frequency intervals, and so processing that was tuned to the dominant source would allow more efficient cancellation. Wan et al. (2014) demonstrated that many (but not all) of the spatial attributes of speech discrimination in the presence of multiple-speech maskers can be described with this type of binaural model. All of these model variations are based on extensions of the EC model. The basic processing implemented by these models is illustrated schematically in Fig. 4.9.

The inputs to the model are the acoustic waveforms arriving at the left and right ears. Each of these waveforms is passed through a bank of contiguous bandpass filters. The signals are represented in both the binaural pathway and the two monaural pathways and are corrupted by time-varying "jitter" in time and amplitude. These values are applied independently in each frequency channel and the equalization and cancellation process occurs in each time-frequency unit independently. A 20-ms sliding time window that is rectangular in shape is applied with an overlap between adjacent time windows of 10 ms.

These binaural models applied to multiple speech sources have not yet been modified to explicitly include IM. When target and masker speech sources are co-located there are no spatial cues to separate masker and target and, depending on the other source separation cues available, source confusions may occur resulting in a significant amount of IM. However, when speech interferers are spatially separated from the target, confusions about whether the target words come from the target source direction or from the masker source direction are greatly reduced, which in turn reduces source confusions and IM. This is illustrated for the case of two speech maskers in Fig. 4.10, which shows the results of applying the short-time EC (STEC) model to conditions with independent maskers on both sides as a function of separation of the maskers from the centered target.

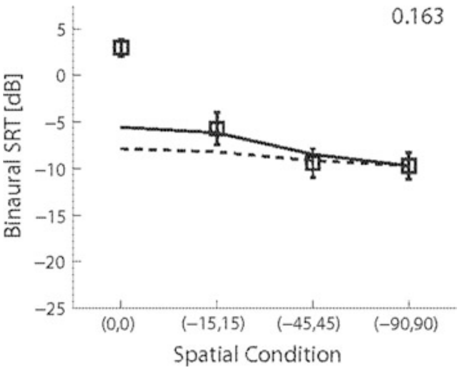

Fig. 4.10 Simulated and measured binaural speech reception thresholds (SRTs) as a function of spatial separation of two speech maskers from the target talker at 0° azimuth. Symbols are the measurements from Marrone and colleagues (2008), and the error bar is one standard error. Predicted values are connected by solid lines (short-term EC model) and dashed lines (steady-state EC model). The number in the upper right corner of the plot gives the value of the Speech Intelligibility Index criterion, which was chosen to match STEC prediction and data in the (–90°, +90°) condition. (From Wan et al. 2014, *The Journal of the Acoustical Society of America*, with permission.)

More specifically, this figure shows the obtained (Marrone et al. 2008) and predicted speech-reception thresholds for conditions in which the speech target was presented from straight ahead of the listener (0° azimuth) and two independent speech maskers were presented from locations symmetrically separated from the target. The predictions of the STEC model are connected by the solid lines. The dashed lines connect predictions from the steady-state EC (SSEC) model without selectivity with respect to time (from Wan et al. 2010). Note that the predicted values were fit to the threshold for the widely separated (–90°, +90°) masker condition (by adjusting a constant in the model). The thresholds for ±15° and ±45° angular separation are captured relatively well by the model, reflecting the ability of the model to describe the spatial aspects of the release from masking. The lack of IM in the model is illustrated by the poor fit for the co-located case where the amount of masking is almost ten decibels greater than in the (–15°, +15°) separation case. This large increase in masking when the sources are co-located is consistent with significant confusions between the speech masker and the speech target. Because of the strong similarity between the targets and maskers (both were CRM sentences), performance in some cases was no better than would be expected simply from attending to the more intense (louder) talker. The resulting threshold of about 4 dB in the co-located condition is consistent with the idea of choosing the target on the basis of its higher level. This illustrates the need to incorporate IM in binaural models of SOS masking. Even when the T/Ms are sufficient to extract target information in a reasonable subset of T–F slices, the difficulty of perceiving/recognizing which samples contain information about the target itself leads to errors.

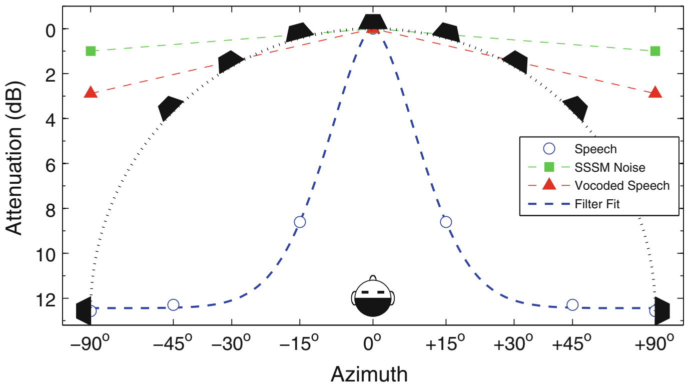

Fig. 4.11 Spatial tuning schematic showing attenuation of off-axis sources due to an attentional filter operating on interaural differences caused by different source locations (azimuth in degrees). The filter is oriented symmetrically around the point corresponding to 0° azimuth (directly in front of the simulated listener) and 0 dB attenuation. The amount of attenuation is assumed to be equal to the spatial release from masking (SRM) from human speech recognition experiments (Marrone et al. 2008) plotted in decibels and the roex filter function is a least-squares fit to the data. Overlaid on the spatial filter plot is a second schematic representing the location and arrangement of the listener and loudspeakers in a typical speech recognition experiment as used to measure SRM. The open circles on the filter function are group mean results for two independent speech maskers; the squares are data obtained using the same subjects and procedures but for two independent speech-shaped speech envelope–modulated noise maskers (also from Marrone et al. 2008) and the triangles are from eight-channel noise-vocoded speech maskers separated by ±600 µs ITDs under earphones, one to the extreme left and the other to the extreme right

The net result of binaural processing may be viewed conceptually as implementing a "spatial filter" that attenuates sounds along the horizontal (azimuthal) dimension (or potentially other spatial dimensions). This perspective was proposed by Arbogast and Kidd (2000), who used a variation of the "probe-signal" method to obtain accuracy and response-time measures that exhibited "tuning" in azimuth in sound field conditions high in IM. The basic idea is illustrated schematically in Fig. 4.11.

In this illustration, a listener is located in the center of a semicircle of loudspeakers from which target and masker sound sources may be presented. This physical layout is illustrated by the sketch of loudspeakers along the dotted-line semicircle; this sketch is not related to the labeling of the axes, which is used for the empirical data plotted as open circles, green squares, and red triangles. These data are all for the case with the target source at 0° azimuth and with interfering sources symmetrically located at the azimuths where the data are plotted (and so as to appear filter-like are mirrored in the two hemispheres). The ordinate denotes the attenuation by the hypothetical "spatial filter." The filter is shown by the smoothed function that peaks at 0 dB/0° azimuth and attenuates sounds off-axis symmetrically around the target location. The arbitrarily chosen filter function has the "rounded exponential" shape often used to represent filtering in the auditory system along the frequency dimension. The values for the filter parameters were obtained from least-squares fits to SOS masking data from Marrone et al. (2008) and those thresholds are plotted as open circles along the fitted function. In the Marrone and colleagues experiment, there were two independent speech maskers that, when separated, were located symmetrically around the target location (one to either side). Conceptually, the attenuation of the filter is proportional to the amount of SRM measured in speech identification experiments; in this case, the data from Marrone and colleagues were obtained using the CRM materials/procedures. The maximum attenuation—equal to the maximum SRM—is about 12 dB. Two other sets of thresholds are also plotted representing results obtained with maskers producing lower levels of IM: one set obtained using two independent speech-shaped speech-modulated noises (also from Marrone et al. 2008) and the other obtained using "distorted" but intelligible eight-channel noise-vocoded speech (Best et al. 2012) separated by ITDs (±600 μs). These thresholds emphasize the point that the amount of "attenuation" of masking (i.e., masking release) that is possible by the attention-based spatial filter is limited by the amount of IM that is present.

4.5 Summary

Early in the history of study of SOS masking, the potential influence of nonperipheral mechanisms was considered by leading auditory and speech scientists. Although the empirical work available at the time often did not support drawing strong conclusions about peripheral versus central components of masking, it is clear from Miller's (1947) work that factors such as the intelligibility of competing speech or the uncertainty of the listening situation (e.g., "improbable vocal effects") motivated the design of his experiments. In his famous article that coined the term "cocktail party problem," Cherry (1953) elaborated several factors that human observers could use to solve the SOS masking problem, some of which fundamentally involved significant processing beyond the auditory periphery. The evidence he presented indicating that listeners perceived only certain attributes of unattended sounds presented to one ear while engaged in the recognition of speech in the contralateral attended ear demonstrated the existence of central effects and encouraged decades of study of the binaural processing of sounds. Perhaps as importantly, though, sophisticated higher-level mechanisms were implicated in Cherry's observations about the importance of the transition probabilities inherent to normal speech. The idea that aspects of natural speech communication—e.g., turn-taking in conversation, sources joining or leaving the auditory "scene," the unpredictable mixing of speech and nonspeech competition—involve the exploitation of predictability (e.g., that "a vast store of probabilities allows…noise or disturbances to be combatted") is an underappreciated observation that has found increasing relevance as tools for studying perception in natural sound fields have been developed. Unambiguous evidence for SOS masking that could not be accounted for by peripheral overlap of sounds was provided by Broadbent (1952a, b), who later argued convincingly for the important role of central factors. The importance of these factors in solving SOS masking problems led Carhart et al. (1969a, b) to propose a separate category of masking, termed perceptual masking, to account for otherwise unexplained results.

Numerous examples of the influence of what is now termed IM may be found in the modern-day literature. That is, reports of large masking effects beyond those that can be attributed to EM are commonplace and variables that lead to perceptual segregation of sources—without accompanying reductions in EM—have been found to produce significant release from masking in SOS conditions. In many instances, clear demonstrations of the role of linguistic variables in producing, or releasing, SOS masking have been reported that cannot be attributed to changes in the peripheral overlap of sounds. Historically, Theories explaining the masking of speech paralleled those of masking in general. Although such theories provide a good account of conditions dominated by EM, they are less successful in accounting for conditions dominated by IM. With respect to the causes of IM, even early work (e.g., Broadbent, 1952b) implicated the important role of failures of selective attention. However, the complex interaction of attention and memory and, particularly, the complications inherent to the comprehension of multiple streams of speech, caution against assigning IM to simple categories or attributing its effects exclusively to any single mechanism or process (cf. Watson 2005; Kidd et al. 2008a; Mattys et al. 2012).

The benefits of interaural differences between target and masker have been the subject of considerable modeling efforts over the years. These models originally were intended to account for the empirical findings from experiments in which tones or speech were masked by noise. As these models developed over time they were adapted to account for some of the spectrotemporal fluctuations of speech maskers and thus allowed the model parameters to vary across frequency channels or even small T–F units. The underlying physiological mechanism that could achieve this fine-grained parameter variation—whether it would respond solely to low-level stimulus features common to T–F units from the same source or would require some higher-level influence—presently is unclear. However, the underlying assumptions of even these refinements of traditional models of binaural analysis do not adequately provide for IM, as discussed in Sect. 4.4 The assumption that only the channels (or T–F units) containing target energy govern performance—and all other channels/units may be disregarded—does not provide for the deleterious effects of those units that are dominated by masker energy. It is clear from studies of SOS masking, however, that humans cannot disregard the nontarget energy in such units that may exert a profound influence on overall performance. Thus, what often could matter the most is not improving the T/M in units with significant target energy as much as it is minimizing masker energy in units where it is dominant. Current modeling approaches may be adapted to account for such circumstances (e.g., the EC model could null locations containing high-IM sources) but the higher-level processes that come into play with such putative mechanisms are quite complex.

Acknowledgements The authors are indebted to Christine Mason for her comments on this chapter and for her assistance with its preparation. Thanks also to Elin Roverud and Jing Mi for providing comments on an earlier version and to the members of the Psychoacoustics Laboratory, Sargent College graduate seminar SLH 810, and Binaural Group for many insightful discussions of these topics. We are also grateful to those authors who generously allowed their figures to be reprinted here and acknowledge the support of the National Institutes of Health/National Institute on Deafness and Other Communication Disorders and Air Force Office of Scientific Research for portions of the research described here.

Compliance with Ethic Requirements

Gerald Kidd, Jr. declares that he has no conflict of interest.

H. Steven Colburn declares that he has no conflict of interest.

References

ANSI (American National Standards Institute). (1997). *American National Standard: Methods for calculation of the speech intelligibility index*. Melville, NY: Acoustical Society of America.

Arbogast, T. L., & Kidd, G., Jr. (2000). Evidence for spatial tuning in informational masking using the probe-signal method. *The Journal of the Acoustical Society of America, 108*(4), 1803–1810.

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

Başkent, D. & Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *The Journal of the Acoustical Society of America, 139*(3), EL51–EL56.

Beranek, L. (1947). Design of speech communication systems. *Proceedings of the Institute of Radio Engineers, 35*(9), 880–890.

Best, V., Marrone, N., Mason, C. R., & Kidd, G., Jr. (2012). The influence of non-spatial factors on measures of spatial release from masking. *The Journal of the Acoustical Society of America, 131*(4), 3103–3110.

Best, V., Mason, C. R., Kidd, G. Jr., Iyer, N., & Brungart, D. S. (2015). Better ear glimpsing efficiency in hearing-impaired listeners. *The Journal of the Acoustical Society of America, 137*(2), EL213–EL219.

Best, V., Mason, C. R., & Kidd, G., Jr. (2011). Spatial release from masking as a function of the temporal overlap of competing maskers. *The Journal of the Acoustical Society of America, 129*(3), 1616–1625.

Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *The Journal of the Association for Research in Otolaryngology, 8,* 294–304.

Beutelmann, R., Brand, T., & Kollmeier, B. (2009). Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences. *The Journal of the Acoustical Society of America, 126*(3), 1359–1368.

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America, 127*(4), 2479–2497.

Broadbent, D. E. (1952a). Listening to one of two synchronous messages. *The Journal of Experimental Psychology, 44*(1), 51–55.

Broadbent, D. E. (1952b). Failures of attention in selective listening. *The Journal of Experimental Psychology, 44*(6), 428–433.

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

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics, 77*(5), 1465–1487.

Brouwer, S., Van Engen, K., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America, 131*(2), 1449–1464.

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

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America, 120*(6), 4007–4018.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2009). Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. *The Journal of the Acoustical Society of America, 125*(6), 4006–4022.

Brungart, D. S., & Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America, 132*(4), 545–2556.

Brungart, D. S., & Simpson, B. D. (2004). Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. *The Journal of the Acoustical Society of America, 115*(1), 301–310.

Buss, E., Grose, J., & Hall, J. W., III. (2016). Effect of response context and masker type on word recognition. *The Journal of the Acoustical Society of America, 140*(2), 968–977.

Calandruccio, L., Brouwer, S., Van Engen, K., Dhar, S., & Bradlow, A. (2013). Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. *American Journal of Audiology, 22*(1), 157–164.

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America, 128*(2), 860–869.

Calandruccio, L., Leibold, L. J., & Buss, E. (2016). Linguistic masking release in school-age children and adults. *American Journal of Audiology, 25,* 34–40.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969a). Release from multiple maskers: Effects of interaural time disparities. *The Journal of the Acoustical Society of America, 45*(2), 411–418.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969b). Perceptual masking in multiple sound backgrounds. *The Journal of the Acoustical Society of America, 45*(3), 694–703.

Carlile, S. (2014). Active listening: Speech intelligibility in noisy environments. *Acoustics Australia, 42,* 98–104.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979.

Clayton, K. K., Swaminathan, J., Yazdanbakhsh, A., Patel, A. D., & Kidd, G., Jr. (2016). Exectutive function, visual attention and the cocktail party problem in musicians and non-musicians. *PLoS ONE, 11*(7), e0157638.

Colburn, H. S., & Durlach, N. I. (1978). Models of binaural interaction. In E. Carterette & M. Friedman (Eds.), *Handbook of perception: Hearing* (Vol. 4, pp. 467–518). New York: Academic Press.

Cooke, M., Lecumberri, M. G., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America, 123*(1), 414–427.

Dirks, D. D., & Bower, D. R. (1969). Masking effects of speech competing messages. *Journal of Speech and Hearing Research, 12*(2), 229–245.

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America, 35*(8), 1206–1218.

Egan, J. P., & Wiener, F. M. (1946). On the intelligibility of bands of speech in noise. *The Journal of the Acoustical Society of America, 18*(2), 435–441.

Ezzatian, P., Avivi, M., & Schneider, B. A. (2010). Do nonnative listeners benefit as much as native listeners from spatial cues that release speech from masking? *Speech Communication, 52*(11), 919–929.

Fletcher, H. (1940). Auditory patterns. *Review of Modern Physics, 12*(1), 47–65.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America, 19*(1), 90–119.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America, 109*(5), 2112–2122.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masker talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America, 115*(5), 2246–2256.

Freyman, R. L., Helfer, K. S., & Balakrishnan, U. (2007). Variability and uncertainty in masking by competing speech. *The Journal of the Acoustical Society of America, 121*(2), 1040–1046.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America, 106* (6), 3578–3588.

Helfer, K. S., & Jesse, A. (2015). Lexical influences on competing speech perception in younger, middle-aged, and older adults. *The Journal of the Acoustical Society of America, 138*(1), 363–376.

Hirsh, I. J. (1948). The influence of interaural phase on interaural summation and inhibition. *The Journal of the Acoustical Society of America, 20*(4), 536–544.

Hygge, S., Ronnberg, J., Larsby, B., & Arlinger, S. (1992). 'Normal hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech and Hearing Research, 35*(1), 208–215.

Iyer, N., Brungart, D. S., & Simpson, B. D. (2010). Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task. *The Journal of the Acoustical Society of America, 128*(5), 2998–3010.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology, 41*(1), 35–39.

Jeffress, L. A., Blodgett, H. C., Sandel, T. T., & Wood, C. L. III. (1956). Masking of tonal signals. *The Journal of the Acoustical Society of America, 28*(3), 416–426.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., et al. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science, 24,* 1995–2004.

Kalikow, D. N., Stevens, K. N., & Elliot, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61*(5), 1337–1351.

Kellogg, E. W. (1939). Reversed speech. *The Journal of the Acoustical Society of America, 10*(4), 324–326.

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

Kidd, G., Jr., Best, V., & Mason, C. R. (2008a). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America, 124*(6), 3793–3802.

Kidd, G., Jr., Mason, C. R., & Best, V. (2014). The role of syntax in maintaining the integrity of streams of speech. *The Journal of the Acoustical Society of America, 135*(2), 766–777.

Kidd, G., Jr., Mason, C. R., Best, V., & Marrone, N. L. (2010). Stimulus factors influencing spatial release from speech on speech masking. *The Journal of the Acoustical Society of America, 128*(4), 1965–1978.

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008b). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–190). New York: Springer Science + Business Media.

Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., et al. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America, 140*(1), 132–144.

Levitt, H., & Rabiner, L. R. (1967a). Binaural release from masking for speech and gain in intelligibility. *The Journal of the Acoustical Society of America, 42*(3), 601–608.

Levitt, H., & Rabiner, L. R. (1967b). Predicting binaural gain in intelligibility and release from masking for speech. *The Journal of the Acoustical Society of America, 42*(4), 820–829.

Licklider, J. C. R. (1948). The influence of interaural phase relations upon the masking of speech by white noise. *The Journal of the Acoustical Society of America, 20*(2), 150–159.

Marrone, N. L., Mason, C. R., & Kidd, G., Jr. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *The Journal of the Acoustical Society of America, 124*(2), 1146–1158.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin, 44*(2), 105–129.

Newman, R. (2009). Infants' listening in multitalker environments: Effect of the number of background talkers. *Attention, Perception, & Psychophysics, 71*(4), 822–836.

Newman, R. S., Morini, G., Ahsan, F., & Kidd, G., Jr. (2015). Linguistically-based informational masking in preschool children. *The Journal of the Acoustical Society of America, 138*(1), EL93–EL98.

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *The Journal of the Acoustical Society of America, 118*(3), 1274–1277.

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America, 120*(6), 3988–3997.

Samson, F., & Johnsrude, I. S. (2016). Effects of a consistent target or masker voice on target speech intelligibility in two- and three-talker mixtures. *The Journal of the Acoustical Society of America, 139*(3), 1037–1046.

Schubert, E. D., & Schultz, M. C. (1962). Some aspects of binaural signal selection. *The Journal of the Acoustical Society of America, 34*(6), 844–849.

Schubotz, W., Brand, T., Kollmeier, B., & Ewert, S. D. (2016). Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. *The Journal of the Acoustical Society of America, 140*(1), 524–540.

Speaks, C., & Jerger, J. (1965). Method for measurement of speech identification. *Journal of Speech and Hearing Research, 8*(2), 185–194.

Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V. A., et al. (2015). Musical training and the cocktail party problem. *Scientific Reports, 5*, 1–10, No. 11628.

Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., et al. (2013). Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *The Journal of the Acoustical Society of America, 134*(4), 3039–3056.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America, 121*(1), 519–526.

Wan, R., Durlach, N. I., & Colburn, H. S. (2010). Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *The Journal of the Acoustical Society of America, 128*(6), 3678–3690.

Wan, R., Durlach, N. I., & Colburn, H. S. (2014). Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments. *The Journal of the Acoustical Society of America, 136*(2), 768–776.

Watson, C. S. (2005). Some comments on informational masking. *Acta Acustica united with Acustica, 91*(3), 502–512.

Webster, F. A. (1951). The influence of interaural phase on masked thresholds. I: The role of interaural time-deviation. *The Journal of the Acoustical Society of America, 23*(4), 452–462.

Webster, J. C. (1983). Applied research on competing messages. In J. V. Tobias & E. D. Schubert (Eds.), *Hearing research and theory* (Vol. 2, pp. 93–123). New York: Academic Press.

Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 255–276). Boston: Allyn and Bacon.