Central auditory processing in the cocktail-party effect

Adelbert Bronkhorst[1,2]

[1]*TNO Human Factors, POB 23, 3769 ZG Soesterberg, The Netherlands*

[2]*Cognitive Psychology Department, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands*

When we try to understand one talker in a group of talkers the capacities of our auditory system are stretched to the limit. Using the superposition of incoming sounds as input, it has to identify the target speech, trace it over time, fill in parts masked by other sounds, and finally convert it to a stream of meaningful information. Research into this "cocktail party" effect has proceeded along different lines that for a long time showed little or no overlap. Well-known for most psycho-acousticians are studies of peripheral effects such as (energetic) masking and binaural unmasking. In this presentation an overview is given of three other research lines that have addressed central processing of complex speech stimuli, and relationships between these lines are discussed. The oldest line looked at the role of attention in the selection of the target speech from all signals entering the ears. A more recent line has focused on the process of separating and piecing together acoustic information across time and space, which is referred to as grouping. In the third line, masking is studied but effects of peripheral (un)masking are factored out so that only the excess masking – referred to as informational masking, remains.

INTRODUCTION

The cocktail-party (CP) effect (Cherry, 1953) has been defined in many ways but the most important elements are (1) that a listener presented with speech from different talkers can focus on a single target talker; and (2) that the listener makes use of differences in spatial position between target and interfering talkers. Psychoacousticians have for a long time looked at the CP effect from the point of view of audibility. This is of course a sensible standpoint: audibility is a prerequisite for all subsequent processing. Furthermore, it appeared that several complex effects were involved in determining this basic variable:

- Because both the target and the interfering speech show strong spectrotemporal fluctuations, audibility itself varies in time in a way that is difficult to predict.
- The frequency spectrum of a sound depends in a complex manner on angle of incidence.
- Due to binaural interaction (unmasking), audibility is increased when the interaural properties of the target sound differ from those of the interfering sound.

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

- Room reverberation smears spectrotemporal fluctuations and decreases the dependence of audibility on spatial position.
- Hearing impairment and the use of hearing aids interact in nontrivial ways with all the above effects.

Over the last decades, research has shed a lot of light on these issues (e.g. Bronkhorst, 2000), but the work has not been finished. Illustrative in this respect is the fact that the modelling of speech intelligibility has progressed relatively slowly. While successful models of the effects of steady-state noise and reverberation have been available for many decades (French and Steinberg, 1947; Houtgast and Steeneken, 1973), and several models for binaural listening have been developed (vom Hövel 1984; Zurek, 1990; Van Wijngaarden and Drullman, 2008), effects of fluctuating background noise were only considered relatively recently (Rhebergen *et al*., 2006) as well as the modelling of hearing impairment (e.g. Beutelmann and Brand, 2006).

Audibility represents, however, only one of the phenomena relevant for understanding the CP effect. In this paper I will address three other relevant phenomena: attention, auditory grouping and informational masking.

An essential aspect of the CP effect is that the listener is able to selectively tune in to one talker, ignoring the others. This is a highly flexible process: when the story of someone else becomes more interesting, attention can be easily and quickly switched to that person. Thus, knowing the acoustical information presented to a listener and the audibility of the speech signals is not sufficient. Attention determines which speech signal is selected for further processing.

Attention is not the only process that helps us to select information. The speech signal generated in a CP environment by the attended talker is highly variable and embedded in similar fluctuating sounds. The target sounds need to be segregated from the other sounds and consecutive sounds must be linked together into a stream. As research into auditory grouping shows (Bregman, 1990; Darwin and Carlyon, 1995; Darwin, 2008), our auditory system uses low-level acoustical features of sounds as well as top-down knowledge of sound properties to perform this difficult task.

A relatively recent line of research with its origins in psychoacoustics has taken a different approach to study effects other than audibility alone. The approach is in principle straightforward: speech intelligibility is either measured in conditions where audibility is not an issue, or compared to results of baseline conditions where audibility is the dominant factor. Any excess masking that cannot be attributed to reduced audibility must be due to additional factors (such as attention and grouping). The term "informational masking" is generally used for this excess masking (Kidd *et al*., 2007). The advantages of this approach are its simplicity and that its results can be expressed in the same variables (scores, SNRs, etc.) as the classic masking studies. It is however still difficult to map the current results on those obtained in studies on attention and grouping because the focus has been more on varying acoustic parameters than on manipulating attention or grouping more directly.

ATTENTION

For anyone who has read Cherry's (1953) paper, it is clear that his interest in the CP effect was not driven by audibility issues. He wanted to know to what degree listeners could focus their attention on one talker while ignoring other auditory input. The "shadowing" paradigm that he developed, in which subject hear target speech (which they have to repeat) through one ear and competing sounds through the other ear, has been used widely to study (auditory) attention, especially during the 1950s and 1960s. It provided input for the various "filter theories" of attention that were developed in that period (Broadbent, 1958; Treisman, 1960; Deutsch and Deutsch, 1963). These theories state that the attentional system works like a filter that can selectively let through or reject input, but they disagree on the degree (and "level") of processing of the information that is ultimately suppressed. The most extreme views are represented by the "early filter" theory of Broadbent (1958), who stated that only low-level features are processed unattended, and the "late filter" theory of Deutsch and Deutsch (1963), who claimed that both attended and unattended information are processed at least up to a semantic level. The finding that listeners can easily focus on the target ear and can be unaware of quite drastic manipulations of the nontarget signal supports Broadbent's view. The well-known results of Moray (1959), showing that about one-third of the listeners noticed it when their own name was presented to their nontarget ear, supports the ideas of Deutsch and Deutsch (1963). A view in-between the two extremes was brought forward by Treisman (1960) in her "attenuation" theory, which proposes that all nontarget information is indeed processed up to a high (semantic) level but with reduced resources and thus more slowly.

It is interesting to delve a bit deeper into the evidence supporting the various theories because it sheds light on the functioning of auditory attention and its interaction with grouping. Some key studies in this field were carried out by Cowan, Wood and colleagues (reviewed in Cowan and Wood, 1997), who took the trouble to replicate a number of classic shadowing studies using modern methods for stimulus control and data analysis. A striking finding presented in Wood and Cowan (1995) is that listeners who noticed that their own name was embedded in the nontarget stimuli showed a clear deficit in their shadowing performance over a period of several seconds following the occurrence of the name. No performance drop was found for listeners who did not notice their name or were presented with other names. Similar results were observed for listeners presented with a fragment of time-reversed speech embedded in normal speech. These findings indicate that high-level processing of target and nontarget information does not occur independently, and strongly suggests that attention was in fact transiently drawn to the nontarget ear when the oddball stimulus occurred. Because the performance deficit occurs relatively late it probably was not caused by the processing of the stimulus itself but perhaps by resources required for storage of the event in long-term memory.

The conclusion that these findings validate Broadbent's (1958) "early filter" theory is however premature. Some – relatively high-level – analysis must be going on to be able to detect and identify stimuli like one's own name and Wood and Cowan's

(1995) results do not show performance deficits at the moment the stimulus occurs. A recent priming study by Rivenez *et al*. (2006) supports the view that such analysis of unattended input indeed takes place. They used a high presentation rate (2 words per second) and – in one experiment – also a secondary task to discourage attention switches to the nontarget ear. Nevertheless they found that response times to a target word were lowered when the same word was presented directly before it to the nontarget ear. It is also interesting to recall the results of an old study by Treisman (1960) who found that when subjects were presented with target and nontarget passages that suddenly switched, some of them reproduced a number of nontarget words directly after the switch and most of these subjects did not notice that they had in fact reproduced words from the wrong ear. The number of errors decreased when the predictability of the passages was reduced. These results suggest that grouping not only precedes allocation of attention but that a grouped stream may summon attention in a bottom-up fashion.

The problem with most of the previous work on auditory attention is that attention was, in fact, poorly controlled. This was already brought forward by Holender (1986) who reviewed a large number of shadowing studies and concluded that all the proposed evidence for semantic processing of unattended information can in fact be explained by uncontrolled switches of attention. Although improved paradigms have been developed more recently (e.g. Rivenez *et al*., 2006), we still know very little about switches driven by bottom-up capture of attention. One way to get more grip on this, is to combine behavioural experiments with direct measures of brain activity using for example fMRI, ERPs or MEG data. A study by Escera *et al*. (2003) provides an example of such results. They used a paradigm where irrelevant auditory stimuli were presented to subjects performing a visual task and found evidence that novel sounds embedded in the auditory stream always summoned attention (reflected by N1 and novely-P3 enhancements) also when these sounds, in addition to being irrelevant, were not identifiable for the subjects (i.e. were perceived as meaningless noise bursts). Another example of such research, in this case focusing more on top-down attentional processes, is a study by Teder-Salejarvi and Hillyard (1998), who collected both behavioural responses and ERPs from listeners performing a spatial attention task. Standard stimuli as well as infrequent target stimuli were presented from 7 loudspeakers divided over an arc of 54˚ in front of the listeners, who only had to respond to targets delivered by one, attended, speaker. Only few errors were made, indicating that attention had a relatively narrow spatial focus (within ±9˚) and the same spatial tuning was found for ERPs occurring around 300 ms post-stimulus. Interestingly, earlier ERPs had a less narrow focus, suggesting that the attentional filter operates in different stages with increasing sharpness.

The question how auditory attention works (and can be modelled) remains important for understanding the CP effect because it is clear that audibility of a speech signal may be necessary but is not sufficient for its ultimate selection for high-level (conscious) processing. Also, the study of exogenous effects is relevant because attention to a certain talker in a CP environment is not only determined by voluntary control but

also by bottom-up capture. Such capture occurs more easily by unexpected sounds and it is not unlikely (though no research has been done that supports this) that part of the difficulties hearing impaired listeners have with hearing aids are caused by attentional effects.

It is a pity that interest in auditory attention has declined markedly since the 1960s and that models of attention (e.g. Treisman, 1991) are almost exclusively based on visual research and thus perhaps poorly matched to the specific properties of the auditory system. Fortunately, the recent interest in multisensory processing has created a "revival" in research on the auditory system and also considers novel issues such as the question to what degree attention is supramodal.

GROUPING

While most audibility studies have taken attention for granted and vice versa, both research lines have paid relatively little attention to a processing stage that also forms a necessary step in converting incoming sound into information that can proceed to a semantic level. Remember that the essence of the CP environment is that the listener is presented with a superposition of speech sounds. Given that the target signal is (mostly) audible in this babble, and that the attentional system is prepared to select this signal for further processing, an additional process is required to make sure that the correct signal is extracted and followed over time. The idea (propagated among others by Bregman, 1990) is that this process groups sounds making use of both acoustic properties of the sounds and knowledge of the specific sound properties. The grouping consists of segregation of sounds that seem to originate from different sources, and the forming of a stream of input ("streaming") attributed to the target source. According to Bregman (1990), the acoustic properties form the basis of *primitive* grouping, which supposedly takes place pre-attentively, while knowledge of the sound properties is associated with *schema-based* grouping that occurs under voluntary attentional control.

A great number of studies (see Bregman, 1990; Darwin and Carlyon, 1995) provide information about cues that are used for primitive grouping. Many of these cues are based on principles that were already proposed about eighty years ago by the Gestalt psychologists, who tried to describe perceptual organization in vision. Some of these principles are: proximity, similarity, good continuation, and common fate. An example of the proximity principle in audition is that a succession of tones will be grouped together more easily when their frequency difference is small than when it is large (Van Noorden, 1977). Good continuation has been for example studied in speech perception, where it was shown that listeners could more easily hear the order of vowels presented in a repeating cycle (which indicated that they were grouped) when these had smoothly varying formant frequencies instead of sudden frequency jumps (Dorman *et al*., 1975). Sounds probably have a common fate when they occur simultaneously and/or originate from the same spatial location. In speech perception, the most important features to which Gestalt principles can be applied are temporal

envelope, fundamental frequency, formant frequency, and spatial location (Bregman, 1990). The principles seem to reflect a single overarching property of the auditory system, namely that it is highly tuned to natural sounds occurring in our environment. In such natural sounds the pitch and natural resonances do not abruptly change – in speech these changes are for example governed by the physiological constraints of our vocal chords and vocal tract. It is also unlikely that a sound source suddenly jumps from one to another spatial location or that sounds occurring exactly at the same time are not generated by the same source.

In order to understand the primitive grouping process, many trade-off studies have been done, where some features are tuned to promote segregation, while others are set to stimulate fusion. Two important results are worthwhile mentioning. One result is that the auditory system does not strictly segregate sounds originating from different spatial location. This was already shown in a seminal study by Broadbent and Lagefodet (1957; but see Darwin and Hukin, 2004), who found that when the two lower formants of a synthetic vowel were presented to different ears, they were nevertheless fused. More recent research looking specifically at the contribution of interaural time delays to segregation also found that these do not provide a strong basis for segregation (Culling and Summerfield, 1995; Darwin and Hukin, 1999). A related result (discussed below) is the finding of Brungart and Simpson (2002) that subjects listening to speech embedded in competing speech can be highly distracted by irrelevant speech presented to their nontarget ear. The second – rather counterintuitive – result is that two sounds that are segregated can nevertheless contribute to a combined percept. Cutting (1976) observed this when he presented the two formants of a synthesized CV syllable dichotically while introducing a difference in fundamental frequency. The listeners identified the correct syllable in more than 75% of the trials (chance performance was 33%), independent of frequency difference, but nevertheless almost always reported hearing two sounds even when the frequency difference was only 2 Hz (see Darwin, 1981 for similar results). These results indicate that auditory grouping operates in a flexible way, disregarding evidence promoting segregation when other cues make it likely that the sounds nevertheless originate from a single source.

Schema-based grouping is more difficult to address than primitive grouping, in particular in speech perception, because many schemata can be used, at the lexical, syntactical and semantic levels, as well as those based on contextual information, and because speech is such an overlearned stimulus that top-down effects are not easily separated from automatic processes. In addition, schema-based grouping is viewed as a top-down process and therefore dependent on attention. Nevertheless, some studies give insight into the interaction between primitive grouping, schema-based grouping and attention. The above example of duplex perception found by Cutting (1976) probably represents competition between primitive and schema-based grouping – the former promoting segregation and the latter fusion. Another example is a study Darwin (1975; cited in Bregman, 1990) who replicated the experiment of Treisman (1960) discussed above, in which listeners shadowed speech that suddenly switched

to their nontarget ear. Darwin varied the switching of the pitch contour of the speech and of its semantic content independently and found that both factors introduced shadowing errors, but that pitch contour seemed to guide attention more strongly than semantic content. Other studies that have looked at top-down effects in auditory grouping have used informational masking paradigms and will be reviewed below (Brungart and Simpson, 2004; Freyman *et al*., 2004). Basically these studies show that top-down information such as a semantic prime or knowledge of the interfering speech can strongly influence grouping.

It may be tempting to see auditory grouping mainly as preprocessing which lines up streams of information so that they are ready to be selected by attention for high-level processing. This one-way view is evidently incorrect given that top-down information can affect the grouping process itself. A more probable view is that primitive grouping actually performs such "dumb" preprocessing at a preattentive level and that attention does not just select one segregated stream but employs various stored schemata to correct and refine the grouping, creating a more plausible stream as output. Both views suppose that primitive grouping occurs pre-attentively, and this is in fact supported by a different line of research which has studied the mismatch negativity (MMN) – an ERP component elicited by a deviance occurring in a stream of standard stimuli (e.g. Schröger, 2005). The MMN occurs independent of whether attention is focused on the stimuli and is therefore thought to reflect automatic, preattentive processing. Although the MMN research has not looked at all primitive grouping cues, it is clear that at least a number of them (e.g. frequency, location, envelope) can elicit MMNs when used to differentiate deviants from standard stimuli.

INFORMATIONAL MASKING

As indicated above, informational masking can be defined as excess masking that cannot be explained by reduced audibility (i.e. by energetic masking). It should be noted (see also Kidd *et al*., 2007) that this definition has limitations because energetic masking itself is not fully understood. Informational masking has been primarily studied using tonal stimuli (see Watson, 2005; Kidd *et al*., 2007 for reviews). Some basic findings of this research are that when target tones are presented together with (or embedded in) a sequential stream of unpredictable distracter tones, thresholds for detecting changes in the target tone are increased dramatically (up to 50 dB). The increase in thresholds also occurs when energetic masking of the target tones is minimized, which means that it is mostly due to the informational component. The unpredictability of the interferers (in other words: the masker uncertainty) is seen as one important factor causing informational masking. Another factor is target-masker similarity (Durlach *et al*., 2003a; Durlach *et al*., 2003b).

A dominant paradigm in the study of informational masking in speech, used primarily by Brungart and colleagues, makes use of the so-called coordinate response measure (CRM) corpus. This corpus consists of phrases of the form "Ready <call sign> go to <colour> <number> now". There are eight call signs (e.g. "Baron"), four colours and eight numbers, resulting in 256 possible phrases, spoken by four male and four

female talkers. Listeners are asked to attend to the talker uttering a specific call sign and to respond both the colour and the number of the following phrase. It appears that they already have a hard time responding correctly when only one interfering talker is mixed with the target talker. While performance is not too different from that for speech in noise when the interferer has a different sex than the target talker, scores for same-sex interferers are about 20% lower and, when the interfering and target speech is spoken by the same talker, about 40% lower (Brungart, 2001). Importantly, the maximum decrease occurs at SNR around 0 dB, where intelligibility for speech in modulated or steady-state noise is already close to perfect. At this SNR, no loudness cues are available that could help the listeners to disentangle target from interfering speech.

The CRM test is so effective in revealing informational masking because target and interfering phrases not only have an identical structure (making grouping schemes based on syntax, semantics and context ineffective) but are also temporally aligned (which hampers streaming based on primitive cues). Follow-up research by Brungart and colleagues has revealed some remarkable and interesting effects. Brungart and Simpson (2002) observed that when listeners not only heard the target-interferer combination in one ear, but also a second interferer in the other ear, performance decreased substantially, even when this second interferer was presented at a much lower level. Interestingly, the performance reduction mainly consisted of confusions between the phrases presented to the same ear. This breakdown of the spatial attentional filter tells us something about the interaction between attention and auditory grouping. Elaborating on the view presented above that attention employs schemata to refine grouping, it seems that Brungart and Simpson's (2002) conditions disturb this process, possibly by exogenously drawing attention to the wrong (contralateral) stream. Primitive grouping by location may then still be able to prevent intrusions from words presented contralaterally, but less resources are available for segregating the ipsilateral streams.

Another relevant study was published by the same authors two years later (Brungart and Simpson, 2004). In this study, masker uncertainty was varied by "freezing" (holding constant) the interfering talker and/or the content of the interfering phrase. It appeared that knowledge of the interfering talker did not help, while knowledge of the phrase did. A related study, in which target uncertainty was varied, was conducted by Ericson *et al*. (2004). They found that, in this case, knowledge of the target talker did improve performance, as well as knowledge of the target position (see also Kidd *et al*., 2005). These finding indicate that primitive grouping and attention do not treat target and interferers in a "symmetric" way. A priori knowledge of primitive properties (such as voice characteristics) can improve selection of target information but cannot enhance suppression of nontarget information. This suboptimal behaviour may well be due to limited resources, which are apparently used to optimize the grouping and selection of target information and not to maximize suppression of information that should be rejected. Such a strategy, in fact, makes sense because a listener generally knows more about the sound he/she wants to listen to than about all

possible interfering sounds. The finding that top-down knowledge of the interferer phrase leads to better performance may have several explanations. One is that when listeners have erroneously selected the interferer phrase they can correct themselves by "replaying" the verbal input using their phonological loop. Another is that listeners can already correct themselves while listening to the phrases at the moment they hear the call sign.

There is another line of research into speech-based informational masking that is also worth mentioning. Freyman and colleagues have used a paradigm where target speech is presented together with interfering speech and where a perceived difference in location between target and interferer is introduced by presenting a delayed or advanced copy of the interfering from the side (Freyman *et al*., 1999). The speech stimuli were nonsense sentences spoken by female talkers and were likely to induce informational masking because same-sex talkers were used for target and interferer (minimizing primitive grouping cues) and because the sentences lacked semantic content (making schema-based grouping more difficult). Adding the copy of the interferer caused a clear performance improvement. This was attributed to a release of informational masking occurring because the spatial difference allowed for better segregation of target and interferer. In a follow-up study, Freyman *et al*. (2004) investigated whether informational masking could be overcome when listeners received information that allowed them to improve their attentional focus on the target phrase. He presented a prime either acoustically (the first part of the sentence, not containing the target word) or visually (a printed version). The results showed clearly reduced masking for both types of priming. The two studies elegantly demonstrate that release of informational masking can be induced both by primitive cues (location) and schema-based processes (knowledge of the target).

DISCUSSION

In this review I have argued that audibility is only one of the factors relevant for understanding the CP effect and that we also need to look at results from research into three other phenomena: attention, grouping and informational masking. As I have tried to point out, these phenomena are highly interrelated and it is not easy to tease apart their separate effects on speech intelligibility. A useful view already proposed by Bregman (1990) is to discriminate between pre-attentive primitive grouping and schema-based grouping which only occurs within attended stimuli and is thus an effortful process drawing on limited resources. Schema-based grouping uses top-down information to refine the grouping and create the most plausible output. In doing that, it can overrule primitive grouping, even if that was based on "compelling" information such as differences in source location.

Although informational masking may be related to various kinds of processing limitations, research has primarily focused on the performance deficits that occur when grouping fails. Such masking can thus be caused either by a lack of primitive grouping cues and/or by the inability to apply appropriate schemata. This interplay between primitive and schema-based grouping is illustrated in many informational

masking studies. Typically, speech material specifically designed to make grouping difficult is used and then a release from informational masking is induced either by introducing primitive grouping cues (such as differences in spatial location or fundamental frequency) or by providing top-down information that enable schema-based grouping (e.g. Arbogast *et al*., 2002, Darwin *et al*., 2003; Freyman *et al*., 2004).

Attention is linked to the process of schema-based grouping but we still know relatively little of this relationship. For example, it seems probable that less attentional resources are required for selecting and applying simple (overlearned) schemata than for schemata that use complex semantics or context, but this has, as far as I know, not been investigated. A subject that has been disregarded largely in auditory research is the interplay between exogenous and endogenous attention. Virtually all studies (exceptions are cueing studies – see e.g. Spence and Driver, 1997) have only tried to manipulate endogenous attention and have ignored possible exogenous switches of attention. Precisely these attention switches may be responsible for a number of effects that have either been attributed to "unattended" processing or have not yet been understood.

REFERENCES

Arbogast, T., Mason, C., and Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086-2098.

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 331-342.

Bregman, A. S. (**1990**). *Auditory scene analysis: the perceptual organisation of sound* (MIT Cambridge, MA: Bradford Books).

Broadbent, D. E. (**1958**). *Perception and communication* (New York: Pergamon).

Broadbent, D. E., and Ladefoged, P. (**1957**). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. **29**, 708-710.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: a review of speech intelligibility in multiple-talker conditions," Acustica **86**, 117-128.

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

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

Brungart, D. S., and Simpson, B. D. (**2004**). "Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," J. Acoust. Soc. Am. **115**, 301-310.

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

Cowan, N., and Wood, N. L. (**1997**). "Constraints on awareness, attention, processing and memory: Some recent investigations with ignored speech," Consc. Cogn. **6**, 182-203.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785-797.

Cutting, J. E. (**1976**). "Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening," Psychol. Rev. **83**, 114-140.

Darwin, C. J. (**1975**). "On the dynamic use of prosody in speech perception," in *Structure and Process in Speech Perception: Proceedings of the Symposium on Dynamic Aspects of Speech Perception*, edited by A. Cohen and S. G. Noteboom (Springer-Verlag, New York).

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset time," Q. J. Exp. Psychol. A **33**, 185-208.

Darwin, C. J. (**2008**). "Listening to speech in the presence of other sounds," Phil. Trans. R. Soc. B. **363**, 1011-1021.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *The Handbook of Perception and Cognition* Vol. 6 Hearing, edited by B. C. J. Moore (London: Academic Press), pp. 387-424.

Darwin, C. J., and Hukin, R. W. (**1999**). "Auditory objects of attention: the role of interaural time-differences," J. Exp. Psychol.: Hum. Percept. Perform. **25**, 617-629.

Darwin, C., Brungart, D., and Simpson, B. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913-2922.

Darwin, C. J., and Hukin, R. W. (**2004**). "Limits to the role of a common fundamental frequency in the fusion of two sounds with different spatial cues," J. Acoust. Soc. Am. **116**, 502-506.

Deutsch, J. A., and Deutsch, D. (**1963**). "Attention: Some theoretical considerations," Psych Rev. **70**, 80-90.

Dorman, M. F., Cutting, J. E., and Raphael, L. J. (**1975**). "Perception of temporal order in vowel sequences with and without formant transitions," J. Exp. Psychol.: Hum. Percept. Perform. **1**, 121-129.

Durlach, N. I., Mason, C. R., Kidd Jr, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003a**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984-2987.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd Jr, G. (**2003b**). "Informational masking: counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**, 368 - 379.

Ericson, M. A., Brungart, D. S., and Simpson, B. D. (**2004**). "Factors that influence intelligibility in multitalker speech displays," J. Aviation Psych. **14**, 311-332.

Escera, C., Yago, E., Corral, M.-J., Corbera, S., and Nuňez, I. (**2003**). "Attention capture by auditory significant stimuli: semantic analysis follows attention switching," Eur. J. Neurosc. **18**, 2408-2412.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90-119.

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

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

Holender, D. (**1986**). "Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal," Behav. Brain Sci. **9**, 1-66.

Houtgast, T., and Steeneken, H. J. M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66-73.

Kidd Jr, G., Arbogast, T. L. Mason, C. R., and Gallun, F. J. (**2005**). "The advantage of knowing where to listen," J. Acout. Soc. Am. **118**, 3804-3815.

Kidd Jr, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2007**). "Informational masking," in *Auditory Perception of Sound Sources,* edited by W. A. Yost, A. N. Popper and R. R. Fay (Springer US), pp. 143-189.

Moray, N. (**1959**). "Attention in dichotic listening: Affective cues and the influence of instructions," Q. J. Exp. Psych. **11**, 56-60.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988-3997.

Rivenez, M., Darwin, C. J., Guillaume, A. (**2006**). "Processing unattended speech," J. Acoust. Soc. Am. **119**, 4027-4040.

Schröger, E. (**2005**). "The mismatch negativity as a tool to study auditory processing," Acustica **91**, 490-501.

Spence, C. and Driver, J. (**1997**). "Audiovisual links in exogenous covert spatial orienting," Percept. Psychophys. **59**, 1-22.

Teder-Sälejärvi, W. A., and Hillyard, S. A. (**1998**). "The gradient of spatial auditory attention in free field: An event-related potential study," Percept. Psychophys. **60**, 1228-1242.

Treisman, A. (**1960**). "Contextual cues in selective listening," Q. J. Exp. Psychol. **12**, 242-248.

Treisman, A. (**1991**). "Search, similarity, and integration of features between and within dimensions," J. Exp. Psychol.: Hum. Perc. Perf. **17**, 652-676.

Van Noorden, L. P. A. S. (**1977**). "Minimum differences of level and frequency for perceptual fission of tone sequences ABAB," J. Acoust. Soc. Am. **61**, 1040-1045.

Van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index" J. Acoust. Soc. Am. **123**, 4514-4523.

Vom Hövel, H. (**1984**). "Zur Bedeutung der Übertragungseigenschaften des Außenohres sowie des binauralen Hörsystems bei gestörter Sprachübertragung," dissertation (RWTH Aachen).

Watson, C. (**2005**). "Some comments on informational masking," Acustica **91**, 502-512.

Wood, N. L., and Cowan, N. (**1995**). "The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel?," J. Exp. Psych.: Learn. Mem. Cogn. **21**, 255-260.

Zurek, P. M. (**1990**). "Binaural advantages and directional effects in speech intelligibility," In: *Acoustical Factors affecting Hearing Aid Performance,* edited by G.A. Studebaker and I. Hochberg (Boston: Allyn and Bacon), pp. 255-276.