Influences of spatial cues on grouping and understanding sound

Barbara Gail Shinn-Cunningham
Boston University Hearing Research Center, 677 Beacon St., Boston, MA, USA, **shinn@cns.bu.edu**

Spatial cues such as interaural time and level differences often have such a weak influence on auditory grouping that they are overwhelmed when opposed by other cues, such as common onset or continuity. In reverberant listening conditions, spatial cues provide relatively little benefit when understanding a sound source in the presence of competing noise. Yet everyday experience shows that listening with two ears is critical for being able to converse at a real cocktail party, a case in which there are usually multiple simultaneous talkers as well as reverberant energy. This apparent paradox can be partially explained by recent studies, which demonstrate that 1) the influence of spatial cues on auditory streaming can be strong even though their influence on simultaneous grouping is weak, and 2) spatial cues play an important role in separating competing sound sources that are easily confused with each other, a benefit that is distinct from the binaural-processing benefits that break down in the presence of reverberant energy. Results suggest that spatial cues are utilized by the auditory system in many different ways, at many different levels. In simple listening conditions (like those examined in traditional studies of binaural processing), spatial cues can help a listener detect sound that might otherwise be masked; but in more complex conditions, spatial cues are critical for properly parsing the mixture of sound into different objects and focusing attention on the source of interest.

1 Introduction

In nearly every setting (other than the laboratory), the acoustic energy that reaches our ears is a mixture of signals coming from many sources and many directions. Despite this complexity, we are incredibly adept at sorting out what sound sources are present in the environment. The spatial cues in the signals we hear play an extraordinarily important role in our ability to sort out and understand sources in complex, natural settings. However, many past studies suggest that spatial cues have at best a weak influence on auditory scene analysis (ASA; the process of sorting the mixture of sound we hear into appropriate "auditory objects," or perceived sound sources). This paper reviews studies from both the literature and our own laboratory that explore the influence of spatial cues and spatial perception in order to resolve the apparently conflicting results of these studies.

2 Spatial cues and segregation

2.1 Evidence for weak effects

Many cues influence how individual elements of sound energy are grouped into auditory objects, including harmonicity, common onset and offset, common modulation, and spatial cues (see the seminal book by Bregman for a review of this topic [1]). A typical approach to studying how these various cues influence grouping and streaming is to pit different cues against one another and then measure which cues "win." In many past experiments in which spatial cues compete against other cues, the spatial cues have very little effect on how a mixture of sound is perceptually organized into perceived objects.

The "double vowel" paradigm is one approach that has been used to test ASA. In these experiments, mixtures of harmonic complexes are constructed in which a pair of vowels may be perceived, but only if the constituent complexes of the sound mixture are appropriately segregated. In most such studies, the default percept (if segregation doesn't occur) is of a single, non-vowel object; only if there is some cue that causes the complexes to be appropriately split into two objects do listeners perceive a pair of vowels. The double vowel paradigm has been used to test the efficacy of many possible segregation cues, including common modulation, harmonicity, common onset, and spatial cues (e.g., interaural time differences or ITDs). Differences in fundamental frequency or in the onsets of sound components in the mixture are very effective in promoting segregation [2-4]. In contrast, spatial cues are insufficient to allow subjects to identify the vowels in the mixture [5, 6] (however, see [7]).

Studies examining sensitivity to ITD also suggest that spatial cues in one component are usually insufficient to cause the component to be segregated from a sound mixture. Sensitivity to changes in the spatial attributes of a single target sound component are reduced when the target component is heard as part of an object composed of many elements; sensitivity is enhanced when other cues, such as harmonicity or common onset, promote hearing the target as a separate object [8, 9]. However, even when listeners perceive the target as an object separate from the other simultaneous sound elements, the spatial cues in the non-target elements causes some interference in target ITD sensitivity [10, 11].

Taken together, these studies suggest that mismatches between the spatial cues in different simultaneous sound elements are generally not influential enough to drive segregation into different sound objects. Instead, other grouping cues usually determine what simultaneous sound elements are heard within an object. While the perceived location of an object is primarily determined by the spatial cues in the elements making up that object, there is still some effect of the spatial information from elements not grouped into that object.

2.2 When spatial cues matter

Although many studies suggest that spatial cues have little influence on simultaneous segregation, studies that emphasize streaming sound elements over time are heavily influenced by spatial location.

The importance of across-time grouping is particularly evident in studies of speech intelligibility when there are multiple, simultaneous talkers. For instance, Darwin and Hukin [12] manipulated speech to control the spatial cues and fundamental frequency in target and carrier phrases, as well as the vocal tract length of the simulated talker. They then asked listeners to report the target word contained in a target carrier phrase during presentation of a competing carrier phrase. Two candidate target words were presented simultaneously during a time-aligned temporal gap present in both the target and competing carrier phrases. They found that listeners reported the target word whose spatial location matched that of the target phrase when other possible grouping cues were placed in opposition to the spatial cues. These results suggest that spatial continuity is an important cue used to organize sound over time (see also [13]). Other studies also support the idea that spatial cues are important for parsing sources over time. Many recent studies of speech understanding in the presence of competing speech show that differences in the spatial cues in the competing sources boost target speech intelligibility (e.g., see [14-17]), a topic considered in Section 3.

These studies show that particularly when considering how sound is organized over time, spatial cues in the signals reaching the listener have a large influence on perceptual organization of sound.

2.3 Nothing is simple

The results reviewed in the previous subsections seem consistent with a fairly straightforward explanation for how spatial cues affect auditory scene analysis:

1) Spatial cues do not influence grouping of simultaneous sources, instead other sound features determine how simultaneous or near-simultaneous sounds are grouped locally in time and in frequency, forming "snippets" of sound.

2) Once a sound snippet is formed, its spatial location is computed, based primarily on the spatial cues in the sound elements grouped into that snippet.

3) Sound snippets are then pieced together across time in a process that relies heavily on perceived location of the snippets.

However, even this is too simplistic a view.

For instance, Darwin and Hukin [18, 19] investigated stimuli in which a target tone could logically fall into one of two streams, one a sequence of repeated tones, one a simultaneous harmonic complex. In the experiments, they measured the degree to which the ambiguous target was heard as part of the harmonic complex. They found that when the spatial cues in the target matched those in the simultaneous harmonic complex, the tone was heard more prominently in the harmonic complex than when the spatial cues were uninformative. Moreover, the way in which trials were blocked influenced how prominently the target was heard in the complex: the same stimulus gave different results, depending on what subjects had been hearing in past trials [18]. They concluded that spatial cues can influence simultaneous grouping when other grouping cues are ambiguous and that top-down listener expectations also influence grouping.

It has long been known that segregation of a complex sound mixture builds up over time as a listener accrues information about the sound pattern they are hearing (e.g., see [1]). However, recent results show that this build up also depends on top-down listener attention: the build up appears to begin at the moment a listener attends to a particular sound mixture or sound [20].

Recent work in our own lab, based on the ambiguous-tone paradigm of Darwin and Hukin, also provides evidence that the way an acoustic mixture is perceptually organized depends on what a listener is attending [21]. Using the two-object (tone sequence, harmonic complex) paradigm, we investigated not only how prominently the ambiguous target influenced perception of the harmonic complex, but also how prominently the target was heard in the repeating tone sequence. We found that there was no predictive relationship between the degree to which the target was "in" one auditory object and the degree to which it was "out" of the other (using the same stimuli with the same listeners, just changing which object the listener was asked to attend). In particular, while spatial cues influenced the degree to which the target was heard in both the across-time tone sequence and the simultaneous harmonic complex, spatial cues were far more influential on the across-time sequence.

Taken together, it appears that spatial cues have some influence on grouping both across frequency and across time. The relative weight given to spatial information, compared to other segregation cues, is much greater when one looks at piecing together sound snippets across time rather than how one determines what simultaneous sound elements comprise a snippet from a single source. However, results also suggest that the way in which listeners parse the world depends on top-down factors: the same stimulus can be segregated differently depending on the recent history of sounds heard by the listener [18], the time over which the listener attends a scene or object [20], and which object in the acoustic scene a listener is attending [21].

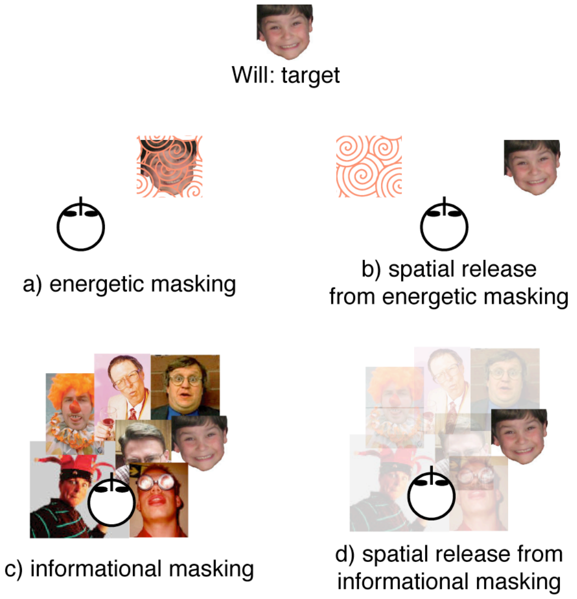
Will: target
a) energetic masking
b) spatial release from energetic masking
c) informational masking
d) spatial release from informational masking

Figure 1: Visual analogues of spatial release from energetic and informational masking.

3 Spatial unmasking

Spatial cues have long been known to play an important role in behavioural tasks when there are competing sound sources. However, the relationship between studies of source segregation and spatial unmasking has not been entirely clear.

Many studies have focussed on measuring how well listeners can detect or understand a target sound in the presence of a competing source. In many conditions, performance improves when the target and competing sources arise from different spatial locations. This “release from masking” due to spatial separation of target and masker is known as spatial unmasking.

Traditional studies of spatial unmasking examined what happens when a target source is presented in a background of statistically stationary noise. In these conditions, the dominant kind of interference between target and masker is that the masker can render the target inaudible; if the target can be heard (is audible), then it is easy to segregate the target and masker, and there is relatively little difficulty detecting the presence of the target or interpreting its content. The problem is that not all of the target can be heard. The visual analogy of this kind of “energetic masking” effect is illustrated in the top left panel of Figure 1. Spatial separation of target and masker can render more of the target perceivable, illustrated in visual analogy in the top right of the figure.

However, in conditions where target and masker are similar to one another (such as speech on speech masking), the problem may be that the target speech is confused with the masker speech [22], not that the target is inaudible. Spatial cues can play a very important role in helping a listener focus attention on the target, throwing out the sensory “clutter” that is the masker and devoting more computational resources to processing the target. The visual analogue of this kind of “informational masking” is shown in the bottom left panel of Figure 1. We believe that the role of spatial cues in this kind of masking is to modulate competition between the target and the masker, reducing interference of the masker, shown by visual analogy in the bottom right of Figure 1.

In order to understand spatial unmasking, it is important to understand how spatial cues can cause release from both energetic and informational masking. In particular, both kinds of masking contribute to perception in almost every setting encountered in daily life; spatial separation of target and masker reduces both forms of masking. Below, we discuss three distinct mechanisms (better-ear acoustics, binaural processing, and spatial attention) that we believe contribute to spatial unmasking in everyday settings.

3.1 Energetic (better-ear) effects

When sources overlap in time and frequency, spatial separation can reduce the amount of peripheral, energetic masking for purely physical reasons. Acoustic interactions with the head reduce the amount of energy received at the far ear when a source is to the side of the listener. For instance, in the top right panel of Figure 1, the right ear receives less energy from the masker simply because the masker is spatially displaced from the target. This “better-ear” effect can produce large improvements in speech intelligibility with spatial separation of target and masker [23, 24], especially for sources that contain significant energy above 2 kHz where the “acoustic head shadow” effect is large (e.g., with birdsong stimuli, where most of the information is between 3-6 kHz; see [25]).

3.2 Binaural processing

When portions of a sound (in time-frequency space) are masked by another sound such that they are inaudible even when listening with the acoustically better ear, binaural processing can yield further improvements in target audibility. This is demonstrated in Figure 2, which shows the output of a simple model of binaural processing, based on the normalized interaural cross-correlation of narrowband left- and right-ear signals as a function of time (e.g., similar to the many traditional binaural models [23, 26]). In the figure, the image intensity represents the cross-correlation value for a given interaural time delay, with different interaural time delays shown on the y axis (see [27] for details of modelling approach).

In Figure 2, when target and masker are both presented from the same location, the cross-correlation output does not change significantly when the tone is turned on. However, if the tone is at a different location than the masker, it causes interaural decorrelation of the masker, producing fluctuations in the running cross-correlation output as a function of time whenever both sources are on simultaneously. In such situations, the listener identifies a temporary decrease in correlation as being caused by an (otherwise inaudible) source at that time and frequency (e.g., see [5]), effectively reducing the energetic masking caused by the masker.

3.3 Spatial attention

The third posited contributor to spatial unmasking works not by increasing target audibility, but by reducing the confusion between target and masker. In vision, physiological evidence for “spatial attention” is well known (see [28]). We posit that spatial attention also operates in audition (and across modalities).

Freyman and his colleagues found that *adding* masker energy that caused the masker to be perceptually displaced from a similar target increased target speech intelligibility [14]. Kidd and colleagues have performed a number of studies in which the amount of spectro-temporal overlap of competing signals is very limited. In such cases, traditional within-band binaural unmasking mechanisms based on detection of interaural decorrelation (illustrated in Figure 2) predict little or no spatial unmasking, as nearly all of the target energy is audible. However, they find that spatial unmasking is prominent when target and masker are statistically similar, but negligible when the masker is steady-state noise (e.g., see [15, 16]). Moreover, ordinary room reverberation, which decorrelates the left- and right-ear signals even when only one source is present, disrupts spatial release from within-band energetic masking, but not spatial release from informational masking [15].

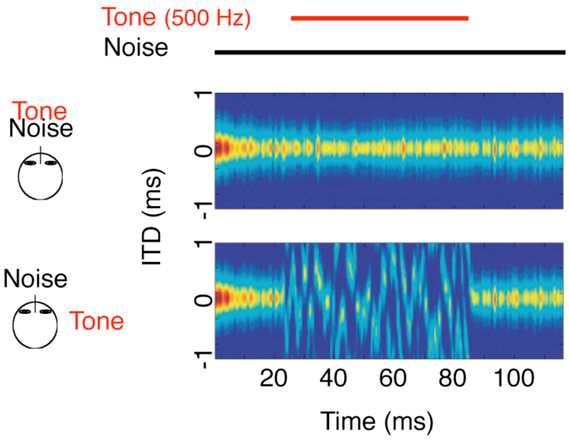

Figure 2: Model cross-correlation output of a 500-Hz channel for a 500 Hz tone temporally centred in broadband noise. Top: tone and noise spatially coincident. Bottom: tone and noise separated.

In our lab, we have seen similar effects using spectro-temporally complex birdsongs [25]. When listeners are asked to identify one of five songs in the presence of steady-state birdsong-shaped noise, there is significant spatial unmasking; however, the entire effect can be attributed to better-ear energy effects (performance for a binaural condition is equal to performance when listeners are presented with the “better ear” stimulus diotically). In contrast, when the masker is a bird chorus (a mixture of unfamiliar bird calls) there is less masking overall (presumably because the masker has fluctuations in energy over time and frequency, reducing the amount of energetic masking), but a larger spatial release from masking. Moreover, when the target and chorus masker are spatially separated, binaural performance is significantly better than diotic better-ear performance. In other words, when target and masker are similar, perceived spatial separation contributes to spatial unmasking, but not when target and masker are dissimilar and easily segregated

Using stimuli with little energetic masking (after [16]), we find that no matter what kind of spatial cues cause the target and masker to be perceived in different locations, spatial release from masking is essentially the same magnitude, after one accounts for the better-ear acoustic advantage [29]. We believe that any cues that lead to differences in perceived location can be sufficient to guide spatial attention, whether the spatial cues are ITDs (which also underlie the traditional within-band spatial unmasking illustrated in Figure 2), interaural level differences, or a full set of realistic spatial cues.

These results consistently show that spatial attention plays an important role in spatial unmasking when target and masker are statistically similar and therefore hard to piece together across time.

4 ASA and spatial unmasking

There is a growing realization that spatial auditory cues play a critical role in parsing the mixture of signals we hear, and moreover, that this contribution has an important and direct impact on our ability to understand sources in the everyday world,

The role of spatial cues on auditory scene analysis can be weak or strong. In grouping simultaneous sound elements, spatial cues have a relatively minor influence. In determining which sound elements belong to the same sound source over time, spatial cues are very influential. Stimulus properties alone do not determine how sound mixtures are organized perceptually; instead, other top-down factors influence ASA, including listener expectations, listener attention, and the kind of object being attended. In short, the role that spatial cues play in ASA cannot be described using a simple, bottom-up process; instead, the internal model we construct of the sources in the environment builds up slowly over time as we accrue information, biased by our expectations as well as the properties of the object we are attending. In this course of this iterative process, spatial cues play an important role.

Traditional studies of spatial unmasking, in conditions where target and masker are statistically dissimilar, do not relate directly to ASA; such studies measure how spatial cues may allow a listener to hear out target elements that were otherwise inaudible due to purely within-frequency channel interactions. With traditional spatial masking, target audibility improves with spatial separation of target and masker both because of acoustic effects at the better ear and through low-level binaural processing mechanisms. However, many studies now show that spatial attention plays a very important role in spatial unmasking: perceived spatial separation of target and masker reduces confusion between sources that are otherwise hard to separate.

In the everyday world, traditional spatial unmasking is less important than in the laboratory: reverberant energy reduces the better-ear advantage and decorrelates the signals reaching the ears, degrading the benefits of binaural unmasking. However, spatial attention is robust in such settings.

We believe that spatial attention is the most important contributor to spatial unmasking in everyday settings. Furthermore, spatial attention is a manifestation of source segregation: spatial attention aids performance by allowing listeners to segregate the target from the masker. By focussing attention on the target location and pulling out the target from the background, the listener can devote more computational resources to processing the target and reduce central interference between the target and masker.

This important link between segregation and spatial unmasking is one that must be developed further. Currently, no models can adequately explain the process of source segregation and predict the benefits of spatial unmasking, which depend on target and masker stimulus properties, characteristics of the listening environment, and listener expectation and knowledge. Many unanswered questions arise related to how we focus attention on sources of interest. How does information from other modalities interact with auditory source cues? How important is a priori knowledge about where and when to listen for a source of interest in allowing us to focus attention? How does ASA affect and interact with the computation of auditory object location? We are currently developing paradigms to explore these questions in detail.

5 Acknowledgements

This work was funded by grants R01 DC05778-02 from the National Institutes of Health and FA9550-04-1-0260 from the Air Force Office of Scientific Research. Kosuke Kawakyu, Adrian KC Lee, Antje Ihlefeld, Gin Best, Erol Ozmeral, Erick Gallun, Satyavarta, Kemal Sen, Andrew Oxenham, Eric Larson, and Simon Carlile contributed to the work and ideas reviewed here. Will Cunningham graciously agreed to serve as “target” (a function he normally serves for Nick Cunningham).

References

[1] A.S. Bregman, '*Auditory Scene Analysis*', MIT Press, Cambridge, MA (1990).

[2] R.W. Hukin and C.J. Darwin, 'Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification'. *Percept Psychophys*, Vol. 52. pp. 191-196 (1995).

[3] P.F. Assmann and Q. Summerfield, 'Modeling the perception of concurrent vowels: vowels with different fundamental frequencies'. *J Acoust Soc Am*, Vol. 88. pp. 680-97 (1990).

[4] A. de Cheveigne, S. McAdams, and C.M.H. Marin, 'Concurrent vowel identification. II. Effects of phase, harmonicity, and task'. *J Acoust Soc Am*, Vol. 101. pp. 2848-2856 (1997).

[5] J.F. Culling and Q. Summerfield, 'Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay'. *J Acoust Soc Am*, Vol. 98. pp. 785-97 (1995).

[6] T.M. Shackleton and R. Meddis, 'The role of interaural time difference and fundamental frequency difference in the identification of

concurrent vowel pairs'. *J Acoust Soc Am*, Vol. 91. pp. 3579-81 (1992).

[7] W.R. Drennan, S. Gatehouse, and C. Lever, 'Perceptual segregation of competing speech sounds: the role of spatial location'. *J Acoust Soc Am*, Vol. 114. pp. 2178-89 (2003).

[8] M.A. Akeroyd, 'Threshold differences for interaural time delays carried by double vowels'. *J Acoust Soc Am*, Vol. 114. pp. 2167-77 (2003).

[9] T.N. Buell and E.R. Hafter, 'Combination of binaural information across frequency'. *J Acoust Soc Am*, Vol. 90. pp. 1894-1900 (1991).

[10] W.S. Woods and H.S. Colburn, 'Test of a model of auditory object formation using intensity and interaural time difference discrimination'. *J Acoust Soc Am,* Vol. 91. pp. 2894-2902 (1992).

[11] M.A. Stellmack and R.H. Dye, Jr., 'The combination of interaural information across frequencies: The effects of number and spacing of components, onset asynchrony'. *J Acoust Soc Am*, Vol. 93. pp. 2933-2947 (1993).

[12] C.J. Darwin and R.W. Hukin, 'Effectiveness of spatial cues, prosody, and talker characteristics in selective attention'. *J Acoust Soc Am*, Vol. 107. pp. 970-7 (2000).

[13] C.J. Darwin and R.W. Hukin, 'Auditory objects of attention: the role of interaural time differences'. *J Exp Psych Hum Percept Perf*, Vol. 25. pp. 617-29 (1999).

[14] R.L. Freyman, U. Balakrishnan, and K. Helfer, 'Spatial release from informational masking in speech recognition'. *J Acoust Soc Am*, Vol. 109. pp. 2112-2122 (2000).

[15] G. Kidd, Jr., C.R. Mason, A. Brughera, and W.M. Hartmann, 'The role of reverberation in release from masking due to spatial separation of sources for speech identification'. *Acustica united with Acta Acustica*, (2005).

[16] T.L. Arbogast, C.R. Mason, and G. Kidd, Jr., 'The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners'. *J Acoust Soc Am*, Vol. 117. pp. 2169-80 (2005).

[17] M.L. Hawley, R.Y. Litovsky, and J.F. Culling, 'The benefit of binaural hearing in a cocktail party: effect of location and type of interferer'. *J Acoust Soc Am*, Vol. 115. pp. 833-43 (2004).

[18] C.J. Darwin and R.W. Hukin, 'Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony'. *J Acoust Soc Am*, Vol. 103. pp. 1080-4 (1998).

[19] C.J. Darwin and R.W. Hukin, 'Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity'. *J Acoust Soc Am*, 102(4): pp. 2316-24 (1997).

[20] R. Cusack, J. Deeks, G. Aikman, and R.P. Carlyon, 'Effects of location, frequency region, and time course of selective attention on auditory scene analysis'. *J Exp Psych Hum Percept Perf*, Vol. 30. pp. 643-56 (2004).

[21] A.K. Lee, B. Shinn-Cunningham, and A.J. Oxenham. 'The missing target: Evidence of a tone's inability to contribute to the auditory foreground'. *Proc. Mid-winter meeting of the ARO*, New Orleans (2005).

[22] N.I. Durlach, C.R. Mason, B.G. Shinn-Cunningham, T.L. Arbogast, H.S. Colburn, and G. Kidd, Jr., 'Informational masking: counteracting the effects of stimulus uncertainty by decreasing target-masker similarity'. *J Acoust Soc Am*, Vol. 114. pp. 368-79 (2003).

[23] P.M. Zurek, 'Binaural advantages and directional effects in speech intelligibility', in *Acoustical Factors Affecting Hearing Aid Performance*, G. Studebaker and I. Hochberg, Editors. 1993, College-Hill Press: Boston, MA.

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

[25] V. Best, E. Ozmeral, F.J. Gallun, K. Sen, and B.G. Shinn-Cunningham, 'Spatial unmasking of birdsong in human listeners: Energetic and informational factors'. *J Acoust Soc Am* (submitted).

[26] H.S. Colburn, 'Theory of binaural interaction based on auditory-nerve data. I: General strategy and preliminary results on interaural discrimination'. *J Acoust Soc Am*, Vol. 54. pp. 1458-1470 (1973).

[27] B. Shinn-Cunningham and K. Kawakyu. 'Neural representation of source direction in reverberant space'. *Proc. IEEE WASPAA,* New Pfalz, New York (2003).

[28] R. Desimone and J. Duncan, 'Neural mechanisms of selective visual attention'. Annual Reviews of Neuroscience, Vol. 18. pp. 193-222 (1995).

[29] B.G. Shinn-Cunningham, A. Ihlefeld, Satyavarta, and E. Larson, 'Bottom-up and top-down influences on spatial unmasking'. *Acustica united with Acta Acustica* (submitted).