Investigating Auditory Reverse Hierarchy Theory (RHT) in Perception and Perceptual Learning Processes

Thesis submitted for the degree of “Doctor of Philosophy”

By

Mor Nahum

Submitted to the Senate of the Hebrew University

February/2009
Investigating Auditory Reverse Hierarchy Theory (RHT) in Perception and Perceptual Learning Processes

Thesis submitted for the degree of “Doctor of Philosophy”

By

Mor Nahum

Submitted to the Senate of the Hebrew University

February/2009
This work was carried out under the supervision of

Prof. Merav Ahissar & Prof. Israel Nelken
Investigating Auditory Reverse Hierarchy Theory (RHT) in Perception and Perceptual Learning Processes

In the past few decades, the hierarchical nature of auditory processing in non-human primates and in humans has been gradually revealed. It is now commonly accepted that auditory information is processed along the auditory pathways in a fine-to-crude hierarchical manner, as in other sensory systems. Although the functions of the various stages of this hierarchy, particularly at its cortical levels, are not well-understood, the auditory hierarchy can be crudely divided into lower and higher representation levels. Lower level representations reliably and selectively encode fine spectro-temporal acoustic features. Thus, at the brainstem level of the superior olivary complex (SOC), inputs from the two ears are compared within narrow frequency bands and with microsecond resolution. In contrast, cortical levels integrate across time and frequency and form more abstract, spectro-temporally broader and perhaps ecologically-relevant categories. One of these higher representation levels is believed to be the phonological representation, which underlies human speech perception.

Along this hierarchy, there is convergence of auditory information across different dimensions. Thus, for example, two very different speech utterances, whose low-level acoustical features are very different, may belong to the same high level category, as the same syllable (e.g. different instances of /ba/). The fact that fine acoustic differences may be encoded at low levels of the auditory hierarchy, but not at its high levels, raises the question whether such differences can be utilized for perceptual discriminations even when they are lost at the high representation levels. Although this question addresses the basic relations between the information available to the auditory system and our ability to use it for conscious perception, it is still unresolved.

Two main different views were related to this question in the past. One, the “unlimited view”, proposed that all the information represented in the low-levels of the auditory system is available for perception. Thus, perception fully utilizes low-level information, which is limited only by the variability of the neuronal responses at lower
representation levels. The unlimited view was successful in accounting for human performance in a broad range of psychoacoustical tasks, but did not take into account considerations that are prevalent under more natural listening conditions, as listener's interest and attention and stimulus uncertainty. According to the “limited capacity” view, utilization of low-level information depends on the behavioral context. Thus, under demanding behavioral conditions, which pose extra load on the limited attentional and perceptual resources, as those with high uncertainty or high attentional load, performance is lower than expected based on the information available at the low representation levels.

In the current work I propose a third alternative, which is based on The Reverse Hierarchy Theory (RHT). RHT had been successful in accounting for the discrepancy between the accurate spatial information available at lower levels of the visual hierarchy and its limited use in fast perceptual discriminations. I now applied its concepts to the auditory domain. It proposes that only high-level representations (such as phonological representations in the auditory domain) are immediately accessible to perception and therefore underlie our initial perceptual experience. Low-level representations (such as high-resolution interaural time differences) are accessible only under specific, privileged, conditions. Hence, in general, low-level information would be available for perceptual discriminations only when high-level representations are essentially equivalent to the low-level representations, whereas high-levels are fully accessible. Therefore when the equivalence fails, perceptual discriminations can fully benefit from low-level resolution only under special behavioral conditions, which allow a search backwards along the "reverse hierarchy" for tracking the most informative low-level population. This search can be successful only when concurrent access to high-levels is not required, and when the informative low-level populations can be detected using consistent cross-trial information.

In order to critically test the predictions of these three views, I measured the utilization of low-level binaural information when discriminating speech sounds embedded in noise, in a variety of behavioral conditions. I compared the behavioral results to those obtained from an “ideal listener” model of the early auditory system, which assumes full utilization of low-level information. Binaural cues, i.e. comparison of fine temporal discrepancies between the signals reaching the two ears in order to better
extract the signals from noise, are known to be calculated relatively early in the auditory system, within narrow frequency bands at the brainstem SOC. The ability to efficiently use them (termed here “binaural benefit”) can therefore serve as a marker for utilization of low-level information.

In Study 1 I have manipulated the similarity of the words within the two-word set (either phonologically similar or phonologically different), the task (either identification or semantic association) and the protocol by which the relevant low-level binaural information was presented (either consistent or randomly interleaved throughout the block). Task difficulty was kept constant, by equating the absolute thresholds obtained under the different conditions. According to the unlimited view, performance should match that of the ideal listener in all conditions. According to the limited capacity view, utilization of low-level information should remain constant under all conditions, since task difficulty is equated. According to RHT, the distinction between phonologically similar and phonologically different words, which is irrelevant for the two other views, is crucial. Phonologically different words have distinctive low-level (acoustical) representations and distinctive high-level (phonological) representations, and therefore equivalent low-level and high-level representations; RHT therefore predicts full use of low-level information, regardless of task or protocol requirements. In contrast, phonologically similar words have distinctive low-level representations, but at the phonological level their representations will have a high degree of overlap. In this case, extracting the more abstract phonological categories causes partial loss of low-level information at the higher representation levels. RHT thus predicts that the benefit from low-level information should match the performance predicted by ideal listener models only in specific protocols that allow backward search to find the informative low-level populations.

The overall pattern of results of Study 1 could be parsimoniously accounted only by the RHT. In case of phonologically different words, for which the task can be resolved using the default high level representations, “ideal listener” levels of utilization was obtained under all conditions. In contrast, efficient use of low-level information was obtained for the phonologically similar words only under a consistent trial-by-trial presentation of the low-level information and when the task did not require concurrent
high-level access to semantic stores. These conditions presumably enabled the backtracking to the low levels of the hierarchy, which is needed in order to perform this fine discrimination.

In **Study 2** I asked whether manipulating task difficulty would affect utilization of low-level binaural cues, as predicted by the *limited capacity view*. To this end, I have manipulated task difficulty in two different forms: by changing the *cognitive load* (i.e. the set size of the words to-be-discriminated) and by changing the *perceptual load* (i.e. the success level required to complete the task). This allowed to test whether posing additional load, either cognitive or perceptual, reduces the efficiency of utilization of low-level cues. The results were in line with the ones found in Study 1: manipulating task difficulty *per se* did not affect utilization of low-level binaural information. Instead, utilization of low-level cues was optimal for the phonologically different words and sub-optimal for the phonologically-similar words (unless the binaural protocol was consistent), irrespective of task difficulty. As with Study 1, the results can be accounted by the RHT, whereas the predictions of the *limited capacity view* or the *unlimited view* do not match the overall pattern of results.

**Study 3** was aimed to test the specific predictions of auditory RHT in perceptual learning. Specifically, I asked whether, as RHT predicts, the implicit variability within the protocol by which the relevant low-level information is presented affects the ability to utilize these cues for discrimination between phonologically similar words, when long-term training is involved. Unlike most perceptual learning studies, the binaural benefit tool used here enables to test the effect of implicit variability within the low-level cues, since subjects’ task is to identify the words rather that to directly identify the binaural configuration. The results obtained from three groups of listeners, each trained using a different amount of variability within the low-level binaural cues throughout practice, showed that, in line with *RHT* predictions, full utilization of low-level information following long-term practice is possible only when the relevant information is consistent or temporally-ordered (in a predictable manner) throughout the block, but not when it is randomly presented in an unpredictable manner. Randomly mixing the relevant cues improves performance, but does not enable access to the relevant low-level cues.
I conclude that the RHT, originally developed to explain visual perception, can account for patterns of information utilization in the auditory domain as well. Moreover, several other auditory phenomena could all be accounted for using the framework provided by RHT. Specifically, context effects in speech perception, the phonemic restoration effect and informational masking, can all be interpreted as resulting from an immediate default access to high representation levels at the cost of low-level resolution, which can be only obtained under privileged conditions.

The fact that similar processing principles underlie the two important sensory systems, the auditory and the visual, makes it possible to draw important analogies between the two. The results of my research suggest that similar defaults and tradeoffs, of reversed relations between the hierarchy of processing and perceptual accessibility, underlie sensory processing in the visual and auditory modalities. Thus, this work is one step towards achieving a parsimonious unified framework of hierarchical sensory processing in general.

Future studies would need to first establish the empirical relevance of RHT to other auditory parameters, such as pitch, spectral processing and sound localization. In this work I have tested only one form of low-level cue, the binaural benefit, but many other low-level cues exist and need to be tested for matching the predictions of the theory. Brain imaging tools, as magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI) that are now intensively used in many human brain studies, could be used to functionally test the timeline of activation of high level and low level brain areas, and to see if it follows the reverse pattern suggested by RHT principles. The first step towards confirming auditory RHT predictions for perceptual learning has also been made in this work. However, other important and non-intuitive predictions for perceptual learning, as the importance of the training context to generalization of learning to less predictable environments, should be critically tested. If established, they could be used in the future to improve existing applied software tools designed to enhance brain function.
Chapter 1: Introduction ................................................................. 3 - 22
Section 1.1: Hierarchical information processing in the auditory modality .......... 3 - 10
Section 1.2: The unlimited access view & the limited capacity account .......... 11 - 12
Section 1.3: The Reverse Hierarchy Theory (RHT) ........................................ 13 - 17
Section 1.4: Binaural benefits for speech perception: a tool for studying information processing in the auditory modality ......................................................... 18 - 19
Section 1.5: Summary, research goals & experimental questions ......................... 20 - 22

Chapter 2: Materials & Methods .......................................................... 23 - 39
Section 2.1: Behavioral experiments ......................................................... 23 - 32
Section 2.2: “Ideal listener” simulation ....................................................... 33 - 39

Chapter 3: Results ................................................................. 40 - 76
Section 3.1: Manipulating task requirements (Study 1) ..................................... 42 - 54
Section 3.2: Manipulating difficulty (Study 2) ............................................... 55 - 65
Section 3.3: Perceptual learning under low-level variability patterns (Study 3) .... 66 - 76

Chapter 4: Discussion ................................................................. 77 - 98
Section 4.1: Summary of results ............................................................... 77 - 79
Section 4.2: Binaural benefits and the Reverse Hierarchy Theory ................. 80 - 82
Section 4.3: RHT & perceptual learning: the importance of the training protocol .. 83 - 89
Section 4.4: Auditory RHT: a broader perspective ........................................ 90 - 94
Section 4.5: Auditory RHT: caveats, future directions & practical implications .... 95 - 98

References ..................................................................................... 99 - 113
Appendices


Chapter 1: Introduction

Section 1.1: Hierarchical information processing in the auditory modality

The auditory hierarchy: low Level vs. high level processing

A large body of anatomical, physiological and functional data that has been accumulated over the past decades suggests that processing of auditory information, much like that of information processing in other modalities (as the visual and the somatosensory modalities, Felleman and Van Essen, 1991; Ungerleider and Mishkin, 1982) is implemented in a hierarchical manner. Going up the auditory hierarchy, processing shifts from local aspects of the stimulus at low levels to more global ones at high levels (Griffiths et al., 2000; Kaas and Hackett, 1998; Kaas et al., 1999; Rauschecker, 1998a; Rauschecker et al., 1995; Romanski et al., 1999; Warren and Griffiths, 2003; Wessinger et al., 2001; Zatorre and Belin, 2001).

This processing hierarchy can therefore be crudely divided into lower and higher processing levels. The analysis of auditory information which is done at the lower levels, up to and including the midbrain station of the inferior colliculus (IC) is relatively well-understood (Figure 1.1.1). Lower levels are considered responsible for extracting basic spectro-temporal features of the auditory signal, as the frequencies comprising it and the temporal patterns of the stimulus within short time scales. Lower level representations therefore reliably and selectively encode fine spectro-temporal acoustic features. For example, at the brainstem level of the Superior Olivary Complex (SOC), inputs from the two ears are compared in the Medial Superior Olive (MSO) within narrow frequency bands and with microsecond resolution (Batra et al., 1997a; Batra et al., 1997b; Blauert, 1997; Jiang et al., 1997; Palmer et al., 2000; Yin and Chan, 1990). Based on these observations, several researchers developed quantitative models that simulate, with relatively good accuracy, the analysis conducted at these lower levels, up to the brainstem level of the SOC (Bleeck et al., 2004; Patterson et al., 1995).
**Figure 1.1.1 Lower levels of the auditory pathway** A highly schematic diagram of the bilateral auditory pathway. The main pathways and nuclei, from the cochlea and up to the auditory cortex, are featured. The levels up to the inferior colliculus (IC) are considered low levels, and processing in these levels is relatively well-understood.

However, the analysis performed at higher levels of the hierarchy, particularly at the cortical levels, is far less understood. Nevertheless, anatomical and functional studies in non-human primates and in humans revealed several distinct cortical areas dedicated to processing of auditory information, that are mainly hierarchically connected. At the early 70’s, Merzenich & Brugge (1973) showed that the primary auditory cortex at the superior temporal plane in composed of two tonotopically-arranged areas (right plot of Figure 1.1.2). At the same time, Pandya & Sanides (1973) made an anatomical distinction
between a granular core area of a primary cortex and a surrounding belt area. In a series of seminal papers in the late 90’s, Kaas & Hackett (1998; 2000; 1999) systematically mapped the different areas of the primate auditory cortex. The authors characterized, based on cytoarchitectonic and functional features, at least four different, hierarchically-organized processing levels, each containing several segregated sub-regions: the primary-like “core” areas that receive their input from the thalamus; surrounding lateral and medial “belt” areas, that receive input from core areas; a parabelt area on the dorsal plane of the superior temporal gyrus; and higher-level areas in the superior temporal sulcus and the frontal lobe (Figure 1.1.2). The nature of the hierarchy with respect to frequency representations was roughly described as follows. Neurons in the primary core areas are tonotopically-arranged, have relatively narrow tuning curves and respond best to pure tones. Neurons in the belt and pare-belt areas respond to a wider range of frequencies and respond better to noise bands than to pure tones (see also Rauschecker, 1998b; Rauschecker and Tian, 2000; Rauschecker et al., 1995). In the past few years researchers have also managed to identify distinct high-level areas that process pitch (Warren and Griffiths, 2003), spectral features (e.g. Metherate et al., 2005) and sound movement (e.g. Griffiths et al., 1998). These findings suggest that a hierarchical nature of representations also characterizes cortical stages, though the nature of "mid level" representations had not been explored yet.

Studies of the human auditory system used mainly imaging techniques as fMRI (functional Magnetic Resonance Imaging) and PET (Positron Emission Tomography), and found similar characteristics in the nature of cortical hierarchy (Belin and Zatorre, 2000; Davis and Johnsrude, 2003; Griffiths et al., 1998; Griffiths and Warren, 2004; Scott and Johnsrude, 2003; Warren and Griffiths, 2003; Warren et al., 2005; Warren et al., 2002; Wessinger et al., 2001; Zatorre and Belin, 2001). For example, Wessinger and colleagues (2001) used fMRI to show similar core-belt organization in the human auditory cortex: while core areas (that are located in Heschl’s gyrus, HG, in humans) responded best to pure tones, surrounding belt areas showed a larger response to band-pass noise bursts (BPN) with the same carrier frequency (Figure 1.1.3). Other studies have shown some evidence for a distinction between “what” and “where” pathways (Arnott et al., 2004; Scott, 2005; Warren and Griffiths, 2003; Zatorre and Belin, 2001;
Zatorre et al., 2002), similar to that found in the visual modality (Ungerleider and Mishkin, 1982).

**Figure 1.1.2 The Auditory Hierarchy** Left: cortical and sub-cortical connections of the primate auditory system; Right: Lateral view of the macaque cerebral cortex; Core, belt and parabelt areas are marked (adapted from Kaas and Hackett, 2000).

Although the overall picture is far from being clear, the converging evidence suggest that more abstract tasks engage higher cortical areas. Higher areas are believed to integrate across time and frequency channels to create spectro-temporally broader, more abstract categories of ecologically-meaningful auditory objects (Chechik et al., 2006; Las et al., 2005; Nelken, 2004; Nelken and Ahissar, 2006; Wang et al., 2005). Indeed, tuning curves in the cortex are wider than those found in the periphery (Winer et al., 2005) and in the Medial Geniculate Body (MGB; Moshitch et al., 2006). There is also some evidence that selectivity for interaural time differences (ITDs) in the cortex is context-dependent (Moshitch et al., 2005). Several studies further suggest subsequent
convergence to longer temporal units in the cortex (Ulanovsky et al., 2003), and therefore imply a temporal hierarchy as well (Griffiths et al., 1998).

**Figure 1.1.3 Hierarchical processing of auditory information in the human cortex, revealed by fMRI** Bandwidth-specific data from 3 subjects. Areas shown in blue represent higher activation by pure tones (PT), yellow represents higher activation by band-pass noises (BPN) and green represents overlap between PT and BPN. It can be seen that PT stimuli activate only the core region, while the surrounding belt areas are only activated by BPN bursts (adapted from Wessinger et al., 2001).

It is important to note that the division between low and high stages of the hierarchy is probably not categorical. Recent evidence that the primary auditory cortex (A1) responds in a highly non-linear, context-dependent manner (e.g. Nelken, 2004; Nelken and Bar-Yosef, 2008), suggest that it is already part of the more abstract, higher level stages. In comparison to the visual modality, A1 may be the "homologue" of area V4 rather than the primary visual cortex (area V1), based on the number of synapses from the peripheral organ, and on the fact that bilateral ablation does not induce deafness (Baran et
al., 2004), but deficits which are more similar to agnosias; Such agnosias are similar in nature to results obtained from ablations of midlevel cortical areas in the visual hierarchy (e.g. Fort et al., 2002).

**Hierarchical processing of speech in humans**

Several lines of research have shown that processing of *speech* signals in humans is also hierarchical. Functional imaging studies of the perception of speech and other complex stimuli in humans provide further evidence for a hierarchy of processing that extends beyond core regions, through the human equivalent of belt, parabelt and beyond (Binder, 2000; Binder et al., 2000; Binder et al., 1997; Davis and Johnsrude, 2003; Demonet et al., 1992; Giraud et al., 2000; Griffiths et al., 2001; Hickok and Poeppel, 2007; Jancke et al., 2002; Price et al., 1996). Whereas speech-specific responses are not seen in A1, a region of left superior temporal gyrus (STG) that is lateral to A1 has recently been shown to be sensitive to language-specific phonological structure (Jacquemot et al., 2003, Blue region in Figure 1.1.4). In a pathway extending anterolaterally from HG (Heschel's gyrus, see Figure 1.1.2), in what is probably the human equivalent of the parabelt, left-lateralized activation is observed in response to stimuli with the acoustic features of phonetic cues (Binder et al., 2000) and to intelligible speech (Scott et al., 2000), but also bilaterally to non-speech stimuli with complex spectrotemporal structures that are analogous in complexity to speech signals, such as frequency modulated (FM) tones, harmonic tones and sounds with changing spectral structure (Scott and Johnsrude, 2003; Wise, 2003, Lilac regions in Figure 1.1.4).

Activation specific to intelligible speech, both single words and sentences, has been reported in the left anterior superior temporal sulcus, in areas beyond intrinsic auditory cortex (Scott et al., 2000; Scott and Wise, 2004, Purple areas in Figure 1.1.4). Cortical areas in posterior middle temporal regions and anterior temporal lobe regions (Binder et al., 2000; Binder et al., 1997; Davis and Johnsrude, 2003; Jancke et al., 2002) may process semantic information. In addition, aspects of verbal working memory have been found to be associated with left posterior STS (Hickok et al., 2003) and supramarginal
gyrus activations (Jacquemot et al., 2003, Pink regions in Figure 1.1.4): this might relate to the need for transient representations in auditory memory that encode the temporal dimension (Scott, 2005).

Figure 1.1.4 Functional responses to speech and candidate stream of processing in the human brain. The lateral surface of the human brain, the colored regions indicate broadly to which type of acoustic signal each temporal region (and associated parietal and frontal region) responds. Regions in blue show a specific response to language-specific phonological structure. Regions in lilac respond to stimuli with the phonetic cues and features of speech, whereas those in purple respond to intelligible speech. Regions in pink respond to verbal short term memory and articulatory representations of speech (adapted from Scott, 2005).

Two main characteristics of the cortical hierarchy: feedback connections and convergence-divergence patterns

Two important and well-documented characteristics of this (and other) processing hierarchy are relevant for our subsequent discussion. The first is the massive feedback connectivity from higher to lower levels of the hierarchy (Bajo and Moore, 2005; Bajo et al., 2006; Maunsell and van Essen, 1983). These massive connections do not seem to strongly affect receptive field properties at the level of single cells, and their functional importance is not well understood.
The second characteristic is its convergence-divergence pattern. This combined pattern of connectivity yields a pattern of specificity and generalization at high-level representations. On the one hand, physically different auditory stimuli, whose low-level acoustical representations are very different, may belong to the same high level category. Presumably, the pattern of bottom-up convergence produces generalization across different instances within the same category. For example, the same tune is perceived by most listeners as the same irrespective of the scale in which it is played. This generalization is traded off with the ability to represent fine within-category distinctions at the high representation levels. On the other hand, the pattern of divergence is such that physically similar stimuli, whose low-level representations are similar, may belong to different high level categories.

This convergence-divergence pattern is particularly relevant to speech perception, in which lower level representations correspond to acoustic representations while the relevant higher level representations are the phonological representations that underlie human speech perception (Binder, 2000; Binder et al., 1997; Davis and Johnsrude, 2003; Hickok and Poeppel, 2007). At the higher levels, two very different acoustic stimuli may be perceived as the same syllable, or at least as belonging to the same high level category (e.g. two speakers pronouncing the syllable /ba/), whereas two acoustically similar sounds may instantiate two separate phonological categories, perceived as quite different (e.g. /ba/ and /wa/ pronounced by the same talker; see Nelken and Ahissar, 2006).

Convergence, or integration, of information across basic spectro-temporal dimensions suggests that fine details are not well-represented at higher levels. Given the difference between high and low-level representations, the relative degree of accessibility of these levels to perception, has crucial impact on what we can perceive. Thus, perception may be limited not only by the information in the system as a whole but also by the hierarchical level at which this information is represented. For example, can we "hear" minute within-category differences which are represented at low but not at high levels due to convergence? Although these questions address the basic relations between the information available to the auditory system and our ability to use it for conscious perception, they are still unresolved. The following sections will discuss three different views regarding this question.
Section 1.2: the unlimited access view & the limited capacity account

The unlimited access view: “ideal listener” levels of performance

Most psychoacoustic studies assume that all available information is accessible. In this view, the only bottleneck to performance is the total information that is available within the auditory pathways. These accounts of human auditory performance were dominated by the methodologies proposed by Green & Swets (1966) and by Siebert (1965; 1968); These methodologies link the output of the (relatively well-understood) low-level representations and psychoacoustic performance with a ‘central processor’ (representing higher processing stations), which optimally utilize low-level information ("ideal listener"). According to this view, perception fully utilizes low-level information, which is limited only by the internal noise in the neuronal responses (Durlach et al., 1986; Green and Swets, 1966; Gresham and Collins, 1998; Pfafflin and Mathews, 1966).

Indeed, results obtained from human listeners in a broad range of psychoacoustic tasks have been successfully accounted for using ideal listener models. These models usually assume two basic processing stages: a low-level neuronal representation of the input, which encapsulates all the information believed to be available at the level of the brainstem (Patterson et al., 1995), and a subsequent decision-making stage (Green and Swets, 1966), which performs statistically optimal decisions based on the full array of low-level activity. Such models successfully accounted for a range of auditory behavioral phenomena, including spectral shape discrimination (Dai and Green, 1993; Dai et al., 1996; Green et al., 1992), comodulation masking release (CMR; Verhey et al., 1999), masking (Delgutte, 1995) and binaural tone detection tasks (Bernstein and Trahiotis, 2003; Bernstein et al., 2006). In partial support of this approach, Shackleton and coworkers (2003) demonstrated that single inferior colliculus neurons in the guinea pig, used optimally, can resolve interaural differences that are as small as those detected by humans.

However, the above-mentioned studies were usually designed to probe best performance, and as such employed simple behavioral tasks, featuring simple auditory stimuli that were presented in a consistent manner, with as little cross-trial uncertainty as
possible. Factors as attention, uncertainty and predictability were not taken into account and this view (termed here “the unlimited view”) may therefore be limited in its ability to account for results obtained under more realistic listening conditions.

The limited capacity account: the use of information depends on task difficulty

In parallel to the conditions that were analyzed with ideal listener models, other studies focused on the impact of demanding task conditions, or excessive attentional load, on performance. These studies showed that under demanding behavioral conditions, performance is lower than expected based on the information available at the low representation levels (Hafter and Saberi, 2001; Lee et al., 1999; Muller-Gass and Schroger, 2007; Treue and Maunsell, 1999; Yi et al., 2004). Though most of these studies were conducted in the visual modality, there are also some compelling examples in the auditory domain. Particularly strong illustrations were provided by masking studies in which the stimuli are designed so that the low-level representations of the target and the masker do not overlap. Still, performance of many listeners is substantially degraded by the masking stimulus, due to uncertainty or confusion between signal and masker, indicating poor use of low-level information under certain aspects of uncertainty (“informational masking”, Brungart, 2001; Durlach et al., 2003a; Freyman et al., 2001; Kidd et al., 2002; Shinn-Cunningham and Ihlefeld, 2004).

Recent conceptualization of such attention-focused studies (e.g. Huang-Pollock et al., 2005; Lavie, 2005; Lavie et al., 2004; Schwartz et al., 2005; Yi et al., 2004) introduced a flexible aspect to the traditional term of attentional load. These updated concepts of “limited capacity” suggest that under low attentional load, low-level information can be fully utilized, whereas under high load, the perceptual system can only process a portion of the relevant low-level information, due to its limited capacity. The term “load” is not accurately defined, but is intuitively associated with task difficulty. Thus, according to the limited capacity account, as long as task difficulty remains the same, the ability to utilize low-level information should not change.
Section 1.3: The Reverse Hierarchy Theory (RHT)

In the current work I propose a third alternative, based on the Reverse Hierarchy Theory (RHT, Ahissar and Hochstein, 1997; Ahissar and Hochstein, 2004; Hochstein and Ahissar, 2002). RHT was originally developed to address the relations between hierarchical processing and perception in the visual modality. Here I claim that the concepts and basic tenets of RHT are applicable to the auditory modality as well. In the following sections I explain the basic concepts of RHT for perception and for perceptual learning.

RHT and conscious perception: immediate “vision at a glance”

RHT proposes that, by default, rapid conscious perception is based on high level representations alone. For example, visual scenes are quickly perceived since they are represented at high levels of the visual hierarchy. We hear phonological contrasts since they are represented at high-levels of the auditory and speech perception hierarchy. Lower level representations are not immediately accessible to our conscious perception (Figure 1.3.1). Explicit conscious perception is therefore initially based on higher cortical areas ("vision at a glance"), and can access low levels for additional details only in a gradual top-down manner ("vision with scrutiny").

This simple assumption implies that our immediate conscious perception reflects only the information stored at higher levels. Assuming that high levels are global, abstract, and represent the "gist" of ecologically-relevant objects and events, it is only this gist that will be immediately experienced. Thus, when we see a house, we immediately tag it as a house, but are not able, without further scrutiny (i.e. without the reverse-hierarchy return to low-level representations), to accurately experience (and consciously notice) its fine details. From a physiological perspective, this constraint stems from the convergence-divergence pattern described above: Activation of a high-level population denoting a specific perceptual category may be the result of many, not necessarily similar, low-level activation patterns, to which we have no immediate access. From an ecological
perspective, this limitation is a byproduct of the need to generalize across different instances of the same object or even of similar objects.

![The visual Reverse Hierarchy Theory](image)

**Figure 1.3.1 The visual Reverse Hierarchy Theory (RHT)** RHT proposes that the forward hierarchy, from simple features at low-level areas, which gradually generalize to form higher level representations of abstract objects and categories, acts implicitly, with explicit perception beginning at high-level cortex (initial “vision at a glance”). Later, explicit perception returns to lower areas via the feedback connections, to integrate into conscious vision with scrutiny the detailed information available there (adapted from Hochstein and Ahissar, 2002).

Does immediate categorization benefit from all low-level information, even though we are not initially aware of these details? RHT’s approach to this question is ecological (Ahissar et al., 2008). Local details retain separate representations along the bottom-up hierarchy when they discriminate between basic high-level categories (e.g. different phonemes); in this case they would be accessible and fully used. On the other hand, even crude information is merged (e.g. different acoustic instances of the same phoneme) when combined into the same category along the processing hierarchy. Thus, there is a perceptual cost for the convergence-divergence structure of the ascending sensory pathways in conjunction with the exclusive accessibility to high levels (Hochstein and Ahissar, 2002). These limitations are specifically revealed when within-category
discriminations, particularly non-practiced ones, are required: in this case high level representations do not suffice, and we need to backtrack along the hierarchy and retrieve the relevant information, which is not immediately accessible.

**Perception of low level details: top-down “perception with scrutiny”**

According to RHT, when local details discriminate between high-level categories (e.g. when two words are very different), discrimination at low and high representation levels is equivalent. But when local details do not discriminate well between the high-level categories (e.g., when words to-be-discriminated are very similar), perceptual discriminations typically fail to fully utilize its low-level information. It can fully benefit from low-level resolution only under special behavioral conditions, which afford direct access to the relevant low-level representations for making the specific discrimination. Such access is gained via a backward search along the “reverse hierarchy” to locate the most informative low-level inputs. This backward search requires time and/or repetitions. Under typical stimulation conditions, in which the signal-to-noise ratio (SNR) is not very good, a successful backward search requires repetition of the same stimuli in a sequence. Thus, when repeated stimuli are consistently used, the backward search can be successful in allocating a more informative input population, resulting in better behavioral performance. When the relevant information changes unexpectedly on a trial-by-trial basis, the relevant low-level population cannot be tracked, and performance would rely on higher level representations. This means that according to RHT, the *protocol* of stimulus presentation should have a large effect on the efficient use of low-level information. A successful backward search also has a cost: while tracking backwards, we lose access to high representation levels and consequently we temporarily lose the benefit of global and ecologically-meaningful perception, afforded by higher level representations (Ahissar et al., 2008). Thus, for example, we cannot have concomitant accurate within-category perception and semantic processing.
RHT & perceptual learning: gaining access and specificity to local stimulus attributes

RHT makes specific and testable predictions about perceptual learning as well. Perceptual learning is usually referred to as the improvement in performance of a perceptual task following a significant amount of practice (Goldstone, 1998). Recent studies suggest that this practice may induce changes at sensory areas rather than only at higher, decision-making stages. Over the past decades, such training was found to improve the performance of humans on a variety of auditory (Amitay et al., 2005; Amitay et al., 2006; Ari-Even Roth et al., 2003; Delhommeeau et al., 2005; Delhommeeau et al., 2002; Demany, 1985; Karmarkar and Buonomano, 2003) and visual (Adini et al., 2002; Ahissar and Hochstein, 1993; Fahle et al., 1995; Fiorentini and Berardi, 1980; Sagi and Tanne, 1994; Schoups et al., 1995; Schwartz et al., 2002) perceptual tasks.

An important factor in training outcome is the protocol of training itself. The impact of this choice, beyond that of the trained task and stimuli, had been systematically explored only recently. These studies found that when subjects were trained with a narrow range of stimuli (e.g. frequencies, speakers or contrasts) there was typically a large degree of improvement yet this improvement was specific to the trained range of stimuli (Adini et al., 2002; Fahle et al., 1995; Fiorentini and Berardi, 1980; Poggio et al., 1992; Shiu and Pashler, 1992; Watson et al., 1976). On the other hand, when a large degree of cross trial stimulus variability was introduced (e.g., roving different frequencies in a frequency discrimination task or different speakers within a block of speech identification task) a similar amount of training often resulted in no benefit (e.g. Yu et al., 2004) or in a small benefit on a broader range of stimuli than that trained (Adini et al., 2004; Amitay et al., 2005; Kuai et al., 2005; Otto et al., 2006). The intermediate case, in which learning of several parameters was done in a blocked or temporally-structured manner (e.g. first block with parameter 1, second block with parameter 2 etc.), usually yielded learning on the trained parameters, which was later transferred to the more variable, less predictable protocol in which the different parameters were roved (Adini et al., 2004; Spiegel and Watson, 1981; Yu et al., 2004; Zhang et al., 2008).
RHT provides a parsimonious account for the strong dependence of perceptual learning on the training protocol. According to RHT, the different protocols induce changes in different loci along the processing hierarchy. Since naïve performance is limited not by low-level information but by its loss at the easily accessible higher levels, improvement following training is contingent upon a gradual top-down gain of access to more informative, lower level populations (Ahissar and Hochstein, 2004). However, acquiring this low level resolution requires top-down “guidance”, which can be provided by consistent presentation of the relevant low-level cues, as detailed above. RHT proposes that training with consistent conditions enables modifications of low-level neuronal representations, and therefore sharpens resolution to local details with stimulus specificity of the lower level populations that underlie it. A similar account is proposed for several different cues presented in a blocked manner (as was found in Kuai et al., 2005; Zhang et al., 2008).

In contrast, training with a more variable protocol does not enable the relevant top-down guidance to low levels. Increasing stimulus variability, so that consecutive stimuli still belong to the same perceptual category, is expected activate different low-level populations in different trials but to activate the same high-level population in all trials. A successful backtracking search can not be achieved, since there is no consistent low-level population that can be gradually tracked. According to RHT, such learning will therefore be based on broad high level representations. It is therefore expected to show only a limited degree of improvement on the one hand, and substantial generalization across low-level features on the other hand (Ahissar and Hochstein, 1997; 2004; Ahissar et al., 2008; Hochstein and Ahissar, 2002; see also Fahle, 2005).
Section 1.4: Binaural benefits for speech perception: a tool for studying information processing in the auditory modality

In order to critically test the predictions of the three views presented above, it was crucial to find a parameter which is best represented at low levels and converges at higher levels of the auditory pathways. The attribute I used in my study is the benefit (in terms of dB) of binaural interactions.

Binaural benefits for identification of speech and tones in noise

In the studies I conducted for my thesis, the measure for utilization of low-level information was the ability to use fine temporal disparities (“binaural interactions”) between the inputs reaching the two ears in order to extract speech from noise (Durlach and Colburn, 1978; Yin and Chan, 1990). In ecological conditions, such time differences arise when the source of the noise has a different azimuth than the source of the speech. These differences have been shown to be highly efficient for the extraction of signals from noise of frequencies below 1500 Hz (Blauert and Cobben, 1978; Dubno et al., 2002). The assumption is that the effective use of the cues provided by spatial separation improves the functional signal-to-noise ratio (Dubno et al., 2002). These small time differences (< 1ms) are usually expressed as interaural phase differences, since they are calculated within narrow frequency bands (Bernstein and Trahiotis, 1992) at the brainstem Superior Olivary Complex (SOC, Jiang et al., 1997; McApline et al., 1996; McApline et al., 2000; Palmer et al., 2000).

Research regarding the benefit of binaural interactions for identification and detection of tones and speech stimuli masked by noise started over 60 years ago (Hirsch, 1948; Licklider, 1948). Typical studies use two different configurations of interaural phase differences, diotic and dichotic. The diotic configuration (also termed N₀S₀) contains no binaural phase information, since identical input (signal + noise) is presented to the two ears. The dichotic configuration introduces interaural phase information for separation between signal and noise. To maximize this information, the noise is identical in the two
ears, while the signal is added with opposite phase to the two ears (termed $N_0S_\pi$). The ability of listeners to utilize phase information is measured by the difference between dichotic and diotic thresholds (termed here binaural benefit).

Using these measures, Hirsch (1948) reported that the binaural benefit for tone detection in noise ($BMLD$, Binaural Masking Level Difference) is in the order of 12-15 dB. Licklider (1948) reported much smaller values for binaural benefit of word identification in noise ($BILD$, Binaural Intelligibility Level Difference), ~3-3.5 dB. Since then, many studies replicated and extended these basic findings (e.g. Carhart et al., 1968a; Carhart et al., 1967; Johansson and Arlinger, 2002; Levitt and Rabiner, 1967a; Levitt and Rabiner, 1967b), reporting a broader range of binaural benefits for speech stimuli, of up to 7 dB. Several models have been proposed to explain binaural effects in speech perception, based on full utilization of low-level binaural information within frequency bands (Levitt and Rabiner, 1967b; Zurek, 1993). These models were successful in explaining several observations, including BILDs being smaller than the corresponding BMLDs of the same base frequency, BILDs being larger for words with more syllables (Carhart et al., 1967) and more masking voices (Carhart et al., 1968b), and the increase of BILDs with decreasing intelligibility levels (Levitt and Rabiner, 1967a; Schubert, 1956).

In the current study I use the magnitude of the binaural benefit for the extraction of speech signals from noise as an index for the degree of use of low-level information in perception. Since binaural interactions are calculated early in the auditory system, and speech identification is based on high-level phonological representations, the magnitude of the binaural benefits can be used as a marker of the ability to use low-level interaural phase information in daily perceptual situations. Moreover, binaural interactions form an implicit marker: naïve subjects performing a task of identification of speech in noise are completely unaware of the presence or absence of the low-level binaural cues, although their discrimination thresholds are greatly reduced in the presence of these cues. In the next section I specify in detail my research goals and the way they were addressed in my study.
Section 1.5: Summary, research goals & experimental questions

In the current study I ask which aspects of stimulus processing are used for perception under a variety of task conditions and protocols. I compare the predictions of three hypotheses, the traditional ones, the unlimited view and the limited capacity account, respectively, and the Reverse Hierarchy Theory (RHT). I use binaural benefits to characterize the conditions under which information calculated at low levels of the hierarchy can or cannot be efficiently used for different perceptual tasks.

Utilization of low-level binaural information: the impact of task and stimulus protocol

In Study 1 I tested the predictions of the unlimited view, the limited capacity view and RHT for utilization of low-level information when extracting speech from noise, in a variety of behavioral conditions. I manipulated the degree of phonological similarity between the words in the test set (phonologically similar or phonologically different), the task required (word identification or semantic association) and the protocol by which diotic and dichotic configurations were presented (in blocked or randomly mixed manner). Task difficulty was retained fixed. In addition, I calculated the expected performance of an “ideal listener”, which has full access to low-level information and makes optimal statistical decisions (see Section 2.2). The unlimited view predicts ideal listener level of performance under all conditions. Since task difficulty (our measure for attentional load) is kept constant, the limited capacity view predicts similar utilization of low-level information under all conditions. According to RHT, full utilization of low-level information is afforded when low-level information is retained at high levels, as is the case with phonologically different words that have non-overlapping high-level representations. However, when discriminating between words with overlapping high-level representations, as in the case of phonologically similar words, attaining access to low levels is required for obtaining optimal binaural benefits.
The overall pattern of results could be parsimoniously accounted for only using the *RHT* logic. The results therefore show that in the auditory modality as well, the default levels accessed are indeed the higher levels. Backtracking to retrieve low-level information (e.g. when the words to-be-discriminated are phonologically similar) can only be done under privileged conditions, that include a consistent trial-by-trial presentation of that low-level information and a task which enables access to the relevant low-level population; Only these conditions afford “ideal-like” utilization of low-level information when it can not be accessed by default. The results of this study are detailed in *Section 3.1* of the *Results* chapter.

**Utilization of low-level binaural information: the impact of task difficulty**

In *Study 2* I asked whether manipulating task difficulty would affect utilization of low-level binaural cues, as predicted by the *limited capacity view*. I manipulated task difficulty in two different forms: by changing the *cognitive load* (i.e. the set size of the words to-be-discriminated) and by changing the *perceptual load* (i.e. the success level required to complete the task). These manipulations allowed me to test whether posing additional load, either cognitive or perceptual, reduces the efficiency of utilization of low-level cues, as proposed by the *limited capacity view*. The results were in line with the ones found in *Study 1*: manipulating task difficulty per-se did not affect utilization of low-level binaural information. Instead, utilization of low-level cues was optimal for the phonologically different words and sub-optimal for the phonologically-similar words (unless the binaural protocol was consistent), irrespective of task difficulty. These results could be naturally accounted for by the *RHT* logic, whereas the predictions of the *limited capacity view* or the *unlimited view* did not match the overall pattern of results. The relevant findings are detailed in *Section 3.2* of the *Results* chapter.
Gradually gaining access to low-level cues: The impact of the training protocol

In Study 3 I assessed the predictions of RHT for perceptual learning of speech in noise. I asked whether, as RHT predicts, the variability of low-level information would affect the ability to utilize binaural low-level information when long-term training is involved. Unlike most perceptual learning studies, the binaural benefit tool used here enabled me to test the effect of implicit variability within the low-level cues, since subjects’ task is to identify the words rather than to directly identify the binaural configuration.

The results obtained from three groups of listeners, each trained using a different protocol of cross-trial binaural variability, showed that, in line with RHT predictions, full use of low-level information following long-term practice is achieved only when the relevant information is consistent or temporally-ordered (presented in a predictable manner) throughout the block. Mixing binaural cues resulted in blockade of access to improved binaural information even following massive training. These results are detailed in Section 3.3 of the Results chapter.
Chapter 2: Materials & Methods

Section 2.1: Behavioral experiments

Participants

In Study 1, a total of 80 subjects were tested (mean age: 24 ± 3 y). In each of the four experiments (I–IV), 20 subjects were tested, 10 in each type of phonological similarity (phonologically different and phonologically similar). Thus, different subjects were tested in the different experiments and different conditions, to avoid effects of task and protocol learning.

In Study 2, a total of 40 subjects were tested (mean age: 24 ± 3 y): 25 subjects in Experiment I and 15 subjects in Experiment II. In this study, each subject performed all conditions in each experiment.

In Experiment I of Study 3 a total of 90 subjects were tested (mean age: 24 ± 3 y), 30 in each protocol (consistent, 1-1 and mixed). In Experiment II, 42 subjects (a subset of subjects from Experiment I), 14 in each group (consistent, 1-1 and mixed, see below) were trained. For the control groups, 30 additional subjects (mean age: 25 ± 4 y) were used as the naïve control group for the /dilen/ - /tilen/ contrast (10 in each protocol). Ten of these subjects (that performed the task in the consistent protocol, see below) were also trained on the /dilen/ - /tilen/ discrimination.

All subjects in all studies were undergraduate students at the Hebrew University of Jerusalem, and were native Hebrew speakers with normal hearing. All subjects gave their informed consent for participation in the study.

Stimuli
Stimuli in all experiments were disyllabic pseudo-words and familiar Hebrew words, all recorded by the same female speaker. Each word had two different instances. Overall root mean square (RMS) and duration were equated for all words. In Study 1, the same word pairs were used in all four experiments: 2 phonologically similar pairs, within which the difference was in a single phoneme (/tamid/ vs. /amid/, /shalom/ vs. /chalom/), and 2 phonologically different pairs, in which words differed in most phonemes (/tamid/ vs. /chalom/, /sikum/ vs. /amid/). In Experiment I of Study 2, the following ten-word sets were used: a set of ten Hebrew digits for the phonologically different condition (/efes/, /ahat/, /shtaim/, /shalosh/, /arba/, /hamesh/, /shesh/, /sheva/, /shmonet/ and /teshah/), and a set of ten familiar words, composed of five phonologically similar pairs, for the phonologically similar condition (/shalom/ vs. /chalom/, /tamid/ vs. /amid/, /banuy/ vs. /panuy/, /tmuna/ vs. /tluna/ and /shanim/ vs. /panim/). For the “set size 2” condition, one set of digits (4 /arba/ and 9 /tesha/) was used as the phonologically different pair, and a pair of similar words (/shalom/ vs. /chalom/) out of the list of ten words was used as the phonologically similar pair. In Experiment II of Study 2, pairs of phonologically similar (/barul/ vs. /parul/) and phonologically different (/dilen/ vs. /talug/) pseudo-words were used. In Study 3 I used the phonologically similar pair /barul/ vs. /parul/ and another pair for control, /dilen/ vs. /tilen/.

The masking noise in all studies was speech noise (Dreschler et al., 2001), played at a constant level of 66 dB SPL (sound pressure level) to both ears. The speech noise used has well-defined spectral and temporal characteristics similar to those typically found in real life speech signals; It was generated from live English speech from the EUROM database (Chan, 1995) in which a female speaker is explaining about the system of arithmetical notation. The original speech signal was split into bands with different frequencies; The bands were then scrambled (to preserve the modulation properties of the original signal but to be completely unintelligible), filtered again, scaled (so that they all have the same spectral density level) and added together to form one signal with a white spectrum and with the original modulation preserved in each of the three frequency-ranges. The resulting noise is a speech-spectrum shaped noise corresponding to female speech, which is representative of normal speech as it preserves both the spectrum and modulation of the speech.
The noise was always identical in both ears. Words were played in two different configurations: diotic \((N_0S_0)\), in which the word was added to the noise in-phase at both ears, and dichotic \((N_0S_\pi)\), in which the word was phase-inverted in one of the ears before it was added to the noise. The duration of the noise was 1.4 s, whereas the duration of the word was 0.8 s. Thus, the noise began 0.3 s before and ended 0.3 s after the word. All stimuli were digitally played by a TDT system III signal generator (Tucker Davis Technologies), and presented to listeners through HD-256 Sennheiser headphones.

**Tasks**

Three different types of tasks were used in the studies. All tasks were performed in a sound-attenuated booth.

**Two-word identification task**

In this task, on each trial, one of two possible words was presented over headphones, masked by noise, and the listener had to press the left/right button on the computer screen whose label matched the played word. Feedback was given after every button press: a positive feedback for correct responses (happy face) and a negative feedback for incorrect responses (sad face). Subjects were instructed to respond as accurately and as quickly as possible.

**Two-word semantic-association task**

In this task, on each trial, one of the two words was presented in noise. Immediately following the auditory presentation, a word was visually presented on the screen for 500 ms. Listeners had to decide whether the acoustically presented word was semantically related to the visually presented word. In each trial, the visually presented word was selected from a set of 20 different words, ten of which were semantically associated to one auditory word and ten to the other word. Subjects had to press the right button
(green: “match”) if it matched the auditory word and the left button (red: “no match”) if it did not. The feedback protocol was the same as for the two-word identification task. Subjects performing these experiments were given a short, 20-trial training session prior to the experiment. Subjects were instructed to respond accurately and quickly. We verified that they did so by measuring their RTs (from the end of the visual presentation until button press). Average RTs were calculated for the 75 trials comprising each assessment, and were further averaged across the diotic and dichotic binaural configurations.

**Ten-word identification task**

In this task subjects were required to discriminate between 10 words. Subjects heard on each trial one of the ten words, masked in noise, and were requested to orally report the word to the experimenter. The experimenter, who sat in the booth with the subject, pressed a green (“correct”) or red (“incorrect”) button following a correct or incorrect response, respectively. For this task, subjects were first given a short practice of 20 trials in which they had to correctly identify the words presented without any masking noise.

**Protocols**

**Protocol for measuring thresholds**

Thresholds for correct identification were measured in all studies using an adaptive staircase procedure (Levitt, 1971). In most experiments (excluding part of Experiment II of Study 2, see below), thresholds were measured using a 3 down–1 up adaptive staircase procedure, converging at 79.4% correct. In Experiment II of Study 2, the “60% correct” condition was measured using another up–down procedure, converging to 61.8% correct. In this method, signal level was decreased after at least two consecutive successes out of every three trials. Signal level was increased after any of the other five combinations of successes and errors out of every three trials.

The level of the masking noise was kept constant while the presentation level of the word was adaptively varied (see left panel of Figure 3.1.1, A). In Study 1, five different
step sizes were used, beginning at 3 mV and switching to smaller steps after every 3 reversals (3, 2, 1, 0.5, and 0.2 mV). Experiment I of Study 1 was initially run with slightly different step sizes (2, 1, 0.5, 0.3, 0.1 mV), but then re-ran with step-sizes similar to the other conditions; since the results were similar to the ones obtained initially, they are not reported here. Study 2 and Study 3 were run with step sizes: 2, 1, 0.5, 0.3, 0.1 mV. Each measurement block was composed of 75 trials for each binaural configuration (see below). Diotic and dichotic thresholds were calculated as the arithmetic mean of signal amplitude in the last five reversals. The binaural benefit was calculated as the difference (in decibels) between the measured diotic and dichotic thresholds (illustrated in Figure 3.1.1, A).

**Binaural protocols**

Three different protocols of presentation of the binaural configurations (diotic and dichotic) were used in the studies.

- **Consistent Protocol**: diotic and dichotic configurations were measured in different experimental blocks of 75 trials each, administered in immediate succession. The order of the sessions was counterbalanced between subjects. In this protocol the low-level binaural information was consistent throughout the block, i.e. either diotic or dichotic but not both at the same block.

- **Mixed Protocol**: diotic and dichotic configurations were randomly interleaved across the block, such that on each trial one of these binaural configurations was chosen uniformly at random. Each experimental block therefore contained both diotic and dichotic configurations and was consisted of 150 trials, 75 for each configuration. Adaptive thresholds for diotic and dichotic configurations were tracked separately throughout the assessment.

- **1-1 Protocol**: diotic and dichotic configurations were ordered throughout the block, such that the diotic configuration was presented on odd trials, and the dichotic configuration was presented on even trials. Each block therefore contained both diotic and dichotic configurations and was consisted of 150 trials, 75 for each configuration.
Adaptive thresholds for diotic and dichotic configurations were tracked separately throughout the assessment.

**Experimental Procedures**

*Study 1 (task & protocol manipulation)*

In this Study, the two-word identification task and the two-word semantic task were administered under two binaural protocols, *consistent* and *mixed*. Each subject was administered one assessment per word pair with each binaural configuration (i.e., 150 trials with each word pair). Each subject performed the same experiment twice, with two different word pairs. Both pairs were either phonologically similar (i.e. /tamid/ vs. /amid/ and /shalom/ vs. /chalom/) or phonologically different (i.e. /sikum/ vs. /amid/ and /tamid/ - /chalom/). The experimental design of *Study 1* is detailed in Table 2.1.1.

**Table 2.1.1 The experimental design of Study 1 (task & protocol manipulation)**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Task</th>
<th>Binaural Protocol</th>
<th>Stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Phonologically Similar</strong></td>
</tr>
<tr>
<td>I</td>
<td>Identification</td>
<td>Consistent</td>
<td>/tamid/ vs. /amid/ &amp; /shalom/ vs. /chalom/</td>
</tr>
<tr>
<td>II</td>
<td>Semantic</td>
<td>Consistent</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>Identification</td>
<td>Mixed</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>Semantic</td>
<td>Mixed</td>
<td></td>
</tr>
</tbody>
</table>

*Study 2 (difficulty manipulation)*

In Experiment I of *Study 2* (“cognitive load”), I used both the 2-word identification task and the 10-word identification task, with sets of phonologically-similar and phonologically-different words. This experiment was performed using only the *mixed*
binaural protocol. Each subject performed all conditions of the experiment, i.e. both set-size 2 and set-size 10 conditions, with both phonologically different and phonologically similar word pairs.

In Experiment II of Study 2 (“perceptual load”), the task was a 2-word identification task, performed with both phonologically similar (/barul/ vs. /parul/) and phonologically different (/dilen/ vs. /talug) pseudo-word pairs. This experiment was performed using both the consistent and mixed binaural protocols, converging to both 60% & 80% correct (using different staircase procedures, see above). Each subject (different subjects from those in Experiment I) performed both the 60% & 80% correct conditions, with both types of pseudo-word pairs. The experimental design of Study 2 is detailed in Table 2.1.2.

Table 2.1.2 The experimental design of Study 2 (difficulty manipulation)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Manipulation</th>
<th>Condition</th>
<th>Task</th>
<th>Binaural Protocol</th>
<th>Stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Set Size 2</td>
<td>Set Size 2</td>
<td>2-word</td>
<td>Mixed</td>
<td>/shalom/ vs. /tesha/</td>
</tr>
<tr>
<td></td>
<td>“Cognitive Load”</td>
<td></td>
<td>identification</td>
<td></td>
<td>/chalom/ vs. /arba/</td>
</tr>
<tr>
<td></td>
<td>Set Size 10</td>
<td></td>
<td>10-word</td>
<td></td>
<td>5 pairs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>identification</td>
<td>Mixed</td>
<td>10 digits</td>
</tr>
<tr>
<td>II</td>
<td>Success Level 60%</td>
<td>60%</td>
<td>2-word</td>
<td>Consistent</td>
<td>/barul/ vs. /dilen/</td>
</tr>
<tr>
<td></td>
<td>“Perceptual Load”</td>
<td></td>
<td>identification</td>
<td>Mixed</td>
<td>/parul/ vs. /talug/</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Study 3 (perceptual learning)
In the two experiments of Study 3, subjects performed the two-word identification task using a phonologically similar pseudo-word pair (/barul/ vs. /parul/). In Experiment I, each group of subjects performed one block of the task using a single binaural protocol: consistent, mixed or 1-1.

In Experiment II, 14 out of the 30 subjects in each group continued practicing on the task, using the same initial binaural protocol, for a total of 7 practice days. Usually, training was done in successive days or with a one day interval between sessions (with the exception of the weekend, which imposed a 2-day break). Ten of the 14 subjects in each group then underwent the testing phases detailed below. On day 8, each group performed the task with a different protocol: for the consistent and 1-1 group, the test protocol was mixed, whereas for the mixed group the test protocol was 1-1. On day 9 ("baseline"), subjects again performed the task in the original training protocol (i.e. the same protocol used for days 1-7). On day 10, subjects performed the task with the original protocol, but discrimination was required between a different pair of phonologically similar pseudo-words: /dilen/ & /tilen/. Throughout training and testing, all subjects performed 3 blocks per day. Each session took ~25 minutes.

As a control, a different group of subjects (30 subjects, 10 per protocol) performed separately the task of discrimination between /dilen/ & /tilen/. The 10 subjects that performed the task in the consistent protocol underwent additional training of 6 more days on discrimination between /dilen/ & /tilen/, similarly to the /barul/-/parul/ training. On day 8, subjects were tested on discrimination between /barul/ & /parul/. These data were used for comparison with the /dilen/-/tilen/ data collected on day 10 of testing. The experimental design of Experiment II of Study 3 is detailed below in Table 2.1.3.

Table 2.2.3 The experimental design of Study 3 (perceptual learning), Experiment II

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Day</strong></td>
<td>1-7</td>
<td>8</td>
</tr>
<tr>
<td><strong>Purpose</strong></td>
<td>Training</td>
<td>Another Protocol</td>
</tr>
</tbody>
</table>
**Data analysis**

**Study 1 (task & protocol manipulation)**

I used univariate analysis with between-subject factors of task (two levels: identification and semantic-association) and protocol (two levels: consistent and mixed), thus comparing results of Experiments I–IV. Binaural benefits and diotic thresholds were separately used as the dependent variables. Data analysis was performed separately for each word set (phonologically similar and phonologically different sets). Results were corrected using the Greenhouse-Geisser correction. Comparison of reaction times (RTs) between the relevant word pairs of the semantic-association task was performed using non-paired two-tailed Student *t*-tests.

**Study 2 (difficulty manipulation)**

In Experiment I, I used ANOVA with within-subject factors of set size (two levels: 2 and 10) and similarity (two levels: phonologically similar and phonologically different). In Experiment II, I used ANOVA with within-subject factors of percent correct (two levels: 60% and 80% correct) and similarity (two levels: similar and different), and a between-subject factor of protocol (two levels: consistent and mixed). Results were corrected using the Greenhouse-Geisser correction.

**Study 3 (perceptual learning)**
In Experiment I, I used Multivariate analysis with dependent factors of diotic threshold and binaural benefit, and an independent variable of protocol (3 levels: consistent, 1-1 and mixed). A post-hoc Scheffe contrast was performed to compare the 3 binaural protocols.

In Experiment II I used a separate analysis for each phase. For the training phase I used separate ANOVAs for diotic thresholds and binaural benefits, with within-subjects factor of days (7 levels) and between-subjects factor of protocol (3 levels). Post-hoc Scheffe contrast was used to discriminate between the 3 protocols. For the transfer across protocol phase I conducted a separate multivariate analysis for each protocol (2 dependent variables: diotic thresholds & binaural benefits), with an independent factor of state (3 levels for the variable protocol: naïve, after 1-1 training and after consistent training; 2 levels for the 1-1 protocol: naïve and after mixed training). For the mixed protocol, I additionally conducted a post-hoc Scheffe analysis for state. In addition, I used 2-tailed Students’ t-tests to compare performance between the last day of training (day 7) and the testing day (day 8). Performance was compared to the first block in the new protocol for each of the trained groups. For the transfer across pairs phase I performed a multivariate analysis (2 dependent variables: diotic thresholds & binaural benefits) with between-subjects factors of protocol (3 levels) and state (2 levels: naïve and trained). I conducted a post-hoc Scheffe contrast for the protocol variable. For the /dilen/ - /tilen/ learning (a separate group of subjects) I performed an ANOVA with factors of aurality (2 levels: diotic thresholds & binaural benefits) and of days (7 levels, day 1 – day 7).
Section 2.2: “Ideal listener” simulation

I used an ideal listener model to calculate performance given access to all low-level information. These calculations were used in Study 1 and Study 2, to test the predictions of the unlimited view. The model consisted of a peripheral stage ending with a binaural cross-correlator (roughly simulating the auditory system up to the level of the SOC), followed by an ideal listener under the assumption of additive Gaussian noise. In the following paragraphs I specify in detail the different stages of the simulation, including and up to the decision-making stage.

Creating the templates

Simulating auditory nerve fibers

The inputs consisted of the waveforms of the words used in our experiments (e.g. /barul/), masked by frozen speech noise of different levels (see Results Section). The first step was to simulate the representations of the inputs at low levels of the auditory processing hierarchy. Due to peripheral processing, the neuronal representations at this stage have less information than the raw acoustic signals; some modeling of the peripheral processing is therefore necessary in order to correctly fit human performance (e.g. Gresham and Collins, 1998). For each word, I simulated two configurations, diotic (N₀S₀) and dichotic (N₀Sₐ). In the diotic configuration, the inputs to the left and right ears were identical. In the dichotic configuration, the noise was the same for both ears, but the signal was inverted in the right ear before adding it to the noise. The activation of the peripheral frequency channels was calculated using the AIM software package (Bleeck et al., 2004), and included filtering, half-wave rectification, compression, and a low-pass filter with a corner frequency of 1200 Hz. In order to control for the effects of low-pass filtering on binaural benefits, I repeated the simulations with a low-pass filter at 500 Hz following half-wave rectification (as suggested by e.g. Bernstein and Trahiotis, 1996). Since the results using the 500 Hz cutoff frequency were not substantially different from those using 1200 Hz as the cutoff frequency, they are not reported here. I simulated the
activation of 32 channels in the frequency range of 100-4000 Hz for 800 ms (the standardized duration of each word). The auditory nerve activity patterns for the left and right ears are displayed in Figure 2.2.1, A & B for two phonologically-similar pseudo-words /barul/ (Figure 2.2.1A) and /parul/ (Figure 2.2.1B), respectively, at a signal to noise ratio of +10dB. These words were used in Experiment II of Study 2. For each word, two patterns are presented: one for the left ear and one for the right ear. The two patterns for the two ears for each word are very similar, except for a phase shift in the fine structure of the activity at the low frequency channels (which cannot be seen at the resolution of the figure). The rapid fluctuations of the activation levels in the low-frequency channels are due to the pitch of the voiced segments in these words. Since the noise was frozen in the experiment, I created an activity pattern for each SNR level used in the experiment. I denote the left and right activity patterns before averaging as ANF_L(f,t) and ANF_R(f,t), respectively, with implicit dependence on the noise level and on the binaural configuration.

The activity patterns were used to generate “energy” and “binaural correlation” templates for each token and each SNR level. The absolute values of the simulated responses were smoothed by averaging over consecutive 10 ms periods, resulting in 80 time samples for each channel. The energy template (f_ener) was the sum of the left and right smoothed activity patterns. In parallel, the auditory nerve activity patterns were cross-correlated, also in 10 ms segments. The binaural correlation template (f_bc) consisted of the correlation coefficient between the two ears at lag 0, for each of the 80 segments of 10 ms composing the stimulus and for each frequency channel. Lag 0 is the lag at which the reduction in binaural correlation is largest in the $N_0S_\pi$ configuration, and is therefore the most informative lag.

Figure 2.2.1C (as well as Figure 2.2.1D) shows the energy (left) and binaural correlation (right) templates for two words (same words used in panels A & B, Figure 2.2.1). The energy template is essentially a smoothed version of the neural activity patterns. The binaural correlation template shows high correlation (close to 1) in those frequency/time combinations that are dominated by the noise (low-level energy activation). In frequency/time combinations that are dominated by the speech signal, the addition of the signal with reversed phase to the two ears causes a large decrease in
binaural correlation. Thus, in the diotic configuration only $f_{ewr}$ is informative (the binaural correlation template is identically 1 for all words). On the other hand, in the dichotic configuration, the correlation template $f_{cc}$ is different for different words, and therefore may be used to enhance the discrimination performance.

Figure 2.2.1 “Ideal Listener” Simulation

(A & B) The auditory nerve activity patterns for the left and right ears for the phonologically similar pseudo-words /barul/ (A) and /parul/ (B) at a SNR of +10 dB. Patterns are calculated at 32 frequency channels between 100–4,000 Hz, at 80 time bins of 10 ms each. Note the difference in patterns, despite the similarity of the words. (C & D) The energy (left of each panel) and binaural correlation (right of each panel) templates for the same pseudo-words, calculated from the auditory nerve pattern in panels (A) and (B). (E) Euclidean (left) and DTW (right) distances calculated for a pair of phonologically different (blue bars) and a pair of phonologically similar (red bars) words. Distances are normalized by standard deviations. Euclidean distances are essentially equal for both pairs, whereas the DTW distance is much smaller for the phonologically similar pair (see text). The use of the DTW distance was needed in order to account for the higher discrimination thresholds of phonologically similar word pairs.
Making optimal decisions

Since the external noise was frozen and was included in the generation of the templates, the only limitation on performance of the ideal listener comes from internal noise. Presumably, the internal noise is due to multiple sources in different levels of the nervous system; it is therefore modeled as a Gaussian noise. My assumption is that the internal Gaussian noise was independent at different times, energy channels and correlation channels. Its variance had a fixed value at all energy channels and a different (fixed) value at all correlation channels.

For uncorrelated Gaussian internal noise, the statistically optimal decision in the case of consistent binaural protocol consisted of selecting the template that was closer (in the least squared difference sense) to the incoming sound (Green and Swets, 1966). In the mixed binaural protocol, the optimal decision required a more complicated decision variable, but which was still essentially based on the least squared difference distance between the incoming sound and the templates (see below). Thus, the main effort in the simulation consisted of determining the distance between an incoming sound and the templates. The actual performance depended on the amount of noise.

I used dynamic time warping (DTW, Myers and Rabiner, 1981; Rabiner and Juang, 1993) to calculate the distance function between the input and the template words. The need for using DTW arose since regular Euclidian distance between the word templates did not correspond to the phonological similarity between them. Thus, using Euclidian distance, words which differed in a single phoneme (e.g. /barul/ - /parul/) were not necessarily closer to each other than words which differed in many phonemes (e.g. /dilen/ - /talug/). DTW is an algorithm that optimally warps the duration of acoustic segments in order to reach the smallest possible distance between the two words. Thus, for example, in Figure 2.2.1 the two phonologically-similar words /barul/ and /parul/ differ substantially when compared frame-by-frame throughout their nominal 800 ms duration. This is due to the fact that in Hebrew, voice onset time in /p/ occurs roughly at the stop release time while for /b/ there is substantial pre-voicing. Therefore, the word token for /barul/ contains a rather long initial closure period during which a voice bar is apparent.
On the other hand, /parul/ essentially starts at voice onset time. As a result, the /arul/ part of /parul/ occurs earlier and is somewhat longer than the corresponding part of /barul/. Computing acoustic distance on a frame-by-frame basis would result therefore in a distance between /barul/ and /parul/ that is as large as the distance between two phonologically-different words, such as /dilen/ and /talug/. The DTW algorithm solves this problem by aligning the acoustically similar frames in the two words. As a result, most of the distance between the two words results from differences in the initial consonant (/b/ vs. /p/). Using the DTW metric made it possible to use different templates of the same stimulus, with various temporal relations between their phonetic segments, all available to the optimal decision of the ideal listener (as done in the experiments).

The effect of the use of DTW on the distance between words is illustrated in Figure 2.2.1E, where the distance between phonologically-different (left) and phonologically-similar (right) word pairs is computed with and without the use of DTW. Whereas the acoustic, low-level distance is essentially the same when directly applying Euclidian distance to the word pairs, it is smaller for the phonologically-similar pair when the two words are aligned with DTW. Concomitantly, without DTW, the ideal listener predicted substantially better discrimination thresholds than the experimentally measured ones in the phonologically-similar cases. Note that although the DTW algorithm enabled a better fit of the thresholds, it did not affect the predicted binaural benefits, since it affected both diotic and dichotic thresholds similarly, compared to using a Euclidian metric.

In practice, given two time series $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_M\}$, where each frame $x_i$ and $y_j$ is a length $k$ vector, the DTW distance between them is the shortest distance between the two sequences, allowing the replication of frames (artificially lengthening the appropriate segment). The DTW distance is computed using dynamic programming: the function $D(i,j)$ is calculated using the recursive formula

$$D(i,j) = d(i,j) + \min([D(i-1,j), D(i-1,j-1), D(i,j-1)]),$$

where $d(i,j)$ is the Euclidian distance between frames $x_i$ and $y_j$. This formula recursively computes the minimum distance between the subsequences $\{x_1, \ldots, x_i\}$ and $\{y_1, \ldots, y_j\}$. The final DTW distance is then $D(N,M)$.

In this case, $X$ was the energy and correlation matrices calculated from the incoming input word, with the additional internal noise. Since each template in our simulation
MATERIALS & METHODS

consisted of two matrices \( f_{ener} & f_{cc} \), we ran the algorithm on the combined distances for both energy and correlation templates, after normalizing their values by dividing each matrix by the variance of its corresponding internal noise. I computed the DTW distance between \( X \) and \( Y_1 \) (template of word 1) and between \( X \) and \( Y_2 \) (template of word 2) (for larger set sizes, see additional explanations below). The input word was then recognized as the word corresponding to the template with the smaller DTW distance. In the actual experiment, each word had two tokens (i.e., two recordings per experiment; see Materials & Methods). For each run of the simulation, one token of each of the two words was selected for generating the sequence of incoming word to be discriminated, and was then compared to all other tokens within the set.

The distance calculations were performed using only the channels \( X \) time combinations in which the SNR was larger than 0 dB when the overall SNR was 10 dB. Although the channels \( X \) times combinations to use were selected based on the energy templates, they were also used for the correlation templates.

The variance of the internal noise in the monaural channels was set to a value that fitted one of the experimental conditions (/barul/ - /parul/, used in Experiment II of Study 2, at -12 dB SNR where performance was 79.4%, measured under the consistent binaural protocol). Similarly, the variance of the internal noise in the binaural correlation channels was set to the value that yielded similar success rate in the dichotic configuration of the same words (~-22 dB SNR). These values were then used for all other conditions. The estimated thresholds for Study 1 (all experiments), Study 2 Experiment I and Study 2 Experiment II (phonologically-different word pair) were all computed with these estimates for the variances. Thus, for Study 1, I used the estimated variance values obtained from /barul/ - /parul/ to compute ideal listener performance for discrimination between all 4 word pairs used in Experiments I & III (/tamid/ - /amid/, /shalom/ - /chalom/, /chalom/ - /tamid/, /sikum/ - /amid/), under both consistent and mixed binaural protocols. In Experiment I of Study 2 I used the estimated variances to account for discrimination between 10 words (both phonologically-similar and phonologically-different; see main text), as well as between another pair of phonologically- different words (/arba/ - /tesha/). In Experiment II of Study 2 I used the same variances to calculate performance for the same pair (/barul/ - /parul/) under a different protocol.
(mixed) and under a different difficulty level (60% correct). Similarly, I calculated performance under the same conditions in this experiment for another pair of phonologically-different words (/dilen/ - /talug). I therefore used the variance values set from a single word pair in a single condition to explain thresholds obtained for a variety of word pairs, 10-word sets, different protocols and different difficulty levels.
Chapter 3: Results

Study 1 and Study 2 were aimed to critically test the predictions of these three views, the unlimited view, the limited capacity view and RHT. In order to do that, in a variety of behavioral conditions I measured the utilization of low-level binaural information, i.e. the ability to use fine temporal cues between the inputs reaching the two ears, in order to extract speech from noise (Durlach and Colburn, 1978; Yin and Chan, 1990). Thus, stimuli were presented in two configurations: diotic, which contains no phase information and dichotic, in which the noise is identical in the two ears, while the signal is added with opposite phase to the two ears; this configuration maximizes phase information for separation between the signal and the noise (Hirsch, 1948; Licklider, 1948). The ability of listeners to use the low-level phase information was measured by the difference between dichotic and diotic thresholds (binaural benefit). In addition, I calculated the expected ideal listener performance in each of these conditions. According to the unlimited view, performance should match ideal listener thresholds in all conditions. According to the limited capacity view, utilization of low-level information should depend on task difficulty (measured by diotic thresholds; Study 2; Table 3.1.1) and should not change when task difficulty remains the same (Study 1; Table 3.2.1).

In order to assess RHT predictions, I used two types of word sets (disyllabic Hebrew words and non-words embedded in speech noise, Dreschler et al., 2001), composed of phonologically different words (differing in most phonemes, as /tamid/ & /chalom/) and phonologically similar words (differing in only one phoneme, as /tamid/ and /amid/), respectively. This distinction is irrelevant for the ability of listeners to use low-level cues according to either the unlimited or the limited capacity views, since the use of phonologically similar or different words by itself does not change the difficulty of the task (which is adjusted adaptively) and does not change the amount of low-level information (as is shown below). However, RHT makes specific predictions for these two cases. Phonologically different words have distinctive low-level representations (since they are acoustically different, and hence the low-level speech parts comprising them are different) and distinctive high-level representations (since they are phonologically
different, and hence high-level representations are distinct enough). Thus, phonologically different words have the property that low-level and high-level representations are equivalent, and therefore, RHT predicts full use of low-level information, regardless of task difficulty (right-most column in Table 3.1.1 & Table 3.2.1). In contrast, phonologically similar words have distinctive low-level representations (as will be demonstrated below for the word pairs used in this study), but at the phonological level, their representations will have a high degree of overlap (since they are phonologically similar). In this case, extracting the more abstract phonological categories causes partial loss of low-level information at the higher representation levels. Therefore, RHT predicts that the benefit from low-level information should match the performance predicted by ideal listener models only in specific protocols that allow backward search to find the informative low-level populations.

**Study 3** was designed to test auditory RHT predictions for perceptual learning. Specifically, I tested the effect of protocol manipulation on the ability to learn to utilize low-level binaural cues. In this study, subjects were therefore trained to discriminate between a phonologically similar pair (/barul/ - /parul/) under different variability patterns of the low-level binaural cues (i.e. under different mixtures of diotic and dichotic configurations). Following training, they were tested on transfer of learning to a different binaural protocol and a different discrimination, of a phonologically similar pair on which they were not trained.
Section 3.1: Manipulating task requirements (Study 1)

In this series of experiments of Study 1, I asked whether binaural benefits can be modified without changing task difficulty, namely, without changing diotic thresholds. I used an ideal listener model (see Section 2.2) to calculate the expected performance. The two free parameters of the model (noise levels in the energy and correlation channels, respectively) were calculated from performance with a single, different set of words (used in Experiment II of Study 2, Section 3.2). Thus, in all the calculations for this study (Study 1) the model had no free parameters.

Experiment I - word identification with no binaural uncertainty

The first experiment was designed to replicate studies that found binaural benefits to match those calculated by ideal listener models. The behavioral task was to identify which of the two words comprising the stimulus set (either phonologically different or phonologically similar) was presented in a given trial. Diotic and dichotic configurations were administered in separate blocks, so that the same binaural configuration was repeated across trials for each threshold measurement (“consistent”), eliminating binaural uncertainty within a block.

Figure 3.1.1 (A and B) shows the average changes of signal level during the adaptive tracks in the diotic (thick lines) and dichotic (thin lines) blocks, for both types of word sets, respectively. The plots denote the estimated signal-to-noise ratio (SNR, the difference between stimulus and noise levels) in decibels, as a function of trial number during the assessment. The initial SNR reflects an experimenter-selected level, but subsequent signal levels were set according to performance. Typically, a steady state level of performance is reached by the 40th trial, reflecting the SNR needed to attain 80% correct (which I use here as the discrimination threshold).
Table 3.1.1 The Success of the Predictions of the Three Models (Unlimited, Limited Capacity, and Reverse Hierarchy Theory) for Experiments I–IV of Study 1

Experimental results included measures of binaural benefits, or difference in sensitivity to noise in conditions when discrimination between either phonologically similar or phonologically different words, which were presented to both ears, was required. A ✓ sign indicates that the experimental result fitted the prediction of the model. A ✗ sign indicates that the experimental results did not fit the prediction of the model. A downward arrow (▼) sign indicates that the model predicts less than ideal listener level of performance. In all other cases, the model predicts an ideal listener level of performance.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Task</th>
<th>Protocol</th>
<th>Predictions of the three models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unlimited</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Similar</td>
</tr>
<tr>
<td>I</td>
<td>Identification</td>
<td>Consistent</td>
<td>✓</td>
</tr>
<tr>
<td>II</td>
<td>Semantic</td>
<td>Consistent</td>
<td>✗</td>
</tr>
<tr>
<td>III</td>
<td>Identification</td>
<td>Mixed</td>
<td>✗</td>
</tr>
<tr>
<td>IV</td>
<td>Semantic</td>
<td>Mixed</td>
<td>✗</td>
</tr>
</tbody>
</table>

* For the limited capacity view, the prediction is that as long as the difficulty is not changed, binaural benefits should remain the same. Since ideal listener levels were obtained for the simplest condition (Experiment I), performance in all other experiments is expected to be similar.

As expected, the diotic threshold for discriminating between phonologically similar words (Figure 3.1.1, B, thick red curve) was higher than that for discriminating between phonologically different words (Figure 3.1.1, A, thick blue curve), since this discrimination is more difficult. However, binaural benefits for both types of word pairs were similar, and reached 9–10 dB (10.2 ± 0.9 dB and 9.1 ± 0.8 dB for phonologically different and phonologically similar sets, respectively; F(1,18) = 0.84, not significant [n.s.]). In both cases, the measured binaural benefits reached the benefits predicted by the ideal listener calculations (see Figure 3.1.2 for full details). Both ideal listener thresholds (Figure 3.1.2, A & C) and binaural benefits (Figure 3.1.2, E) matched well the experimental ones for both phonologically similar and phonologically different word
The deviations between modeled and measured values were on average 1.5 dB (compare filled and empty bars of Figure 3.1.2, E), while measured binaural benefits themselves were ~9 dB on average. Moreover, the dichotic thresholds were mainly determined by the change in correlation, with the energy channels below detection threshold (data not shown).

The large binaural benefit obtained with the phonologically similar words might seem surprising given their perceived similarity. However, as shown by the ideal listener calculations, low-level representations of the phonologically similar pair are distinct enough and contain informative binaural cues (Figure 2.2.1 in Section 2).

These results show that, in line with previous reports, under these simple conditions, binaural benefits reach ideal listener levels both when discriminating between phonologically similar words and when discriminating between phonologically different word pairs. These results are therefore consistent with all three views (Table 3.1.1, Experiment I).
RESULTS

Phonologically-Different
Exp. I - Word Identification with no binaural uncertainty

Phonologically-Similar
Exp. II - Semantic task with no binaural uncertainty

Exp. III - Word Identification with binaural uncertainty

Exp. IV - Semantic task with binaural uncertainty

SNR (dB)

Binaural Benefit (dB)

Identification semantic Identification semantic simulated consistent mixed consistent mixed simulated
Figure 3.1.1 Results of Study 1, Experiments I–IV

Left: results using phonologically different word pairs (blue). Right: results using phonologically similar pairs (red). (A - H) The dynamics of the adaptive threshold assessment as a function of trial number (averaged across subjects ± SEM, n = 10 for each of the eight conditions). The level of the signal was modified in relation to subject’s performance (following a three down–one up adaptive procedure). Illustrations of diotic (thick curves) and dichotic (thin curves) thresholds, which are calculated as the mean of last five reversals (see Materials & Methods), are marked by dashed lines in (A). Thresholds are denoted in decibel SNR. Binaural benefits (vertical arrows in all panels) are calculated as the difference between the diotic and dichotic thresholds. (A & B) Experiment I: the identification task with no binaural uncertainty (consistent protocol). (C & D) Experiment II: the semantic task with no binaural uncertainty (consistent protocol). (E & F) Experiment III: the identification task with binaural uncertainty (mixed protocol). (G & H) Experiment IV: the semantic task with binaural uncertainty (mixed protocol). (I and J) A summary of the average binaural benefits obtained in Experiments I–IV (filled shaded bars), and the benefits calculated by an ideal listener model (open bars; see Figure 3.1.2), for phonologically different (left, [I]) and phonologically similar (right, [J]) pairs.

Experiment II - semantic task with no binaural uncertainty

In Experiment II, I manipulated the nature of the behavioral task without modifying its absolute level of difficulty. Semantic processing was not necessary in Experiment I, in which listeners were asked only to discriminate between the two words. Thus, in Experiment I, listeners could have used any low-level acoustic cue that differentiated between the two stimuli. I now wanted to ensure that listeners would process word meaning, as they typically do in more ecological conditions. In Experiment II, I therefore used a semantic-association task in which participants were asked to determine whether a visually presented word is semantically related to the auditory word, which was chosen from the same two-word set used in Experiment I. Visual presentation was brief, and subjects were instructed to respond immediately after stimulus presentation, imposing temporal constraints on the behavioral task (see below). The visually presented word in each trial was randomly selected from a large word set, inducing cross-trial variability in the association required and, hence, forcing semantic processing anew in every trial. Yet,
low-level acoustic information was identical to that of Experiment I since the same two-word auditory sets were used, and the diotic and dichotic configurations were administered in separate blocks (consistent).

Introducing the semantic requirement did not affect task difficulty, as measured by absolute diotic thresholds, either for the phonologically different (-15.3 ± 0.8 dB for the semantic-association task; -16.9 ± 0.8 dB for the identification task; $F(1,36) = 0.46$, n.s.; compare Figure 3.1.1, A & C) or for the phonologically similar pair (-8.9 ± 0.5 dB for the semantic-association task; -9 ± 0.4 dB for the identification task; $F(1,35) = 3.9$, n.s.; compare Figure 3.1.1, D & B). However, its impact on binaural benefits greatly differed between these conditions. When the semantic task was performed with the phonologically different pair, binaural benefits remained as large as those of an ideal listener as measured in Experiment I (10.9 ± 1 dB compared with 10.2 ± 0.9 dB for the identification task; no effect of task: $F(1,36) = 1.5$, n.s.; Figure 3.1.1, C & I). However, when the task was performed with the phonologically similar pair, dichotic thresholds were elevated, i.e., binaural benefits decreased (4.1 ± 0.9 dB compared with 9.1 ± 0.8 dB for the identification task; effect of task: $F(1,36) = 5.3$; $p < 0.03$; Figure 3.1.1, D & J).

The differences between performance with the phonologically similar and phonologically different sets cannot be attributed to differences in response times (RTs), as those were the same for the two word pairs used (672 ± 66 ms and 670 ± 112 ms for the phonologically similar and phonologically different pairs; $t$-test: $t = -0.13$, $df = 17$, n.s.).

The finding that binaural benefits remained equivalent to those of an ideal listener when the semantic task involved phonologically different words is in line with the unlimited view, which predicts full use of low-level information. However, this account cannot explain the failure of an ideal listener model to account for binaural benefits in the case of phonologically similar words. Since absolute diotic thresholds were not increased, there is no basis on which to assume an increase in perceptual or cognitive load. Moreover, had an increase in attentional load occurred with no impact on absolute thresholds, it should have reduced the ability to use binaural cues for both pair types. Thus, the ideal listener levels of binaural benefits for phonologically different words, but poorer benefits for phonologically similar words, are inconsistent with both the unlimited
view and with the limited capacity view, but are in line with RHT predictions (Table 3.1.1, Experiment II).

**Experiment III - word identification with binaural uncertainty**

In Experiment III, I asked whether introducing uncertainty in the low-level binaural configuration affects the use of binaural cues. I used the same word sets and the same identification task as in Experiment I. However, the diotic and dichotic configurations were randomly interleaved across trials (“mixed”). This manipulation therefore caused the low-level binaural cues required for correct performance to vary from trial to trial. Yet, the higher-level phonological and semantic representations as well as the definition of task demands were identical to those of Experiment I.

As expected, absolute diotic thresholds were not affected by this binaural variability, either for the phonologically similar pair (-9.8 ± 0.5 dB and -9 ± 0.4 dB in Experiments III and I, plotted in Figure 3.1.1, F & B, respectively; effect of protocol: $F(1,35) = 0.6$, n.s.) or for the phonologically different pair (-16 ± 0.8 dB compared with -16.9 ± 0.8 dB, plotted in Figure 3.1.1, E & A, respectively; effect of protocol: $F(1,36) = 0.19$, n.s.). However, the use of binaural cues for discriminating between phonologically similar words was substantially (~ 6 dB) smaller than that predicted by the ideal listener model (3.6 ± 0.8 dB compared with 9.1 ± 0.8 dB for the consistent protocol; $F(1,36) = 7.96$, $p < 0.01$): This difference stems from discrepancies in dichotic thresholds, rather than diotic thresholds, between the simulation and the experiment. In contrast, for the phonologically different words, binaural benefits remained equivalent to those of the ideal listener (9 ± 1 dB and 10.2 ± 0.9 dB in the mixed and consistent protocols, respectively; no effect of protocol: $F(1,36) = 0.5$, n.s. see Figure 3.1.2, right). The specific ideal listener calculations are detailed in the following section.
Figure 3.1.2 Comparing the results of the ideal listener model to the experimental results of Study I

Graphs compare simulated (empty bars) and experimental (filled bars) thresholds (A - D) and binaural benefits (E & F) for both the phonologically different (/tamid/ vs. /chalom/; blue bars) and phonologically similar (/tamid/ vs. /amid/; red bars) word pairs, under both consistent (left; Experiment I) and mixed (right; Experiment III) binaural protocols. (A & B) Diotic thresholds; (C & D) dichotic thresholds; (E & F) binaural benefits. Note the difference between simulated and experimental binaural benefits for the phonologically similar words, measured under the mixed binaural protocol (red bars of [F]).

Ideal listener simulation results for Experiment III
In the mixed binaural protocol, the ideal listener had two sets of templates for each word, one for the diotic configuration and one for the dichotic configuration. The optimal decision was not the least distance, but rather a complex function of the distances between the actual word that occurred and both templates. However, at the SNR levels relevant to the experiment, the templates for the diotic and for the dichotic configurations were so different that the optimal decision was well approximated by a sequential decision, in which the binaural configuration was detected first and then an optimal decision was performed using the appropriate templates for the selected binaural configuration.

Using the same $f_{\text{ener}}$ and $f_{\text{cc}}$ internal noise variances that were calculated for the consistent binaural protocol (Experiment I), the simulated results are essentially identical to the ones obtained for the consistent binaural protocol (see right side of Figure 3.1.2). This is in fact a proof that the suboptimal procedure used in the case of the mixed binaural protocol indeed reached optimal performance. The reason is that since the procedure used to determine these thresholds was suboptimal, these thresholds were only upper bounds - using the optimal procedure may have resulted in lower thresholds. However, the theoretical performance of an ideal listener in the mixed binaural protocol could not be better than its performance in the consistent binaural protocol, and therefore the thresholds in the consistent binaural protocol formed a lower bound on the thresholds in the mixed binaural protocol. Since the thresholds determined by the suboptimal procedure (upper bounds for the mixed binaural protocol) were essentially equivalent to those determined by the ideal listener in the consistent binaural protocol (lower bounds for the mixed binaural protocol), it follows that the use of the suboptimal procedure closely approximated the performance of the optimal procedure in the mixed binaural protocol.

The main result of this calculation is the finding of much larger binaural benefits than the ones actually obtained experimentally for phonologically-similar, but not for phonologically-different, word pairs in the mixed binaural protocol (compare filled and empty bars of Figure 3.1.2, F; see Figure 3.1.1 where these data are replotted). Thus, introducing variability of the informative low-level information across trials disabled listeners from reaching ideal listener levels of binaural benefits when discriminating
between phonologically similar words, but not when discriminating between phonologically different words. The results of this experiment pose an even greater challenge to the limited capacity view, since not only measurable (diotic) thresholds remained the same, but also introspective task demands were exactly as in Experiment I (Table 3.1.1, Experiment III). In post-test questionnaires, listeners failed to report any information regarding the binaural configuration, indicating that they were not aware of this low-level variability.

**Experiment IV - semantic task with binaural uncertainty**

In Experiment IV, I combined the two types of manipulations. Subjects were asked to perform a semantic-association task (similar to the one in Experiment II) while diotic and dichotic configurations were randomly interleaved (i.e. variable) within the block (as in Experiment III).

The results of this experiment (Figure 3.1.1, G & H) were similar to those of Experiments II and III. Thus, having the two constraints together yielded the same results that each of them produced separately. Absolute diotic thresholds were similar to those of Experiment I for both phonologically different pairs (−15.2 ± 1.2 dB and −16.9 ± 0.8 dB in Experiments IV and I, respectively; interaction of task × protocol: \(F(1,36) = 1.4\), n.s.) and for phonologically similar pairs (−7.6 ± 0.7 dB and −9 ± 0.4 dB for Experiments IV and I, respectively; no significant interaction of task × protocol: \(F(1,35) = 2.9\), n.s.). However, binaural benefits were similar to those of Experiments II and III. They matched those of the ideal listener (as measured for Experiment III) for the phonologically different word set (10.7 ± 1.1 dB; no significant interaction of task × protocol: \(F(1,36) = 0.21\), n.s.), and were significantly poorer than the ideal listener prediction for the phonologically similar words (4.7 ± 0.9; significant interaction of task × protocol: \(F(1,36) = 12.7\), \(p < 0.005\); see Table I, Experiment IV). As in Experiment II, RTs were kept below 1 s, and did not differ significantly between phonologically different (722 ± 70 ms) and phonologically similar (723 ± 95 ms) pairs (\(t\)-test: \(t = -0.14, df = 17\), n.s.).
RESULTS

Replication of Study 1 with other word pairs

In order to verify that this set of results systematically characterizes the manipulations I introduced and is not specific to the two word pairs used in Study 1, I fully replicated Study 1 with two other word pairs (/shalom/ vs. /chalom/ and /tamid/ vs. /chalom/), and obtained similar results (Figure 3.1.3).

Thus, for Experiments I & III (identification task under consistent and mixed binaural protocols, respectively), diotic thresholds were similar for both the phonologically different pair (-25 ± 1.1 dB vs. -25 ± 0.7 dB under consistent and mixed protocols, respectively; effect of protocol: F(1,35) = 0.07, n.s.) and the phonologically similar pair (-17.7 ± 0.5 dB vs. -18.6 ± 1.2 dB under consistent and mixed binaural protocols, respectively; F(1,35) = 0.16, n.s.). Binaural benefits were relatively large for both pairs under the consistent protocol (Experiment I; phonologically-different: 7.2 ± 1.5 dB; phonologically-similar: 5.2 ± 0.5 dB). Note that although binaural benefit was a bit smaller for the phonologically-similar pair, it still matched that simulated by the ideal listener model (see below & Figure 3.1.3, E & F). However, under the mixed protocol (Experiment III), binaural benefit was reduced for the phonologically-similar pair (1.9 ± 0.3 dB; F(1,35) = 16.8; p < 0.0005), but remained at the ideal listener level for the phonologically-different pair (7.7 ± 0.5 dB, F(1,35) = 0.17, n.s.).

For the semantic-association tasks (Experiments II & IV), diotic thresholds were similar to those measured in the identification task (compare panels C & D to panels A & B, respectively, Figure 3.1.3) for both word pairs. For the phonologically different pair they averaged to -23 ± 1 and -23 ± 0.8 dB in the consistent (Experiment II) & mixed (Experiment IV) protocols (Figure 3.1.3, C), and the phonologically similar pair to -17.9 ± 0.8 and -16.3 ± 0.6 dB for these protocols, respectively (Figure 3.1.3, D; no effect of task: F(1,35) = 1.5, n.s.). Statistically, there was no significant interaction of task X protocol for both the phonologically similar pair (F(1,35) = 2.1, n.s.) and phonologically different pair (F(1,35) = 0.99, n.s.). Binaural benefits remained ideal-like only for the phonologically different pair (9.6 ± 0.7 & 8.2 ± 0.9 dB for the consistent and mixed binaural protocols; no effect of task: F(1,35) = 2, n.s.; no significant interaction of task X
RESULTS

protocol: F(1,35) = 0.9, n.s.). However, for the phonologically similar pair, binaural benefits decreased (2.8 ± 0.4 & 2.6 ± 0.4 dB for consistent and mixed protocols, respectively; effect of task: F(1,35) = 4.4; p < 0.05; interaction: F(1,35) = 14.4, p < 0.001).

Figure 3.1.3 Replication of the results of Study 1, Experiments I-IV, with two other pairs of words (/sikum/ vs. /amid/ and /shalom/ vs. /chalom/)
Left - for the phonologically-different word pair (blue). Right - for the phonologically-similar pair (red). (A - D) Diotic and dichotic thresholds measured for Experiments I-IV of Study 1 (averaged across subjects ± SEM, N=10 for each of the 8 conditions). Binaural benefits are the differences between the two thresholds.
(A & B) Identification task with no binaural uncertainty (consistent protocol; left, Exp. I) and with binaural uncertainty (variable protocol; right, Exp. III). (C & D) Semantic-association task with no uncertainty (consistent protocol; left, Exp. II) and with uncertainty (mixed protocol; right, Exp. IV). (E & F) A summary of the average binaural benefits (the difference between diotic and dichotic bars in each panel) obtained in Experiments I-IV (filled shaded bars), and the benefits calculated by an ideal listener model (empty bars).

Summary of the results of Study 1

In Study 1 I found that, in line with the unlimited view, full use of binaural information can be obtained with both phonologically similar and phonologically different word sets. However, the unlimited view fails to predict binaural benefits for phonologically similar words when low-level cross-trial uncertainty is introduced. A similar drop in utilization of low-level information is found when semantic processing is required. These failures cannot be explained by limited capacity models either (e.g. Allport, 1980; Bundesen, 1990; Lavie, 2005; Lavie et al., 2004), since these manipulations did not increase task difficulty, as reflected by the unchanged diotic thresholds (Experiments II - IV), and were in some cases transparent to participants (Experiment III). Table 3.1.1 summarizes the predictions and results of the three views for Experiments I - IV.
Section 3.2: Manipulating task difficulty (Study 2)

In Study 1, I manipulated explicit (Experiment II) and implicit (Experiment III) task requirements without modifying task difficulty, and assessed the impact of these manipulations on binaural benefits. In Study 2, I designed manipulations that were aimed at modifying task difficulty (diotic thresholds) in order to assess whether this type of change affects the use of binaural cues, as predicted by the limited capacity view.

Experiment I - manipulating set size (“cognitive load”)

In this experiment, I increased the cognitive load of the task by increasing stimulus set size. This manipulation (increasing “memory set size”) has been shown to increase the cognitive load both in the visual (e.g. Shiffrin and Schneider, 1977) and in the auditory (Poltrock et al., 1982) domains. I expected that diotic thresholds would increase and tested the resulting effects on binaural benefits. In the new condition with high cognitive load, the presented word on a given trial was selected from a set of ten words rather than two words. The sets were composed of either phonologically different (ten different words) or phonologically similar (five pairs of similar words) words. I used the mixed binaural protocol with randomly interleaved diotic and dichotic trials (as was used in Experiments III and IV of Study 1).

Table 3.2.1 The success of the predictions of the three models (unlimited, limited capacity, and RHT) for Experiments I & II of Study 2

Experimental results included measures of binaural benefits, or difference in sensitivity to noise in conditions when discrimination between either phonologically similar or phonologically different words, which were presented to both ears, was required. Notations as in Table 3.1.1. Note that the unlimited view fails in its prediction only for the phonologically similar words, as in Study 1; the limited capacity view predicts sub-
ideal performance in cases of increased difficulty, whereas the experimental results are different. RHT predicts ideal listener levels for phonologically different words under all conditions, but only under the consistent protocol for the phonologically similar words, regardless of task difficulty.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Task</th>
<th>Condition</th>
<th>Protocol</th>
<th>Predictions of the three models</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>60%</td>
<td>Consistent</td>
<td></td>
<td>Unlimited</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Similar</td>
</tr>
<tr>
<td>I</td>
<td>80%</td>
<td>Consistent</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>II</td>
<td>60%</td>
<td>Mixed</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td>Mixed</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

As expected, increasing the set size from two to ten significantly increased diotic thresholds for both the phonologically different set (from -21.8 ± 0.6 dB to -16.7 ± 0.2 dB SNR; Figure 3.2.1, A vs. C) and the phonologically similar set (from -18 ± 1.1 dB to -8 ± 0.2 dB SNR; Figure 3.2.1, B vs. D; F(1,18) = 39.1, p < 0.00001). A larger increase in thresholds was found for the phonologically similar set (a significant interaction of set size × similarity: F(1,18) = 71.8, p < 0.00001). However, binaural benefits did not significantly change (no effect of set size F(1,18) = 0.09, n.s.). They were quite large for the phonologically different words (5.8 ± 0.3 dB and 6.6 ± 0.7 dB for set sizes of ten and two, respectively, Figure 3.2.1, E), matching those of the ideal listener (the difference between simulated and experimental binaural benefits was ~0.4 dB for the “set size 2” case; see Figure 3.2.1, E and Figure 3.2.2 for details). They were smaller for the phonologically similar words (3 ± 0.5 dB and 2.3 ± 0.3 dB for set sizes of ten and two,
RESULTS

respectively; Figure 3.2.1, F), and did not reach the values predicted by the ideal listener model. In this case, simulated binaural benefits were 3.7 dB and 6 dB larger than the ones measured experimentally, for both “set size 2” and “set size 10” conditions, respectively. The differences stem from reduced simulated dichotic thresholds in both cases, while diotic thresholds were similar to the experimental ones (Figure 3.2.2). Note that the “set size 10” condition was simulated by calculating the DTW distances between the input word and each of the 10 templates of the words comprising the set, with the input word being recognized as the template with the smallest DTW distance from it.
Figure 3.2.1 Results of Study 2, Experiment I Left: results using phonologically different words (blue). Right: results using phonologically similar words (red). (A - D) The dynamics of the adaptive threshold assessment as a function of trial number (averaged across subjects ± SEM, n = 25). Notations as in Figure 3.1.1. Vertical arrows denote binaural benefits. All measurements were done using the mixed binaural protocol. (A & B) An identification task using a set size of two words. (C & D) An identification task using a set size of ten words. (E & F) A summary of the average binaural benefits obtained in the experiment (filled shaded bars) and the benefits calculated by the ideal listener model (open bars; see Figure 3.2.2), for the set size 2 and set size 10 conditions.

Thus, binaural benefits reached ideal listener levels for phonologically different words, but failed to reach these levels for phonologically similar words, regardless of set size (a significant effect of phonological similarity, $F(1,18) = 51, p < 0.0001$; no significant interaction of set size $\times$ similarity, $F(1,18) = 3.7, \text{n.s.}$).

The results of this experiment show that although increasing the cognitive load (by increasing the set size from two to ten) yields the expected increase in diotic identification thresholds, it does not change binaural benefits. This experiment thus clearly dissociates between task difficulty and the ability to use low-level information, and its results are therefore inconsistent with limited capacity models, but are in line with RHT predictions (Table 3.2.1, Experiment I).
Figure 3.2.2 Comparing the results of the ideal listener model to the experimental results of Experiment 1 of Study 2 Results are compared for set sizes of two (left) and ten (right) of phonologically different and phonologically similar word pairs. Notations as in Figure 3.1.2. (A & B) Dotic thresholds; (C & D) dichotic thresholds; (E & F) binaural benefits. Note the difference between simulated and measured binaural benefits for the phonologically similar pair (red bars).
Experiment II - manipulating success level ("perceptual load")

In this experiment, I increased the perceptual load by modifying the adaptive procedure to a procedure that converges at approximately 60% rather than 80% correct (Levitt, 1971). Subjects reported that this protocol “felt more difficult,” presumably due to the lower SNRs at which most stimulus presentations occurred. I asked whether this change in difficulty affects binaural benefits. I calculated ideal listener performance for both levels of difficulty and compared them to the measured binaural benefits.

First, I replicated Experiments I and III of Study 1, using the original adaptive procedure converging at 80% correct, using other word pairs (/barul/ vs. /parul/ and /dilen/ vs. /talug/, respectively). Indeed, when the task required identification and was administered with the consistent binaural protocol (with separate measurements of the diotic and dichotic thresholds, as in Study 1, Experiment I), binaural benefits reached the ideal listener levels (see below), of 9–10 dB, for both word sets (10.5 ± 0.7 dB and 9.2 ± 0.8 dB for the phonologically different and phonologically similar pairs, respectively; Figure 3.2.3, A & B). However, only the phonologically different set yielded similar benefits under the mixed binaural protocol, when diotic and dichotic trials were randomly interleaved (9.2 ± 0.7 dB compared with 4.6 ± 0.6 dB obtained with the phonologically similar pair; Figure 3.2.3, C & D), fully replicating Experiments I and III of Study 1 (see Table 3.2.1, Experiment II).

I then asked whether a similar pattern of binaural benefits would be found with the adaptive protocol converging to approximately 60% success in the task, rather than to 80% success. As expected, diotic thresholds for both phonologically similar and phonologically different pairs were lower for the 60% correct condition (Figure 3.2.3, E-H) compared with the 80% correct condition (analysis of variance [ANOVA]: percent correct: *F*(1,23) = 52.6, *p* < 0.00001; similarity: *F*(1,23) = 18.2; *p* < 0.0005; and between-subjects factor of binaural protocol: *F*(1,23) = 1.9, n.s.). Moreover, binaural benefits obtained with 60% correct had the same pattern, and did not significantly differ from those obtained with 80% correct (*F*(1,23) = 0.43, n.s.). They were large for both sets
under the consistent binaural protocol ($7.8 \pm 0.8$ dB and $9.7 \pm 1$ dB for phonologically different and phonologically similar pairs, respectively). Yet, only the phonologically different set yielded similar binaural benefits with the mixed binaural protocol ($10.5 \pm 1.5$ compared with $3.2 \pm 1.2$ obtained with the phonologically similar pair). Thus, there was a significant effect of protocol ($F(1,23) = 8.9, p < 0.008$) and a significant interaction between similarity and protocol ($F(1,23) = 14.9, p < 0.001$).
RESULTS

Phonologically-Different

80% success with no uncertainty

80% success with uncertainty

60% success with no uncertainty

60% success with uncertainty

Binaural Benefit (dB)

consistent mixed simulated consistent mixed simulated

I J
RESULTS

Figure 3.2.3 Results of Study 2, Experiment II

Left: results using phonologically different word pairs (blue). Right: results using phonologically similar pairs (red). (A - H) The dynamics of the adaptive threshold assessment for identification of word pairs as a function of trial number (averaged across subjects ± SEM, n = 15). Notations as in Figure 3.1.1. Vertical arrows denote binaural benefits. (A & B) The adaptive protocol converging to 80% correct identification with no uncertainty (i.e., using the consistent binaural protocol). (C & D) The adaptive protocol converging to 80% correct identification with uncertainty (mixed binaural protocol). (E & F) The adaptive protocol converging to 60% correct identification with no uncertainty (consistent binaural protocol). (G & H) The adaptive protocol converging to 60% correct identification with uncertainty (mixed binaural protocol). (I & J) A summary of the average binaural benefits obtained in the experiment (filled shaded bars), and the benefits calculated by an ideal listener model (open bars; see Figure 3.2.4).

Ideal listener calculations for Study 2, Experiment II

I calculated the “ideal listener” performance for these conditions as well. Discrimination between phonologically similar words under the consistent binaural protocol at 80% correct was used for calculating the variances in neural activity (these were the values used at all other ideal listener calculations in this work; once these values were set, the model had no free parameters). Using the values obtained, I then calculated ideal listener performance for all other conditions. To simulate the other conditions in Experiment II, I matched the simulated percent correct levels with the experimental ones; thus, the reported SNRs are the ones that yielded either 60 or 80 percent correct.

The results of the simulation in both 60 & 80% correct levels matched nicely the experimental results of the phonologically-different pair, for both consistent and mixed binaural protocols (blue bars of Figure 3.2.4). Simulated diotic, as well as dichotic thresholds were similar to the experimental ones (Figure 3.2.4, A-D). Binaural benefits (Figure 3.2.4, E-F) were thus predicted to be large (10-11 dB) in all cases, as was indeed measured in the actual experiment (9-10.5 dB). The resultant maximal difference between simulated and experimental binaural benefits was therefore ~1.8 dB on average for the phonologically-different pair.

The simulation results also matched the ones obtained for the phonologically-similar pair, when measured under the consistent binaural protocol (red bars of Figure 3.2.4, left). Note, however, that this condition, discrimination between /barul/ and /parul/ in
80% success was the ones used for setting the 2 free parameters of the simulation, monaural and binaural variances. Thus, the fact that for 80% success thresholds and binaural benefits matched the experimentally measured ones is not surprising; however, the same values yielded similar results for the 60% success case as well, which was not used in the fitting process. In both cases, simulated binaural benefits fitted the measured ones, of about 9-10 dB.

Figure 3.2.4 Comparing the results of the ideal Listener model to the experimental results of Experiment II of Study 2 Results are compared for performance levels of 60% and 80% correct for both phonologically different (blue bars) and phonologically similar
RESULTS

(red bars) pseudo-word pairs, measured under consistent (left) and mixed (right) binaural protocols. Notations as in Figure 3.1.2. (A & B) Diotic thresholds; (C & D) dichotic thresholds; (E & F) binaural benefits.

In contrast, the ideal listener simulation results for the dichotic condition of the phonologically-similar pair differed substantially from the experimental ones when measured under the mixed binaural protocol, by an average of 5.6 dB for both 60 & 80% success (red bars of Figure 3.2.4, D). As a result, the ideal listener simulations predicted substantially larger binaural benefits (9 & 10 dB for 60 & 80% success; Figure 3.2.4, F) than the experimentally measured ones (3.2 & 4.6 dB, respectively).

To summarize, the ideal listener model accounted for performance in all conditions that required discrimination between phonologically different words pairs, but only for the performance in the consistent binaural protocol when discriminating between phonologically similar words.

Summary of the results of Study 2

The two experiments of Study 2 show different manipulations that affect task difficulty and yet do not affect the use of binaural cues. The finding that increased difficulty does not decrease the use of low-level information indicates that, in contrast to the limited capacity view, attentional load is not the bottleneck for our ability to use low-level information. Similarly to Study 1, the ideal listener simulation results of Study 2 show that in case of discrimination between phonologically-similar words under a mixed binaural protocol, an ideal listener obtains much larger binaural benefits than the ones measured experimentally. This trend is kept for both larger set sizes (Experiment I) and different percent correct levels (Experiment II). Table 3.2.1 summarizes the predictions and results of the three views for Experiments I and II of Study 2.
Section 3.3: Perceptual learning under low-level variability patterns (Study 3)

In perceptual learning, the amount of improvement and its transfer to untrained stimulus parameters has been recently shown to be affected by the variability within the stimulus set throughout training. Thus, introducing larger variability during training (e.g. roving different frequencies in a frequency discrimination task) is expected to show limited improvement on the trained parameters but larger generalization (e.g. to untrained frequencies), due to changes at higher processing levels (Adini et al., 2004; Kuai et al., 2005; Nygaard and Pisoni, 1998; Otto et al., 2006; Yu et al., 2004); In contrast, training with consistent stimulus-response mapping and limited stimulus parameters is expected to reach the fine-tuned low level representations and therefore result in larger specific improvement but only limited transfer to untrained parameters (Adini et al., 2002; Fahle et al., 1995; Fiorentini and Berardi, 1980; Poggio et al., 1992; Sagi and Tanne, 1994; Schoups et al., 1995; Watson et al., 1976). The intermediate case, in which learning of several parameters is done in a blocked or temporally-structured manner (e.g. all frequencies are ordered in an increasing manner throughout the block), yields learning in some cases, which transfer to the more variable (roved) conditions (Kuai et al., 2005; Zhang et al., 2008).

Such observations are consistent with the RHT view, of top-down guided learning, which starts at higher levels for naïve subjects and gradually progresses to lower levels if an informative low-level population can be traced (Ahissar and Hochstein, 1997, 2004; Ahissar et al., 2008). According to RHT, only consistent or blocked presentation of the relevant cues enables such gradual low-level access throughout learning, while more variable conditions may not enable the identification of the most informative low-level cues, and would therefore be based on high-level representations.

In the current study (Study 3) I aimed to test these specific predictions of auditory RHT. I used the same marker for low-level access used in Study 1 & Study 2, the magnitude of the binaural benefits for a speech-in-noise discrimination task. In the previous studies I used this marker to show that, in line with RHT predictions, when discriminating between phonologically similar words (e.g. /tamid/ vs. /amid/), utilization
of low-level binaural information critically depends on the protocol of binaural cross trial variability. Full utilization is afforded when this information is consistent throughout the block, while only partial utilization is afforded when the information is randomly interleaved throughout the block.

I now used the same tool in the context of perceptual learning: three groups of subjects were first tested on a two-word discrimination task under three protocols which differed in the amount of implicit binaural variability (Experiment I). Data of two of the groups (consistent, with no variability and mixed, with randomly interleaved trials) were previously used in Study 1 & 2. The third group (‘1-1’) was chosen since it had characteristics which combine the first two protocols: binaural information was presented within a mixed protocol, but in a fully predicted manner. Odd trials were diotic, and even trials were dichotic. This protocol therefore enabled to dissociate between the predictability of the information (which was fully predictable) and its consistency across consecutive trials (information was not the same on consecutive trials). This experiment was aimed to assess the effect of protocol on utilization of low-level information for discrimination between phonologically similar words.

In the second part, Experiment II, I added training. Half of the subjects from each group (14/30) continued to train for a total of 7 days on the same two-word discrimination task using the same protocol they began with. Following training, transfer of learning to an untrained protocol and an untrained word pair was tested. This procedure enabled to test the effect of the protocol on learning to implicitly use low-level cues.

In both experiments, the auditory stimuli we used were disyllabic Hebrew pseudo-words (/barul/ and /parul/ for most phases; /dilen/ and /tilen/ for part of the testing phase) embedded in speech noise (Dreschler et al., 2001), presented under both diotic and dichotic configurations (see Section 2.1). Subjects’ task was to identify which of the two words composing the stimulus set was presented in a given trial (e.g. /barul/ or /parul/). Thresholds for ~80% correct discrimination were measured using a 3 down-1 up staircase procedure (Levitt, 1971).
Experiment I: The effect of protocol on the use of low-level information in session I

In Experiment I diotic and dichotic thresholds for word identification in noise were measured under three different protocols of the binaural information:

- ‘Consistent’, in which diotic and dichotic configurations were measured in separate blocks, so that the same binaural configuration was repeated across trials for each threshold measurement.
- ‘Mixed’, in which the diotic and dichotic configurations were randomly interleaved throughout the block.
- ‘1-1’, in which the diotic configuration was presented on odd trials, while the dichotic configuration was used on even trials.

Each subject performed the task with a single protocol, and subjects were unaware of the manipulation of low-level variability. Whereas the first two protocols (consistent and mixed) have been previously tested in Study 1 & 2, the 1-1 protocol tested whether a blocked presentation of the relevant cues, such that they can be easily tracked but still vary across the block, is sufficient for utilization of the low-level information.

Figure 3.3.1 (top) depicts the estimated signal to noise ratio (SNR - the difference between stimulus and noise levels) in dB, as a function of trial number during the first assessment of each of the 3 protocols. Typically, a steady state level of performance was reached by the 40th trial, reflecting the SNR needed to attain 80% correct (which I used here as the discrimination threshold).

As seen in Figure 3.3.1, diotic thresholds did not significantly differ between the three protocols (F(2,87) = 0.36, p = 0.7, n.s.). On the other hand, binaural benefits, calculated as the difference between diotic and dichotic thresholds (Figure 3.3.1, bottom), significantly differed between protocols (F(2,87) = 7.2, p < 0.002; vertical arrows in Figure 3.3.1, top). They were largest for the consistent protocol (7.8 ± 0.5 dB, blue bar), and remained ~5 dB for the other two protocols (5.2 ± 0.8 and 4.9 ± 0.4 dB for 1-1 and mixed protocols, respectively). A post-hoc Scheffe contrast between the three protocols showed that the consistent protocol is significantly different from both the 1-1
RESULTS

protocol \((p < 0.01)\) and the mixed protocol \((p < 0.005)\), while the two later conditions are not significantly different from one another \((p = 0.95, \text{n.s.})\).

**Figure 3.3.1 The effect of protocol** The dynamics of the adaptive threshold assessment as a function of trial number under each of 3 stimulus protocols differing in binaural cross-trial consistency (averaged across subjects ± SEM, \(N=30\) in each protocol group): consistent (blue, left), 1:1 (green, middle) and mixed (red, right). The level of the signal was modified adaptively (following a 3 down - 1 up staircase procedure). Threshold measurements are denoted in dB Signal-to-Noise Ratio (SNR). Estimated binaural benefits, calculated as the difference between the diotic and dichotic thresholds, are marked as vertical arrows in all panels. Bottom: binaural benefits (averaged across subjects ± SEM). Binaural benefits were significantly larger under the consistent protocol although task difficulty (diotic thresholds) did not differ between protocols.

The results of Experiment I show that the implicit manipulation of low-level information, which did not change the overall task difficulty (as measured by the diotic thresholds), affected the degree that low-level binaural cues were used. Thus, replicating the previous results, efficient use was only found when diotic and dichotic thresholds were measured separately and the information within the block was fully consistent.
Extending these findings, when the information was fully predictable temporally, yet inconsistent across consecutive trials (“1-1”), utilization was again sub-optimal, as in the *mixed* protocol. This shows that in order for naïve subjects to fully benefit from low-level information, it has to be consistently presented across consecutive trials. Presumably, naïve subjects cannot track this simple (but implicit) consistency in order to guide the top-down search for the relevant information for each configuration. In Experiment II, I asked whether long term training would remove this constraint.

**Experiment II: Learning to use low-level cues with low-level variability**

In Experiment II I tested the predictions of auditory RHT with respect to perceptual learning; specifically, I asked whether the implicit protocol affects the improvement gained following multi-session training on the two-word discrimination task. According to RHT, training would improve utilization of low-level information only when the relevant low-level population can be tracked and accessed. Under the *consistent* protocol, this can be achieved quickly within the first assessment which consists of several dozen trials. Under the *mixed* protocol, tracking of the relevant information is difficult since no “pointer” is provided for the relevant population when the binaural information varies randomly between two options. The *1-1* case, in which naïve subjects do not efficiently use all low-level information, is expected to allow this use following training: since the temporal structure of the two configurations throughout the blocks is simple (i.e. a “blocked” presentation), it may gradually be detected by the top-down search mechanisms.

Fourteen of the subjects of each group kept training on the same two-word identification task (discrimination between /barul/ and /parul/), using the same protocol they were initially assigned to. The three training groups (‘*consistent group*’, ‘*1-1 group*’ and ‘*mixed group*’) were trained for 7 days, and were later tested for generalization across protocols and across word pairs (see **Section 2.1** and **Table 2.1.3** for details).
Training and the utilization of low-level binaural information (Days 1 - 7)

Figure 3.3.2 (upper panels) depicts the average thresholds on each day of training. Diotic thresholds improved for all three groups to a similar extent, showing that significant learning did occur for all three protocols (effect of days: F(6, 216) = 49.6, p < 0.00001; no significant effect of protocol: F(2, 36) = 1.2, p = 0.3, the interaction is not significant as well). Thus, for the consistent group (blue plot), diotic thresholds improved in ~ 4dB overall, from -11 ± 0.4 dB SNR to -15.4 ± 0.5 dB SNR. Similarly, for the 1-1 (green plot) and mixed (red plot) groups, thresholds improved from -11.4 ± 0.7 dB SNR to -14.5 ± 0.5 dB SNR and from -10.8 ± 0.4 to -14.1 ± 0.5 dB SNR, respectively. Data for individual subjects is shown in Figure 3.3.3.
RESULTS

Signal-to-Noise Ratio (SNR). Bottom: binaural benefits, calculated as the difference between the diotic and dichotic thresholds above. Left: consistent group; Middle: 1-1 group; Right: mixed group. Subjects on the consistent and 1-1 group were tested in day 8 on the mixed protocol (red bars); Subjects in the mixed group were tested on the 1-1 protocol (green bars). Results are shown as averages (N = 14 for days 1-7; N = 10 for day 8) ± SEM.

Binaural benefits (Figure 3.3.2, lower panels) significantly changed during learning as well (effect of days: \( F(6,216) = 6.9, p < 0.0001 \)), but this change mainly resulted from the improvement in the binaural benefit for 1-1 group (Figure 3.3.2, B; effect of protocol: \( F(2,36) = 7.8, p < 0.005 \)), which increased from 5.4 ± 0.5 dB to 8.4 ± 0.6 dB from 1\(^{st}\) to 7\(^{th}\) day of training. For the mixed group (Figure 3.3.2, C), binaural benefits increased throughout training to a much lesser extent (~1 dB), from 5 ± 0.5 on the 1\(^{st}\) day to 6.1 ± 0.6 dB on the 7\(^{th}\) day. For the consistent group (Figure 3.3.2, A), binaural benefits were large and similar in the 1\(^{st}\) and 7\(^{th}\) days (7.2 ± 0.3 and 7.6 ± 0.5 dB for 1\(^{st}\) and 7\(^{th}\) days, respectively). The major increase in binaural benefits for the 1-1 group occurred between the 1\(^{st}\) and 2\(^{nd}\) days, in which binaural benefits were 7.6 ± 0.5 dB. The increase continued until the 5\(^{th}\) day of training. By the end of training, binaural benefits and dichotic thresholds were similar for both the consistent and 1-1 groups. A post-hoc Scheffe contrast showed that binaural benefits in the mixed group were significantly lower than those in the consistent (\( p < 0.02 \)) and 1-1 (\( p < 0.005 \)) groups, while the two later groups did not significantly differ (\( p = 0.93 \), n.s.).
Figure 3.3.3 Individual data for Experiment II, days 1-8 Diotic (solid lines) and dichotic (dashed lines) thresholds obtained throughout training (days 1-7) are shown for each of the 14 subjects in each protocol group: consistent (top graph, blue), 1-1 (middle graph, green) and mixed (bottom graph, red). For each subject, the threshold in each day is calculated as the average of thresholds obtained in all three blocks performed that day. Thresholds are denoted in dB Signal-to-Noise Ratio (SNR).
Learning was transferred across protocols

Following 7 days of practice, 10 of the 14 subjects in each group performed the same discrimination with a different binaural protocol than the one they were trained with. Subjects in the ‘consistent’ and ‘1-1’ groups performed the task in the mixed protocol, while subjects in the ‘mixed’ group performed the task in the 1-1 protocol. In order to assess generalization across protocols, performance was compared to both naïve and trained performance of subjects trained with the tested protocol.

Learning under the consistent and the 1-1 protocols fully transferred to the mixed protocol, resulting in improvements both in diotic thresholds (-15.7 ± 1 & -14.7 ± 0.9 dB SNR following consistent and 1-1 training, respectively; F(2,47) = 12.1, p < 0.0001 compared to naïve performance) and binaural benefits (8.5 ± 1 & 7.8 ± 0.7 dB, respectively; F(2,47) = 11.3; p < 0.0005). Moreover, thresholds and binaural benefits were similar to those obtained on the last (7th) day of practice on each protocol (t-tests: p = 0.4 & p = 0.3 for consistent and 1-1, respectively; compare days 7 & 8 for left and middle panels, Figure 3.3.2). The magnitudes of the effects, for both diotic thresholds and binaural benefits, were somewhat larger for the consistent (post-hoc Scheffe contrast: p < 0.001) than the 1-1 (p < 0.01) protocol.

Learning with the mixed protocol showed transfer in diotic thresholds to the 1-1 protocol (-14.6 ± 0.8 dB SNR; F(1,37) = 13.1, p < 0.002 compared to naïve performance), but no significant improvement was found for binaural benefits (7.8 ± 0.9 dB; F(1,37) = 2.7, p = 0.11, n.s.). Thresholds and binaural benefits did not significantly differ from those obtained on the last (7th) day of practice (t-tests: p = 0.045, n.s. given Bonferroni correction; compare days 7 & 8 for right panel, Figure 3.3.2).

Improvement was specific to the trained word pair

The same 10 subjects were subsequently tested on an untrained pair with a different phonological contrast (/dilen/-/tilen/), with the protocol they trained. Their data were
compared to that obtained from a group of naïve subjects, performing the same discrimination with the same protocol (see Section 2.2, Materials & Methods).

As with the /barul/-/parul/ word pair for naïve subjects (see Figure 3.3.1), diotic thresholds did not differ between the three protocols (compare naïve performance on Figure 3.3.4; effect of protocol: $F(2,56) = 0.36$, $p = 0.7$, n.s.), whereas binaural benefits did ($F(2,56) = 5.05$, $p < 0.01$). They were larger for the consistent protocol compared with the 1-1 (post-hoc Scheffe contrast: $p < 0.03$) and mixed ($p < 0.03$) protocols. Thus, the effect of low-level consistency is also relevant for this new word pair.

Figure 3.3.4 Learning remained largely specific to the trained word pair Thresholds (average ± SEM; top; filled bars: diotic thresholds, shaded bars: dichotic thresholds) and binaural benefits (bottom) for discrimination between the untrained pair /dilen/ & /tilen/. Performance is shown for naïve ($N = 10$ per protocol) and trained ($N = 10$ per protocol) subjects for all three protocols (bar colors as in Figure 3.3.2). Note the difference in thresholds and binaural benefits between consistent and the other two protocols for naïve subjects, similarly to what is found for /barul/-/parul/ (refer back to Figure 3.3.1).
Following training on /barul/ - /parul/, there was only a minor difference between naïve and trained performance (on /dilen/ - /tilen/) in diotic thresholds (-14.4 ± 0.5 & -15.3 ± 0.5 dB SNR for naïve and trained, respectively; effect of state: F(1,56) = 4.3, p < 0.05), which mainly stem from the reduced thresholds for the consistent protocol. Binaural benefits did not increase following training (F(1,56) = 0.02, p = 0.9, n.s.; Figure 3.3.4). Thus, for the untrained pair, use of binaural information did not improve following training on a different pair, for all three protocols. To a large extent, learning is therefore specific to the phonetic contrast of the trained pair.

Since diotic thresholds for naïve performance are about 3 dB lower for the /dilen/ - /tilen/ discrimination than for the /barul/ - /parul/ discrimination, it was not clear whether the low thresholds for /dilen/ - /tilen/ following /barul/ - /parul/ training are the result of learning or not. Moreover, it was not clear whether improvement in these thresholds, being so low to begin with, could occur following training. In order to verify the specificity of learning to the trained pair, I therefore performed a complementary control study, in which the reversed pattern was tested: i.e. subjects were trained on the /dilen/ - /tilen/ discrimination using the consistent protocol, and tested for transfer to discrimination between /barul/ and /parul/.

![Training on /dilen/-/tilen/, consistent protocol](image)
Figure 3.3.5 Specificity of training on /dilen/ - /tilen/ Subjects were trained on discrimination between /dilen/ - /tilen/ for 7 days, under the consistent protocol. On the 8th day, performance on /barul/ - /parul/ was tested (see Materials & Methods). Top: diotic (solid line) and dichotic (dashed line) thresholds (average ± SEM; N = 10) obtained during training: Bottom: corresponding binaural benefits. Results of the /barul/ - /parul/ discrimination are presented in light blue, for both naïve subjects (taken from Experiment 1, here in day 1) and following /dilen/ - /tilen/ training (day 8).

The results obtained from a group of 10 subjects (see Section 2.2, Materials & Methods) showed that learning did take place: following training, diotic thresholds for discrimination between /dilen/ and /tilen/ improved, from -14 ± 0.5 to -16.6 ± 0.6 dB SNR (F(6,48) = 6.4, p < 0.005), whereas binaural benefits remained large and constant throughout training, as is expected for the consistent protocol, in the order of 8 dB (F(6,48) = 1.2, p = 0.3, n.s.; Figure 3.3.5). In contrast, no significant generalization to the untrained pair (/barul/ - /parul/) occurred: thresholds for the /barul/ - /parul/ discrimination, following the /dilen/ - /tilen/ training remained ~4 dB higher than those for /dilen/ - /tilen/ (see day 8 on Figure 3.3.5), and did not significantly differ from those of the naïve group (t-test: p = 0.8 & p = 0.4, n.s, for diotic thresholds and binaural benefits, respectively.; compare light blue bars on days 1 & 8 in Figure 3.3.5). These results strengthen the conclusion that learning is specific to the trained phonological contrast.

Summary of results of Study 3

The results of Study 3 showed that following multi-session training on the discrimination between phonologically similar words, /barul/ and /parul/, similar and significant improvements in absolute (diotic) thresholds were obtained following training on both 3 protocols (consistent, 1-1 and mixed). In contrast, binaural benefits increased, already by the second day of training, only for the 1-1 protocol, for which naïve binaural benefits were non-optimal (as for the mixed protocol). For the remaining two protocols,
binaural benefits remained unchanged throughout training, being either low (for the mixed protocol) or high (for the consistent protocol) throughout training.

The large binaural benefits for the consistent and 1-1 protocols fully transferred to the mixed protocol following learning. In contrast, for all three protocol groups, learning was found to be specific to the trained word pair and did not transfer to an untrained pair with a different phonetic contrast.
Chapter 4: Discussion

Section 4.1: Summary of results

Auditory information is processed in a hierarchical manner. Local features are analyzed at lower level stations and higher level representations are more global and abstract. In this study I tested the hypothesis that the relations between stimulus processing along the auditory hierarchy and auditory perception follow the concepts of the Reverse Hierarchy Theory (RHT; Ahissar and Hochstein, 1997, 2004; Hochstein and Ahissar, 2002).

Towards this goal, I tested the efficiency of the use of low-level binaural information for high-level tasks which involved the extraction of speech signals from noise in three studies. In the first two studies, Study 1 and Study 2, I contrasted the predictions of three views, the unlimited view (Green and Swets, 1966; Siebert, 1965; Siebert, 1968), the limited capacity view (Allport, 1980; Bundesen, 1990; Lavie, 2005) and the Reverse Hierarchy Theory (Ahissar and Hochstein, 1997), each yielding specific (and different) predictions regarding the ability to use low-level information. I compared the experimental results obtained from a large group of listeners to those obtained from a theoretical “ideal listener” model, for which all low-level information is fully utilized. In Study 3 I assessed the predictions of RHT for perceptual learning of speech in noise by testing the impact of the variability of low-level binaural information on the ability to learn to use these cues more effectively.

I found that when the set of stimuli was composed of phonologically very different words, binaural benefits were optimal, matching those predicted by the ideal listener model under a variety of tasks demands (requiring either mere identification or additional semantic processing; Study 1, Experiments I & II), task difficulty levels (requiring either 60 or 80% correct performance; Study 2), and for different binaural protocols (cues were either consistent or randomly interleaved throughout the block; Study 1, Experiments III
DISCUSSION

& IV). However, when exactly the same conditions were administered with phonologically similar pairs, differing only in a single phoneme (e.g. /tamid/ vs. /amid/), binaural benefits were substantially lower than those predicted by the ideal listener model under most conditions and were highly sensitive to the test conditions. In this case, the only condition in which subjects achieved ideal listener level of performance was an identification task (i.e. no requirement for semantic processing) performed under a consistent binaural protocol. The difference found between phonologically similar and different words in utilization of low-level information could not be explained in terms of differences in available low-level binaural information, since the ideal listener model explicitly accounts for these differences. Moreover, the binaural benefits predicted by the ideal listener model (and hence by the unlimited view) were achieved for the phonologically similar pair in Experiment I of Study 1 and in Experiment II of Study 2.

In Study 3 I employed a multi-day training to test the effect of stimulus presentation protocol on the ability to improve in utilization of low-level binaural cues for the discrimination between two words comprising a phonologically similar pair. I found that even following several days' practice, full utilization of low-level cues was obtained only when the relevant information was consistent or temporally-ordered (presented in a predictable manner, with odd diotic trials and even dichotic trials) throughout the block, but not when it was randomly interleaved (mixed) throughout the block. Once obtained, the improvement in binaural benefits was generalized to the less certain (mixed) conditions as well. However, it was highly specific to the trained contrast regardless of protocol.

My conclusion is that only RHT accounts for the overall pattern of the results. While the results for the phonologically different pairs (Study 1 & Study 2) are fully accounted for by the unlimited view, which predicts ideal listener levels of performance under all conditions, the results for the phonologically similar pairs, showing mostly sub-optimal utilization of low-level cues, contrast with this view. The results of Study 2, which dissociate between task difficulty and the ability to use binaural cues, rule out limited capacity models of attention according to which performance is expected to be limited by attentional load (i.e. task difficulty) per se. The constraints on the use of low-level
 information should therefore be formulated in terms of the properties of the stimulus set, rather than in terms of behavioral difficulty or general attentional demands. RHT, which concretely addresses the relations between utilization of low-level information for perception and the nature of representations of the stimuli along the processing hierarchy, parsimoniously account for this set of results: The general auditory RHT account is detailed in Section 4.2. The results of Study 3 further validate RHT predictions that cross-trial variability in the informative low-level population precludes access to low-levels even with massive training. The auditory RHT account for the perceptual learning case is detailed in Section 4.3.

Since this is the first study to assess the validity of RHT predictions in the auditory modality, in the remaining sections of the Discussion I provide a detailed description of auditory RHT in general (Section 4.2) and in relation to perceptual learning (Section 4.3), followed by the RHT account for several auditory and speech perception phenomena (Section 4.4). A description of the caveats, the open questions and future research directions concludes the Discussion (Section 4.5).
Section 4.2: Binaural benefits and the Reverse Hierarchy Theory

Hierarchies and reverse hierarchies in sensory processing

The basic tenets of RHT (Ahissar and Hochstein, 1997; 2004; Hochstein and Ahissar, 2002) are the presence of a local-to-global hierarchy of stimulus representations, and the presence of massive feedback connections throughout this hierarchy. Feedback connections are well established throughout the brain (Bajo and Moore, 2005; 2006). There is also an increasing amount of evidence for an auditory processing hierarchy in which lower stations represent acoustic features of sounds, whereas higher stations represent sounds more abstractly (Kaas and Hackett, 1998; Rauschecker, 1998b; Romanski et al., 1999; Warren and Griffiths, 2003; Wessinger et al., 2001; Zatorre and Belin, 2001). Along this hierarchy, acoustic fidelity is presumably gradually replaced by ecologically relevant representations (Chechik et al., 2006; Las et al., 2005; Nelken, 2004; Nelken and Ahissar, 2006; Wang et al., 2005). In analogy to the visual system, low-level representations are determined by the physical (acoustic or visual) nature of the stimulus, and high-level representations converge across different low-level representations that denote the same objects or events.

Anatomically, the lower, acoustic levels may roughly correspond to the stages up to, and including, the inferior colliculus (IC, e.g. Nelken, 2004), whereas the more abstract levels, though less well understood, may correspond to cortical areas. For example, according to some recent imaging data, cortical areas ventral (“belt”) and posterior (“parabelt”) to A1, and portions of the superior temporal sulcus (STS), process temporal and spectral feature combinations that may be related to phoneme discrimination (Binder, 2000; Demonet et al., 1992; Hickok and Poeppel, 2007; Price et al., 1996; Scott and Johnsrude, 2003). Cortical areas in posterior middle temporal regions (Binder, 2000; Binder et al., 2000; Binder et al., 1997; Davis and Johnsrude, 2003; Jancke et al., 2002) may process semantic information.
The tradeoff between global perception and perception of details

RHT asserts that perception and perceptual learning are based, by default, on stimulus representations at higher levels of the processing hierarchy, which are immediately accessible to perception. This functional structure allows rapid, and yet crude, evaluation of meaningful objects and events. The finding that binaural benefits utilize all low-level information in the case of phonologically different words tested and using the given simulation parameters is therefore consistent with RHT assertions. This is because the phonological representations of phonologically distant stimuli are as informative as the low-level (acoustic) representations, and therefore, can be used to achieve the performance level suggested by ideal observer models.

However, in the case of phonologically similar words, the phonological representations of the two words are close and largely overlapping. This leads to information loss about the acoustic differences between them (which are as large as those of the phonologically different words, as is shown by the large binaural differences generated by an ideal listener; see Section 3.1), since much of the acoustic difference between the two words is irrelevant at the phonological level and is therefore not explicitly represented (e.g. Darwin and Hukin, 1999, see Figure 2.2.1). To discriminate between the two words, it is necessary to access the discriminative features that are represented at lower, acoustic representation levels. These features depend on the binaural configuration (Ahissar et al., 1992; Stecker et al., 2005); they are energy cues in diotic trials and correlation cues in dichotic trials, which are presumably coded in different low-level representations (Ahissar et al., 1992; Casseday et al., 2002). According to RHT, access to the appropriate lower level representations requires a backward search down the auditory hierarchy, since there are a number of possibly informative low-level representations. For example, there are monaural pathways through the ventral and dorsal cochlear nuclei, binaural pathways through the medial and lateral superior olive nuclei (SOC), and pathways through the nuclei of the lateral lemniscus, all of which reach the inferior colliculus and remain partially segregated there (see Casseday et al., 2002). RHT postulates that the backward search for the specific low-level
neural population that best represents the discriminative acoustic features is difficult. In particular, it is gradual, and cannot be conducted from scratch in every trial; RHT suggests that this search is aimed at identifying a population that is consistently informative across several trials (Ahissar and Hochstein, 2004).

This logic therefore accounts for the substantially reduced binaural benefits in the case of phonologically similar pairs when binaural conditions vary in an uncertain (mixed) manner across trials: optimal performance presumably requires access to low-level populations that vary from trial to trial. Since identification of the most discriminative population requires several stimulus repetitions, a successful backward search can be achieved only in the consistent protocol.

Listeners’ limited ability to use binaural information in the semantic-association task, even when the binaural configuration is consistent across trials, can also be accounted for by this logic. Comprehending the visually presented word, which immediately follows the auditory presentation and changes on a trial-by-trial basis, requires access to higher, semantic representation levels on every trial and interferes with the backward search for informative low-level representations. Therefore, the requirement for semantic processing prevents access to low-level representations, and thus limits the use of binaural information in the case of phonologically similar words.
Section 4.3: RHT & perceptual learning: the importance of the training protocol

Learning to use binaural cues: consistent vs. mixed protocol

According to RHT, practice-induced improvement of discrimination between similar stimuli is the result of a gradually gained access to more informative, lower level populations (Ahissar and Hochstein, 2004). The cascade of learning is therefore top-down guided, starting from the high levels (the levels accessed by default by naïve performers, as was also shown by e.g. Furmanski et al., 2004; Mukai et al., 2007; Schwartz et al., 2002; Sigman et al., 2005) and reaches low levels as training progresses, if the discrimination requires fine resolution (Ahissar and Hochstein, 2004).

However, RHT asserts that reaching low levels throughout training would improve performance only if the following two conditions are provided. First, that it is the loss of this low-level information that was the limiting factor for naïve performance to begin with (as is the case for discrimination between phonologically similar words), and second, that the relevant low-level population could be correctly tracked throughout training. Allocating the right population depends on the consistency of the presentation of the relevant information (implicit or explicit) throughout the block. Therefore, a major prediction of RHT is that perceptual learning of fine details may not be attained without blocked presentation of the relevant cues. This prediction was indeed validated in several previous investigations, showing intensive training protocols with a narrow range of stimuli applied in a consistent manner (Adini et al., 2002; Fahle et al., 1995; Sagi and Tanne, 1994; Schoups et al., 1995; Schwartz et al., 2002). On the other hand, increasing stimulus variability, so that consecutive stimuli still belong to the same perceptual category (i.e. same high-level representation) is expected to activate different low-level populations in different trials, and would therefore not enable the relevant low-level population to be tracked. If no informative low-level population can be consistently tracked, RHT predicts that performance will be based on broad high-level representations, which lack resolution for fine discriminations. Such learning is therefore expected to show only a limited degree of improvement on the one hand, and substantial
DISCUSSION

generalization across low-level features on the other hand (Ahissar and Hochstein, 1997; 2004; Ahissar et al., 2008). This prediction was too confirmed in several recent investigations (Adini et al., 2004; Otto et al., 2006; Yu et al., 2004).

The results of Study 3 are also consistent with this view. In Study 3 I have manipulated the consistency of the presentation of the binaural cues throughout training, and tested its effect on the improvement in utilization of binaural information for discrimination between similar words. Unlike the direct measure of consistency usually employed in perceptual learning studies (e.g. testing roving the contrast in a contrast discrimination task, Adini et al., 2004), here I used the magnitude of the binaural benefit as a direct marker for low-level access. This marker made it possible to test the impact of implicit variability of the utilization low-level cues since listeners were unaware of this manipulation, according to their introspection, and it was not directly relevant for the task since binaural convergence occurs well before phonological contrasts are represented.

The results show that this implicit manipulation yielded similar effects to the ones obtained for explicit manipulations. Thus, subjects trained using a consistent protocol (separate blocks for diotic and dichotic trials) showed maximal utilization of binaural information throughout training, already from the first session. In contrast, subjects trained using a mixed protocol of the binaural cues showed no increase in utilization of binaural information throughout training, despite similar improvement in diotic thresholds to that obtained using the consistent protocol.

The fact that the loss of low-level micro-second binaural resolution is the limiting factor for the discrimination between the phonologically similar pseudo-words /barul/ and /parul/ was already shown in Study 1 & Study 2. This means that access to low levels is needed in order to obtain improvements in binaural benefits. Under the consistent presentation of binaural cues, such low-level access is possible already to naïve performers, from the very first session. Since the relevant cues for each binaural configuration are consistently repeated across trials, the correct top-down “pointer” for the relevant low-level population could be correctly allocated, and the backward search could be performed. When training is administered using the consistent protocol, diotic and dichotic thresholds improve, but to a similar extent, resulting in a constant and maximal binaural benefit throughout practice. The consistent protocol therefore enables
low-level access very quickly, well within the first session, and hence full utilization of low-level cues from the start.

The RHT reasoning posits that a mixed protocol, which employs trial-by-trial uncertainty in the binaural configuration, forces the use of high levels and does not enable tracking of the relevant low-level population. It therefore results in sub-optimal binaural benefits for the naïve listeners. Multi-day training performed under this protocol results in similar improvements of diotic and dichotic thresholds, keeping the binaural benefit constant and sub-optimal. In order for the backward search to be initiated and executed successfully, a top-down “pointer” to guide the search to the relevant population needs to be allocated. Under trial-by-trial uncertainty, such pointer cannot be allocated, not even following massive training. Since the relevant population can not be tracked, performance remains at high levels throughout training and does not reach lower levels, which are needed in order to obtain maximal binaural benefit.

Although this result is consistent with several others obtained for visual (Kuai et al., 2005; Otto et al., 2006) and auditory (e.g. Amitay et al., 2005) stimuli, a recent paper by Parkosadze and colleagues (2008) suggests, using perceptual learning with bisection stimuli, that learning under roving could occur providing extensive (in their case > 18000 trials) training. It is therefore possible that many more training trials would eventually result in improvements in binaural benefits in the case of Study 3. Whether improvement began when top-down mechanisms detected the two reference distances as distinct or slow bottom-up mechanisms were involved is hard deduce from that study, and it would require further testing in future studies.

Learning using temporally-patterned cues

The results obtained with the 1-1 protocol are the most surprising and challenging ones. In the 1-1 protocol, the variability is introduced in a single trial window, with odd trials being diotic and even trials being dichotic. Thus, on the one hand, each trial is different from the one preceding it. On the other hand, this structure, which repeats throughout the blocks (and throughout the entire training), is fully predictable. While naïve subjects could not fully utilize binaural information in this case, resulting in initial
binaural benefits similar to the ones obtained for the *mixed* protocol, trained subjects reached, already by the second day of training, maximal utilization of the binaural information, as in the *consistent* protocol. Temporally-structuring the implicit data in a predictable manner therefore afforded similar levels of efficiency as in the *consistent* protocol within very few training sessions. For effective learning to take place (i.e. reach low levels), the information should therefore be made *predictable*, even when the predictability is not explicitly apparent to the subjects.

The RHT account in this case is not straightforward. When several parameters are simultaneously involved throughout training, RHT asserts that an effective way of training would be to present the relevant cues in a *blocked* manner, with each cue presented in a separate block. Blocked presentations provide useful top-down expectations, or attentional pointers, which guide the backward search, such that the low-level populations can still be tracked and accessed. However, what is considered a useful “block size” is not well-defined in RHT. In the case of a *consistent* protocol, a block composed of 75 trials is enough to induce low-level access for naïve performers. Here, however, the information is not “blocked” but rather temporally-ordered throughout the block, such that each trial has a unique and predictable identity.

Still, in order for the RHT’s higher-to-lower (or easy-to-difficult) cascade of learning to take place, a top-down cue is needed in order to guide learning to lower levels. The results suggest that such a cue is provided in case of the *1-1* protocol, despite the switch in binaural configuration every other trial. It could be that the high-level discriminability between the two configurations, diotic and dichotic, provided the top-down cue: in the initial SNR chosen for the experiment, the clarity of the signal was larger in the dichotic than in the diotic case. The two configurations were therefore distinct enough (although not explicitly so); this distinctiveness along with the repeating temporal pattern presumably enabled the top-down search and hence substantial learning. This reasoning may imply that had different initial SNRs were used for the two configurations, such that the diotic and dichotic configuration would not have been distinct enough, learning would not be afforded in the temporally-patterned case.

A recent conceptualization proposed by Zhang and colleagues (2008) further elaborates the RHT concept for the temporally-patterned case as well. Zhang et al. (2008)
base their suggestion on their results with perceptual learning using a visual contrast discrimination task. The authors found that learning using multiple contrasts took place when these contrasts were presented in a fixed sequence (i.e. with a temporal pattern) with a consistent rhythm, but not when they were presented in random order (roving). Learning with roving only occurred if the stimuli used were either very different from one another, or if each stimulus was assigned a distinct semantic tag. According to Zhang et al.’s “stimulus tagging model”, for multi stimuli learning to occur, a conceptual tagging of each stimulus needs to take place, in order to switch attention to the appropriate perceptual template. In line with RHT, this model emphasizes top-down influence in perceptual learning, but further stresses that this influence could be conceptual or semantic. Stimulus temporal patterning with rhythm may enable learning by providing the unique tag for each roving stimulus, which enables the backward search required by RHT.

The stimulus tagging model therefore suggests that the 1-1 protocol employed in Study 3 enabled the correct tagging of each configuration, which initiated the top-down search. While the results of Zhang et al. (2008) are mainly consistent with what I found here, they report that an effective block size for learning is composed of a minimum of 5-8 consecutive trials. Aside from the obvious differences between the two studies (auditory vs. visual, different tasks, etc.), the number of different stimuli used was different: while they used four different contrasts, I interleaved only two binaural configurations. In fact, Zhang et al. (2008) report substantial learning even under roving when only two contrasts, which are very different from one another, are used.

I conclude that while the consistency of the implicit information is a sufficient condition for full utilization of binaural information, it is not a necessary one. Learning to use these cues can be done when the cues are interleaved, provided that the cues are different from one another and are presented in a temporally-patterned manner: this presumably enables the correct “tagging” of each stimulus, to initiate the backward search. However, a more specific description of the protocols which enable learning in this case should be added to RHT.
Transfer of the improvement across protocols

The results of Study 3 further show that once acquired, using either the consistent or the 1-1 protocols, the improved binaural benefits transfer to the mixed (and hence less predictable) case as well. In contrast, training using the mixed protocol induces some improvement in thresholds when tested using the 1-1 protocol, but not a significant improvement in binaural benefits.

RHT does not address this question directly, and should thus be extended to include this scenario. These results suggest that once the backward track has been systematically performed, the benefit gained transfers to more uncertain conditions. According to RHT, this transfer requires that the same cues that were used in training would be useful also for the testing phase, and that the same context is maintained between sessions (Ahissar et al., 2008). The results of Study 3 may imply that these requirements were met here: since the protocol manipulation was implicit, the same discrimination cues that were sharpened during training were required for all three protocols. Thus, the attentional pointer which guides the (implicit) search could be used in the less predictable case, generating similar binaural benefits in a single mixed session that followed either a consistent or 1-1 session.

These results are in line with recent results obtained in the visual domain (Adini et al., 2002; Zhang et al., 2008). For example, Zhang et al. (2008) demonstrated that following perceptual learning on a contrast discrimination task using temporal structures, performance was undisturbed even by a single roving session that occurred over 4 hours after training or by extended roving training. Thus, performance transferred to the roved condition. Once perceptual learning has been completed and the stimulus information consolidated, improved performance could no longer be reversed by extended roved training. Using the authors’ “stimulus tagging model” terminology, it is likely that once multiple stimuli have been properly tagged after temporally-patterned training, stored stimulus information can be accurately and efficiently retrieved to guide visual discrimination regardless of the stimulus temporal context. The time periods mentioned by Zhang and colleagues (2008) are also in line with the results reported here: the generalization to the roving condition only occurred if the roved session was
administered over 4 hours after the training sessions. In Study 3, at least one day separated between the end of learning and testing (see Section 2.1), so the improvement gained could be evident for the mixed case as well.

**Specificity of learning to the trained pair**

The final aspect in the results of Study 3 is the relative specificity of learning to the trained pair (/barul/ - /parul/), regardless of protocol. Thus, for all training groups, learning generally did not transfer to an untrained pair (/dilen/ - /tilen/; though see detailed results in Section 3.3). Training on the reverse pattern showed similar specificity to the trained pair: thus, training for 7 days to discriminate /dilen/ from /tilen/ did not transfer to discrimination between /barul/ and /parul/.

These results therefore imply that the learning effects observed in this study are not generalized effects, as learning how to perform the task or the experimental requirements (often referred to as “procedural learning”). Instead, the effects seem highly specific to the discrimination between the exact contrasts used in training. Throughout training, subjects therefore learned specific cues that help dissociate /barul/ from /parul/: these cues are presumably not efficient for discrimination between other phonetic contrasts.
Section 4.4: Auditory RHT: a broader perspective

In the auditory modality, the consequences of our immediate percepts being exclusively based on high-level representations are even more dramatic than those in the visual modality. Thus, when we listen to music we can identify the tune and tag it. However, typically we can not explicitly access the information which is implicitly used for this identification (e.g. to decide whether two subsequent notes are going "up" or "down"). When we hear a speech sound (e.g. a syllable) we are not consciously aware of the formant frequencies and transitions that compose it, but at the same time have no problem categorizing it or even repeating it.

Applying the basic principles of RHT to the auditory modality enables to shed new light on several well-known phenomena in the auditory literature. These include, among others, context effects in speech perception and informational masking. In the following section I provide a parsimonious explanation to these phenomena under the umbrella of Auditory RHT.

Context effects in speech perception

The well-documented context effects in perception of speech stimuli can be explained using RHT principles. In everyday conversational situations, the overall context usually provides prior information that limits the possible word set to words that are semantically related, but are typically phonologically dissimilar. Thus, we are likely to be required to discriminate between “day” and “night” in a conversation, but not between “day” and “bay”, which are phonologically, but not semantically, close. According to auditory RHT, the auditory system performs these ecologically-prevalent discriminations efficiently, as an ideal listener would, regardless of the attentional load imposed by the conversation.

Indeed, when it comes to speech perception, there is a large benefit to relying on context and memory cues, rather than on the incoming bottom-up signal. The segmentation problem, the problem of separating the incoming acoustic energy into meaningful units as words, is just one problem which occurs when relying solely on low-level cues to decipher speech. The large variability in the speech signals, both in the
signals of the same speaker as a result of effects such as co-articulation, and in the spectrums of different speakers, would prevent us to derive similar concepts had we relied solely on the bottom-up input. In that sense, context enables the clear perception of speech, and even more so in noisy environments. In line with this reasoning, several studies showed that top-down processes in speech perception can help us segment the acoustic signal, as well as to recognize phonemes and words. Miller and Isard (1963) and Salasoo and Pisoni (1985), for example, showed that words are understood better when presented in context, than when presented as isolated items in a list. This effect is greater when stimuli are presented in background noise.

However, according to RHT there is an additional cost, or tradeoff, to relying solely on the context, which is represented at high levels: the context can sometimes help us, but at the same time, using high-level representations disables concurrent access to low levels which store the fine details. In those cases that require finer phonological discriminations and optimal performance cannot be provided by the broad high-level representations, a different process than solely using high-level context occurs. Thus, for example, when the speaker might say either /day/ or /bay/ outside of context, we are likely to ask “what?”, “forcing” the speaker to repeat, perhaps at a higher signal level, which improves SNR. In parallel, an implicit attempt to apply a backward search to find more discriminative low-level representations is made. A successful backward search requires a relatively specific expectation (/day/ vs. /bay/), another repetition of the same condition, but cannot be done when high levels are engaged (e.g. when concurrent comprehension is required). Thus, when attention is directed to higher levels (e.g. by requiring to process semantic meaning), low-level information is not accessible, even when it can facilitate perception.

Evidence for such tradeoff has also been reported in some of the speech perception literature. The “phonemic restoration effect” of Warren (1970), for example, demonstrates nicely the superiority of the context and meaning over low-level input. Warren presented listeners a single sentence (“the state governors met with their respective legislatures convening in the capital city”), in which the first /s/ in “legislatures” was replaced by a cough. None of the subjects could identify the correct position of the cough in the sentence, and neither of them noticed the missing /s/. In terms
of RHT, this example shows that when one utilizes higher level cues, as context and expectations, one cannot capture low-level modifications of the input, presumably represented at lower levels of the hierarchy.

From the opposite direction, there are numerous examples showing that the context can alter our low-level perception. In a research by Ladefoged & Broadbent (1957), the authors tape-recorded multiple versions of the sentence “Please say what this word is”, with each version having characteristic cadence; different versions of the sentence sounded as if they were spoken by different speakers. Listeners heard the different versions of the sentence, and following each presentation heard a single word uttered either with the same cadence as the immediately preceding sentence or with one of the other five possible cadences. Listeners frequently misidentified the last word when its cadence differed from that of the preceding phrase. For instance, the same physical utterance was heard as “bit” when preceded by one sentence version and as “bet” when preceded by another. This provides a strong demonstration of context effects in speech perception, which alter our fine-grained perception. Again, the RHT-based interpretation is that the hierarchical tradeoff does not enable simultaneous low-level and high-level access.

To summarize, RHT proposes that this tradeoff reflects the dynamic balance, characteristic of perception in general, between the breadth of perception at a given moment, and the efficiency in retrieving detailed stimulus information. The hierarchical structure of representations allows fast, though crude, perception of meaningful and ecologically-prevalent events, and yet full utilization of the information stored, in a system which has relatively slow time constants. However, these broad, abstract high-level representations retain only information which is relevant for crude, phonologically-based, interpretations. Attaining detailed low-level acoustic information requires slow internal search processes, which can be successfully completed only under privileged, typically non-ecological, conditions. Thus, these less-prevalent discriminations are either fast or use all low-level information, but not both.
Informational masking & RHT: high-level uncertainty effects

Auditory RHT can also account for a number of results obtained in studies which characterized the limits of high-level auditory selection mechanisms. In these studies, also termed “informational masking” studies, the masking stimulus is designed so that it is perceived as similar to the target stimulus and is therefore partially overlapping with the target stimulus at high levels, but is acoustically non-overlapping with the target and therefore well-segregated at early, acoustic, low levels of the hierarchy (Durlach et al., 2005; Durlach et al., 2003a; Durlach et al., 2003b; Kidd et al., 2002; Kidd et al., 2003; Kidd et al., 2005b; Lutfi et al., 2003; Tang and Richards, 2003). This is in contrast to the traditional energetic masking studies, in which the masker overlaps with the target signal already at the auditory periphery (Srinivasan and Wang, 2008). Informational, unlike energetic, masking, seems to occur at higher levels, as a result of attentional confusion due to the similarity or uncertainty, and not directly as a result of peripheral factors (Durlach et al., 2003a).

In a typical informational masking study, listeners are required to identify a target signal (either tone or speech signal) that is masked by either randomly-varying multi-tone complexes or other speech stimuli, that do not contain energy at the relevant signal band (Durlach et al., 2005; Durlach et al., 2003b; Gutschalk et al., 2008; Kidd et al., 2005a; Kidd et al., 2005b). Under these conditions, many listeners are confused by the high level masking and erroneously select the masker rather than the target stimulus. Most of the informational masking studies show, in addition, that the effects of release from informational masking (due, for example, to spatial separation of the signal and the masker) far exceed the ones seen under pure energetic masking (Arbogast et al., 2002; Brungart and Simpson, 2002; Freyman et al., 2001; Freyman et al., 1999; Kidd et al., 1994; Kidd et al., 1998). Informational masking studies, as opposed to energetic masking ones, also show large inter-individual differences in measured thresholds (Tang and Richards, 2003).

According to RHT, such high-level masking results from confusions and similarity at the global, high level. In order not to be confused, listeners should retrieve the relevant
DISCUSSION

low-level information: at lower levels, the target and the masker are well-segregated (since there is no peripheral masking). However, the conditions under which these experiments are performed do not enable backtracking to the appropriate low level representations. The difficulty in back-tracking to lower levels may stem, in some of the studies, from the need to globally process the stimulus before selecting low-level populations, as when listeners are required to process the meaning or the name of the target speaker (e.g. Brungart, 2001); This requirement is similar to that posed in the semantic-association task of Study 1 (see Section 3.1). In other studies, the difficulty in backtracking is probably due to the cross-trial variability, or uncertainty in the properties of the masker (Durlach et al., 2005; Oh and Lutfi, 1999); similar difficulty was found in the randomly-interleaved (mixed) protocol in my experiments, which forced high-level use since the relevant low-level population could not be tracked. A proposed RHT scheme which incorporates informational masking is depicted in Figure 4.4.1.

To summarize, the auditory RHT account seems to explain the large pool of “informational masking” results, by which masking occurs at higher rather than lower, peripheral levels.

Figure 4.4.1 Hierarchies and Reverse Hierarchies in the Auditory Modality A hypothetical description of the auditory feedforward and feedback hierarchies. In the feedforward hierarchy, lower levels perform spectro-temporal analysis of the incoming auditory signal. A1 presumably performs some extraction of the relevant low-level cues, while higher areas perform stream formation. The highest cortical areas probably contribute to our overall auditory experience, of continuous speech, musical tunes etc. The feedback reverse hierarchy can use top-down connections to extract the relevant low-level information (modified from Shamma, 2008).
Section 4.5: Auditory RHT: caveats, future directions and practical implications

In the current work I have established the basic principles of the Reverse Hierarchy Theory, initially designed to account for visual perception, in the auditory modality. This work therefore suggests that similar defaults and tradeoffs characterize the relations between processing hierarchies and perception in the visual and the auditory modalities (Figure 4.5.1). Applying RHT concepts to the auditory modality makes it possible to establish a parsimonious account of hierarchical processing in the brain in general. This work is one step towards achieving such a unified framework.

However, this study points to the similarities between the systems in broad strokes only; the specific implementations of these mutual principles may vary greatly between the two systems, due to the large differences between them. These differences include the different physical dimensions in the stimuli dealt with by each system (spatial dimension in the visual system, temporal dimension in the auditory system), the differences in the stations along the pathway (e.g. A1 is not analogous to V1; Fort et al., 2002) and the fact that the functional distinction between the “what” and “where” (dorsal and ventral, respectively) pathways in the auditory modality is highly debated (Belin and Zatorre, 2000; Cohen and Wessinger, 1999). Thus, while the results of my study imply that cortical systems for different sensory modalities may share principles of functional organization, the interpretation of such shared principles should be done carefully, while considering the unique dimensions of each modality. Further research is needed in order to validate this analogy and in order to provide a unified framework of processing principles in the brain in general.

In the following paragraphs I propose directions for future research. These should mainly focus on deepening our understanding of auditory RHT, using different acoustical parameters and different methodologies, on the one hand, and using the binaural benefit tool as a means to probe the efficiency of auditory processing on the other hand.
DISCUSSION

Figure 4.5.1 Schematic description of the local-to-global processing hierarchies (a) An example of the visual hierarchy (adapted from: Hochstein & Ahissar, 2002) and (b) an example of the auditory hierarchy (adapted from: Ahissar et al., 2008).

Deepening our understanding of auditory RHT

Initially, future studies would need to establish the empirical relevance of RHT to other auditory parameters, as pitch, spectral processing and sound localization. In this work I have tested only one form of low-level cue, the binaural benefit, but many other low-level cues exist and need to be tested for matching the predictions of the theory. Similarly, complex auditory stimuli other than speech could be used; these include complex tunes, frequency modulated (FM) and amplitude modulated (AM) sweeps, that have complexity similar to that found in speech signals, and the like.

Methodologically, brain imaging tools, as magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI) could be used to functionally test RHT predictions. MEG is known for its excellent temporal resolution (Hamalainen et al., 1993), and has therefore been extensively used in auditory research, whereas fMRI enables to localize specific brain areas with relatively high spatial accuracy (see Logothetis, 2008). Several recent results in the visual modality obtained using fMRI are
in line with RHT’s top-down cascade of perception. These imaging tools could be used separately or combined to test the timeline of activation of high level and low level auditory brain areas, and to examine whether it follows the reverse top-down pattern suggested by RHT.

**Characterizing the deficits of specific impaired populations**

Using binaural interactions as a tool, I have shown here that in the auditory system can be highly efficient in utilizing the available low-level binaural information. However, when it is required to perform discriminations that are ecologically less likely, between similarly-sounding words, there is a tradeoff between accuracy and speed of processing. The normal auditory system is therefore not always efficient in using all low-level information in the system. It would be interesting to further examine whether different populations that suffer from various central auditory processing deficits (rather than peripheral ones), exhibit similar pattern of efficiency and inefficiency of utilization of binaural information. For these populations, the binaural benefit may therefore serve as a tool for characterizing the efficiency of auditory processing, compared to that found in the general population.

The results of **Study 3** have further shown that in the general population, certain perceptual learning procedures could be used to improve utilization of these cues. Specifically, in line with RHT predictions, when training is done under consistent or temporally patterned conditions, utilization of low-level cues is maximized, and the improvement is fully transferred to the roved condition. Such understanding could be used in the near future to improve existing training procedures, designed to enhance the brain’s perceptual abilities. These training procedures use perceptual learning principles in order to affect the plastic brain, and were used for improving abilities in the general adult population, as well as in special or impaired populations (e.g. dyslexics, see Merzenich et al., 1996). Training procedures similar to the ones used here, that emphasize the importance of the stimulus protocol, may be applied to general and
specific populations and by that enhance auditory abilities by modifying central representations. Of course, this is just a first step. Further research is needed in order to establish the relevance of these results outside the specific laboratory context, to less predictable environments.
References


REFERENCES


REFERENCES

REFERENCES


Myers, C. S., and Rabiner, L. R. (1981). A comparative study of several dynamic time-
warping algorithms for connected word recognition. The Bell System Technical Journal
60, 1389-1409.
Neurobiol 18, 413-417.
system: the role of primary auditory cortex. In Dynamics of speech production and
343-354.
Front Neurosci 2, 107-113.
Percept Psychophys 60, 355-376.
Soc Am 106, 3521-3528.
colliculus to binaural masking level differences created by inverting the noise in one ear.
J Neurophysiol 84, 844-852.
operculum in rhesus monkey and its projection pattern. Z Anat Entwicklungsgesch 139,
127-161.
Perceptual learning of bisection stimuli under roving: slow and largely specific. J Vis 8, 5
1-8.
peripheral auditory processing: A modular architecture and a software platform. J Acoust
Soc Am 98, 1890-1894.


REFERENCES


REFERENCES


REFERENCES


