ACTIVE LISTENING: SPEECH INTELLIGIBILITY IN NOISY ENVIRONMENTS

Simon Carlile
School of Medical Sciences and The Bosch Institute, University of Sydney, Sydney, Australia
simonc@physiol.usyd.edu.au

Attention plays a central role in the problem of informational masking, a key element of the cocktail party problem, itself described more than 60 years ago. This review considers recent research that has illuminated how attention operates, not only on the auditory objects of perception, but on the processes of grouping and streaming that give rise to those objects. Competition between endogenous and exogenous attention, the acoustic and informational separability of the objects making up an auditory scene and their interaction with the task requirement of the listener all paint a picture of a complex heterarchy of functions.

INTRODUCTION

Here we briefly review some experiments that have contributed to our understanding about listening in noisy environments: The so-called "cocktail party problem" (CPP). Concurrent talkers at a cocktail party will mask each other not only in terms of their energetic overlap but in the way in which a listener can extract meaning from the ensemble of sounds. Central to this process is how the bottom-up mechanisms of grouping segregate the acoustic elements associated with each sound and then how streaming these groups over time contribute to the formation of the auditory objects of our perception. Over the last decade, research has increasingly pointed to the important role of attention in overcoming informational masking. We will consider some of the evidence that attention not only acts on auditory objects but can modulate some of the "primitive" processes of grouping and stream formation. These advances have significantly shifted the focus from "hearing in noise" to listening as an active cognitive process and complement the development of ideas more generally in speech recognition and semantic processing [1].

ENERGETIC AND INFORMATIONAL MASKING IN THE COCKTAIL PARTY PROBLEM

When sounds are occurring concurrently (or even in close temporal proximity) the perception of any one sound can be interfered with by the other sounds. In general, this is referred to as masking and its study has a long history in hearing research. Over the last couple of decades or so, masking has come to be classified as energetic or informational masking. Energetic masking is probably the most well understood (although this understanding is incomplete on a number of levels [2]): When one sound is sufficiently loud that it dominates the output of a processing channel, it will mask a second quieter sound as it is unable to influence the output of the channel. Often termed peripheral masking, this could be conceived of as the motion of the basilar membrane being dominated by a high intensity sound so that the target sound makes no appreciable impact on the output of the cochlea. Psychoacoustically, this sort of phenomenon has been modelled as the output of a critical band energy detector and the ability to predict the presence of a target [3, 4].

Informational masking is most often described as the component of masking that cannot be accounted for by energetic masking. On the one hand this is a simple and parsimonious explanation but on the other, it is not very helpful in understanding the sources of such masking. What has become clearer over the last decade or so is that informational masking can involve interactions at many stages of processing. These include the segregation of spectral components associated with a particular sound, the perceptual grouping and streaming of those components to form an auditory object, spatial and non-spatial attentional control, working memory and other aspects of executive and cognitive functions. The study of informational masking goes back to the mid-1970s although hints as to its effects can be seen in the analysis and discussions of many papers leading up to that time. A splendid and detailed review of the history and early work on informational masking can be found in [5]. That review also considers in detail the work involving multi-tone complexes and the respective roles of target and masker uncertainty in generating informational masking. In this short review we will be more concerned with informational masking as it applies to speech masking and its application to understanding the cocktail party problem.

It has long been recognised that the segregation of a talker of interest from other background talkers is a challenging task. Colin Cherry coined the term the "cocktail party problem" in his seminal paper in 1953 [6]. In a break with the dominant, signal detection based research themes of the time, his paper was focussed on the roles of selective attention in speech understanding, the "statistics of language", voice characteristics, the effects of temporal binaural delays and the costs and time course of switching attention. He makes a very clear distinction between the sorts of perception that are studied using simple stimuli used to study energetic masking and the "acts of recognition and discrimination" that underlie understanding speech at the cocktail party. In this most prescient of papers, Cherry foreshadows much of the work that has now come to dominate research into informational masking and auditory scene analysis as it applies to speech intelligibility in noisy environments. Despite these penetrating insights, most of the work over the last half of the 20th Century continued to be dominated by bottom-up approaches focussed more on energetic masking effects and binaural processes resulting in masking release (see [7, 8] for excellent reviews of much of this work). Notably though, Bronkhorst describes how others had noted that speech interference of speech understanding seemed to amount to more than the algebraic sum of the spectral energy. Indeed, as early as 1969, Carhart and colleagues had referred to this as “perceptual masking” or “cognitive interference” [9].

Right at the turn of the century, Richard Freyman and colleagues reported an experiment that demonstrated that differences in the perceived locations of a target and maskers (as opposed to actual physical differences in location) produced significant unmasking for speech but not for noise [10]. Such a result was not amenable to a simple bottom-up explanation of energetic masking – Freyman appropriated the term “informational masking” and this work led to a large number of studies which have systematically looked at what was driving this speech on speech masking. When the target and the masker both originated from the front of the listener the masking was higher when the masker was speech than when it was noise, particularly at unfavourable signal-to-noise ratios (SNRs) [10]. This indicated that the masker talker was contributing a level of informational masking over and above the energetic masking associated with its SNR. Informational masking was reported to be its highest with two competing talkers and with further increases in the number of the talkers the mixture becomes increasingly more dominated by energetic masking [11]. The exact number of talkers at which masking is maximised probably relates to the speech material and the nature of the talkers but does suggest a relatively small limit on the number of competing streams in informational masking. Similarities between the target voice and the competing talkers was also shown to markedly increase informational masking [12, 13] but again this effect did not increase from 2 to 3 competing maskers when corrected for SNR. Interestingly, listening monaurally to three talkers of the same gender as the target talker (i.e. high similarity) produced less masking than if one of the talkers was a different gender. This “odd sex” distractor effect indicated that the informational masking is not simply mediated by the extent of the similarity between the target and the maskers - a point we will return to later.

Varying the actual locations of target and maskers will result in changes in the relative levels of the targets and maskers in each ear. These level changes result from differences in the interactions of the sounds from each source with the head and pinna of the outer ear. Presumably, an advantage in hearing out the target of interest could simply result from attending to the ear with the most favourable SNR (see Figure 1): so called “better-ear” listening.

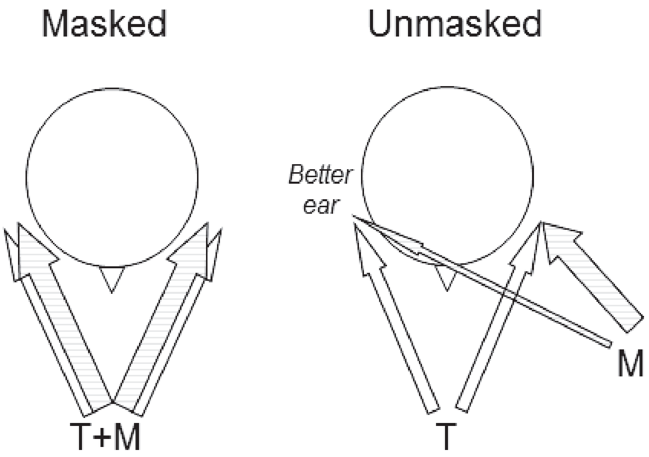

Figure 1. Spatial release from masking can be demonstrated by comparing the speech reception threshold (SRT) obtained with the target and the masker co-located (Masked - usually in front) with the SRT obtained when the masker[s] are moved to another location (Unmasked). The SNR increase at the "better ear" is indicative of how much of the masking release can be attributed to energetic unmasking.

To examine the effects of actual difference in location between target and maskers, Kidd and colleagues [14, 15] compared the speech reception thresholds (SRT) for a target and a masker collocated in front to the SRT obtained with the masker at 90º to the right. They used interleaved frequency channels of modulated noise band speech or band filtered noise to manipulate the energetic interactions between the target and the maskers. In summary, they found that the “better ear” effect could account for around 7 dB of unmasking when the masker was a noise but an average of 18 dB unmasking was found when the masker was another talker. This suggests that the spatial separation of targets and masker provided a much greater advantage for informational compared to energetic masking. In this experiment the modulated noise band speech would have produced target and masker voices that sound quite similar, producing quite high levels of informational masking.

Another strategy employed to “hear out” a talker of interest, particularly against a background of other talkers, is to take advantage of the amplitude modulation of the maskers to “glimpse” the target talker during the intervals where the SNR is favourable. Consonant recognition in modulated noise was found to be well predicted by the proportion of the target “glimpsed” over a -2 dB to +4 dB “local” or instantaneous SNR range [16]. In a clever binaural experiment, Brungart and Iyer [17] presented diotically over headphones the better ear glimpses available with a target in front and symmetrically placed maskers. Such a diotic paradigm maximised the SNR but eliminated the perception of difference in location of the target and maskers. They found that the glimpses, even though they only appeared transiently in one or the other ear, provided a significant unmasking. In that experiment the gender of the maskers and target were different so that the informational masking was relatively low (i.e. the listeners could already easily tell the talkers apart). By contrast, Glyde and colleagues [18] used the speech materials from the LiSN-N speech test [19] where the amount of informational masking could be varied. They found that speech intelligibility was significantly worse in the diotic glimpsing condition when compared with natural binaural listening and the magnitude of the difference was larger when there was more informational masking. Together these results suggest that the perception of the difference in location was adding significantly to the unmasking and that better ear glimpsing was effective mainly for energetic rather than informational masking.

Consistent with the above, a number of experiments have shown that informational masking is not about audibility. An analysis of the sorts of speech identification errors for speech masked by other talkers shows that, more often than not, the errors relate to a word spoken by a masker rather than a guessing error (e.g. [12, 14]). This shows that not only are the maskers audible but they are intelligible and it is their attribution to the target "stream" that is compromised [see also below]. Familiarity with the target talker (i.e. knowing who to listen for) provides an advantage [11, 12] as does knowing where to listen [20] or when to listen [21] (see also [22]) although the same does not appear to be the case for the maskers [23]. Both auditory, and visual cues about "where" and "when" to listen can be very effective, even in the absence of information about the target content [24].

Maskers in a foreign language also produce informational masking, although less so compared to maskers from the listener's native language, or second language in the case of bilinguals [25-27] (but see [28]). Again, this is masking that is over and above that produced by the energetic interactions between the sounds. Informational masking is also still present but somewhat reduced if the speech from the masker talker is reversed in time [28] but this may be complicated by the increased forward masking because of the phonetic structure of the speech used [25]. A most important point however, is that regardless of the extent of masking produced by these maskers, it appears that intelligibility by itself is not a requirement to produce some level of informational masking. This might suggest a quite different process compared to that discussed above where incorrect but intelligible words are attributed to the target talker.

A ROLE OF AUDITORY ATTENTION

Cueing "where" or "what/who" to listen for reduces informational masking indicating an important role for a top-down focus of attention towards the source or voice of interest. In fact, the word confusions discussed above, suggest that informational masking might be due to a failure of attention towards the voice of interest.

In general, attention is thought of as a process of biased competition involving (i) bottom-up (exogenous) attention driven by such qualities as salience and novelty and (ii) top-down (endogenous) attention driven by cognitive steering based on specific task requirements [29]. The ability to focus and report the talker in one ear in Cherry's experiment is a good example of endogenous attention while noticing the change in gender in the unattended ear represents brief exogenous control. The odd-sex distractor effect found by Brungart and colleagues [13] could represent a bottom-up driven change in the focus of attention that then manifests itself as a task related error in reporting the target conversation.

Returning to the original work of Cherry, listening dichotically with a different talker in each ear, largely eliminates energetic masking, or at least the energetic masking attributable to interactions on the basilar membrane. This would provide two highly segregated channels through which the listener can attend. Cherry (1953) reported that while attending to one ear and reporting on the information presented at that ear, the information in the other ear was completely forgotten to the extent that listeners were even unaware that the language of the talker had been changed or that the speech was reversed in time. Interestingly, some statistical properties of the masker talker were sufficiently salient as to be noticed, such as a change in the gender of the talker or the replacement of the talker with a 400 Hz tone. In experiments conducted nearly 50 years later, the listening task was made harder by introducing a masker talker in the same ear as the target [30]. Under these conditions, almost any masker in the other ear was able to produce substantial masking of the target talker. Furthermore, the strength of the contralateral interferer was related to the level of the speech like fluctuations in the spectral envelope - an effect that was thought to engage some form of "preattentive central auditory processing mechanism … that interferes with a listeners ability to segregate speech signals presented in the opposite ear" ([31], p301). There are also a range of listening conditions where the characteristics of the talkers in the unattended ear can direct attention to a target in the attended ear (see for e.g. [32, 33]). This suggests that some level of lexical processing is carried out, even in the absence of attention. These sorts of mechanisms might also explain the masking effects of reversed speech. Such a masker will have the same fluctuations as forward speech but is otherwise unintelligible. This is also consistent with Cherry's observation that significant changes in the unattended speech did come to the attention of the listener, presumably because of some inherent salience. We will return to the issue of what makes speech "speechy" in the context of informational masking later.

GROUPING AND STREAMING SOUND ELEMENTS INTO OBJECTS

Although our understanding of auditory attention is not as mature as say with visual attention, it has been argued that, in line with what is known about visual attention, attention is applied to a perceptual object [34, 35]. For instance, if we perceive an orange rolling across the table, the perceptual elements will include the edges and contours of the shape, the colours, the textures, the motion etc. While these are all encoded and to some extent processed separately, this collection of features are then bound together to form the percept of the object - the orange. In common parlance, an auditory object might be considered to be a particular source of sound - a talker of interest, an instrument in an ensemble or a specific environmental source such as a car. Object formation likely involves an interaction of processes that include the segregation and encoding of the many separate features that we use to distinguish between sounds, as well as an analysis of those features that enables a categorical association between a sound and some meaning [36].

So what are the relevant features for the auditory system? As a sound of interest usually occurs on a background of other sounds, at any point in time, the pattern of stimulation of the inner ear is a multiplexed representation of the sum total of the sounds. Most naturally occurring sounds are spectrally sparse which means that, unless there are very many sounds competing, despite their concurrency, a significant proportion of each sound is on average, not masked by other sounds. So the first challenge for the auditory system is identifying which elements in the pattern of stimulation relate to which sounds - this is the basic problem of auditory scene analysis (see [37] for a foundation work in this area and [38, 39] for relatively recent and quite accessible reviews).

In summary, the auditory system employs mechanisms that exploit the acoustic characteristics of physically sounding bodies to parse out the elements related to different concurrent sounds. For instance, all the acoustic elements that turn on or off at the same time are likely to come from a common source, as are those that are harmonically related or that modulate synchronously in amplitude. These are referred to as acoustic grouping cues, which work on relatively small time frames to segregate and then group the sonic elements of the different sounds [40].

The ability to link together or “stream” these small segments over longer time frames also relies on similar principles of plausibility. For instance sequential groups which come from the same spatial location, have the same or similar fundamental frequencies or spectral profiles or that are changing progressively are all likely to have come from the same source. Such groups are then perceptually linked together into a stream that becomes associated with the auditory object (e.g. [41]). Continuity of voice [12], location [42, 43], prosody and talker characteristics [41], amongst other things, facilitate the streaming of one talker against multiple background talkers. Moreover, over time, continuity also enhances spatial selectivity for a particular target stream [42] indicating that the effects of selective attention appear to build up over seconds. Reverberation has been shown to reduce speech intelligibility and is associated with a degradation in the temporal coding of the fine structure and, to a lesser extent, the envelope interaural time difference (ITD) cues to spatial location [44; 45, 46]. Indeed, the individual differences seen in such listening conditions with normally hearing listeners appear to be related to the fidelity with which the auditory brainstem encodes the periodic temporal structure of sounds [45, 46]. Assuming that the differences in the locations of the target and maskers are the important cues in maintaining the integrity of the target stream, then the fidelity of the location cues must be playing a role in maintaining the spatial continuity supporting attentional selectivity and streaming. For the spatial continuity to be effective, however, the spatial cues must also be encoded and transmitted with sufficient fidelity within the auditory nervous system. This may also provide some clues to the nature of the problems underlying the failure of the hearing impaired listener to solve the cocktail problem where the encoding of fine temporal structure is also compromised.

In the context of the cocktail party scenario, attention also needs to be switched from one talker to another in the course of conversational turn-taking: i.e. there is an intentional break in continuity in order to follow what the next talker is saying. This requires a switch in both non-spatial and spatial attention to a new voice at a new location. Using a dichotic listening paradigm Koch and colleagues [47] found a substantial cost of intentional switching of attention between ears - particularly in the context of reaction time and accuracy of a simple cognitive task applied to the information provided by the target talker. By varying the time between cueing and stimulus, their data also suggests that there is a substantial "inertia" in the auditory attention switching which does not seem to take advantage of early visual cueing for preparation. A recent study [48] examined a group of school age children with significant hearing difficulties in the classroom but no negative audiological findings, auditory processing disorder (APD) diagnosis or other generalised attentional disorder. Using a speeded syllable identification in a variant of a probe-signal task [49], these children were found to have a deficit in attentional reorientation in time. In trials where the target did not occur at the expected time, sensitivity to the target took several seconds to recover, several fold longer than matched controls. This increased inertia in attentional control would have made it very difficult for this group of children to follow the normal turn-taking in the conversation and is consistent with the observations of Koch et al (2011).

When speech streaming breaks down, either as a result of perturbation of the continuity cue or as a result of intentional switching, listeners are likely to attribute words spoken by the masker talkers to the target talker - the classic finding of informational masking. Again, the segregation of the acoustic elements, their perceptual grouping and recognition of the words is not the problem. It is their incorrect attribution to the target talker. Through this lens it is easy to understand how the similarity between concurrent talkers can have such a profound effect on the amount of informational masking. In the early studies discussed above, it was found that spatial separation, or even the perception of a difference in the locations of the target and the maskers, significantly decreased confusion errors (presumably as a result of improved streaming) and thereby produced a significant reduction in informational masking.

The idea that attention is applied to a perceptual object has the important consequence that processing of the object as a whole is in some way enhanced and not just a single feature or features. The advantage of spatial separation was seen even when attention was directed to a non-spatial feature like the timbre of the voice rather than location [50]. In another study [51] subjects were asked to attend to one of two competing speech streams based on their location or pitch. The continuity of the task-irrelevant (non-attended) feature was shown to still influence performance in an obligatory manner.

Recent work has also suggested that object and stream formation is not a simple hierarchical process that provides the objects for selection by attention. The demands of the task performed also have an effect on the formation of objects, particularly where the grouping between various acoustic elements is ambiguous. This is seen particularly in the interactions between the identification of "what" and "where" attributes of an object. A sound's location can be determined unambiguously by the acoustic cues at ears if those cues cover a wide range of frequencies. Grouping should determine which acoustic elements are associated and therefore contribute to the calculation of location and in turn, location is then used to stream the information from a particular source (see [52]). While this idea implies a simple hierarchical processing (grouping then localisation then streaming), experiments using ambiguous sound mixtures suggest more complex interactions. If an acoustic element could contribute to more than one object, the sound mixture is then ambiguous. The contribution a particular element makes to the spectro-temporal content of an object can depend on the focus of attention and the strength of the relative grouping and streaming cues. More surprisingly, however, if an element is not allocated to the attended (foreground) object, it is not necessarily allocated to the background object, that is it gets 'lost' [53]. Likewise, the relative contribution of an ambiguous sound element to the determination of "what" or "where" varies according to the task of the listener (judge "what" or judge "where") and demonstrates considerable individual differences [54]. On one hand, the locations of two sound sources whose components were spectro-temporally intermixed could be reliably estimated on differences in interaural time difference cues to location - that is they can be segregated and localised. On the other hand, spatial segregation and localisation was unable to support identification in the absence of other grouping cues [55].

While many of the experiments discussed above demonstrate the relative strength of the different grouping and streaming cues, in difficult listening situations, the segregation of a target talker is still very challenging. In summary, informational masking could result from (i) ambiguity in the sound mixture and a failure to properly segregate and group the spectral components associated with the target; (ii) disruption of continuity that supports successful sequential streaming of the grouped elements associated with the target of interest; (iii) an error in selecting the appropriate object or stream resulting from high levels of similarity between the cues available for continuity or to (iv) sustain selective attention on the appropriate stream i.e. where saliency in the masker drives the focus of attention away from the target. Of course there are many other factors that also come into play here such as semantic context [27, 56], working memory (e.g. [57-59]), visual cues such as lip reading (e.g. [60, 61]) etc.

The combination of well controlled psychophysical paradigms together with vital imaging (particularly MEG) have driven some spectacular advances in understanding the neural basis of the formation of auditory objects and streams (review: [62]; e.g. [63-66]) and the potential role of temporal coherence in the binding of features (review [67]). Likewise, great strides have been made in understanding the recruitment of auditory spatial and non-spatial attention systems (e.g. [68-70]) and the attentional modulation of activity at different cortical levels [66]. Unfortunately space limits more discussion of these fascinating issues although the interested reader could start with the selection of recent reviews and references above.

CONCLUDING REMARKS

Hearing science and audiology owes much to the fundamental and pioneering work of Harvey Fletcher and colleagues in the early decades of the 20th century. The development of the articulation index (AI) and later the speech transmission index (SII) are founded on the basic assumptions that the intelligibility of speech is related to the instantaneous signal-to-noise ratio within the critical bands up to 6 kHz to 8 kHz. This "bottom-up" approach to understanding speech in noise was very successful in predicting the effectiveness of the telecommunications systems for which it was originally intended. As the attention of researchers turned to more complex maskers such as competing talkers, these energetic masking explanations became less adequate in explaining the extent of masking interactions or the masking release that was afforded by differences between the target and maskers. This informational masking pointed to more complex and cognitive levels of interference that went far beyond the spectro-temporal interactions of the sound energy associated with multiple sources.

This shifted the focus from a single channel problem such as understanding a voice on a telephone line to one of auditory scene analysis – itself a very ill posed problem in a mathematical sense. Research over the last few decades has revealed how the auditory system exploits the physical characteristics of naturally sounding bodies to parse the multiple concurrent sources that most often comprise the listening environment. More critically, this is not just a passive or automatic process but can be influenced by endogenous or "top-down" attentional control. Grouping of the features associated with a sound provides the perceptual object that becomes the focus of attention which in turn is modulated by task requirements and the executive intentions of the listener.

Thus hearing becomes listening – an active process involving a heterarchy of functions feeding forward and feeding backward, weighing the evidence on its saliency, reliability and relevance. Now more than 60 years on, we may finally be within striking distance of Cherry's original goals of understanding the "acts of recognition and discrimination" that enable those critical interactions at the cocktail party.

ACKNOWLEDGEMENTS

Virginia Best, Gavin Lin and Anthony Wakilicz are warmly acknowledged for their careful review and helpful comments on an earlier version of this manuscript.

REFERENCES

[1] Heald S. L. M. & Nusbaum H. C., Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience* **8**:35-35 (2014)

[2] Durlach N. I., et al., Note on informational masking (L). *J Acoust Soc Am* **113**:2984-2987 (2003)

[3] Fletcher H., Auditory patterns. *Reviews of Modern Physics* **12**:0047-0065 (1940)

[4] Green D. M. & Swets J. A., *Signal detection theory and psychophysics* (John Wiley and Sons, New York) (1966)

[5] Kidd G., Jr., Mason C., Richards V., Gallun F., & Durlach N., Informational masking. Auditory perception of sound sources, *Springer Handbook of Auditory Research*, eds Yost W, Popper A, & Fay R (Springer US), Vol 29, pp 143-189 (2008)

[6] Cherry E. C., Some experiments on the recognition of speech with one and two ears. *J Acoust Soc Am* **25**:975-979 (1953)

[7] Bronkhorst A. W., The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica* **86**:117-128 (2000)

[8] Ebata M., Spatial unmasking and attention reated to the cocktail party problem. *Acoustical Science and Technology* **24**:208-219 (2003)

[9] Carhart R., Tillman T. W., & Greetis E. S., Perceptual masking in multiple sound background. *J Acoust Soc Am* **45**:694-& (1969)

[10] Freyman R. L., Helfer K. S., McCall D. D., & Clifton R. K., The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am* **106**:3578-3588 (1999)

[11] Freyman R. L., Balakrishnan U., & Helfer K. S., Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J Acoust Soc Am* **115**:2246-2256 (2004)

[12] Brungart D., Simpson B. D., Ericson M., & Scott K., Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J Acoust Soc Am* **110**:2527-2538 (2001)

[13] Brungart D. S., Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* **109**:1101-1109 (2001)

[14] Arbogast T. L., Mason C. R., & Kidd G.,Jr., The effect of spatial separation on informational and energetic masking of speech. *J Acoust Soc Am* **112**:2086-2098 (2002)

[15] Arbogast T. L., Mason C. R., & Kidd G., Jr., The effect of spatial separation on informational masking of speech in normal-hearing and hearing impaired listeners. *J Acoust Soc Am* **117**:2169-2180 (2005)

[16] Cooke M. P., A glimpsing model of speech perception in noise. *J Acoust Soc Am* **119**:1562-1573 (2006)

[17] Brungart D. S. & Iyer N., Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *J Acoust Soc Am* **132**:2545-2556 (2012)

[18] Glyde H., et al., The effects of better ear glimpsing on spatial release from masking. *J Acoust Soc Am* **134**:2937-2945 (2013)

[19] Cameron S., Dillon H., & Newall P., Development and evaluation of the listening in spatialized noise test. *Ear and Hearing* **27**:30-42 (2006)

[20] Kidd G., Jr., Arbogast T. L., Mason C. R., & Gallun F. J., The advantage of knowing where to listen. *J Acoust Soc Am* **118**:3804-3815 (2005)

[21] Varghese L. A., Ozmeral E. J., Best V., & Shinn-Cunningham B. G., How visual cues for when to listen aid selective auditory attention. *JARO* **13**:359-368 (2012)

[22] Kitterick P. T., Bailey P. J., & Summerfield A. Q., Benefits of knowing who, where, and when in multi-talker listening. *J Acoust Soc Am* **127**:2498-2508 (2010)

[23] Freyman R. L., Helfer K. S., & Balakrishnan U., Variability and uncertainty in masking by competing speech. *J Acoust Soc Am* **121**:1040-1046 (2007)

[24] Best V., Ozmeral E. J., & Shinn-Cunningham B. G., Visually-guided attention enhances target identification in a complex auditory scene. *JARO* **8**:294-304 (2007)

[25] Rhebergen K. S., Versfeld N. J., & Dreschler W. A., Release from informational masking in time reversal of native and non-native interfering speech (L). *J Acoust Soc Am* **118**:1274-1277 (2005)

[26] Cooke M., Lecumberri M. L. G., & Barker J., The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *J Acoust Soc Am* **123**:414-427 (2008)

[27] Brouwer S., Van Engen K. J., Calandruccio L., & Bradlow A. R., Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *J Acoust Soc Am* **131**:1449-1464 (2012)

[28] Freyman R. L., Balakrishnan U., & Helfer K. S., Spatial release from informational masking in speech recognition. *J Acoust Soc Am* **109**:2112-2122 (2001)

[29] Knudsen E. I., Fundamental components of attention. *Ann Rev Neurosci* **30**:57-78 (2007)

[30] Brungart D. S. & Simpson B. D., Within-ear and across-ear interference in a cocktail party listening task. *J Acoust Soc Am* **112**:2985-2995 (2002)

[31] Brungart D. S., Simpson B. D., Darwin C. J., Arbogast T. L., & Kidd G., Jr., Across-ear interference from parametrically degraded synthetic speech signals in a dochotic cocktail-party listening task. *J Acoust Soc Am* **117**:292-304 (2005)

[32] Rivenez M., Darwin C. J., & Guillaume A., Processing unattended speech. *J Acoust Soc Am* **119**:4027-4040 (2006)

[33] Rivenez M., Guillaume A., Bourgeon L., & Darwin C. J., Effect of voice characteristics on the attended and unattended processing of two concurrent messages. *Euro J Cog Psych* **20**:967-993 (2008)

[34] Alain C. & Arnott S. R., Selectively attending to auditory objects. *Frontiers in Bioscience* **5**:D202-D212 (2000)

[35] Shinn-Cunningham B. G., Object-based auditory and visual attention. *Trends in Cognitive Sciences* **12**:182-186 (2008)

[36] Griffiths T. D. & Warren J. D., What is an auditory object? *Nat Rev Neurosci* **5**:887-892 (2004)

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

[38] Darwin C. J., Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**:1011-1021 (2008)

[39] McDermott J. M., The Cocktail Party Problem. *Current Biology* **19**:R1024-1027 (2009)

[40] Darwin C. J. & Carlyon R. P., Auditory Grouping. *Handbook of Perception and Cognition, Volume 6: Hearing*, ed Moore B (Academic Press, Orlando, Florida), pp 387-424 (1995)

[41] Darwin C. J. & Hukin R. W., Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J Acoust Soc Am* **107**:970-977 (2000)

[42] Best V., Ozmeral E. J., Kopco N., & Shinn-Cunningham B. G., Object continuity enhances selective auditory attention. *PNAS* **105**:13174-13178 (2008)

[43] Best V., Shinn-Cunningham B. G., Ozmeral E. J., & Kopco N., Exploring the benefit of auditory spatial continuity. *J Acoust Soc Am* **127**:EL258-EL264 (2010)

[44] Ruggles D. & Shinn-Cunningham B., Spatial Selective Auditory Attention in the Presence of Reverberant Energy: Individual Differences in Normal-Hearing Listeners. *JARO* **12**:395-405 (2011)

[45] Ruggles D., Bharadwaj H., & Shinn-Cunningham B. G., Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *PNAS* **108**:15516-15521 (2011)

[46] Ruggles D., Bharadwaj H., & Shinn-Cunningham B. G., Why middle-aged listeners have trouble hearing in everyday settings. *Current Biology CB* **22**:1417-1422 (2012)

[47] Koch I., Lawo V., Fels J., & Vorlaender M., Switching in the cocktail party: Exploring intentional control of auditory selective attention. *J Exp Psych-Human Percept Perf* **37**:1140-1147 (2011)

[48] Dhamani I., Leung J., Carlile S., & Sharma M., Switch attention to listen. *Nature: Sci Reports* **3**:1-8 (2013)

[49] Wright B. A. & Fitzgerald M. B., The time course of attention in a simple auditory detection task. *Percep Psychophys* **66**:508-516 (2004)

[50] Ihlefeld A. & Shinn-Cunningham B. G., Disentangling the effects of spatial cues on selection and formation of auditory objects. *J Acoust Soc Am* **124**:2224-2235 (2008)

[51] Maddox R. K. & Shinn-Cunningham B. G., Influence of Task-Relevant and Task-Irrelevant Feature Continuity on Selective Auditory Attention. *JARO* **13**:119-129 (2012)

[52] Best V., Gallun F. J., Carlile S., & Shinn-Cunningham B. G., Binaural interference and auditory grouping. *J Acoust Soc Am* **121**:1070-1076 (2007)

[53] Shinn-Cunningham B. G., Lee A. K. C., & Oxenham A. J., A sound element gets lost in perceptual competition. *PNAS* **104**:12223-12227 (2007)

[54] Schwartz A. H. & Shinn-Cunningham B. G., Dissociation of perceptual judgments of "what" and "where" in an ambiguous auditory scene. *J Acoust Soc Am* **128**:3041-3051 (2010)

[55] Schwartz A., McDermott J. H., & Shinn-Cunningham B., Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *J Acoust Soc Am* **132**:357-368 (2012)

[56] Iyer N., Brungart D., & Simpson B. D., Effects of target-masker contextual simillarity on the multi-masker penalty in a three talker diotic listening task. *J Acoust Soc Am* **128**:2998-3010 (2010)

[57] Rudner M., Ronnberg J., & Lunner T., Working memory supports listening in noise for persons with hearing impairment. *J Am Acad Audiol* **22**:156-167 (2011)

[58] Rudner M., Lunner T., Behrens T., Thoren E. S., & Ronnberg J., Working memory capacity may influence perceived effort during aided Speech Recognition in Noise. *J Am Acad Audiol* **23**:577-589 (2012)

[59] Ronnberg N., Rudner M., Lunner T., & Stenfelt S., Assessing listening effort by measuring short-term memory storage and processing of speech in noise. *Speech, Language and Hearing In Press* (2014)

[60] Helfer K. S. & Freyman R. L., The role of visual speech cues in reducing energetic and informational masking. *J Acoust Soc Am* **117**:842-849 (2005)

[61] Mishra S., Lunner T., Stenfelt S., Rönnberg J., & Rudner M., Seeing the talker's face supports executive processing of speech in steady state noise. *Frontiers in Systems Neuroscience* **7** (2013)

[62] Shamma S. A. & Micheyl C., Behind the scenes of auditory perception. *Current Opinion in Neurobiology* **20**:361-366 (2010)

[63] Du Y., et al., Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. *Cerebral Cortex* **21**:698-707 (2011)

[64] Ding N. & Simon J. Z., Emergence of neural encoding of auditory objects while listening to competing speakers. *PNAS* **109**:11854-11859 (2012)

[65] Ding N. & Simon J. Z., Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. *J. Neurosci.* **33**:5728-5735 (2013)

[66] Golumbic E. M. Z., et al., Mechanisms underlying selective neuronal tracking of attended speech at a "Cocktail Party". *Neuron* **77**:980-991 (2013)

[67] Shamma S. A., Elhilali M., & Micheyl C., Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* **34**:114-123 (2011)

[68] Lee A. K. C., et al., Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience* **6**:190-190 (2012)

[69] Larson E. & Lee A. K. C., The cortical dynamics underlying effective switching of auditory spatial attention. *Neuroimage* **64**:365-370 (2013)

[70] Larson E. & Lee A. K. C., Switching auditory attention using spatial and non-spatial features recruits different cortical networks. *Neuroimage* **84**:681-687 (2014)