Detection and Recognition Threshold of Sound Sources in Noise

Tjeerd Andringa (tjeerd@.ai.rug.nl), Carina Pals (c.pals@ai.rug.nl)
Auditory Cognition Group, Department of Artificial Intelligence, University of Groningen
Bernoulliborg, Nijenborgh 9, 9747 AG Groningen, the Netherlands

Abstract

This study examines detection and recognition thresholds for environmental sounds in the presence of noise, with and without specific expectations about the targets. Human listeners were presented with a selection of everyday sounds masked by noise at initially increasing and later decreasing local signal-to-noise ratios. Participants had to indicate if they either detected or recognized the sound in the masking noise. The targets were selected from a database of environmental sounds. A previous perceptual similarity experiment on this database led to a separation in three classes that might be interpreted as predominantly tonal, pulsal, or noisy. We therefore separated the targets in these three groups. The resulting pattern of detection and recognition thresholds, with and without previous exposure to the target, suggest an 8 dB benefit for recognition and a 2dB benefit for detection. For repetitive sounds the detection benefit increased to 7 dB. The overall pattern of results provides support for the suggestion that sound recognition may be a combination of checking the presence of expected targets and a signal driven search in case of an unexpected target. The detection and recognition thresholds, suggest that human auditory perception might indeed employ different strategies for detecting tonal, pulsal, and noisy sounds.

Keywords: Auditory perception; environmental sound; sound source recognition.

Introduction

Humans detect and recognize sound sources to interpret and act on events in their environment. However, surprisingly little research has been aimed at understanding environmental sound recognition. Some interesting studies in this field are Gaver (1993), Ballas (1993), Gygi, Kidd, & Watson (2007), and Guastavino (2007). Research on human everyday listening (defined conform Gaver, 1993) can provide a broader and more integrated understanding of auditory perception in the real world. One of the aims of this research is to discover which algorithms facilitate quick and adequate sound source recognition.

Shinn-Cunningham (2008) argues that top-down attention influences auditory object (percept) formation and auditory object recognition. Therefore a theory of everyday sound recognition must not only explain environmental sound recognition, it must also include the effects of acoustic, contextual, or knowledge-based priming. Chiu (1995) concluded that priming of environmental sounds is predominantly perceptual and not influenced by only naming the source. This leads to the question of how previous exposure can influence priming.

We can hypothesize that different attentional states correspond to different detection and recognition strategies. For example, when listeners know which sounds to expect they may use detailed expectations and *check* these. In contrast, a sound for which no useful a priori expectation exists might force a demanding memory *search* through all possible sound classes that can explain the input. Thresholds shifts are familiar to experimenters that are often exposed to stimuli masked by noise multiple times. They can detect these targets better than “naïve” listeners. However there has been little research into this phenomenon, especially not for environmental sounds.

Experimental Paradigm

We describe an experimental paradigm for studying human (environmental) sound perception and recognition. It is designed to estimate detection and recognition thresholds with and without prior knowledge, so that it can be used to separate listening modes due to expectation differences.

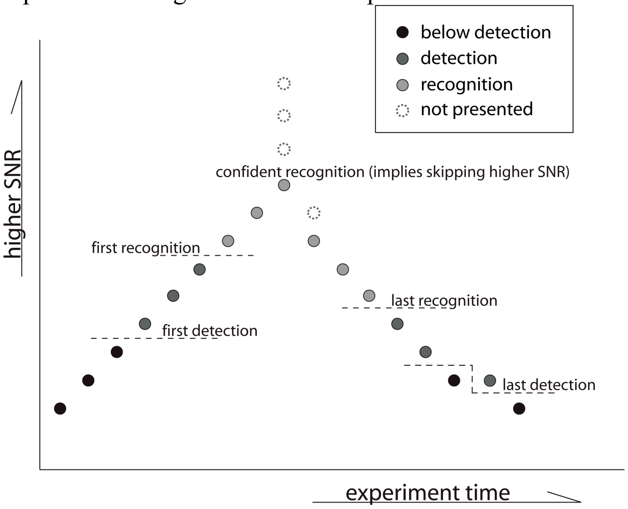

Figure 1 Overview of the experimental design. The 10 different targets were presented in random alternating order. Each sound is first presented with a gradually increasing target-to-noise ratio. After presentation without noise, the target-to-noise ratio decreases again, until the target can no longer be recognized or detected.

The experimental procedure is depicted in figure 1. It involves the presentation of random alternating individual target sounds, referred to as the signal, masked by broadband noise. At every presentation in the first phase of the experiment the signal-to-noise ratio (SNR) of each target increases (detection becomes easier) while in phase two the SNR decreases (detection becomes more difficult). In the first phase the listener does not know the target so a mental search for a suitable class to explain the input is required. In the second phase the participant has heard the target sound without any noise and the participant can use this information to check whether or not this specific target is present or not. For each experimental item the participant is asked whether or not a deviation from noise is detected and whether it is possible to recognize the deviation.

This experimental paradigm can be used to identify robust and informative features for environmental sound recognition by measuring the detection and recognition thresholds for a wide variety of common sounds, and by analyzing the relation between unmasked signal content or signal type and associated thresholds.

Environmental sounds

The sounds used in the current experiment are a subset of those used in a recent perceptual similarity study by Gygi et al. (2007). The full set of 100 short sounds compiled by Gygi are of an unusual composition because they are selected to be both minimally controlled as well as representative for the typical environmental sound production processes described by Gaver (1993). Since they have been derived from a range of Internet sources little is known of the sounds. The only information is the general type of the sound source and the fact that they appear to be dominated by a single source, which is representative for many sounds in realistic environmental conditions.

The study of Gygi et al. (2007) showed a division in three groups of sound sources, which Gygi described as harmonic sounds, discrete impact sounds, and continuous sounds. Although Gygi correlated the perceptual results to a large number of objective signal descriptors he could not identify definite acoustical features that determine the categorization of sounds in these three categories. However, Gygi's results suggest that the composition in terms of tonal, pulse-like, and noise-like components determines much of the perceptual distance between sound sources.

In a previous study we (Andringa 2008) used the energy fractions of tone, pulse, and aperiodic (noisy) energy to predict perceptual similarity between sounds from Gygi's set. We hypothesized three different processing routes, because these three basic signal types each require a different combination of spectro-temporal evidence. A pulse is short and broadband and requires a combination of information over many frequency channels and a very short temporal integration interval. A tone requires a very narrow frequency range, but a longer temporal interval. Aperiodic noise requires both a range of frequency channels as well as a longer temporal interval. The qualitative differences between these three classes which leads to different integration strategies required to detect and track signal evidence might account for Gygi's main results.

Local SNR in Sound Recognition

We use the SNR in dB to measure the detection and recognition thresholds in noise. Recognition thresholds have been studied extensively for speech recognition. Allen (1994) provides an overview of Harvey Fletcher's work at Bell labs (1920 – 1950), which shows that the probability of correct recognition of nonsense syllables depends exclusively on the local (in time and frequency) SNR, rather than the energy spectrum of the speech signal. We expect this to be true for environmental sounds too. Therefore, we constructed our stimuli with the maximum local SNR as a measure of the difficulty of the task.

It is important to stress a fundamental problem with the determination of the local SNR for pulse-like signals. A local SNR is defined as the ratio between the instantaneous power per channel (or range of channels) for the target signal and the masking noise. The instantaneous power is a moving average of the instantaneous signal energy (the square of the excitation). For tonal components we can average over a number of periods to arrive at an instantaneous power-value that is independent of the individual oscillations and phase effects. For pulses the notion of instantaneous power cannot be defined in this manner, because there is no sensible integration time-constant apart from an infinitely small one, which corresponds to no integration at all. As a consequence it is impossible to choose a single time-constant that is ideal for all types of sounds.

Because no experimentally validated integration time-constant for pulse-like signals has been estimated, we use a temporal integration strategy that is more suitable for tonal and noise-like signals. This integration strategy is likely to underestimate the actual SNR for pulse-like signals, which leads to lower thresholds. We use the integration method described in Andringa (2008) with a frequency-channel dependent integration time-constant equal to two times the channels best period, which is the inverse of the channel center-frequency, with a minimum of 5 ms for all center frequencies above 500 Hz. The diverse thresholds described in this paper can help to identify the strategies and associated parameters for (pulse-like) environmental sounds.

The next section describes the stimuli and experimental procedure in more detail. The results and discussion sections address the relation between top-down processing and signal content in terms of noise-like, tonal, and pulse-like contributions.

Method

Participants

Twenty-seven students of the University of Groningen, who reported to have no hearing problems, participated in the experiment.

Stimuli

We used a subset of the set used by Gygi et al. (2007) in their study of similarity and categorization of environmental sounds. A total of thirty sounds was selected from the set of 100 sounds, ten from each of the three categories. Each set of ten was chosen so that half of the sounds was close to the center of the class in the MDS analysis and the other half is evenly distributed over the areas towards the borders between the categories. The selection favored the more typical and familiar sounds in the database (conform Ballas 1993). The resulting set is listed in table 1.

Noise-like center	**Tonal center**	**Pulsal center**
Airplane	Baby (s)	(Hand)clap (s)
Car start (s)	Cat	Clock (s)
Train	Rooster (s)	Glass break (s)
Waves (s)	Siren (s)	Ping pong (s)
Wipers	Whistle (s)	Type writer (s)
Noise-like periphery	**Tonal periphery**	**Pulsal periphery**
Bubbles (s)	Bells	Bowling ball
Electric saw	Cows	Cymbal (s)
Horse running (s)	Laugh (s)	Footsteps (s)
Sneeze	Phone	Icedrop
Zipper	Sheep (s)	Keyboard (s)

Table 1: Description of sounds from the three classes, separated by closeness to the cluster center. The indication (s) indicates the second example of Gygi's sounds.

The selected target sounds were mixed with pink noise to create the experimental stimuli. The average loudness of the masking noise was kept constant. Consequently the loudness of the target sound was adjusted to reach the desired maximum local SNR. This method was chosen to prevent the loudness of the noise from becoming a predictor for the target sound or sound type. Additionally, all stimuli were changed to a standard duration of four seconds to prevent signal duration as a predictor of the source. Likewise the point in time at which the target started varied per noise level. All stimuli were created with a different subset of pink noise, to avoid any local patterns in the noise from becoming a predictor for the target sound or sound type.

The maximum local SNR was calculated as follows. For the noise segment as well as the target sound a cochleogram was generated (Andringa 2002, Andringa 2008), as can be seen in figure 2. The cochleogram has a (almost) logarithmic frequency axis and energy in dB. The time-frequency matrix of cochleogram energy levels of the noise was subtracted from the energy matrix of the target signal, to result in a matrix of local SNR values for each time-frequency point. The maximum value in this local SNR matrix is referred to as the maximum local SNR of the signal. Each target sound was amplified to achieve the desired maximum local SNR, subsequently the amplified target and noise signals were added to create the noisy stimuli.

For each of the 30 target sounds, a set of 22 noisy stimuli was created, each with a different maximum local SNR. These 22 maximum local SNR values were chosen per target sound by listening to each noise masked stimulus and guesstimating the detectability and recognizability to choose the scope of stimuli. As a result the target sounds vary in the set of SNR values. In a pilot experiment, the scope of SNR values for each target sound was adjusted where needed. The step in maximum local SNR between the noisy stimuli was not constant throughout the whole range of 22 noisy stimuli. Around the guesstimated detection and recognition thresholds for each of the target sounds the intervals in SNR between the noisy stimuli were 1dB, while in less interesting regions the intervals were 3 or even 5 dB.

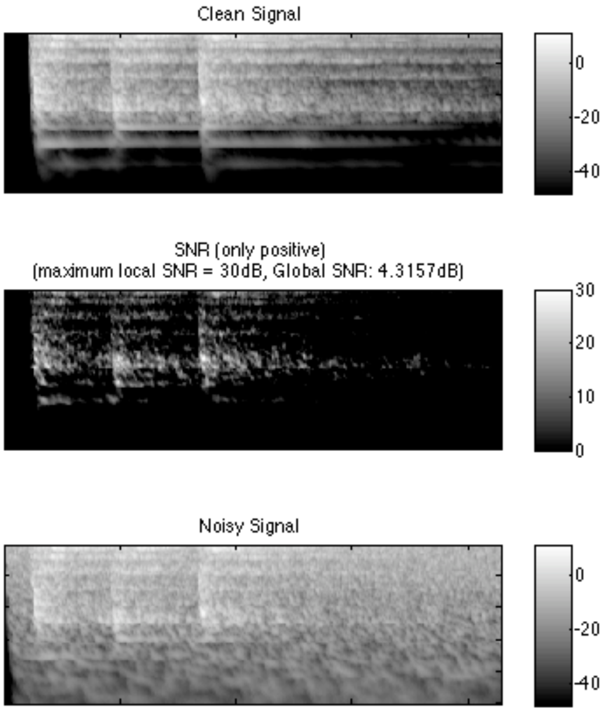

Figure 2 Cochleograms for cymbals. The frequency axis is logarithmic and ranges from 50 Hz to 6 kHz. The time axis spans 1 seconds.

Equipment

The stimuli were presented with closed-back headphones (Sennheiser HD 215) at comfortable listening levels well above the ambient noise level. The experiment was run on a laptop and presented to the participants using a customized Matlab graphical user interface.

Procedure

For each participant 17 sounds were selected randomly from the total set of 30 sounds. Ten were used as targets, six as filler sounds, and one was used as an example before the measurements started. At the start of the experiment, participants were presented with a short description detailing how to interact with the interface, followed by three example items. After these examples, the actual experiment started. The flow of the experiment is illustrated in figure 3. The participants needed on average 50 minutes to complete the experiment.

Items in the experiment consisted of noisy stimulus of a four seconds followed by at least one question. The participants were always asked to indicate what they heard: nothing but noise, something unrecognizable, or something they might have recognized. In case the participant reported to have heard nothing but noise, the experiment proceeded to the next item. When the participant reported to have heard something unrecognizable, the next question was to indicate whether the sound heard in the noise was noise-like, tonal, pulse-like, or a combination. Examples of these were given in the introduction phase. When the participant thought to have recognized the target, he/she was asked to type the source and give it a confidence score ranging form 1 (guess) to 3 (certain). After the questions, the participant could press 'next' to proceed to the following item. All answers were recorded.

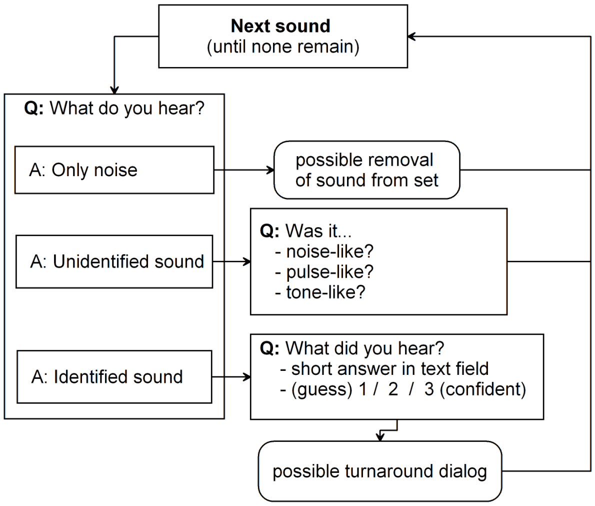

Figure 3 Flow of the experiment

Each next sound was either a target (70% probability) or a filler (30% probability). In case of a target, one of the source types was selected randomly from the list of targets. The resulting random sequence of target and fillers avoids the effect of participants searching for evidence of a specific target. This is expected if all stimuli for a target sound were to be presented consecutively. Filler sounds serve the same purpose. The fillers also ensure a more varied set of noise levels to avoid loss of attention, which may happen at stages of the experiment when the target sounds are initially undetectable. Filler stimuli were chosen randomly from the range of the available noisy filler stimuli, excluding those consisting of mostly noise, and the ones most easily recognizable.

Although the order of the different target sound types was random, the noisy stimuli for each specific target sound were presented in a fixed order. The noisy stimuli were first presented with increasing maximum local SNR. After confident recognition by the participant, or after presentation of the noisy stimulus with the highest maximum local SNR, the stimuli were presented with decreasing maximum local SNR until the target sound was no longer detected. The order in which the noisy stimuli were presented for a single target sound is illustrated in figure 1.

For each target sound the five noisy stimuli with the lowest maximum local SNR were skipped in phase 1 of the experiment to speed up the experiment. When a participant recognized a target sound with confidence rating 3 or had reached the version with the best SNR without a confident recognition, a popup dialog was displayed. This dialog stated the name of the sound source, allowed the participant to listen to a clean version of the target sound (without masking noise), and asked if the participant had recognized the sound correctly. This information was recorded and also verified afterwards. After the presentation of the clean signal one SNR level was skipped, the next noisy stimulus to be played was the one with the second-lower maximum local SNR.

In the second phase of the experiment the time between presentations of each target is in the order of a minute and the memory for the targets must be at least be that long. In this phase a target sound was not immediately removed from the set when it was no longer detected at a particular maximum local SNR level,. Instead, the next time the target sound was chosen for presentation, the noisy stimulus with the same maximum local SNR level was presented to ensure that the participant consistently could not detect the target in the masking noise. If the sound was detected the second presentation, the target sound remained in the set and the next lower maximum local SNR would be presented the next time the target sound was selected. The second time a participant reported not to detect a particular target sound, whether at the same or a lower maximum local SNR, that target sound was removed from the set. The target sound was also removed when the lowest maximum local SNR in the set was reached. The experiment terminated when the last of the target sounds was removed from the set.

Results

The difficulties associated with the computation of a “true” local SNR, as outlined in the theoretical background, can be accounted for by normalizing the experimental results to the first detection threshold. The presentation of relative thresholds is followed by a discussion of the absolute thresholds, and their relation to the local SNR computation.

Relative thresholds

Figure 4 shows the average thresholds of each of the three sound categories according to Gygi (2007) over all subjects. The results are normalized to the threshold of first detection (FD) to discount absolute threshold differences.

The three classes show the same pattern: the threshold of first recognition (FR) is highest; the threshold of first detection (FD), last recognition (LR), and last detection (LD) are similar, with the thresholds of LD a bit smaller than those of LR. This pattern proves the existence of a recognition threshold benefit between FR and LR of about 8.0(mean)±5.0(SD) dB due to previous exposure to the target. The FR threshold is 9.0±4.0 dB above the FD threshold. The thresholds difference between LD and LR is on average 2.6±2.1 dB. The noisy and tonal sounds show a similar pattern of threshold mean and the associated standard deviation.

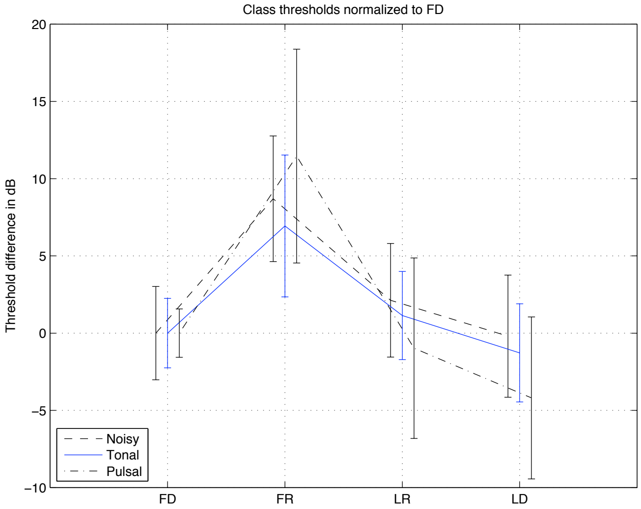

Figure 4 Relative thresholds (means with 1 SD error-bars)

The pulsal sounds deviate from tonal and noisy sounds: while the spread in FD is smallest for pulsal sounds it becomes larger than the other classes for FR, LR, and LD. Furthermore, the pulsal thresholds for FR are a bit higher, while the thresholds for LR and LD are lower. However, the increased spread in FR is an artifact caused by the cymbal sound, which was not recognized with maximal certainty by a number of participants. This led both to a high FR threshold and an increased spread.

The increased spread of LR and LD for pulsal sounds is associated with repetitive sounds like bells, a clock ticking, ice dropping in a glass, and keyboard and typewriter presses. For these sounds the threshold shift is -7.7±0.7 dB for LD and -4±3 dB for LR. The large threshold shift and the small spread in LD of repetitive sounds suggest a high and consistent benefit from previous exposure.

This result is consistent with the hypothesis that previous exposure leads to the activation of specific detection processes that are able to utilize repetitions of (known) target constituents. This entails a combination of an active search and a check for the reoccurrence of recent stimuli. Since the target is repeated in decreasing SNR's, only the unmasked signal constituents refresh the memory of the target. Note that the clock is the only repetitive sound that is perfectly regular, the other sounds are less regular or even irregular. The identical benefit suggests a bias toward known signal constituents instead of a bias toward specific rhythms.

To conclude, both detection and recognition benefit from previous exposure. Recognition benefits in the order of 9 dB SNR, detection of not-repeating targets benefits marginally (less than 2 dB SNR), but consistently, from previous exposure. In case of repetitions the recognition benefit increases to more than 7 dB compared to FD. The additional benefit of repetition in LD is about 5 dB.

Absolute thresholds

The absolute thresholds are meaningful although they should be interpreted with some care due to the difficulties in computing the local SNR. The ranking of the absolute thresholds between the three categories, as estimated with Friedmans test, is significant (p<0.001).

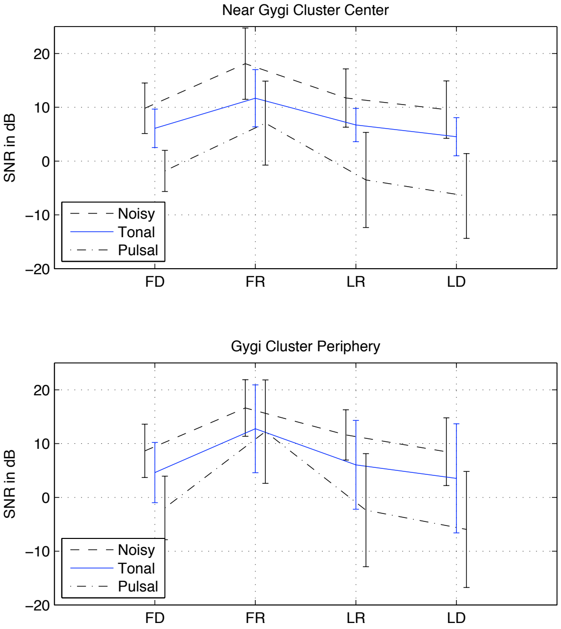

Figure 5 Absolute thresholds (means with 1 SD error-bars), separated according to the division in table 1

Figure 5 shows the absolute thresholds separated according to the division in table 1. The upper panel shows the results for the sounds near the threshold centers and the lower panels shows the results for sounds closer to the category borders. The upper panel shows a clear distinction between the three classes, while the lower panel shows a nearly identical response for all classes. Pulsal sounds have the lowest threshold, which is consistent with the use of temporal integration strategy that underestimates the SNR of pulses compared to tones by about 5±3 dB for FD. The difference between the tonal and the noisy thresholds in the upper panel may have a similar explanation. On the other hand the estimation of a noisy target in noise might also be more difficult than the estimation of qualitatively different target such as tones or pulses in noise.

Compared to the sounds in the cluster periphery, the cluster-center sounds show a larger standard deviation for all thresholds and especially for the tonal category center sounds. In combination with the minimal interclass difference of the class periphery sounds this supports the interpretation that class periphery sounds are a mixture of contributions of tones, pulses, and noises and as such shift to the mean of the detection thresholds of all classes.

Discussion

The results show prominent threshold differences due to previous exposure and a relation between absolute threshold and signal content in terms of pulsal, tonal, and noisy contributions.

The average first recognition threshold of the target sounds lies on average 9 dB above the first detection threshold and 8 dB above the last recognition threshold. This values corresponds to a factor 8 in energy and a factor of almost 3 in distance to the source when assuming a normal inverse square root decay of sound energy with distance. However this estimate presupposes a known but unexpected target sound. With a correct a priori hypothesis, like one based on recent exposure, the source can be recognized wherever it can be detected. The associated evolutionary benefit forms supporting evidence of the usefulness of the two strategy approach hypothesized in the introduction.

The reason why the LD threshold is a bit lower than the FD threshold might be related to a low-level expectation, which leads to a threshold shift: expected sounds need less conclusive evidence than unexpected sounds if the expectation counts as part of the evidence.

The qualitative differences between LD and LR, in the form of smaller threshold shift and much larger spread for LR then for LD, suggests a partial decoupling of detection and recognition; in a number of cases, information derived well above the LD threshold, and therefore useful for detection, could not be coupled to the correct recognition result.

The high benefit of repetition may be a simple statistical process: even when most repetitions are masked by the noise, a few repetitions will be more detectable whenever the fluctuations in the noise mask less of the target. The depth of the noisy fluctuation in vivo are not known, but in the auditory model used to determine the local SNR, the fluctuations have a frequency (i.e. cochlea position) dependent standard deviation of 2 to 3 dB. The 5 dB repetition benefit corresponds therefore to about 2 standard deviations of the noisy fluctuations. Assuming a Gaussian distribution with a mean at the LD threshold without repetition, only 5% of the repetitions will be detectable at a SNR of LD – 2 std. This might nevertheless be sufficient to detect a target with enough repetitions. This can explain the value of the threshold shift. If the proposed mechanism is correct, the average LD value will be a function of the number of repetitions and the standard deviation of noise. This expectation can be checked by further experiments.

The value of the absolute thresholds depends on the temporal integration time. The natural auditory system might rely on multiple time-constants, or on a single (possibly frequency dependent) integration strategy. The current experiment provided clear absolute threshold distinctions for tonal, noisy, and pulsal sources, which provides further evidence for the expectation that the auditory system uses a separate processing/detection route for each type. Although the local SNR estimation might not be perfect, it is much more informative than a global SNR. For example, a global SNR is sensitive to the amount of silence around the target. The local SNR measure used in this paper does not suffer from this effect. Furthermore, the use of a local SNR raises valid the question whether the local SNR must be defined differently for tones, noises, and pulses.

Finally, the consequences for modeling auditory cognition can be profound. In the first place the threshold shifts suggests a double strategy involving both signal-based searching, with thresholds corresponding to first detection, and a checking mode, with a threshold corresponding to last detection. Secondly, the results provide further support for a perceptual separation in tonal, pulsal, and noisy components.

References

Allen, J.B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing* 2(4), 567--577.

Andringa, T.C. (2002). *Continuity Preserving Signal Processing.* PhD dissertation, University of Groningen.

Andringa, T.C., & Violanda, R. (2008). The texture of natural sounds. *In proceedings of acoustics '08*, (pp. 3141-3146). Paris, France.

Ballas, J.A. (1993). Common Factors in the Identification of an Assortment of Brief Everyday Sounds. *Journal of Experimental Psychology: Human Perception and Performance, 19(2)*, 250-267.

Chiu, C.Y.P., Schacter, D.L. (1995) .Auditory Priming for Nonverbal Information: Implicit and Explicit Memory for Environmental Sounds. *Consciousness and Cognition,* vol. 4 pp. 440-458

Gaver, W.W. (1993). What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. *Ecological Psychology, 5(1)*, 1-29.

Guastavino, C., (2007). Categorization of environmental sounds. Canadian journal of experimental psychology vol. 61 (1) pp. 54-65

Gygi, B., Kidd, G.R., & Watson, C.S. (2007). Similarity and categorization of environmental sounds. *Perception and Psychophysics, 69(6)* 839-855.

Hawkins, J.E., & Stevens, S.S. (1950). The masking of pure tones and of speech by white noise. *Journal of the Acoustical Society of America*, 22, 6-13

Nábělek, A., & Robinson, P. (1982). Monaural and binaural speech perception in reverberation for listeners of various ages. *Journal of the Acoustical Society of America*, 71, 1242-1248.

Shinn-Cunningham (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences* vol. 12 (5) pp. 182-186