INTRODUCTION

The Mind sees and the Mind hears; the rest is blind and deaf

(Epicharmus 45O B.C.E.)

A pressure wave pounding on the eardrum is translated into a percept of a car approaching from behind allowing for a timely avoidance. This transformation is arguably one of the primary roles of sensory processing: from the partial and ambiguous information provided by the sensory organs into a coherent, ethologically – relevant description of the sound sources in the environment. A major challenge to this process is the fact that in real life multiple sound sources are often concurrently active. The waveforms they emit reach our ears all mixed together with information possibly irrevocably lost. The task of the brain is to reverse this effect and decompose the sound mixture back into its components. Formulated in this way, however, the problem is mathematically ill-posed – without further constraints, there is no unique solution to the problem of recovering individual sound sources from the composite sound reaching the ear. Nevertheless, we have the introspection that we hear the different sources separately. Moreover, the processes by which the brain performs the separation to sound sources are mostly automatic and do not require explicit effort.

Auditory Scene Analysis (ASA), coined by Bregman (1990), is the term currently used for various schemes that bridge the gap between the physical sound mixture and the introspective percept of hearing separate sources. The question is how a complex spectro-temporal pattern*, the auditory scene*, is segregated and grouped into separated perceptual units, *objects* – entities with a clear relation to things or events in the

environment ('streams' in Bregman's terminology). Bregman and others used carefully controlled behavioral setups, and translated their experimental observations into a list of gestalt-like grouping rules for sounds: proximity, similarity, good continuity and common fate (reviewed in Bregman, 1990, Cooke and Brown, 1993, Darwin, 1997, Moore and Gockel, 2002, Carlyon, 2004). Using similar sound sets, extensive neurophysiological research has been dedicated to identifying brain processing schemes that might underlie behavioral results regarding sound segregation (for reviews, see Carlyon, 2004, Snyder and Alain, 2007, Ciocca, 2008).

Bregman's work has also led to the development of the field of Computational Auditory Scene Analysis (CASA), the attempt to further translate the grouping rules suggested in ASA into computer-based procedures that parse a composite auditory signal into separate sources. Most of these systems include an initial sound decomposition stage inspired by mammalian mechanisms of peripheral hearing, leading to a time-frequency representation of sounds. Then, components of the representation are grouped to form objects extracted from the sound. In the final stage the extracted objects are usually re-synthesized, and then the reconstructed signal is compared to the original (for example, Vercoe and Cumming, 1988, Duda et al., 1990, Mellinger, 1991, Brown, 1992, Ellis, 1996, Smaragdis, 2001). However, both the grouping rules and their implementations remain a collection of heuristics for auditory object segregation in specific circumstances that may fail to generalize. The main problem seems to arise when translating simple, isolated principles inferred from highly constrained psychoacoustic stimuli such as sine-tones and gated white noise to the messier domain of real sounds and sound mixtures. To date, the more principled ways of extracting auditory objects in general enough settings are elaborate computational schemes that are biologically-inspired. For instance, Elhilali and Shamma (2008) combined design principles of the peripheral, the sub-cortical and the cortical auditory system to a procedural cascade that has been successfully applied to segregate both mixtures of speech by different speakers and two-tone sequences.

Not only is there a lack of general procedures, the answer to what constitutes an auditory object is in itself controversial. One option would be to identify auditory objects with material objects in the world, or with events that are related to material objects in the world (e.g. Pasnau, 1999, Matthen, 2010, Nudds, 2010). But, the auditory system does not have direct access to the information regarding the physical sources, only to the energy flux that reaches the ears. Therefore such definition cannot deliver knowledge about the processes that guide ASA. There have been a number of formal attempts to define auditory objects in precise terms. In an influential paper, Griffiths and Warren (2004) laid out four principles for an auditory object: it should correspond to a thing in the sensory world, its analysis requires separating the information related to the object and that related to the rest of the world, it should generalize across different physical realizations of the same auditory object, and it should generalize across sensory modalities. The current usage of the term in research is definitely less restrictive: it is usually partially consistent with the first two principles, but does not require the last two. Possibly, the contemporary working definitions of auditory object represent an earlier stage of transformation of the sensory data relative to the objects as defined by Griffiths and Warren. The practical outcome is that often the auditory object is defined heuristically, as something that fulfills the requirements of being a separate object according to a certain perceptual grouping, or alternatively as the collection of all the sound elements actually emitted from a material object in the world.

Predictive framework for ASA

In the now popular prediction-based formulation of sensory perception (for example, see review at Clark, 2012), perception proceeds via the indirect reconciliation of an internal representation with a perceived sensory signal from the external world. The notion of the sensory systems is as predictive mechanisms that actively generate predictions of sensory inputs applying an internal or generative model. Theories typically apply concepts from Bayesian inference and postulate that the 'purpose' of perception is to generate testable hypotheses about the causal structure of the external

world, based both on prior knowledge and current sensory input (Gregory, 1980). The focus of sensory perception is accordingly shifted from efficiently representing the current state of the outside world (e.g. Attneave, 1954, Barlow, 1961, Atick, 1992) to extracting from the perceived input the predictive information that enables to extrapolate the current knowledge to future sensory input (Bialek et al., 2001, Friston, 2005, Clark, 2012).

Following Winkler et al. (2009), I specifically propose the predictive framework as a general scheme for auditory scene analysis. My suggestion is to identify auditory objects with predictive models, constructed on the basis of regularities extracted from sounds. The basic idea is that natural sounds have underlying regular patterns that govern their evolution in time. By detecting these regularities in the sensory input the brain can construct predictive models and generate predictions of future auditory inputs based on recent experience. The validation of a perceptual organization of a sound into objects is thus performed directly - rather than testing models against the current environment, models can be tested with incoming sensory input. As a result, instead of using perceptual criteria, under the predictive framework an auditory object is defined directly by the physical properties of sounds. In the predictive formulation, the processes of auditory scene analysis are inherently dynamic - the auditory system continuously searches for regularities that govern the evolution of the acoustic signal in time, on multiple levels and timescales. A representation of detected regularities serves as a possible interpretation of the acoustic input, and acts as a hypothesis regarding the appropriate description of the auditory scene. By extrapolating a detected regularity into the future, specific predictions of future sounds are generated and compared to incoming observations. When discrepancies are identified between a prediction and observations, they can trigger an update of the interpretation, possibly inducing the percept of a new auditory object in the scene. Thus, in our formulation, the prediction errors play a key role in the process of auditory scene analysis. By dynamically monitoring the agreement between the auditory inputs and the internal models, the

prediction error outlines auditory objects and can guide the construction of more fitting auditory objects.

From this perspective, the basis of segmenting complex acoustic mixtures into objects is the identification of regularities in the sensory input. Whereas the general principle may be applied in a different way to other modalities, for the auditory modality the temporal domain is most relevant - sounds unfold in time, and regularities concerning the evolution of sounds in time are potentially a prominent source of information. A significant computational toolbox is available for detecting and modeling regularities in time, in great part due to advances in compression techniques for the transmission and storage of acoustic signals. In the last ten years, there has been enormous progress in applying general principles such as dimensionality reduction (Ellis, 1996) and redundancy reduction (Smaragdis, 2001) to the realm of acoustic signals. Here, I apply the Information Bottleneck (IB) method introduced by Tishby et al. (2000) as a general principle for extracting regularities in different sound domains. Formulated in information theoretic terms, the IB method provides a principled way to extract relevant information from one random variable with respect to another random variable. In the studies reported here, I applied the IB method for predictive modeling of physical sounds using the statistics of the waveforms alone. I was thus able to test the conceptual approach with various sounds by applying very minimal general assumptions that are applicable in many real-world acoustic scenarios. Moreover, by employing the predictive framework to sets of sounds actually used as stimuli in auditory experiments, I was able to assess the products of the implementation against experimental results ranging from behavior to single unit responses from the auditory cortex. The results surveyed in this work illustrate the explanatory power of prediction errors that measure the deviation of the acoustic signal from an inferred statistical rule as well as the representation of prediction error in the electrical activity of human and animal brains.

The ascending auditory system

The predictive framework to Auditory Scene Analysis may help answering a longstanding unresolved question in the field of auditory processing: what are the features of sounds that neurons in Primary Auditory Cortex (A1) encode?

Auditory processing in the brain is most commonly conceptualized mainly as a 'bottom-up' process - a hierarchy of transformations on sensory data that starts with the concrete and advances to the more abstract. Multiple functional pathways that run from the periphery to high cortical areas first encode low-level physical cues (e.g. spectral content), then more abstract qualities (e.g. pitch). Anatomical structures along the pathway are often assigned a specific transformation of the input from earlier stations. Thus, the ascending pathway is associated with the emergence of new sensitivities at the single neuron level that are absent or less pronounced in earlier stations.

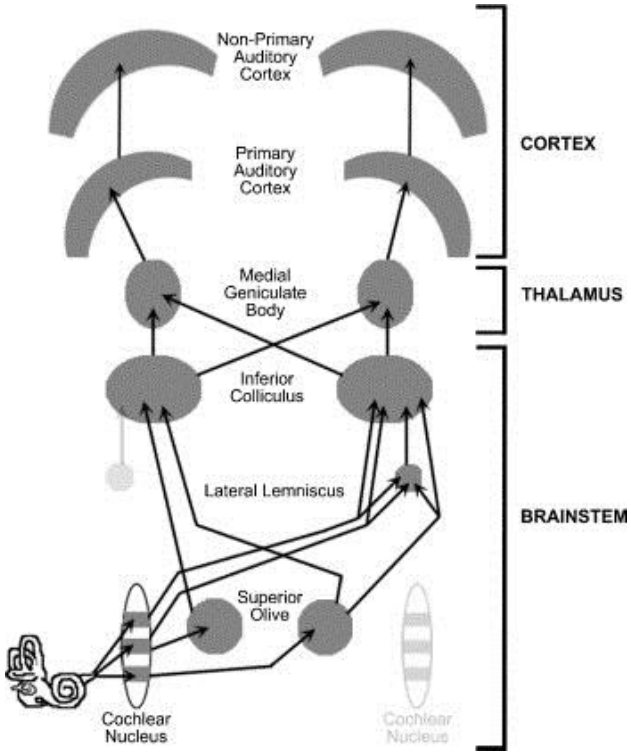

Figure 1. A schematic representation of the mammalian ascending auditory pathway. The transformations of sensory data that take place in the periphery and the brainstem structures (up to and including the Inferior Colliculus) are well understood. The nature of the transformations that occur in the Auditory Thalamus and Primary Auditory Cortex is still disputed.

This ‘box and arrow’ approach was successfully applied to modelling the lower auditory pathway, from the cochlea to the Inferior Colliculus in the midbrain (see Fig. 1). In the cochlea the sound wave is decomposed into its spectral components, creating an orderly representation of frequency that is largely preserved from the auditory nerve fibers, up to and including the auditory cortex. In the brainstem stations that follow, response patterns can often be accounted for by short term spectro-temporal features. These include ‘primary like’, ‘chopper’ and ‘onset’ response types in the ventral cochlear nucleus, ‘type II’ and ‘type IV’ response types in the dorsal cochlear nucleus, and sensitivity to interaural time and level differences in the middle and lateral superior olive. The relatively good understanding of these stations is reflected in the existence of computational models that can reproduce much of the diversity of brainstem neuronal responses to sounds. The last station in the brainstem - the Inferior Colliculus (IC) – is an obligatory station of the auditory pathway. All the separate brainstem pathways (from the Cochlear Nucleus, the Superior Olive and the Laminar Lemniscus converge at the IC (see Fig. 1). In addition, each IC receives input from its contralateral counterpart, descending inputs from the auditory cortex and even non-auditory (e.g. somatosensory) inputs. From the IC the auditory information is projected to the auditory thalamus and then to the cortex. The variety of neuronal responses in the Inferior Colliculus is very high. IC neurons integrate information from essentially all lower processing streams while representing the various sound features calculated previously with a resolution equivalent to that of the earlier stations.

At the other end, neuronal responses in the stations beyond primary auditory cortex are less closely related to the physical structure of sounds, but seem to be linked to high-order features that are more abstract and generalize across many spectro-temporal realizations. In the cat for example, the posterior auditory field (PAF) has neurons with frequency-invariant best levels, possibly encoding level in a frequency-independent way (Phillips et al., 1994) and sound source location, in particular frontal locations, is represented in the secondary auditory field at the sock region of the anterior ectosylvian sulcus (FAES or AES, Las et al., 2008). In primates, neurons in the lateral

intraparietal area (LIP) show selectivity to sound locations, while also encoding stimulus identity (Gifford III and Cohen, 2005). A possible 'pitch area' was identified in the marmoset monkey near the anterolateral border of the primary auditory cortex (Bendor and Wang, 2005). In some higher auditory area, such as the auditory region in the ventral-lateral prefrontal cortex (VLPFC), neurons seem to be less sensitive to the physical structure of natural sounds than to high-level features such as stimulus class (Averbeck and Romanski, 2006), perhaps reflecting the encoding of stimulus category in these regions.

Human fMRI data also seem to be roughly consistent with a hierarchical processing model. Patterson et al. (2002) located a pitch area beyond A1 in the lateral half of the Heschl gyrus, and identified activation produced by melodies higher up, outside the core area. Deouell et al. (2007) reported an explicit, non-attentive representation of sound location in the Planum Temporale, a non-primary auditory region. Similarly, the 'where and what pathways' model (Romanski et al., 1999, Rauschecker and Tian, 2000, Zatorre et al., 2002), suggests that above A1, a posterior pathway extracts sounds spatial properties ('where') whereas an anterior pathway extracts other high features of the sound source ('what'). In accordance with this description, the representation of speech becomes more abstract in processing stations above A1 (Scott and Johnsrude, 2003). All these features represented in cortical areas beyond A1 are complex non-linear transformations of the physical properties of the sound, generally computed over longer timescales and spectral ranges.

Object representation in A1

Mid-level auditory processing consists of the auditory thalamus (MGB –medial geniculate body) and the primary auditory cortex (A1), the focus of this work. Albeit extensive research of neuronal responses at these structures, findings do not seem to fit in the hierarchical progression from concrete to abstract feature encoding. Strikingly, encoding in the cortex of each of the separate short-term spectro-temporal properties usually becomes worse, not better, than in earlier stations (Schnupp et al., 2011b). By

contrast, there are very few, if any, high-level features of sounds whose representation has been shown to emerge in primary auditory cortex.

One possibility is that instead of looking for a single feature that is represented by neurons in A1, we should look for certain special combinations of such simple features. Wang et al. (2005) reported especially strong and sustained responses in single neurons from the marmoset A1 that were evoked by 'best stimuli' –a subset of sounds that were tuned specifically to each neuron over several auditory dimensions. The preferred stimuli used in the experiments were optimized in parallel along multiple auditory dimensions including carrier frequency, sound level, temporal modulation frequency and spectral bandwidth. It is probable that behaviorally relevant natural sounds are also non-trivially defined by a conjunction of many such properties. In this view, the reason why neuronal responses in A1 appear to be non-selective in most experiments is related to the use of non-optimal artificial stimuli.

The notion that the relevant auditory dimensions for describing cortical responses are non-trivially related to the physical dimensions of the sound is further strengthened by the results of Nelken and colleagues (Bar-Yosef et al., 2002, Las et al., 2005, Chechik et al., 2006, Nelken and Bar-Yosef, 2008). Responses from the cat A1 were recorded when complex natural sounds were presented singly and in mixtures. Single neurons in A1 tended to respond to mixtures in the same way they responded to one of the individual components of the mixture. In many cases, the response was dominated by one of the components, very often by a low-level component rather than by the acoustically dominant one. Notably, the same neurons responded to the acoustically-dominant component when presented alone. These results suggest that the auditory cortex responds to specific entities that are non-trivially categorized in an auditory multidimensional space (e.g. by frequency, amplitude and frequency modulation patterns, presentation rates). Furthermore, these entities potentially determine the response in A1 even when other, acoustically dominant, components are simultaneously present. In subsequent work, Chechik and Nelken (2012) calculated the amount of

information in cortical responses concerning two aspects of this stimulus set: the spectro-temporal patterns of the sounds, and the abstract entities present in the stimuli (such as a bird chirp, echoes, and ambient noise). A1 neurons conveyed on average three times more information about the abstract auditory entities than about short-term spectro-temporal features of the acoustics.

Thus, the electrophysiological data suggest that responses of neurons in A1 do not relate in a straightforward way to proposed auditory dimensions, including those represented by neurons in the stations below A1. Rather, neural responses seem to be better understood in terms of auditory objects. This is also consistent with the identification of neural correlates of sequential auditory object segregation ('streaming', see Bregman, 1990) of behavior in the primary auditory cortex (A1) of mammals (Fishman et al., 2001, Kanwal et al., 2003, Fishman et al., 2004, Micheyl et al., 2005) and in the avian homolog, the field L (Bee and Klump, 2004, 2005, Itatani and Klump, 2009).

In this work, my hypothesis is that the primary auditory cortex plays a central role in solving the complex problem of auditory scene analysis. In particular, I expect that responses of A1 neurons will reflect the segregation of the auditory scene into perceptual components. Moreover, I suggest that the predictive framework of ASA is the appropriate tool to investigate role of cortical neurons in the perceptual organization of sounds. The framework directly links between auditory objects and regularity extraction, and thereby relates to another line of research that describes the sensitivity of A1 neurons to statistical regularities in the recent history of the sound. One well documented effect of sound history on neuronal response is adaptation - the decrease in neuronal responses due to continuous or repetitive stimulation. As early as Condon and Weinberger (1991) this decrease was shown to be specific; in their experiments, a repetitive presentation of a tone for a period of 7 minutes resulted in a decrease of up to 70% in the responses of cortical neurons to the same tone whereas responses to nearby tones were not affected. The introduction of the oddball paradigm into single-neuron studies by Ulanovsky et al. (2003) led to ample evidence of the

intricate dependence of the neural response on the recent past of the sound. In a typical oddball experiment, many repetitions of one pure tone frequency are interspersed with rare presentations of another, nearby frequency. The characteristic result is that response size is inversely related to the probability of tone presentation. Commonly presented tones show adaptation, with time constants on the order of 10–100s, but rare tones that could be very close in frequency (10% or even 4% away from the common tone) show less or even no adaptation. This effect, termed "stimulus-specific adaptation" (SSA, Ulanovsky et al., 2003) has been extensively studied in various paradigms to reveal a sensitivity to sound probability in the auditory cortex of rats (Taaseh et al., 2011) cats (Ulanovsky et al., 2003, Ulanovsky et al., 2004, Zhang et al., 2005) and primates (Malone et al., 2002, Fishman and Steinschneider, 2012). A recent study by Yaron et al. (2012) has shown that A1 responses are sensitive not only to the overall rarity of the tone but to the detailed structure of sound sequences over time scales of minutes.

Data thus suggest that in order to uncover the functional role of A1 the experimental design should include non-isolated, complex sounds that have a statistical structure and resemble real world auditory scenes. A1 responses to such complex sounds are related in studies both to the representation of auditory objects and to the probability of a sound as derived from the recent acoustic history. The two research directions intersect in the predictive framework for auditory scene analysis.

In this work, I applied a data-driven approach and explicitly tested the relevance of the predictive formulation of ASA with the following neural data:

1. Extracellular single unit responses from human auditory cortex to the presentation of complex sounds including music and speech as well as random sequences (Chapter 3);
2. MEG extracranial responses to unattended rudimentary auditory scenes that include noise and pure tones (Chapter 4);

3. Local field potentials (LFP) and multi-unit activity (MUA) from rat auditory cortex recorded during the presentation of a musical piece with relatively simple statistics (Chapter 6);

When my analysis of the neural data also yielded results not strictly confined to the predictive framework, these are described in the respective chapters.

In addition, in Chapter 5 I implemented the predictive framework to an experimental sound set used in studies of timbre recognition. I tested the products of the implementation against the behavioral results. This work constitutes a concrete application of the theoretical formulation to actual experimental data.

The results presented in this work suggest that the auditory cortex performs non-linear, context dependent processing of sounds that relies extensively on the statistics of the auditory signal and can be understood through the concept of predictive auditory objects. The results point to the necessity of the concept of auditory object for a better understanding of auditory processing of sound in the cortex, and highlight a key role for prediction errors in this scheme. The predictive formulation is shown to be applicable to concrete experimental scenarios and fruitful in the interpretation of both neuronal responses and human behavior.

GENERAL METHODS

Predictive modeling of an acoustics signal- The Gaussian IB

'Predictive' in the present framework, unless otherwise stated, refers to detecting dependencies within a series of sounds and extrapolating these dependencies to future incoming stimuli. To generate predictive models from the statistics of a sound pressure wave, I used here the Information Bottleneck (IB) method developed in Tishby et al. (2000). The IB method provides a principled way to extract relevant information from one random variable with respect to another random variable. The core of the method is formulated as an optimization problem using information terms: given two random variables *X* and *Y* with a known joint distribution, find *T*, a reduced representation of *X* that conserves maximal information on some target variable *Y*. The representation *T* is given by the conditional probability $p(t|x)$ obtained by minimizing the IB-Lagrangian:

$$\mathcal{L}[p(t|x)] = I(X;T) - \beta I(T;Y),$$

where $I(X;T)$, $I(T;Y)$ are the mutual information terms and β a positive Lagrange multiplier (Tishby et al., 2000)

In my application of the IB framework (Chapters 5, 6), *X* and *Y* are past and future samples of the acoustic signal. The signal is assumed to be stationary. In consequence, the joint probability, $p(past, future)$, is well-defined. The optimal representation *T* contains the information from past samples that is relevant for predicting future samples. T effectively forms a 'bottleneck' between past and future due to the dual requirement it satisfies - maximal mutual information with the future and minimal mutual information with the past (Fig. 1).

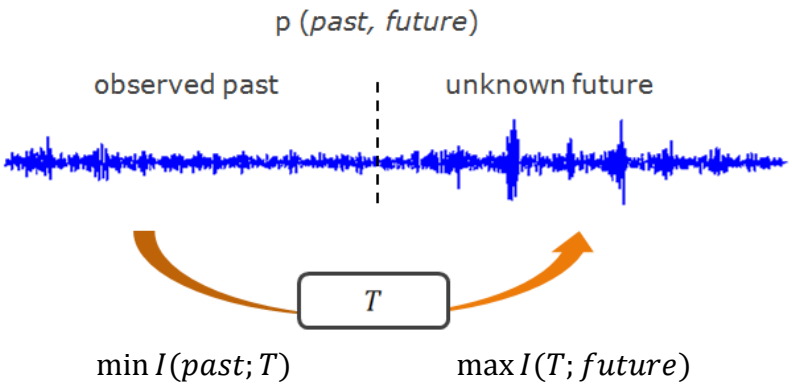

Figure 1. The predictive Information Bottleneck. The elements in the predictive IB formulation: given a sound wave with a stationary join $\min I(past;T)$ probability, $p(past, future)$, the goal is to find the most reduced representation of the signal's past samples, *T*, that best predicts the signal's future samples.

The information bottleneck approach has been mainly applied to categorical variables, with a discrete *T* that represents (soft) clustering of *X* (e.g. Slonim and Tishby, 2000, Dimitrov and Miller, 2001, Friedman et al., 2001). Solving the IB equations for continuous *T* is generally difficult. However, Chechik et al. (2005) introduced a general analytical solution for the case that *X* and *Y* are jointly multivariate Gaussian variables. In this case the optimal representation *T* is obtained by a linear transformation of X, $T = AX$. The rows of the projection matrix *A* are given by weighted left eigenvectors of the normalized regression matrix with non-zero eigenvalues. The normalized regression matrix is obtained by normalizing $\Sigma x|y$, the covariance of sound samples conditioned on future samples by Σxx, the unconditional covariance of sound samples:

$$\Sigma x|y\Sigma xx^{-1} = \mathrm{I} - \Sigma xy(\Sigma yy^{-1})\Sigma yx(\Sigma xx^{-1})$$

Where $\Sigma xx, \Sigma yy, \Sigma xy, \Sigma yx$ denote unconditional covariance matrices. Thus, under a Gaussian assumption, the only sound statistics required to generate optimal predictive models for a sound is its autocovariance matrix. All the unconditional covariance matrices needed for the eigenvector calculation are extracted directly from this autocovariance matrix. Notably, the set of eigenvectors that solves the Gaussian Information Bottleneck is the same basis used in canonical correlation analysis (CCA)

introduced already in 1935 (Hotelling, 1935) to find linear combinations of two random variables which produce maximal correlations between them. In CCA the eigenvectors form the orthogonal basis in which the correlation matrix between the pairs of canonical variables (the linearly transformed *X* and *Y*) is diagonal, and the correlations on the diagonal are maximized. There are differences in the weights used for the different eigenvectors in the two methods (for details, see Chechik et al., 2005), but the link to CCA offers an additional intuition regarding the nature of the optimal representation *T*.

ROC analysis

Receiver operating characteristic (ROC) analysis was used to verify that differences between observed distributions could support classification of the type performed in psychoacoustic experiments. In signal detection theory, an ROC analysis is used to quantify the ability to discriminate between two hypotheses supported by noisy data. The classification of a certain trial to either hypothesis is based only on the value of a noisy measurement obtained in the trial. One way of describing a classification rule is by specifying which trials it assigns to one of the classes (the trials assigned to the other class are the complementary set). Often, the classes are treated in a non-equivalent way: assignment to one class can either be a *correct detection* or an error (*false alarm*), while assignment to the other class is considered as a *correct rejection* or as a *miss*. Note that the probability of a correct detection and the probability of a miss sum to 1, and similarly the probability for a false alarm and the probability for a correct rejection sum to 1. We can thus describe each classification rule as a point on the *ROC plane*, where the ordinate is the probability of correct detection under this rule and the abscissa is the probability of a false alarm under this rule. The lemma of Neyman and Pearson (Neyman and Pearson, 1992) shows that the optimal decision rule, ensuring maximal probability of correct detection under a bound on the rate of false alarms, is based on the *likelihood*- the ratio of the probabilities to observe the measurement given the trial was from one or the other class. A criterion is selected so that trials whose likelihood is

larger than the criterion are assigned to one category and trials whose likelihood is smaller than the criterion are assigned to the other. By varying the criterion on the likelihood we obtain a set of points that lie on the *ROC curve*, a monotonic convex curve that depicts the optimal tradeoff between false alarms and correct detections for the observed distributions. The area under the ROC curve (AUC) then corresponds to the probability of correct classification expected from an ideal observer based on these distributions in a 2-alternative, 2-interval forced choice psychophysical task (Green and Swets, 1966, Fawcett, 2006). The AUC was estimated in this work by the trapezoidal area under the empirical ROC curve. This measure tends to be slightly biased upward when estimated from data (Gordon et al., 2008).

In many applications of ROC analysis, the criterion is set on the observed value of the measure rather than on the likelihood; hence all values that are larger than the threshold are assigned to one class and all values smaller than the threshold are assigned to the other class. Importantly, if the relation between likelihood and the value in not monotonic, such decision rules are not necessarily optimal. The resulting curve in the ROC plane may be concave and the area under the curve evaluating performance will be necessarily smaller than would have resulted from a criterion on the likelihood. In this work I used criteria based on the likelihood unless explicitly stated otherwise. Their performance was indeed better than criteria on the value, although the differences were not substantial.

Data Analysis

Statistical tests were considered significant at the 0.05 level unless explicitly stated otherwise. Stricter significance levels were used when appropriate to correct for multiple comparisons. Variability is reported as the mean ± standard deviation, unless explicitly stated otherwise. Correlation was tested for significance using a t-test, with p-value computed by transforming the correlation to create a t-statistic having n-2 degrees of freedom, where n is the number of data points. More details regarding specific data analysis methods are presented in the results sections.

THE LIMITATIONS OF LINEAR ANALYSIS

Evidence from unit recording in human auditory cortex

* R. Malach, R. Mukamel and I. Fried (for affiliations see the publication below) conducted all recording sessions.

*** This chapter is an extended version of:*

Bitterman Y, Mukamel R, Malach R, Fried I, Nelken I., **Ultra-fine frequency tuning revealed in single neurons of human auditory cortex.** Nature. 2008 Jan 10; 451(7175):197-201.

Spectral sensitivity in humans

The basic operation performed in the auditory periphery is the decomposition of sounds to different frequency bands. The auditory information reaching the central nervous system of mammals is organized tonotopically ('by frequency'), and the tonotopic organization is kept throughout the ascending auditory pathway, at least up to and including primary auditory cortex (Howard III et al., 1996, Schnupp et al., 2011). In vision and somatosensation, the resolution of the peripheral sensors to a large degree determines overall behavioral discrimination capabilities. For example, two-point discrimination in vision is limited by the size of the receptive fields of individual photoreceptors, which is eventually determined by photoreceptor size. However, in the auditory system, just-noticeable differences (JNDs) in frequency in well-trained subjects may be ~30 times smaller than the presumed bandwidth of the peripheral filters ('critical bands', typically about 1/6 octave in humans, as measured in psychoacoustical tests. Moore, 1982). Electrophysiological correlates of critical bands have been suggested (Evans, 1977, Ehret and Merzenich, 1988, Ehret and Schreiner, 1997), and

frequency JNDs can be derived from models of the auditory periphery by integrating information over a large population of peripheral neurons (Heinz et al., 2001). However, there are currently no reports of a significant population of single neurons whose bandwidth corresponds to the behavioral JNDs in frequency. There is no a-priori necessity for an organism to have such neurons – the high sensory selectivity could be represented implicitly, for instance by the association of different motor outputs with the somewhat different population response patterns, evoked by discriminable tone frequencies. Does the high frequency resolution expressed behaviorally have explicit neural representations? And if so, can the same high frequency resolution explain the response patterns to complex sounds?

Experimental setup and data analysis

Patients, electrode implantation, and electrophysiological recordings

The methodology of single-unit recordings in humans relevant to this work is described in details Mukamel et al. (2005). I will specify here the relevant information for our study. The extracellular single unit recordings were obtained from four patients with pharmacologically intractable epilepsy, implanted with intracranial electrodes to identify seizure focus for potential surgical treatment (Fried et al., 1999). Electrode location was based solely on clinical criteria. All patients had electrodes placed bilaterally in Heschl's gyri, loci of the auditory cortex (Talairach coordinates Patient 1: [Right 45.5, Posterior 19.3, Superior 10.2], [Left 43.2, Posterior 12.6, Superior 10.3]; Patient 2: [Right 38.8, Posterior 12.4, Superior 12.3], [Left 41.2, Posterior 18.2, Superior 10.4]; Patient 3: channels 1:8 were from Right middle superior temporal [42.47 right, 1.01 anterior, 10.54 superior], channels 33:40 were from right posterior superior temporal [39.18 right, 23.14 posterior, 16.68 superior], channels 57:64 were from Left superior temporal [48.53 left, 7.03 posterior, 14.01 superior]); Patient 4: [Right 40.95, Posterior 0.83, Inferior 0.35]. Each electrode terminated in a set of nine 40 µm platinum-iridium microwires. Signals from these microwires were recorded at a 14 kHz

for Patient 1 and 28 kHz for Patients 2, 3 and 4. The raw signal was band-pass filtered between 1Hz and 9 kHz and recorded using a 64-channel data acquisition system. To verify electrode position, CT scans following electrode implantation were co-registered to the preoperative MRI using Vitrea@ (Vital Images Inc.). Patients provided written informed consent to participate in the experiment. The study conformed to the guidelines of the Medical Institutional Review Board at UCLA.

Experimental protocol

In each experimental session, patients 1-3 were presented twice with the same audiovisual movie clip, consisting of 8:40 minutes of an un-edited audio-visual segment of the famous western "The Good, The Bad, and The Ugly" (starting from minute 38:25 in the original film). Both clip presentations were shown in succession, with a 5- to 10-minute rest period in between. The patients' task was to follow the plot. Patients 2, 3 were also presented with a sequence of low resolution random chords lasting 3.5 minutes accompanied by random visual textures displayed at a rate of 4Hz. The random chord stimuli used here were similar to those used in previous studies (deCharms et al., 1998, Schnupp et al., 2001). Each chord had three frequencies. Tone duration was 100 ms with 10 ms linear onset and offset ramps. Tones were randomly chosen from a table of 41 frequencies equally spaced along a logarithmic axis from 100 Hz to 10 KHz, so that neighboring frequencies were 1/6 octave apart. Patient 4 was presented with a sequence of higher-resolution random chord stimulus lasting 5 minutes – tone duration for this stimulus was 50 ms with 10 ms onset and offset ramps and the frequency table included 108 frequencies equally spaced along the same logarithmic axis from 100 Hz to 10 KHz, so that neighboring tones were 1/18 octave apart. The artificial stimuli were designed to evenly sample the spectral range of the movie soundtrack. A typical section from the higher-resolution random chord stimulus is depicted in Fig. 1 along with a short excerpt from the movie soundtrack.

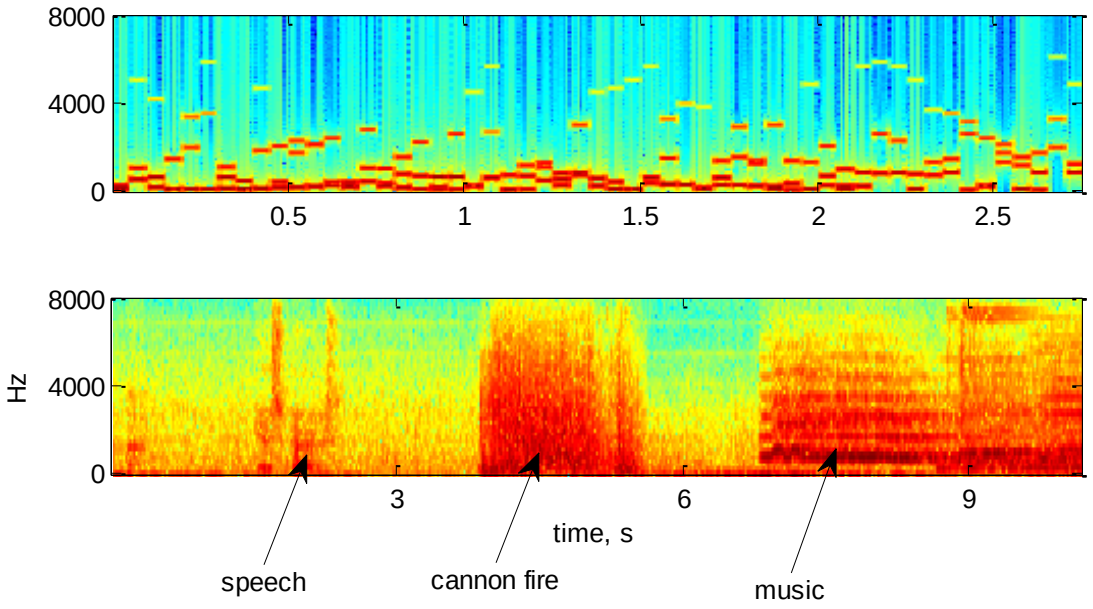

Figure 1. The stimuli. Spectral content of an excerpt from the high resolution random chord stimulus (**top**) and from the soundtrack (**bottom**). The 10 seconds from the soundtrack depicted here include short bits of speech, background noise and music (see **arrows**).

The stimulus ensemble thus included both artificial stimuli and more structured stimuli, including speech and music. Two recording sessions included both kinds of stimuli. Data were acquired in ten sessions (two sessions with Patient 1, three sessions with Patient 2, four sessions with patient 3, and one session with patient 4). All sessions were conducted at the patients' quiet bedside using standard laptop screen and speakers. Sound intensity was set to a comfortable hearing level.

Spike detection and cell selection

For Patient 1, the signal from each microwire was filtered above 300Hz and at multiples of 60Hz. Potential spikes were detected by thresholding the filtered signal. Potential spikes were then sorted based on template matching the first two principal components. For Patients 2, 3 and 4, the raw data was band-pass filtered between 300 and 3000Hz and sorted as in Quiroga et al (2005). Analysis is based on a total of 95 units recorded from the four patients (20 units from patient 1, 21 from patient 2, 38 from

patient 3 and 16 from patient 4). In total, 48 units were tested only with the soundtrack, 14 units with both the soundtrack and the random chords (1/6 octave resolution), and 33 units with only the random chords (17 with the 1/6 octave resolution and 16 with the higher, 1/18 octave resolution).

ROC Analysis

For each unit the empirical spike count distributions elicited by different frequencies were calculated in a relevant response window, starting at chord onset and ending between 50 to 200 ms after chord onset. The counting window was selected so that it matched the period of elevated firing rate in the PSTH of a unit's response to its best frequency. The Receiver Operating Characteristic (ROC) curve was computed for discriminating between the best frequency and each of the other tested frequencies (see Chapter 2). The spike count elicited in a single trial was assigned to the best frequency if the likelihood of that count (the ratio of the probabilities to observe the spike count given the best frequency and the comparison frequency, see Chapter 2) was above the classification threshold. For each threshold, the correct decision was the fraction of best frequency trials correctly assigned to the best frequency, and the false alarm rate was the fraction of trials of the other frequency incorrectly assigned to the best frequency. The area under the ROC is an estimate of the probability of correct discrimination between the two frequencies in a two-alternative, two-interval forced choice paradigm. Discrimination threshold was set at 70.7%, corresponding to the threshold tracked by the standard 2-down 1-up adaptive procedure typically used in auditory psychophysics (Banai and Ahissar, 2004). When the resolution of the tones actually tested in the experiment was too low, we estimated the distribution of responses to "intermediate" untested frequencies from the responses of the unit to frequencies that were presented. Thus, if p_1 is the observed distribution of counts in response to frequency f_1 and p_2 corresponds to f_2, the interpolated distribution $\lambda * p_1 + (1-\lambda) * p_2$ with $0 \le \lambda \le 1$ was taken to represent the distribution of responses

of the unit to a frequency *f* such that $\log(f) = \lambda * \log(f_1) + (1-\lambda) * \log(f_2)$. Thresholds were estimated as the smallest frequency interval that could be discriminated using these artificially constructed distributions (probability of correct discrimination above 70.7%), and expressed as the interval divided by the geometric mean of the two frequencies.

Estimation and validation of Spectro-Temporal Receptive Fields (STRFs)

The linear response function for each unit in response to the soundtrack ('natural STRF') was computed using the software package STRFpak (Theunissen et al., 2001). The spectro-temporal representation of the stimuli that served as input to the calculation was generated by a simulation of the filtering of the auditory periphery implemented by the AIM software package (Bleeck et al., 2004). The STRF can be interpreted as an estimation of the neuron's most efficient stimulus (in the time-frequency domain), or as the inferred spectro-temporal impulse response of the neuron (Nelken, 2002). Under the assumption that neurons integrate signal energy linearly, STRFs can be used to predict the responses of neurons to a new stimulus by convolving the STRF with the spectro-temporal representation of the stimulus (Eggermont et al., 1983, Machens et al., 2004). The models were validated by computing the model on a training set and estimating their predictive power on a test set. For this purpose, the models were trained on 8 of the 9 minutes of the sound track. For each minute of the soundtrack, responses were predicted using both a natural STRF calculated from the responses of the unit to the rest of the soundtrack and an artificial STRF calculated from spike-triggered averaging the same unit responses to the random chords. For each minute of random chord stimuli, predictions were calculated using both an artificial STRF calculated from the remaining random chords responses, and a natural STRF based on the responses of the same unit to the full soundtrack. The similarity between the predicted response given by the STRF and the actual response, both smoothed with a 121 ms hamming window, was quantified by the correlation coefficient between them, as in the studies of Theunissen and his collaborators (Theunissen et al., 2000,

Theunissen et al., 2001). Many units were tested with the soundtrack but not with random chords (48/62). In order to test the assumption that natural STRFs were simply noisier versions of the artificial STRFs, we generated synthetic artificial STRFs ('synthetic STRFs') for these units whenever their natural STRF of these units had a clear dominant excitatory region. The synthetic STRFs consisted of a 2-D Gaussian filter centered at best frequency (see Fig. 2) with a spectral standard deviation of 1/6 octave and a temporal standard deviation of 12 ms. Using different parameters, including optimizing the spectral and temporal widths for each neuron separately, did not change the conclusions of the analysis based on the synthetic STRF.

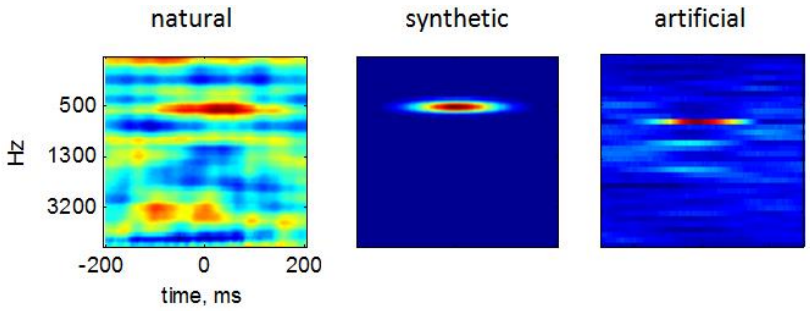

Figure 2. Synthetic STRFs. **Left -** A 'natural STRF': the linear response function calculated from the responses of one unit to the movie soundtrack. **Middle** - a 'synthetic STRF' consisting of a 2-D Gaussian filter in time and frequency, centered at the excitatory region of the natural STRF of this unit. The synthetic STRFs were constructed in case a unit was presented with the soundtrack, but not with the random chords. The minimal structure - a single excitatory region - resembles the typical artificial STRF calculated from responses to the random chords. **Right** - an example of an artificial STRF calculated from the responses of another unit.

Responses to artificial sounds

Ultra-fine frequency tuning

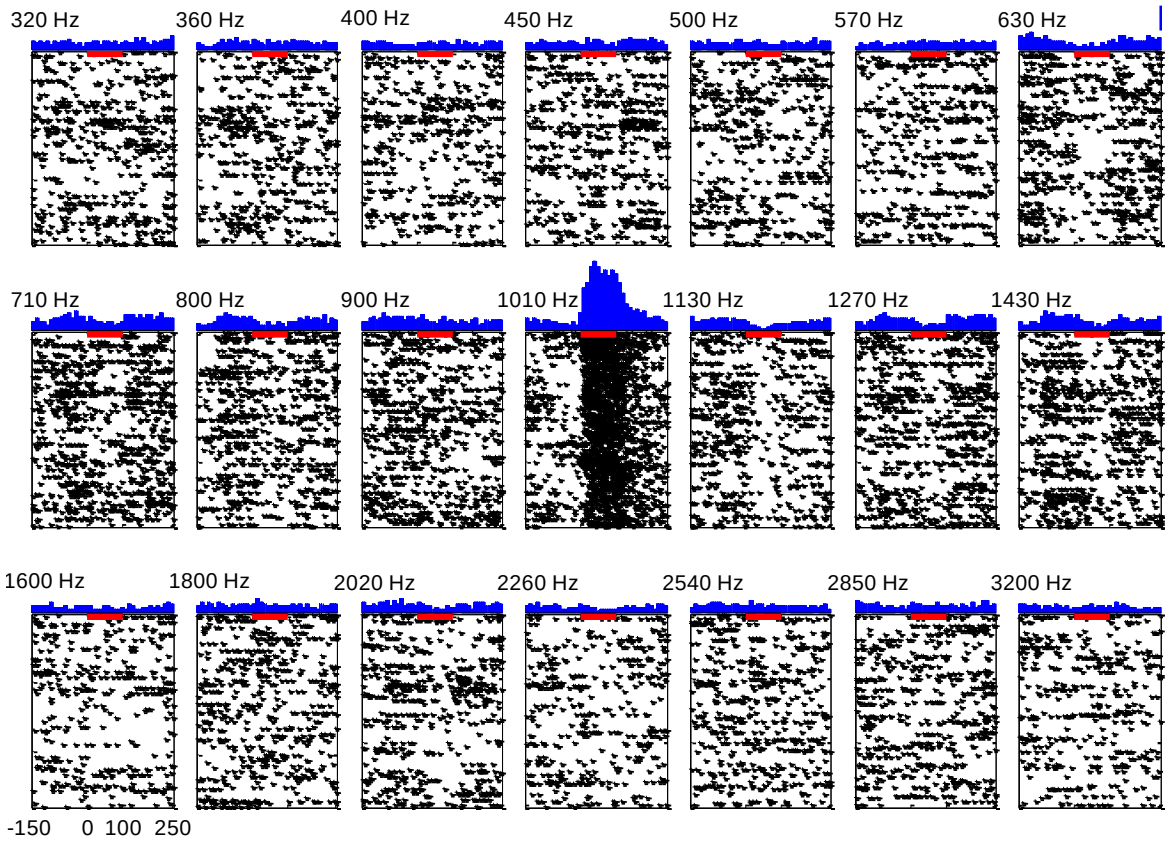

Figure 3 Selective responses to the Random Chord stimulus. Raster plots of responses of one unit to chords containing the frequency specified at the top of each panel (270 repetitions in each panel) and peri-stimulus time histograms (PSTH; bin width 10 ms) based on these raster plots (the scale line at the top right PSTH corresponds to a firing rate of 16 spikes per second). Red bars marks 100 ms (duration of one chord) from the beginning of the response to the preferred frequency. The unit responded significantly only to tone bursts at 1010 Hz (maximum firing rate: 47 spikes per second). The frequency table contained 20 additional frequencies (beyond those represented here, both below 320 Hz and above 3200 Hz); no other frequency elicited significant responses.

Figure 3 displays the raster responses of one unit to the different frequencies in the random chord stimulus with a resolution of 6 tones/octave. Each frequency appeared simultaneously with two other frequencies selected essentially randomly that therefore

produced a constant background rate on average. Only one of the 41 possible frequencies elicited excitatory responses above the mean rate in this unit. Furthermore, when a tone burst of that frequency appeared in the random-chord stimulus, a sustained response outlasting stimulus duration was elicited with high reliability. The lack of excitatory response to the two adjacent frequencies implies that this unit was more selective than the resolution of the frequency table used - 6 tones/octave. Thus, the bandwidth of this unit was substantially narrower than 1/6 octave, the peripheral filter bandwidth of humans.

Out of 31 units from two patients presented with random-chord stimulus at a resolution of 6 tones/octave, 27 had narrow, well-circumscribed frequency response area. About half (14/31) showed reliable responses to tone bursts at a single frequency only, with no consistent excitatory response to any other frequency. Thirteen additional units responded to two to three adjacent frequencies. The rest (4/31) exhibited more complex response patterns. The resolution of the frequency table we used for this stimulus was thus too coarse to directly measure the spectral bandwidth of most units.

A high-resolution random chord stimulus, with 18 tones per octave, was presented to a third patient. Of 16 units recorded in this patient, 14 exhibited highly elevated firing rate in response to a single frequency, with additional weaker, although significant responses to only one or two adjacent frequencies. The average bandwidth of these units can be conservatively estimated at about 1/12 octave, in agreement with the results from the lower resolution random chord stimulus presented above (Fig. 4a). Figure 4b displays typical Spectro-Temporal Receptive Fields (called 'artificial STRFs' below) derived from responses to the random chord stimulus by spike-triggered averaging. The best frequencies, defined for each unit as the frequency that elicited maximal response, ranged from 250Hz to 2 kHz in this population (Fig. 4c).

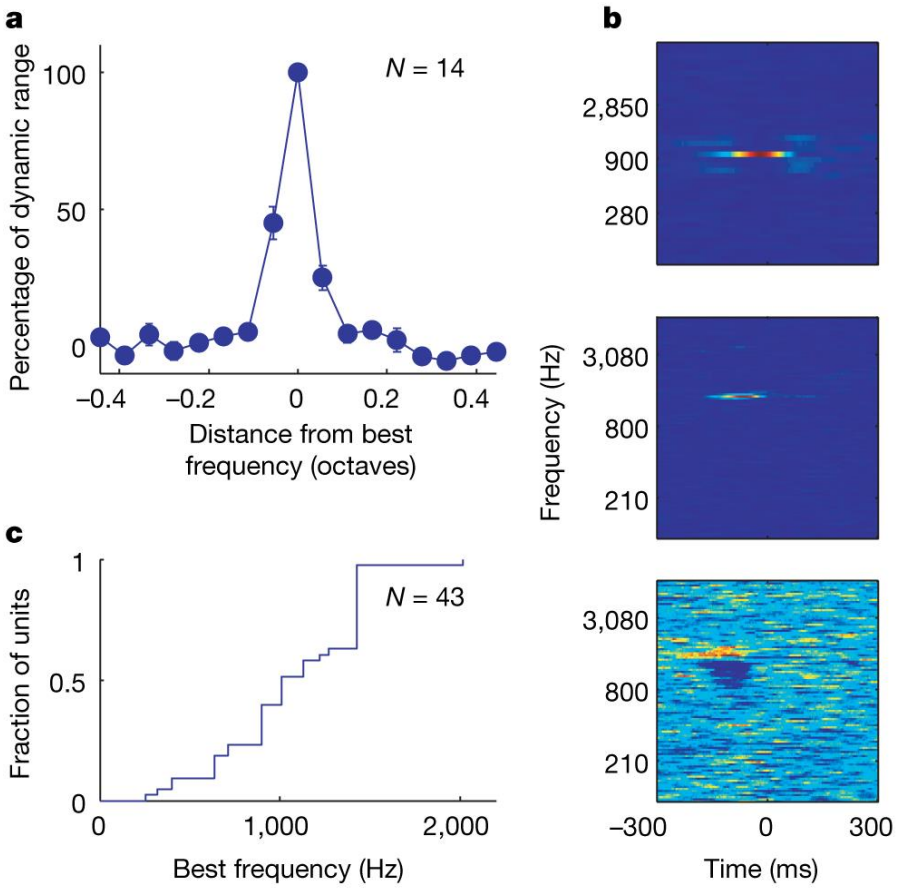

Figure 4. Frequency tuning in the responses to the random chord stimuli. **a**. Mean tuning curve: responses of each unit (mean spike count in a 50 ms bin) to the frequencies around the unit's best frequency were normalized and aligned on best frequency. The average is across all units showing excitatory responses to the high resolution random chords (N=14). Error bars indicate SEM. **b**. STRFs of three units estimated from the responses to the random chord stimuli. STRFs are expressed as excess firing rate relative to the mean. The top panel shows a unit tested with the six-tones-per-octave resolution that responded to a single frequency (color scale saturation: 2.5–39 spikes per second). The middle panel shows a unit tested the 18-tones-per-octave resolution that responded predominantly to a single frequency (color scale saturation: 1-32 spikes per second). The bottom panel shows a unit with complex tuning (color scale saturation: 0-3.4 spikes per second). **c.** Cumulative distribution of the best frequencies of 43 units that had a clear excitatory peak.

It is generally accepted that the frequency tuning curve of the auditory periphery in humans has a width of about 1/6 octave (Moore, 1982). Therefore, when presented with the random chord stimuli, the great majority of auditory cortical neurons showed substantially better frequency selectivity than auditory nerve fibers.

Single unit responses match behavioral performance

In order to study frequency discrimination performance based on the responses of units in single trials, a receiver operating characteristic (ROC) analysis was performed. We compared the empirical spike count distributions elicited by the different frequencies and determined the lowest discrimination threshold for each of the 47 units tested with the random-chord stimuli. Performance was quantified by the area under the ROC curve and discrimination threshold was set at 70.7% (see methods). In more than 60% of the excitatory cells (25/41) discrimination was above threshold for the smallest possible frequency difference tested, the spectral resolution of the stimulus (20/27 units tested at a resolution of 1/6 octave and 5/14 units tested at a resolution of 1/18 octave), as illustrated for one unit (Fig. 5).

For these units, we linearly interpolated spike count distributions to simulate possible distributions at intermediate frequencies that were not actually sampled in the stimulus (see methods). Thresholds were again estimated by the smallest frequency interval that could be discriminated using these intermediate distributions. The maximal slopes of the frequency response curves that determine the frequency selectivity are bounded from below by linear interpolation. Thus, the thresholds estimated in our procedure are underestimates. Even so, this procedure revealed units that had discrimination thresholds that matched and even exceeded behavioral performance of naïve human subjects (see Fig. 5e, Banai and Ahissar, 2004).

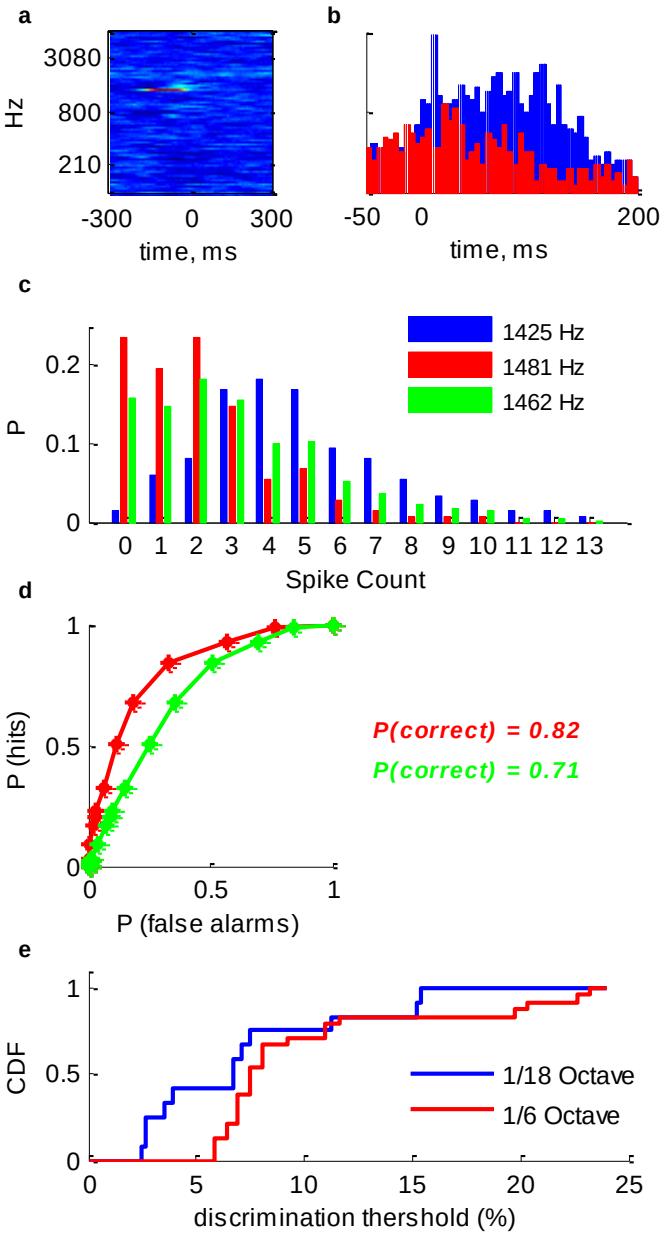

Figure 5. Frequency discrimination based on single trial responses. **a**. STRF of an excitatory unit estimated from the responses to the high resolution random chord stimulus (color scale saturation: 6 to 36 spikes per second). **b**. PSTHs (bin width: 5 ms); **blue** is the response to the best frequency (1425 Hz) and **red** is the response to the adjacent frequency (1481 Hz), scale: 0 to 40 spikes per second. **c**. Empirical spike count distributions based from the actual responses of the unit to its best frequency (**blue**), to the adjacent frequency (**red**) and a weighted sum of the two distributions estimating the responses to an intermediate frequency (1462 Hz, **green**). The ordinate represents the probability P of observing each spike count. **d**. ROC curves generated from pairs of distributions in **c** characterizing discrimination performance. **Red** – between 1425 Hz and 1481 Hz (interval: 3.9%). **Green** – between 1425 Hz and 1461 Hz (interval: 2.5%). The area under the curve estimates the probability of discriminating between the two frequencies in a 2-interval, 2 alternative forced choice task. The probability of correct classification is indicated to the right of the panel with the same color convention. **e**. Cumulative distribution of frequency just-noticeable differences for units tested with random chord stimuli at six tones per octave (**red**, N=24) and 18 tones per octave (**blue**, N=15).

Responses to natural sounds

Do units respond as narrow spectral filters also when presented with natural sounds? We analyzed responses elicited by nine minutes clips from the soundtrack of the feature film "The Good, the Bad and the Ugly", shown twice in each recording session. The soundtrack contained approximately equal-duration segments of dialogue, music and background noise (mainly carriages, street sounds and gun shots). The average firing rate was not significantly different between responses to the random chord stimuli and responses to the soundtrack (paired t-test, $t=1.04$, $df=13$, $p>0.3$), suggesting that the soundtrack was, on average, as successful as the artificial random chords in driving the neuronal response.

A straightforward comparison of the reproducibility between the two sound sets was not possible. While the soundtrack was presented twice in each recording session, the random chord stimuli were presented only once, and specific chord combinations were typically not repeated. In order to compare the response reproducibility between the two sounds, we considered auditory segments similar not based on a similarity in their spectral content, but based on a similarity in the neural response they elicited. During the presentation of random chords, many units had elevated firing rate when their BF was present in the chord, regardless of the other components simultaneously present. Thus, the responses to the random chords effectively consisted of two classes – high firing rate when the best frequency was present in the chord, and low background firing rate when the best frequency was not present. A similar class of preferred events evoking high firing rate in response to the movie was generated by identifying all the 100 ms long sound segments that elicited an elevated response (>92% of maximal response as measured with total spike count) in the first run of the movie. The coefficient of variation (CV) was used as a measure of the reproducibility of the neuron's response to its preferred stimuli. The CV calculated for the responses to chords that included the BF was compared with the CV calculated for the responses to the preferred

events in the second run of the movie. The CVs of the two ensembles were similar, indicating that reproducibility of responses was comparable between the two contexts.

The methods of Theunissen et al. (2000) were used to estimate STRFs from responses to the soundtrack (called 'natural STRFs' below), using generalized reverse correlation techniques. The exquisite spectral filtering clearly apparent in the artificial STRFs was partially lost – natural STRFs were noisier and appeared to have richer structure (Fig. 6a). Nevertheless, there were similarities between natural and artificial STRFs estimated for the same unit. For units recorded with both stimuli, the best frequency of the artificial STRF and the best frequency of the natural STRF were highly correlated ($r=0.63$, $df=13$, $p=0.016$; Fig. 6b). This is in agreement with the general finding that the best frequency is largely independent of auditory context (Woolley et al., 2006).

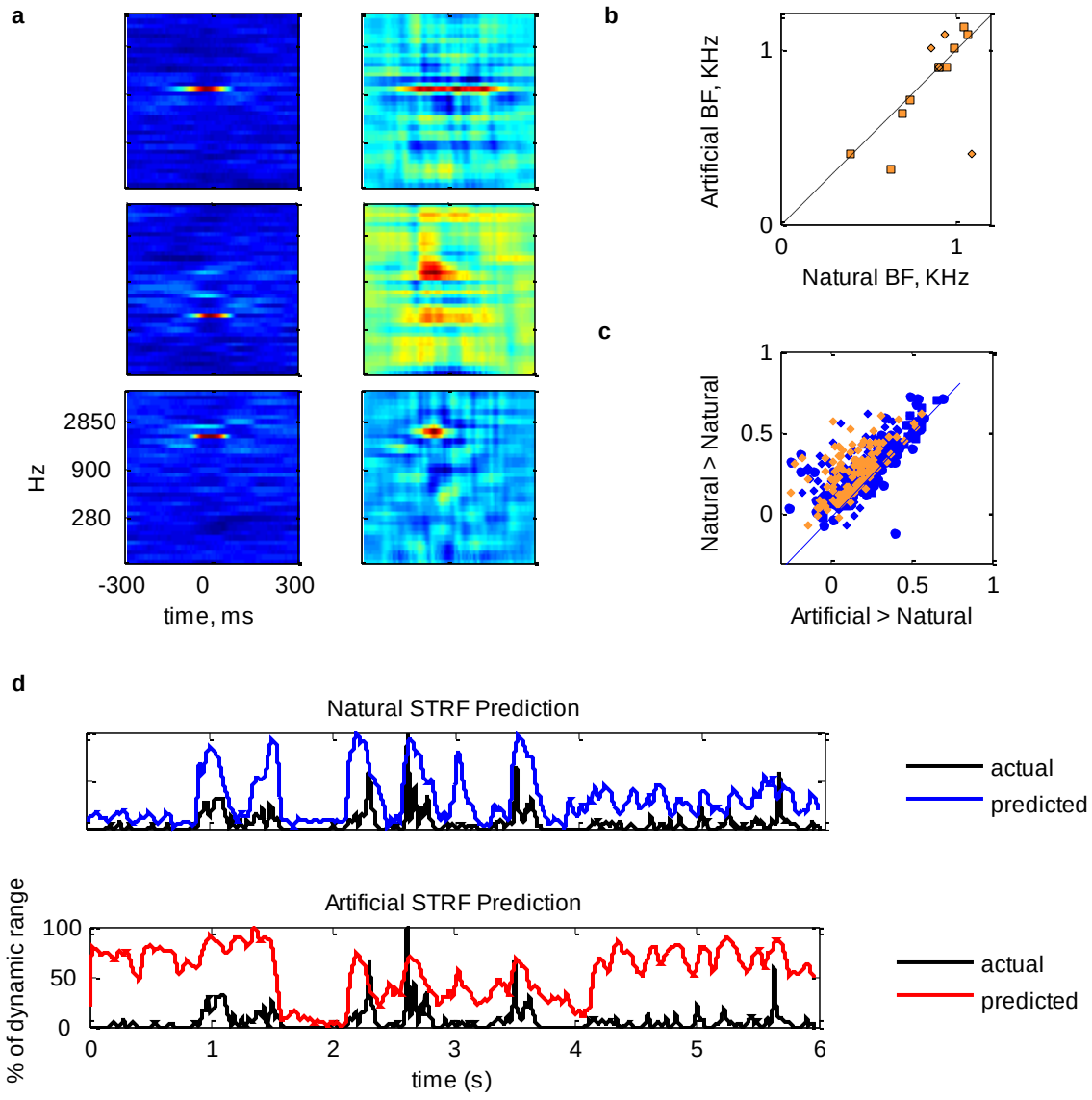

Figure 6. Natural versus artificial STRFs. **a**. STRFs of three units based on responses to the random chord stimulus (**left**) and to the soundtrack (**right**). **b.** Best frequency (BF) of artificial STRFs plotted against best frequency of natural STRFs (N=14). Different symbols represent different sessions (electrode positions). **c**. Correlations between predictions and actual responses to one-minute segments from the soundtrack. Abscissa: using artificial STRFs (**orange**, 14 units) or synthetic STRFs (**blue**, 31 units). Ordinate: using natural STRFs. **d**. Predictions to one minute of the soundtrack by natural (**top**) and artificial (**bottom**) STRFs plotted with the actual response of the unit. Both real response and prediction are smoothed with a 121 ms hamming window. Correlation coefficients were 0.62 and 0.23, for the natural and the artificial STRFs respectively.

The first- and second-order statistics characterizing the soundtrack were fully sampled by the random-chord stimuli (as verified by comparing the joint distribution of spectral and temporal modulations in the two stimulus ensembles) and the calculation of the natural STRFs corrected for second-order correlations in the stimulus (Theunissen et al., 2000). Thus, if the neurons integrated their spectro-temporal input linearly, natural and artificial STRFs should have been essentially equivalent. However, the soundtrack also contained higher-order spectral correlations, the effects of which on the STRFs could become apparent if the neurons had significant nonlinearities. These effects could be the reason for the additional structure in the natural STRFs.

We addressed this point by estimating the predictive power of the STRFs within and across context (defined here by the sound ensemble used to estimate the STRF - random-chord stimuli or film soundtrack). If artificial STRFs can predict responses to the soundtrack as well as (or better than) natural STRFs, then the differences in structure between the STRFs can be attributed to the limitations of the analysis. Alternatively, if natural STRFs can predict responses to the random chord stimuli as well as artificial STRFs, the simple spectral filter models should be treated as simplifications of actually more complex response patterns. In both cases, it can be concluded that the non-linear mechanisms that are not captured by the STRFs have only a small effect on the neuronal responses. Alternatively, if each STRF predicts the responses to new sounds from the same ensemble used to estimate the STRF better than the STRF derived from the other sound ensemble, then it can be inferred that there are significant non-linearities in the responses, with the natural sounds possibly engaging different processing mechanisms than those engaged by the artificial sounds.

For units recorded with both artificial and natural stimuli, predicted responses to one minute segments of the soundtrack were calculated with both the artificial and natural STRF (the natural STRF was estimated without using the responses to the segment whose responses were predicted). The predictive power of the STRFs was quantified by the correlation coefficient between the prediction and the actual response of the unit.

To have an absolute scale for the size of the correlations, correlations between the responses to two presentations of the soundtrack were calculated: they were 0.3 on average. Therefore STRF predictions were not expected to have substantially higher correlations with the responses (Hsu et al., 2004). The predictive power of the artificial STRFs on the soundtrack was notably low - 0.13 ± 0.14 (mean correlation ± std), about 40% of the expected maximum. More importantly, correlations were significantly higher within context: a natural STRF typically predicted the actual responses to a soundtrack segment (not used in its estimation) better than did an artificial STRF: 0.25 ± 0.14 (mean ± std, Fig. 6c), over 80% of the expected maximum correlation. A three-way analysis of variance (ANOVA) on STRF type x predicted segment x neuron showed a highly significant main effect of STRF type ($F(1,229)=72$, $p<<0.01$). The same general result was obtained when narrowband filters fitted to the excitatory area of each unit were used instead of the artificial STRFs (see above).

In a complementary analysis we compared the predictions of the responses to one minute of random chord stimulus using natural and artificial STRFs. The artificial STRFs were estimated omitting the responses to the minute being predicted. Since the random chord stimulus was presented only once, comparison between models was only possible without an independent estimate of the maximum achievable correlation. Predictive power was again better within-context: 0.42 ± 0.18 when using artificial STRFs compared to 0.25 ± 0.15 when using natural STRFs (3-way ANOVA on STRF type x predicted segment x neuron, main effect of STRF type, $F(1,107)=39$ $p<<0.01$). As natural STRFs were substantially less successful in predicting the responses to the random chord stimuli than artificial STRFs, the conclusion is that the richer structure of the natural STRFs is at least in part a reflection of additional nonlinear processing mechanisms that are not engaged by artificial stimuli but do shape the responses to the natural stimuli.

Conclusions

Frequency selectivity in human auditory cortex

The results demonstrate that frequency tuning in the human auditory cortex is substantially narrower than that typically found in the auditory cortex of non-human mammals (except bats, but see below an important exception in human data). Using pure tones under the commonly used barbiturate anesthesia, the tuning width at suprathreshold levels was found to be about one octave in cats (Read et al., 2001) and about a third of an octave on average in rats (Gaese and Ostwald, 2001). Comparisons of tuning between awake and anesthetized animals within the same species have repeatedly shown that bandwidths are wider in the awake preparation, as in cats (Qin et al., 2003, Moshitch et al., 2006) and rats (Gaese and Ostwald, 2001). Surveys of tuning in the auditory cortex of the awake macaque reported bandwidths that were typically half to one octave (Recanzone et al., 2000), and either very narrowly tuned neurons were rare (Recanzone et al., 2000) or bandwidths were wider than a seventh of an octave (Schwarz and Tomlinson, 1990). In the only other report of the frequency tuning of a unit in human auditory cortex (Howard III et al., 1996), the width at half-height was at least one octave. The frequency tuning derived from STRFs is typically somewhat narrower than that derived from pure tone responses, but seems to be wider than the data shown here. For example, in deeply anaesthetized cats, the STRF width was about half an octave (Miller et al., 2002). Thus, in mammalians species, the typical selectivity of cortical neurons was worse, not better, than that found in the auditory periphery. An important exception is the recent findings (Bartlett et al., 2011) in the awake marmoset monkey. Frequency tuning finer than the typical tuning of the marmoset auditory nerve was found for a large portion of the neurons recorded in primary auditory cortex (27%). Tuning was substantially finer than previously reported cortical data obtained from anesthetized animals, including from anesthetized marmosets, and comparable to that reported in this study. Further investigation is needed in order to determine whether the broader tuning observed in previous animal studies results from the use of

anesthesia during physiological recordings or from species differences. Thus, with the caution required by the small sample reported here, we propose that in contrast with studies in anaesthetized animals, the spectral selectivity characteristic of awake human auditory cortex is substantially better than that of the auditory periphery; furthermore, it is possible that this property is expressed in other primate species, but it is definitely absent in rodents and carnivores.

These results bear relevance to the apparent paradox of frequency hyperacuity demonstrated repeatedly in human psychoacoustics. Subjects with normal hearing, even untrained, successfully detect spectral differences that are substantially narrower than the bandwidth of single auditory nerve fibers. Banai and Ahissar (2004) report detection thresholds of about 4% in the general population when tested with a roving two alternative forced-choice frequency discrimination task. Our results demonstrate that frequency differences smaller than 3% could be reliably detected from single trial responses of single units. Thus, the responses of one of these cortical neurons could, in principle, underlie behavioral performance comparable to Banai and Ahissar's (2004) results on a single-trial basis. Consistent with our results, Tramo et al. (2002) reported that bilateral lesions of human auditory cortex cause significant elevations in frequency discrimination thresholds, suggesting a functional role for the electrophysiological findings reported here. Remarkably, the thresholds after bilateral auditory cortex lesions reported by Tramo et al. (2002) were about 10-20%, consistent with the peripheral tuning width in humans. We therefore suggest that the neural responses we observed in human auditory cortex reflect a readout of information available in the activity of large neuronal ensembles in subcortical stations, and that A1 is necessary in order for this readout to be performed, resulting in the behavioral hyperacuity of frequency discrimination in humans.

It is not clear why a low-level cue such as frequency is represented so explicitly in single neurons of human auditory. There is evidence that frequency discrimination in humans is correlated with a number of cognitive skills, including language abilities (Benasich and

Tallal, 2002), working memory (Banai and Ahissar, 2006), sensorimotor coordination (Jacoby et al., 2012) and learning capabilities (McArthur and Bishop, 2005). More speculatively, the high frequency resolution could be related to the unique use of acoustic signals such as music in humans, but more research is needed to clarify this puzzle.

Generalization between contexts is limited

Most importantly for this thesis, the results in this chapter suggest that stimulus encoding is not entirely determined by frequency selectivity. Based on the high, explicit sensitivity of the units to the spectral content of the artificial random chords, one might expect it would be easy to predict their responses to other complex auditory stimuli. The fact that STRFs exhibited superior predictive power when tested with sounds that belong to the ensemble used to estimate them (artificial/natural), suggests that non-linear mechanisms participate in shaping the neuronal responses (Theunissen et al., 2000). In the context of the random chords, the human cortical neurons seem to be straightforward encoders of frequency, and to represent this feature with a higher resolution than the periphery. If there are nonlinear mechanisms in the encoding of the artificial sounds, they were not captured with the artificial STRF and they were not those that appeared in the response to the complex sound from the soundtrack, with its significant higher order statistics. The findings therefore highlight that linear spectro-temporal models provide only partial descriptions of the cortical processing in real-world auditory scenes. However, the data were not sufficient to conclude the nature of the nonlinear encoding of natural sounds in the cortex.

THE OBJECT POTENTIAL

Evidence from MEG

* Andre Rupp, *Biomagnetism Section, Department of Neurology, University of Heidelberg, Germany, conducted all the psychoacoustic and MEG experiments*

Tone in noise as a basic auditory scene

The results surveyed thus far demonstrate that in the auditory cortex, responses to ethologically meaningful auditory scenes engage mechanisms that are not reflected in the responses to artificial sounds. While complex auditory scenes remain by and large outside the auditory laboratory, there has been considerable research into the psychophysics and electrophysiology of a very simple auditory scene: a pure tone ('target') in the presence of noise ('masker'). The relevance of such sounds to 'real life' auditory processing is clear: most of the sounds that sources emit in the environment reach our ears accompanied by noise. The two components of the sound differ physically - in bandwidth, intensity as well as timing and duration. Perceptually, once the tone is loud enough to exceed detection level it stands out as a distinct 'tone object' segregated from the 'background noise'.

Classic experiments traced the threshold for detection of a pure tone in the presence of a masker ('masked threshold') while attributes of the noise were manipulated. Figure 1, adapted from Moore (1999), describes the masked threshold as a function of the bandwidth of a noise band centered on the tone frequency, here 1 kHz. The spectrum level of the noise was kept constant in these experiments, so increasing its bandwidth increased the overall masker energy. For bandwidths less than 500 Hz, an increase in the masker bandwidth also caused an increase in thresholds, as would be expected from

the increased overall energy. When the bandwidth was increased beyond 500 Hz, masked thresholds did not increase further. This implies the existence of a spectral band around the tone frequency – the 'critical band' – that affects tone detection. Importantly, this behavior can be parsimoniously and fully accounted for using principles of signal detection theory without invoking concepts of auditory scene analysis (ASA, see chapter 1, and Schnupp et al., 2011b). The terminology of ASA is nevertheless invoked here because the introspective sensation when listening to such sounds is that of a simple auditory scene - under the appropriate conditions, subjects perceive two objects, discriminated by the quality of 'pitch' that marks the tone as separate from the masker.

Comodulation Masking Release (CMR) and its electrophysiological correlates

The masking results described above pertain to the use of a Gaussian masker. What happens when the amplitude of the background noise is coherently modulated across different frequency bands (a 'modulated' masker)? Above a masker bandwidth of 50 Hz, tone detection thresholds decrease with increasing bandwidth, and they continue to drop even when the masker's bandwidth exceeds the critical bandwidth (Fig. 1). This phenomenon is called Comodulation Masking Release (CMR; Hall et al., 1984).

Neural correlates of CMR have been found early in the auditory processing pathway (Pressnitzer et al., 2001, Neuert et al., 2004). Even at the cochlear nucleus, neurons were shown to 'lock' to the noise amplitude modulations, with responses that followed the slow envelope modulations of the sound. In many neurons, this locking was reduced when coherently modulated energy was added in distant frequency bands, making it easier to detect the responses to the tone.

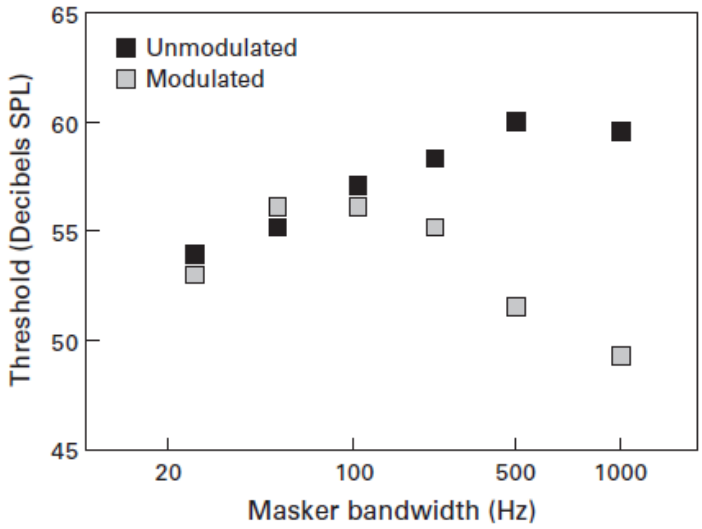

Figure 1. Tone in masker – behavioral results. Detection thresholds for a target tone of 1 kHz in noise for coherently modulated and unmodulated noise. Thresholds are plotted against the bandwidth of the noise. For unmodulated noise (**black**), the threshold increases with the bandwidth until the critical bandwidth is reached. In contrast, for modulated noise (**gray**), detection level drops with the noise bandwidth above 100Hz tones, even above the critical band. This effect is called Comodulation Masking Release (CMR). Adapted from Moore (1999).

Las et al. (2005) described the responses to such modulated maskers and tones using intracellular recordings in the cat A1. Many neurons showed locking of the membrane potential to the amplitude modulations of the masker. The locking increased with masker bandwidth. When a low-level tone was added to the noise, the locking of most of the cortical neurons was markedly diminished ('locking suppression'). Locking suppression was found in A1 for low-level tones, even at levels below the neuronal threshold in silence (Fig. 2). The effect of locking suppression had a special timing structure, starting at the second noise cycle after tone onset (~75 ms after tone onset).

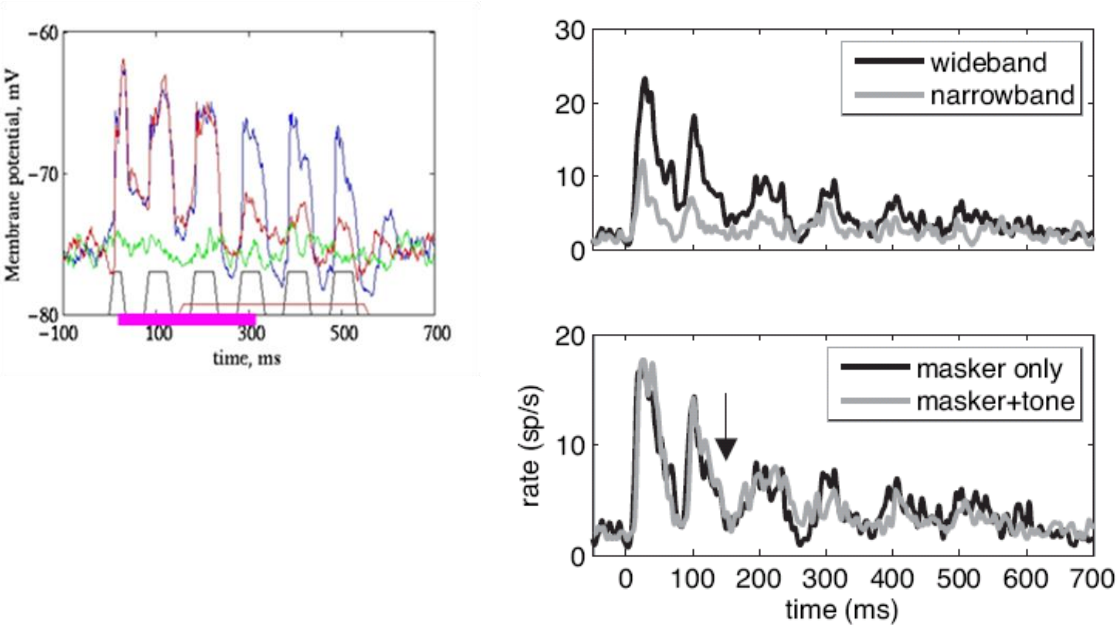

Figure 2. Electrophysiology of CMR. a. The membrane potential of this neuron from cat A1 showed locking to amplitude modulations of the noise (**blue**) that was considerably reduced when a low-level tone was added (**red**). Presented alone, the same tone did not elicit any notable response (**green**). Schematic presentation of stimulus patterns is shown at the bottom of the panel. Adapted from Las et al. (2005). **b.** Population responses in cat A1. Multiple extracellular recordings (45 units) were averaged together to approximate the responses that are expected to drive population signals such as MEG. **Top**: responses to wideband modulated maskers were larger than responses to narrowband modulated maskers. **Bottom**: adding a tone to a modulated masker reduced, but did not abolish, envelope locking (41 units tested at SNRs comparable to those used in the MEG experiment) starting with the 2nd cycle after tone onset (a**rrow**). Adapted from Rupp et al. (2007).

In a subsequent publication (Rupp et al., 2007) the authors generated a pseudo 'population response' from these data by pooling together responses from 41 neurons that were tested with different tone and masker combinations. Tone levels selected for this figure were comparable to those used in the current MEG experiment. The results presented in Fig. 2b are an approximation for the expected average activity in primary auditory cortex in response to maskers and tones of the kind presented in the

experiment. The intracellular data predicted that with populations signals, like the magnetic fields measured in MEG, locking to amplitude modulations of the noise should increase with noise bandwidth and be suppressed, but not eliminated, with the addition of a low-level tone. Furthermore the population locking suppression should start at the 2nd noise cycle following tone onset. In this work I studied MEG signals elicited by these same stimuli, in order to uncover correlates of CMR in human cortex, and to test the predictions of the single unit data.

Experimental setup and data analysis

Human listeners

Twenty-five subjects (12 female and 13 male, aged 21 to 43 (29.5 ± 6.4, mean ± std), 24 right-handed) participated after giving informed consent. No subject reported any history of peripheral or central hearing disorders. All subjects were familiar with MEG recording sessions. All psychoacoustic and MEG data were collected by Andre Rupp, Biomagnetism Section, Department of Neurology, University of Heidelberg, Germany.

Stimuli

The maskers were either wideband (900 Hz) noise or a narrowband (90 Hz) noise, both with a central frequency of 500 Hz and a duration of 525 ms. Maskers were gated at the onset and the offset with a 10-ms linear window and were either unmodulated (the 'unmodulated' condition) or amplitude-modulated (the 'modulated' condition). As in Las et al. (2005), a trapezoidal modulation pattern (10 ms linear onsets and offsets) was used for six trapezoid cycles, the first cycle being 25 ms long and the remaining cycles 50 ms long with 50 ms of silence between each period. The test signal was a 500-Hz pure tone with a duration of 275 ms including 10-ms linear ramps at onset and offset. The tone was switched on 250 ms after masker onset to separate the neuromagnetic onset responses to the tone from the onset responses to the overall presence of sound (see Fig. 3a for a schematic representation of the different maskers with the tone). For all four variants of the noise (modulated vs. unmodulated and wideband vs. narrowband)

the level of the test tone was set to +5 dB and +15 dB above the average detection threshold derived from five subjects in a pilot experiment. The stimulus set was comprised of 12 different sounds – 4 masker conditions each presented in 3 variants, either alone ('noise alone') or with a +5/+15 dB tone.

Stimuli were generated digitally at a sampling frequency of 48 kHz. All noise bursts were created as "running noise" sounds to eliminate effects due to the fine structure of the signal. D/A-conversion was performed using an RME Audio soundcard (RME Audio, Haimhausen, Germany) connected to a PC. Sounds were delivered diotically to the subjects with a custom-made sound list processor via ER-3 (Etymotic Research Inc., Elk Grove Village, IL) earphones connected to 90-cm plastic tubes and foam earpieces. The inter-stimulus interval (offset to onset) was set to 1000 ms plus a random offset ranging from 0-50 ms. A Brüel and Kjær sound level meter (Naerum, Denmark) was used to set the overall level of the stimuli to 70 dB SPL.

Data acquisition

The gradients of the magnetic field were recorded with a whole-head 122-channel gradiometer system (NeuromagElektaOy, Helsinki, Finland) inside a magnetically shielded room (IMEDCO, Hägendorf, Switzerland). Data were sampled at a rate of 1000 Hz and low-pass filtered at 330 Hz online. Before data recording, four indicator coils were attached to the subject's scalp and digitized relative to anatomical landmarks to determine the position of the head relative to the gradiometers. During the recording sessions, subjects sat in an upright position and watched a silent movie of their own choice. They were instructed to concentrate on the movie and to ignore the sounds presented over the earphones. Data were acquired in two separate sessions. Each MEG registration, for both modulated maskers and unmodulated maskers, lasted about 60 minutes.

Data analysis

Averaging with artifact monitoring was carried out such that sweeps exceeding a peak level of 8,000 fT/cm or a slope of 800 fT/cm per sample were rejected. For each stimulus condition, about 350 single sweeps covering a range from 100 ms before stimulus onset to 1500 ms after stimulus onset were averaged. Linear trends were subtracted based on 350 ms quiet periods before stimulus onset and after its offset (-400:-50 ms and 775:1125 ms relative to stimulus onset).

To extract absolute changes in the magnetic signal regardless of signal polarity, all signals were aligned according to the polarity of their initial response to the noise onset. Polarity was determined separately for each condition and each sensor location by the sign of the averaged signal across all subjects 70 ms after noise onset. This polarity was then used to align the signals from that condition and sensor location for all subjects.

Locking to amplitude modulations was quantified by a Fourier analysis in a time window that included the response to the last three noise cycles (300-600 ms after noise onset). Locking suppression was investigated by comparing the fine structure of the stereotypical response pattern to each noise cycle. The response locked to each noise cycle began with an $N_a m/P_a m$ pair – a negativity that peaked ~30 ms after cycle onset, followed by a positive peak ~7 ms later (see Fig. 5b). The magnitude of the $N_a m$ of the masker alone condition was compared to the $N_a m$ elicited by the masker plus tone. The timing of the $N_a m$ peak (31±1 ms after the cycle onset) and the timing of the field minima (7±1 ms before and after the $N_a m$ peak) were determined for each noise cycle in the responses to the noise presented alone. The fields were linearly interpolated between these minima to determine $N_a m$ size. The dependence of $N_a m$ size on tone presence was analyzed using a 4-way ANOVA (tone level X subject X sensor position X noise cycle). For the graphical presentation of $N_a m$ components in Fig. 6, the first step was to interpolate the magnetic signal by a 3^{rd} order polynomial fitted to the 6 ms before the first $N_a m$ minimum and to the 6 ms following the second minimum. The

interpolated field for each modulation cycle was then subtracted from the magnetic signal to extract the N_am signals displayed in Fig. 6 below the traces.

Unless otherwise stated, the analyses in this chapter were based on the full set of gradiometers available in all recordings (N=108). All statistical tests were considered significant at the 0.05 level, unless explicitly stated otherwise.

Psychoacoustics

After MEG recordings were completed an adaptive two-alternative forced-choice task including a two-down one-up tracking rule to estimate the 70.7% correct point on the psychometric curve was applied to determine thresholds for all four stimulus conditions. Subjects were instructed to listen to four noise bursts separated by 0.7 s. The signal was contained in either the 2nd or the 3rd noise burst. At the beginning of each block the tone level was set such that the signal was clearly audible. Visual feedback was given after each manual response. The step-size was set to 1.12 dB. Stimuli were generated using the same equipment and scripts used for the MEG-recordings. All psychoacoustic measurements were carried out by Andre Rupp, Biomagnetism Section, Department of Neurology, University of Heidelberg, Germany.

Psychoacoustic thresholds of human listeners

Figure 3b summarizes the psychoacoustic CMR thresholds and standard errors for the detection of the 500 Hz tone embedded in wideband or narrowband noise. A 'modulation' (modulated versus unmodulated) x 'bandwidth' (wideband versus narrowband) ANOVA for dependent variables revealed a highly significant modulated-unmodulated effect (modulated vs. unmodulated masker: $F(1,24)=610$, $P<<0.01$) as well as a true CMR, a highly significant decrease in the threshold with increase in the bandwidth ($F(1,24)=194$, $P<<0.01$).

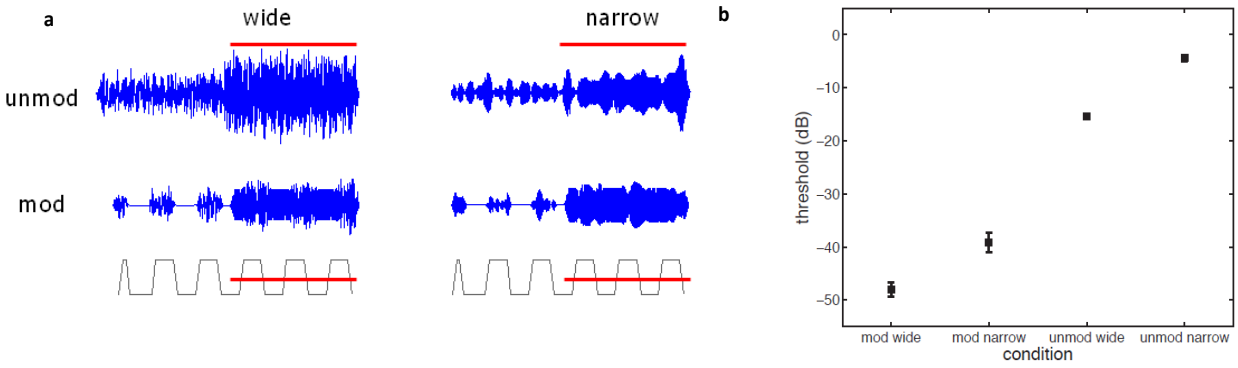

Figure 3. a. The sounds: schematic representation and waveforms for the different combinations of maskers and tone used in the experiment. The tones (represented by the **red line**) were added to noise maskers that were either wideband (**left**) or narrowband (**right**), unmodulated (**top**) or with amplitude modulation (**bottom**, modulation cycles represented in gray). **b. Psychoacoustics:** mean thresholds and standard errors obtained in this experiment for the different maskers (N=25). Stimulus names follow the same conventions as in **a**.

Locking and locking suppression

Figure 4 shows the grand average of the all gradiometer responses (N=108 sensor positions x 25 subjects) to all 12 conditions, organized according to the type of masker (modulated vs. unmodulated and broadband vs. narrowband noise): noise alone (light gray), with a tone added at +5 dB (dark gray) and with a tone added at +15 dB (black). The responses included a prominent noise-evoked N_1m locked to the onset of the masker. In most conditions, the onset of the tone embedded in noise evoked a negativity locked to the onset of the tone as well. Following the offset of the masker, there was a sustained field that decayed slowly back to baseline.

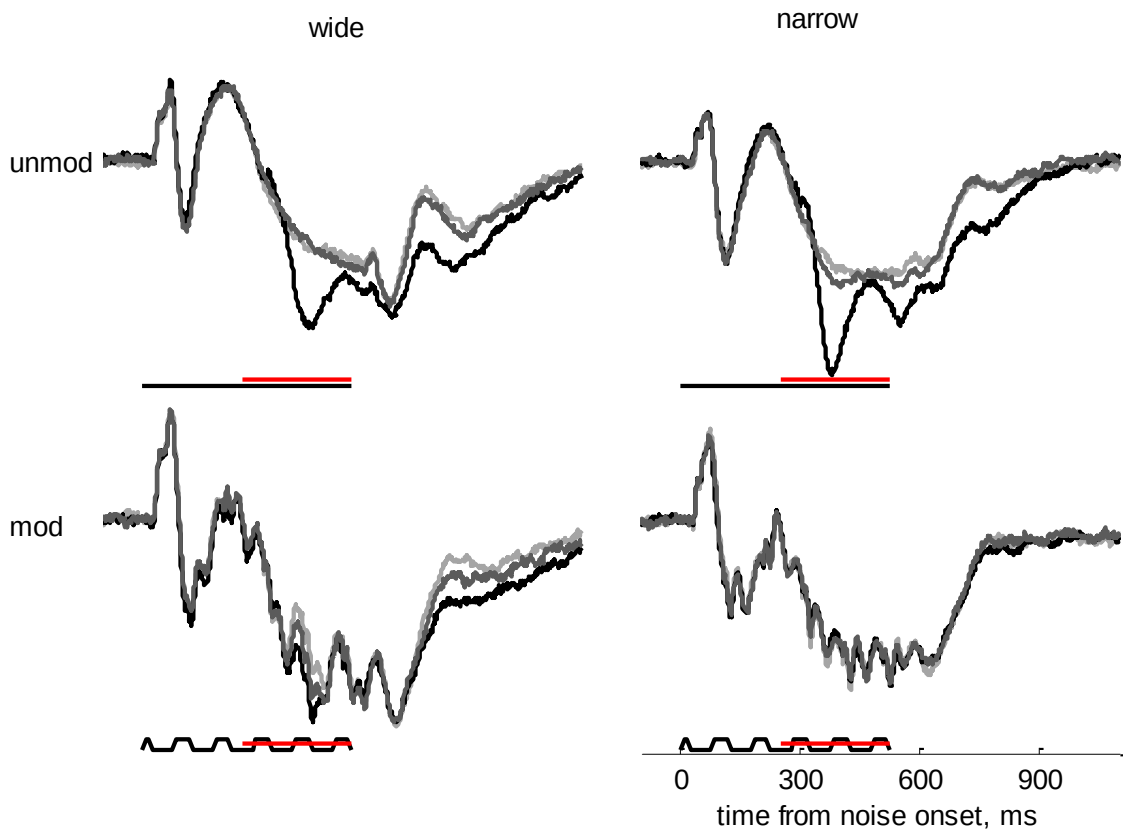

Figure 4. MEG response profiles. Grand Averages of the gradiometer responses across all subjects (N=25) and all sensor positions (N=108) for noise alone (gray) and with a +5 dB tone (**dark gray**) or with a +15 dB tone (**black**). Standard error for these average traces was smaller than the width of the line. Responses are organized according to masker type, in the same layout as in Fig. 3a. Response amplitude in arbitrary units, uniform across panels. The lines at the bottom of each panel present the stimulus schematically: masker (**black**) and tone (**red**).

The slow components of the responses appear similar for all masker types. When the noise was modulated, clearly identifiable fast modulations locked to the cycles of amplitude modulation appeared in the responses on top of the slower signals. Figure 5a shows a magnified view of the average responses to the modulated wideband masker 300-600 ms after noise onset. This time window contains the response to the last three noise cycles after the initial response to the noise is diminished. A repeating pattern locked to the amplitude modulation ('Locking') is easily identifiable. A fast Fourier transform was applied on the gradiometer signals in this 300 ms window for a subset of

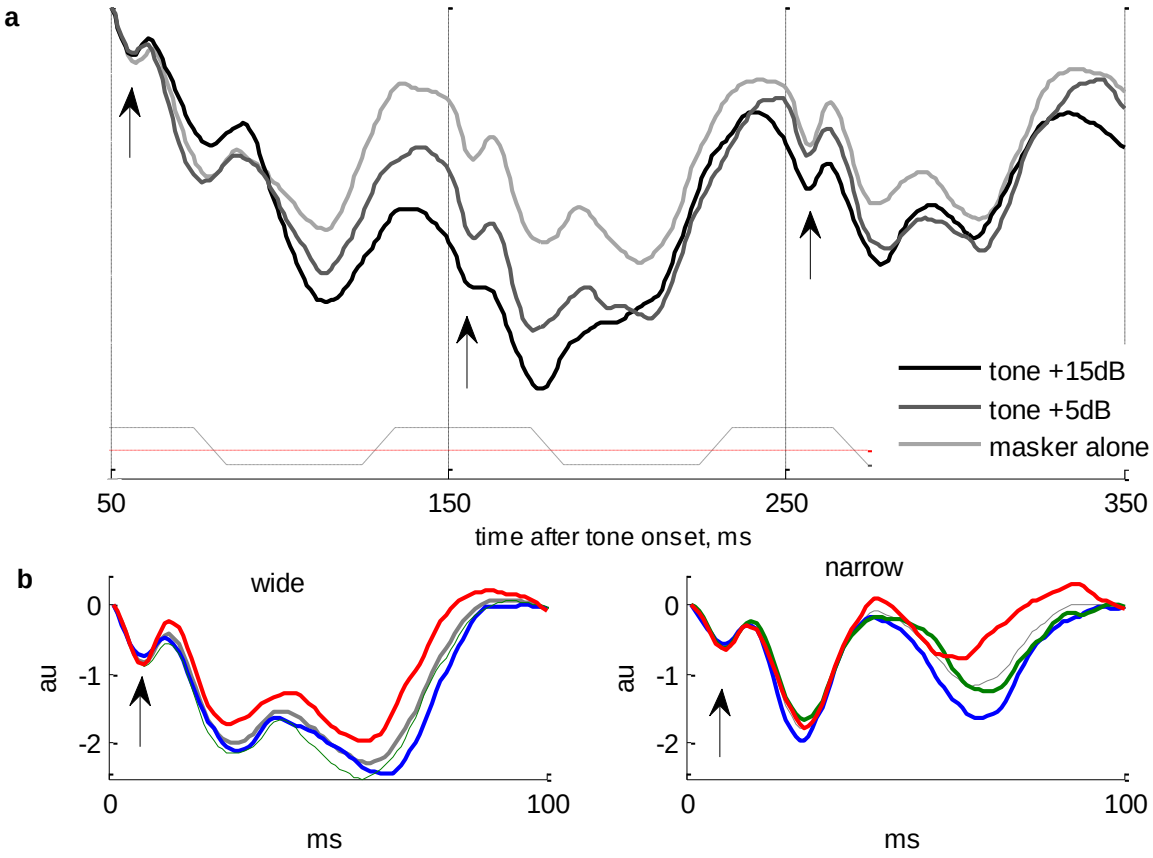

Figure 5. Locking to the amplitude modulation of the noise. a. Magnified view of the response to the last three modulation cycles of the wideband noise (average response across all subjects and sensor positions, N=25*108): masker alone (**light gray**), with a tone added at + 5 dB (**dark gray**) and at +15 dB (**black**). Linear trends were subtracted from each average trace individually. Traces were smoothed with a 6 ms moving average window. Dotted lines present the stimulus schematically in this time window: masker amplitude modulations (**black**) and tone (**red**). **b.** Stereotypical response (in arbitrary units, identical in both panels) to the last 3 noise cycles for wideband (**left**) and narrowband (**right**) noise. Each cycle (**colored lines**) is plotted after subtraction of linear trends, along with the average across cycles (**gray line**). Arrows mark the N_am components.

temporal sensors (7-12 AP, 1-11 ML) where responses were generally larger. The portion of variance captured by the third Fourier component was used as a measure of locking to the 10 Hz modulations. When the wideband modulated noise was presented

alone, 20% ± 0.4% of the variance (mean ± sem) across all sensor positions and subjects was captured by the third Fourier component. This was slightly reduced with tone addition - 18% ± 0.4% and 17% ± 0.4%, (mean ± sem), for the +5 dB and +15 dB tone respectively. These differences were highly significant (3-way ANOVA on tone level X subject X sensor position, main effect of tone level, $F(2,4605)=29$, $p<<0.01$), with post-hoc comparisons confirming that all pairwise differences were significant. Results when the entire sensor set was used confirmed a locking suppression that increases with tone level (16.5% ± 0.2%, 14.2% ± 0.2% and 13.6% ± 0.2% (mean ± sem) for noise alone, noise with +5 dB tone and noise with +15 dB tone respectively).

For both narrowband and wideband noise, the response locked to the amplitude modulations of the noise began with a stereotypical pattern consisting of $N_a m/P_a m$ pair - a negativity that peaked ~30 ms after the cycle started followed by a positivity that peaked ~7 ms later (see Fig. 5b). In the presence of a tone, the peak of the $N_a m$ component was affected, as illustrated in Fig. 6, which depicts responses recorded from a single temporal sensor (9 AP, 10 LM) in two subjects. The time frame displayed contains the responses to the last four noise cycles, with and without tone. The $N_a m$ components were identified in all cycles as the peak that occurred 31±1 ms following cycle onset. Below each trace, the locked components extracted from the signals are shown, with the $N_a m$ components peak times that were used in the analysis marked with an arrow. To isolate the $N_a m$ from other components, an estimate of the magnetic field in the absence of the presumed $N_a m$ component was subtracted from the original signal (see methods). The $N_a m$ components appear below each panel with and without the addition of the tone (black and gray lines respectively). The $N_a m$ component just before the onset of the tone was not affected by the tone, as expected. The $N_a m$ evoked by the 1^{st} noise cycle following tone onset was also essentially as large as that evoked by the tone alone. However, the $N_a m$ component evoked by the 2^{nd} noise cycle following tone onset was reduced in the presence of the tone.

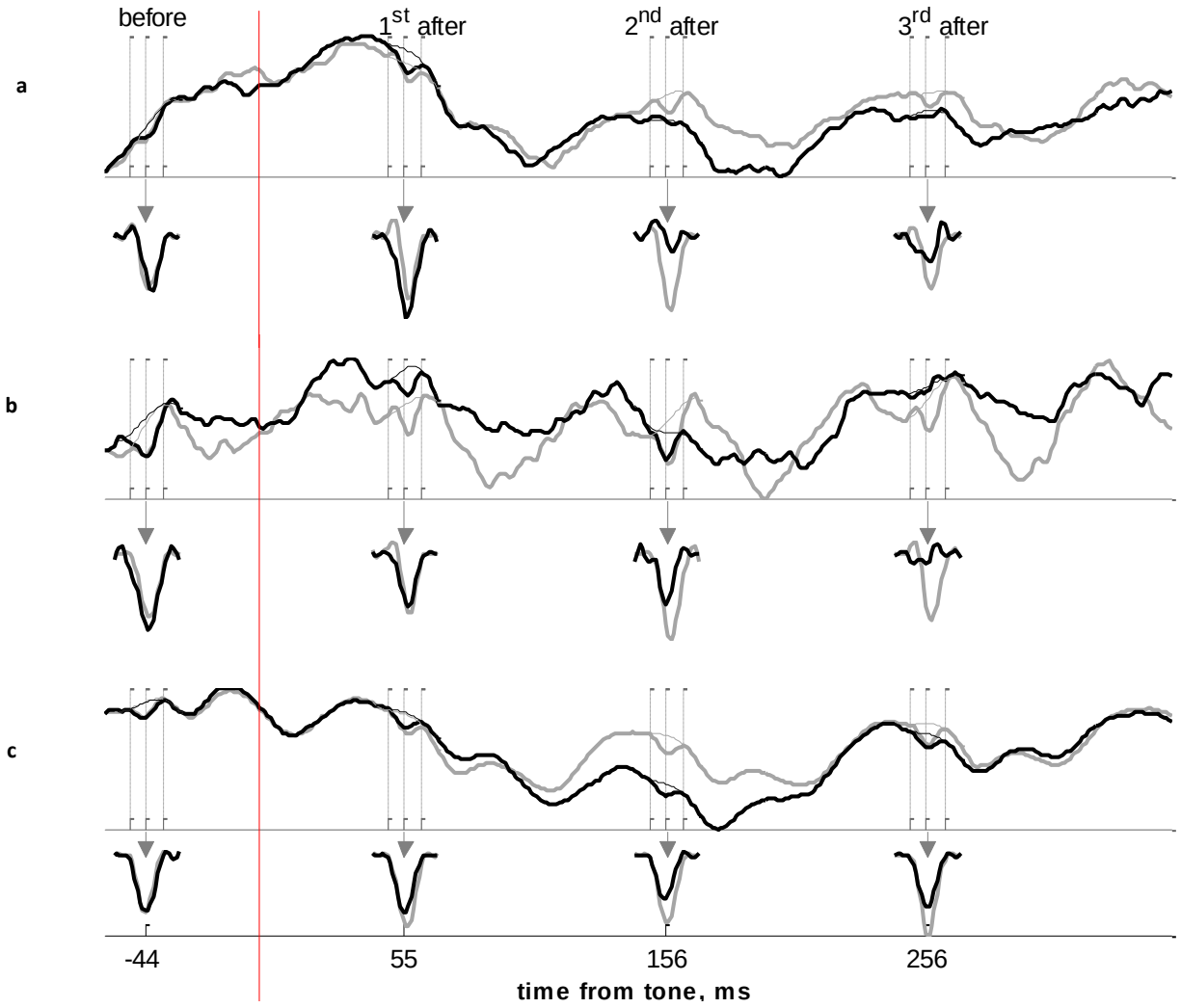

Figure 6. Locking suppression. N_am components in response to the last four cycles of the modulated wideband noise are illustrated in the recordings of a single temporal sensor (T8, position: 9 AP, 10LM) from two individual subjects (ma and mh, **a** and **b** respectively) and in the grand average across all subjects. Averaging was performed on a subset of temporal gradiometers (**c**, N=25 subjects*42 gradiometers, positions: 7-12 AP, and 1-4 8-11 LM). Each panel contains the average trace (**top**) and the N_am components (**bottom**) estimated by subtracting from the gradiometer average (**thick line**) the interpolated field at the same time (**thin line**, see methods): for masker alone (**gray**) and with a tone added at +15 dB (**black**). Red line marks tone onset. Dotted lines mark maxima and minima times used in the statistical analysis. The effect of the tone was most consistent in the 2nd and 3rd cycles after tone onset.

Figure 6c shows the average across all subjects. Population results largely followed the single-sensor data: the N_am evoked by the noise cycle before tone onset was unchanged by the tone. The N_am evoked by the 1st noise cycle following tone onset was affected less than the N_am evoked by later noise cycles. Statistical tests confirmed a significant effect of the addition of the tone on the magnitude of the N_am component for both the wideband noise (4-way ANOVA tone level X subject X sensor position X noise cycle, main effect of tone level, $p=0.0017$, $F(2,26172)=6.4$) and the narrowband noise ($p<<0.01$, $F(2,26180)=13.8$). Post-hoc comparisons verified that the significant decrease occurred for the +15 dB tone in the 2nd noise cycle and 3rd noise cycle both in the narrowband and in the wideband conditions. Although the N_am of the 1st noise cycle after tone addition followed the same trend, and decreased with tone addition, this reduction was not significant for either the narrowband or the wideband conditions ($p>0.05$). The results regarding the 1st noise cycle should be interpreted with caution. Unfortunately, the response to the tone started at about the same time (~50 ms after tone onset, ~25 ms after cycle onset). It is therefore difficult to say whether the effects at the 1st noise cycle after tone onset were due to the specific responses to the tone, or whether they represented a modified N_am response to the noise.

Figure 7 depicts the topographical distribution of the N_am component on the scalp. The signal was maximal in the temporal electrodes and a right hemisphere dominance is apparent and consistent for all noise cycles after tone onset, especially in the noise alone condition.

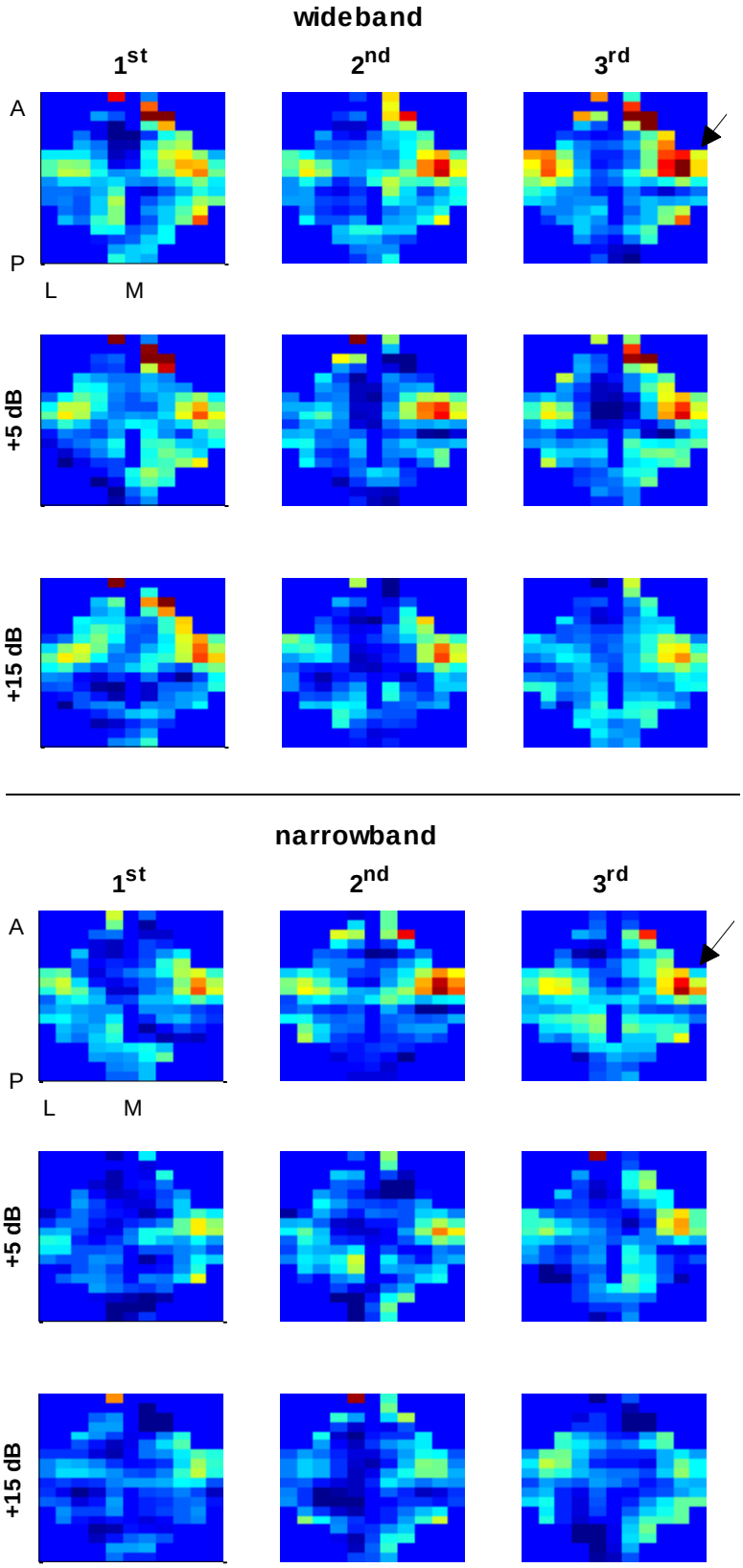

Figure 7. Right hemisphere dominance of envelope locking. Scalp distribution of N_am magnitude for wideband noise (**top** panel) and narrowband noise (**bottom** panel), organized by the order of noise cycle after tone onset (**columns**) and by condition (**top**: noise alone, **middle**: +15 dB tone, **bottom**: +5 dB tone). Scalp distribution was calculated by averaging the Nam magnitude across all subjects, and smoothing with a 3X3 moving window (the product of two Hanning windows along the two axes). Color scale was normalized for each masker condition (identical across all 9 panel of the condition). Arrow marks the temporal electrodes of the right hemisphere.

Effects of tone on longer-latency potentials

In addition to the effects of tone on the N_am, there was a much larger and slower 'tone-difference response' that could be isolated by subtracting the response to the noise alone from the response to tone with noise.

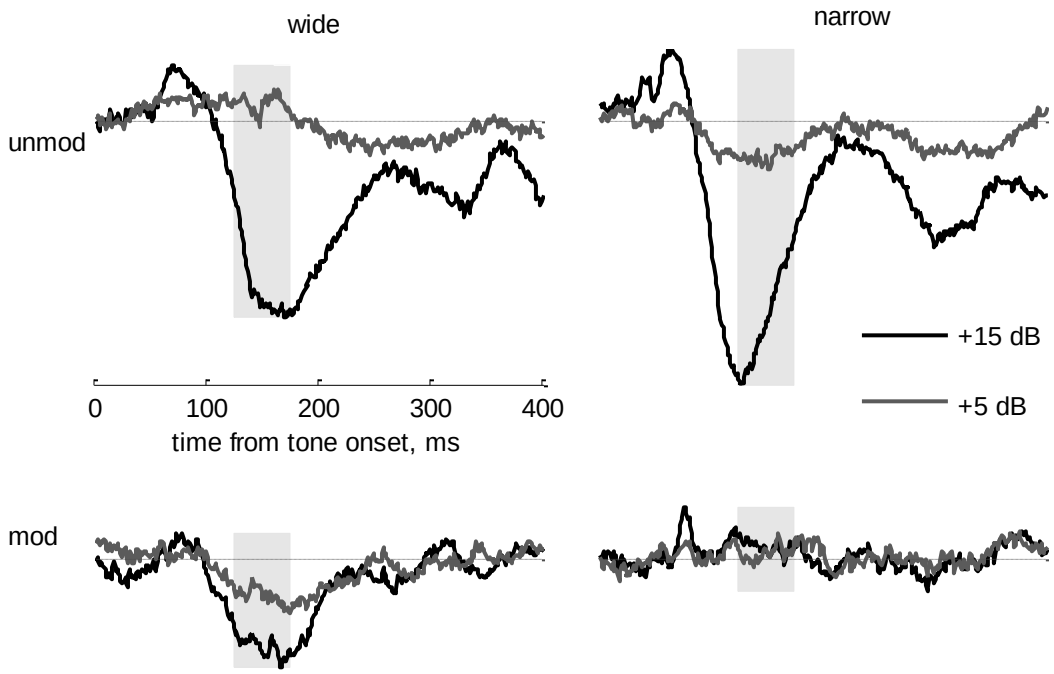

Figure 8. Tone-difference response. The result of subtracting the average response for noise alone from the average response when the noise was presented with tone for each of the four different maskers (N=25 subjects*108 sensors). **Black:** the +15 dB tone. **Gray**: the +5 dB tone. Same layout as in Fig. 3a. Gray areas mark the time window used in the statistical analysis.

The tone-difference response started ~75 ms after tone onset and lasted for more than 150 ms. In the unmodulated conditions (Fig. 8, top row) this response was very prominent for the +15 dB tone. On the other hand, for the +5 dB tone, this response was either significantly diminished (the narrowband noise) or entirely missing (the wideband noise). In contrast, in the modulated condition, the tone-difference signal was present for both tone levels in the wideband condition and was absent for both tone levels in the narrowband condition.

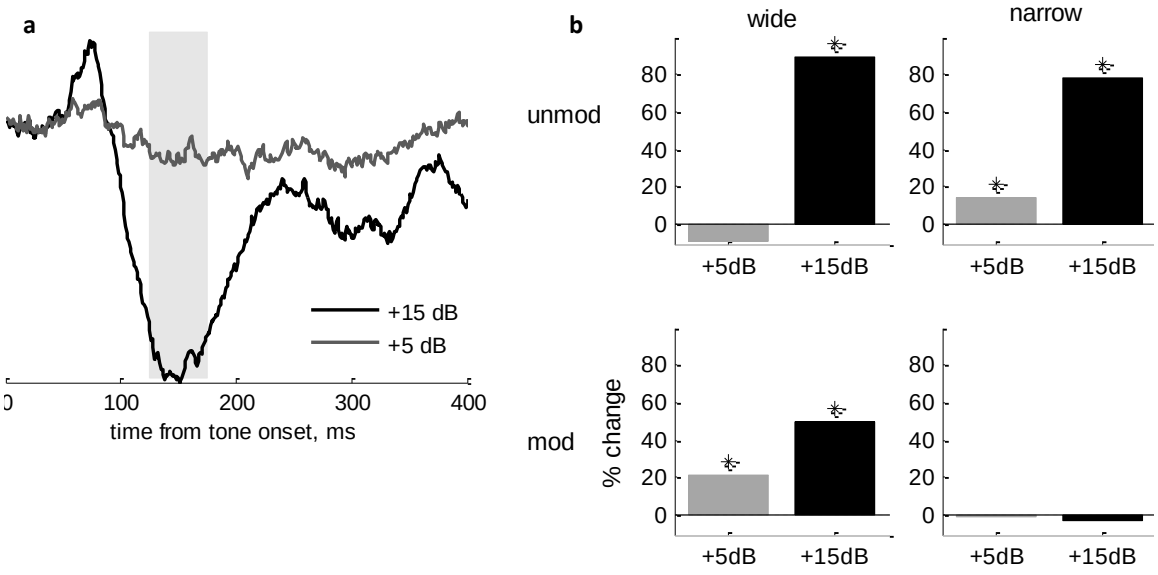

Figure 9. Tone-difference response dependence on masker. a. The temporal window of maximal response selected for analysis. The four different panels in Fig. 8 were averaged across all masking conditions (N=25 subjects*108 sensors*4 maskers), for each tone level separately. The temporal window chosen for analysis (**gray**) coincides with the peak of this average signal. **b.** The change (%) in mean peak magnitude when a tone was added in comparison to the values attained for the masker alone in the same time window for the +15 dB (**black**) and +5 dB (**gray**) tones. Asterisks mark significant increases in magnitude when the tone was added in comparison to the masker alone (3-way ANOVA on tone presence X subject X sensor position, main effect of tone presence, performed separately for each masker and tone level, $p<0.05$).

A statistical test of these results was performed by comparing the average signal in a relevant time window; namely, 125-175 ms after tone onset. This time window corresponded to the peak of the tone-difference response averaged across all conditions (see Fig. 9b). The results summarized in Fig. 9b confirmed the effects suggested by Fig. 8. For both unmodulated maskers, there was an increase of more than 75% in the response in this time window when the +15 dB tone was presented (3-way ANOVA on tone presence X subject X sensor position, main effect of tone presence, $p<<0.01$, $F(1, 5653)>300$, for both wideband and narrowband noise). However, when

the +5 dB tone was presented, the negativity was either absent (wideband noise, mean signal slightly less negative with tone) or much smaller (narrowband noise, mean increase in negativity with tone ~14%, $p<<0.01$, $F(1, 5653)=16$).

When the noise was modulated, results depended on bandwidth. For the narrowband masker no effect was found (3-way ANOVA as above, $p>0.19$): values in the presence of either tone level were indistinguishable in this time window from the noise alone condition. Conversely, for the wideband modulated masker, the negativity induced by the tone was significant for both the +15 dB and +5 dB tone (3-way ANOVA as above, $p<<0.01$ and $F(1, 5704)=85, 21$ for +15 dB and +5 dB tone respectively). The negativity locked to the higher level tone in the modulated wideband noise was smaller than in the unmodulated conditions (only a 50% mean increase in negativity). However, post hoc tests confirmed the effect of the lower-level tone was largest in the wideband modulated noise among all masker conditions (mean negativity ~22% larger with tone), and even significantly larger than the effect of the lower-level tone in the narrowband unmodulated noise.

Figure 10 depicts the topographical distribution of the tone-difference response magnitude. Whenever this signal appeared, it was maximal in the temporal electrodes. The wideband modulated masker showed significant right hemisphere dominance, but the other maskers (narrowband modulated, wideband unmodulated) showed no difference between hemispheres, or even a significant left hemisphere dominance (narrowband unmodulated). Data were insufficient for drawing conclusions from these distinctions.

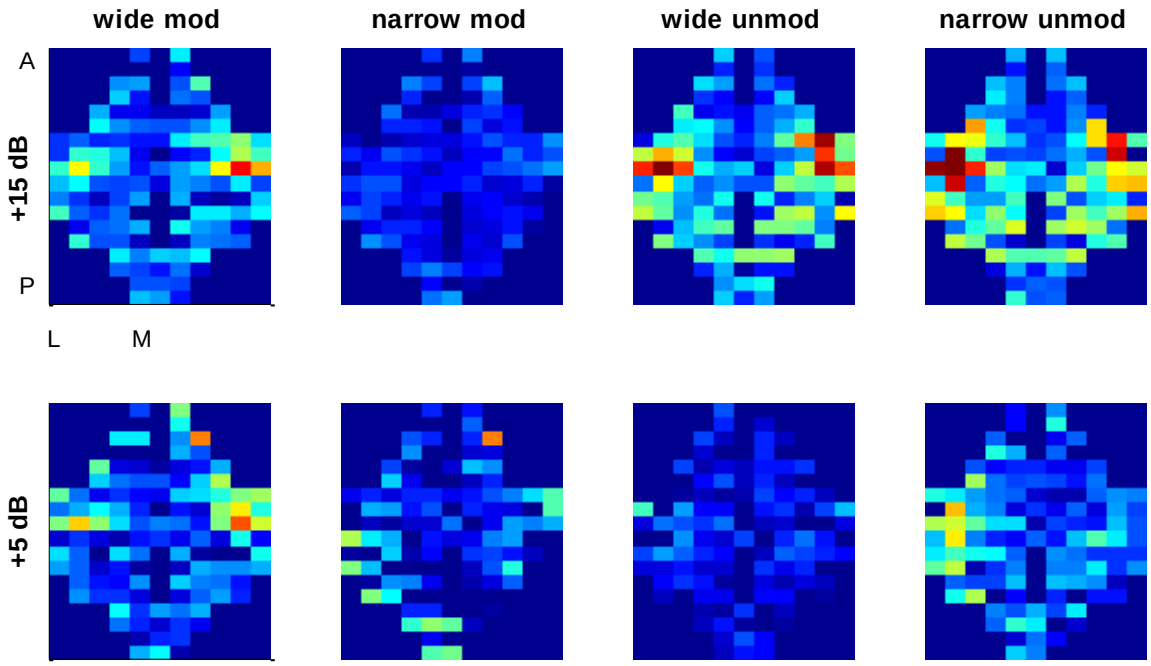

Figure 10. Temporal electrodes show maximal tone-difference response. Scalp distribution of the tone-difference response in all 4 noise conditions for the +15 dB tone (**top**) and the +5 db tone (**bottom**). Temporal foci were apparent in all responsive conditions. Color scale is identical for each row. Saturation at the top row was set to twice that of the bottom row to aid visualization.

The same general trends appeared in the sustained response to the tone (Gutschalk et al., 2004), but the relatively short duration of the tone (275 ms) made it difficult to separate the sustained response from the onset response. Similar response patterns occurred after the offset of the sound. Figures 11 and 12 show the results for the difference response in the 400 ms time window following tone onset: the wideband modulated masker exhibited highly significant negativity for both tone levels which was comparable to the unmodulated maskers for the higher tone level, and the largest among all conditions for the lower-level tone level. Notably, in contrast to the onset response, in this time window the addition of the +15 dB tone also had a significant effect on the response to the modulated narrowband masker (Fig. 12 a,b).

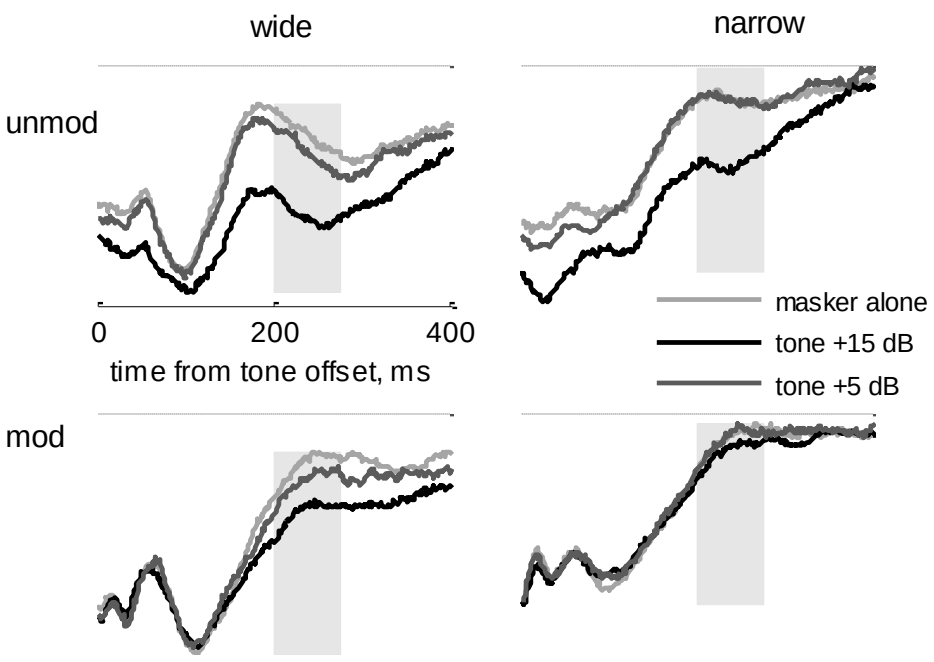

Figure 11. Offset fields. A magnified view of the magnetic fields following tone offset. Grand averages of the gradiometer responses for each of the four different maskers across all subjects (N=25) and all sensor positions (N=108) for noise alone (**gray**) and with a tone added at +5 dB (**dark gray**) and +15 dB (**black**). The gray area marks the time window used in the analysis. Scale is identical in all panels.

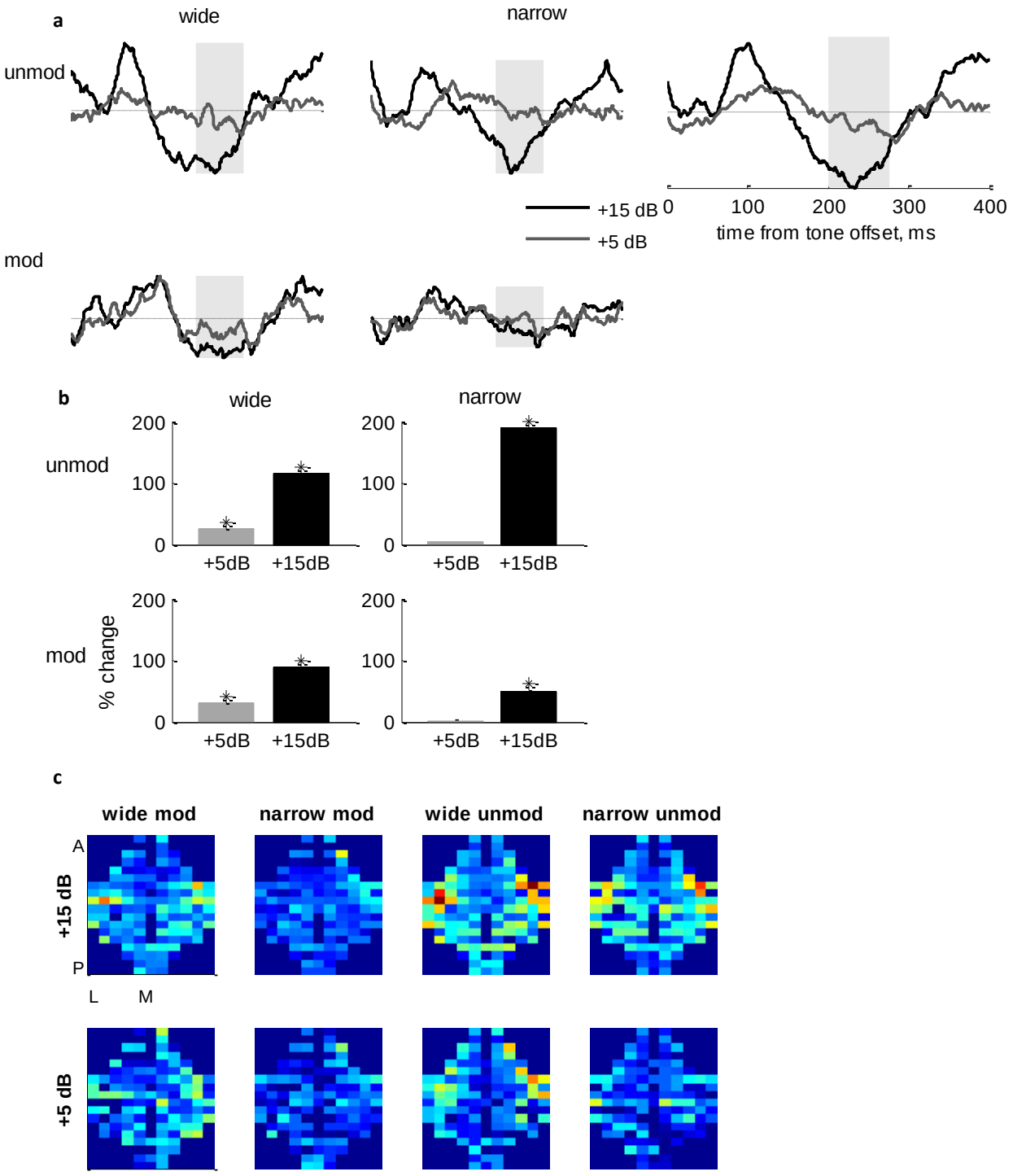

Figure 12. Offset difference response. a. The offset response averaged across all maskers (**right**) and for each masker variant (**left**) are shown along with the temporal window chosen for analysis (**gray**). Linear trends were subtracted. **b.** The change (%) in mean peak amplitude in relation to the noise alone condition is shown for both tone levels. **c.** Scalp distribution of the signal mean peak amplitude for all maskers for both tone levels. Color scale is identical for each row, and saturation for the top row was set to twice that of the bottom row. See Figures 8-10.

An 'object potential'?

Psychophysically, the CMR measured in the current study was comparable to earlier reports (Hall et al., 1984); i.e., the thresholds for tone detection decreased considerably when the masker was modulated (modulated-unmodulated difference) but also with increased masker bandwidth (true CMR).

The auditory evoked magnetic fields recorded from the same subjects in the different conditions showed two main effects. First, 'locking' - in response to modulated noise, the gradiometer signal followed the amplitude modulations of both the narrowband and the wideband maskers. The initial response to each modulation cycle of the noise was marked by a N_am component – a negativity locked to cycle onset with a peak latency of ~30ms. Importantly, when a tone was added to the modulated noise, this response was suppressed both for a tone 15 dB above masked threshold and for a tone only 5 dB above masked threshold. The temporal dynamics of the suppression mirrored to some extent the electrophysiological findings in single units from cat auditory cortex in that the maximal reduction in the magnitude of the N_am occurred only at the 2^{nd} noise cycle after tone onset. Nevertheless, there were small effects as well on the N_am to the 1^{st} noise cycle after tone onset. These small effects might have been related to the response to the tone that occurred in the same time window, or could represent a genuine small modification of the N_am response to the noise. The current experimental design could not distinguish between these two options.

In addition to the locking suppression of the early responses, a much larger late negativity following tone onset was identified. This N_1m-like response component had a longer latency (~75 ms) than the Nam and was long-lasting (~175 ms). For the higher tone level it appeared for modulated wideband noise and unmodulated noise as well. For the lower-level tone, it was largest in the wideband modulated noise condition, and therefore was consistent with the psychophysical preference for tones over wideband modulated maskers.

At the offset of the maskers and tones, the effect of tones generally mirrored the effects at tone onset, with one major difference. At the offset, the addition of the higher level tone to the modulated narrowband noise had a significant effect (compare Fig. 9b and Fig. 12b). Previous studies have shown that the onset, sustained and offset responses of cortical neurons represent different features of sounds, including tone frequency and intensity (Takahashi et al., 2004, Wang et al., 2005), spatial location (Campbell et al., 2010), and vocalization identity (Qin et al., 2008). Interestingly, in the ferret auditory cortex the offset responses of A1 neurons were shown to be more informative about timbre than those of neurons in other fields (Walker et al., 2011). In contrast to locking suppression, the corresponded to findings from the cat, the two components detected at tone onset and tone offset, did not have a correlate in the neural data from cat auditory cortex (Las et al. 2005, Rupp et al. 2007).

Notably, the response to tone onset that was significant for unmodulated maskers as well as for modulated wideband noise, was entirely missing in the narrowband modulated masker. This was true for both the higher- and the lower-level tones. The tone levels used in the experiment were normalized relative to the psychoacoustic masked threshold for each masker condition separately. Behaviorally, the ability to detect that a change has occurred when the tone was added was thus equalized across maskers. This suggests that this neural response did not correlate with tone detection. The account of the subjects themselves regarding the perceptual experience indicated an alternative interpretation. The tone in the narrowband modulated masker seemed to have a different perceptual status than the tone in the other maskers. The addition of the tone in this condition was marked by a qualitative change in the timbre and pitch of a sound already present (the masker). In contrast, when the tone was added to the other maskers (the unmodulated maskers/ the wideband modulated masker) it introduced a perceptual quantitative change – the tone appeared as a new separate entity and the resulting percept was that of two auditory objects (noise + tone) as opposed to one auditory object (the masker alone). I therefore suggest that the difference response illustrated in Figs 8, 9 and 10 is the magnetic signature of the

emergence of a new *object* (here, the tone) in the auditory scene. When the onset of the tone is marked by perceptual *qualitative* change in auditory objects that are already present (e.g. a change of timbre), and not the addition of a new object, this difference response, locked to the tone onset, does not emerge.

IMPLEMENTING THE PREDICTIVE FRAMEWORK

A psychoacoustic application

**Joint work with Daniel Pressnitzer, Trevor Agus and Clara Suied, CNRS & Laboratoire de Psychologie de la Perception & École Normale Supérieure, Paris, France.*

Voices and Instruments

Recognition of timbre is considered to be a computationally intricate problem. The American National Standards Institute (ANSI) definition of timbre describes it as "the attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar." Timbre potentially has a large number of different relevant physical dimensions. The timbre of the human voice is obviously of major importance to human subjects, and a number of studies have shown specific sensitivity of brain responses to the human voice (e.g. Belin et al., 2000, Levy et al., 2003, Murray et al., 2006). In this chapter, I apply the predictive framework to the problem of recognition of human voices among many comparable sounds emitted by musical instruments. This work was prompted by experiments conducted by Agus and Suied (Agus et al., 2010, 2012). Agus et al. used reaction-time and gating techniques with a set of natural sounds that included voices to map listener performance on a sound recognition task.

The set of natural sounds consisted of 156 recordings of single notes having 13 timbres, including both musical instruments and the human voice (sung vowels). For each timbre, notes at all 12 semitones between A3 and G#4 were included. In the experimental design, the sounds were divided into a distractor group (that comprised piano and six

wind instruments: bassoon, clarinet, oboe, saxophone, trumpet, and trombone) and three target categories:

1. voice - male voice singing vowels /a/ or /i/
2. percussion - marimba and vibraphone
3. strings - violin and cello

Sounds were truncated to a duration of 250 ms. Loudness and duration were equated across categories (for more details, see Agus et al., 2012). Individual sounds had significant differences, particularly at sound offsets, as illustrated in Fig. 1. For the analyses I performed on this set, the sounds were downsampled from a sampling rate of 44.1 KHz to a sampling rate of 8.82 KHz.

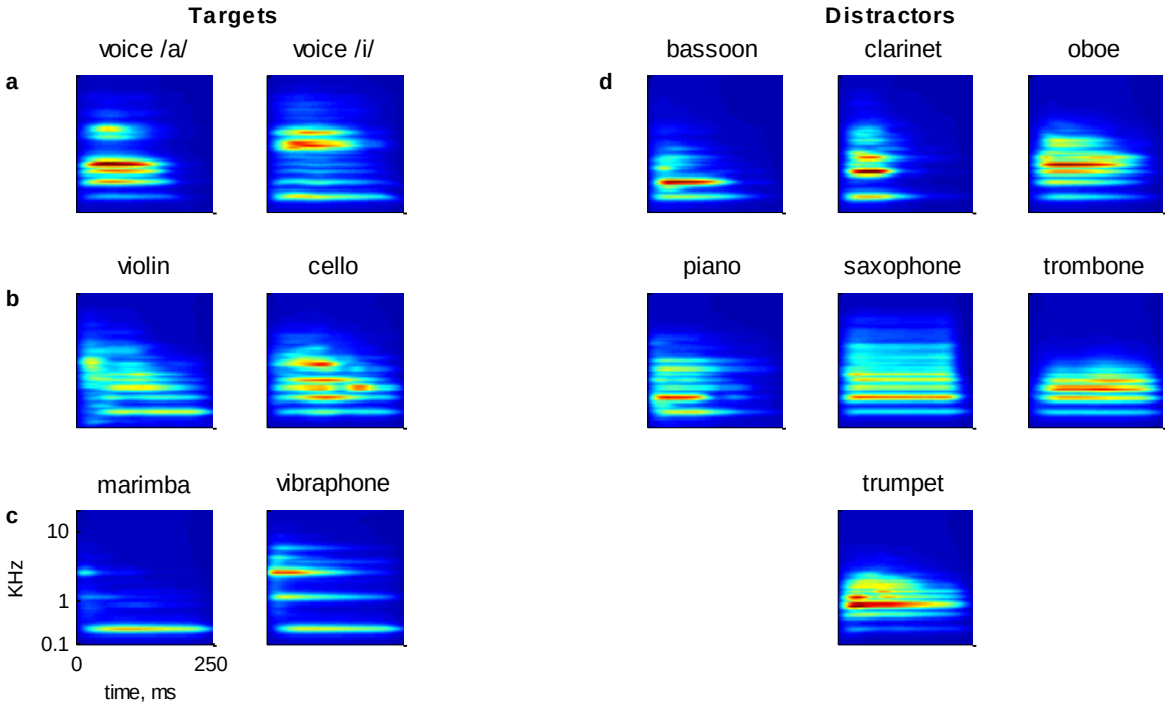

Figure 1. The sounds. A representative sample of the stimuli used in the psychoacoustic experiment. The three target categories on the left (**a**: voice, **b**: strings, **c**: percussion) were each presented interspersed with the same set of distractors presented on the right (**d**). Each panel shows a spectro-temporal excitation pattern (STEP, see Moore, 2003) obtained by simulating cochlear filtering using the Gammatone filter-bank from 100Hz to 20KHz implemented by the AIM software package (Bleeck et al., 2004). All sound types are represented by the pitch D4. In the experiments, pitches from A3 to G#4 were used for each sound type.

Psychoacoustical results

All psychoacoustic experiments were carried out by Trevor Agus and Clara Suied, the lab of Daniel Pressnitzer, CNRS & Laboratoire de Psychologie de la Perception & École Normale Supérieure, Paris, France

Experiment 1

Listeners performed a go/no-go task. On each trial a single sound was presented: either a target (voice, percussion or strings, Fig. 1 a-c) or a distractor from a different group of musical instruments (Fig. 1d). Listeners were asked to respond to targets as quickly as possible and ignore interspersed non-target, "distractor" sounds. No feedback was provided. As a control, listeners also performed a simple 'go' task in which only targets were presented and listeners were asked to respond as quickly as possible on every trial.

Subjects had near-perfect accuracy for all categories, with voices showing the lowest error rate and strings the highest (Fig. 2). Reaction times (RTs) mirrored this order: voices were identified with shortest RT, percussion instruments ranked second, and the string instruments had the longest RTs. The most extreme difference in RT (voice vs. string) was 105 ms, which is a large effect for RT paradigms (Agus et al., 2012). In the control experiment, where subjects reacted to all sounds presented, voices were detected marginally slower than percussion or strings. Thus, an explanation that voices were simply sounds that were detected faster could be rejected.

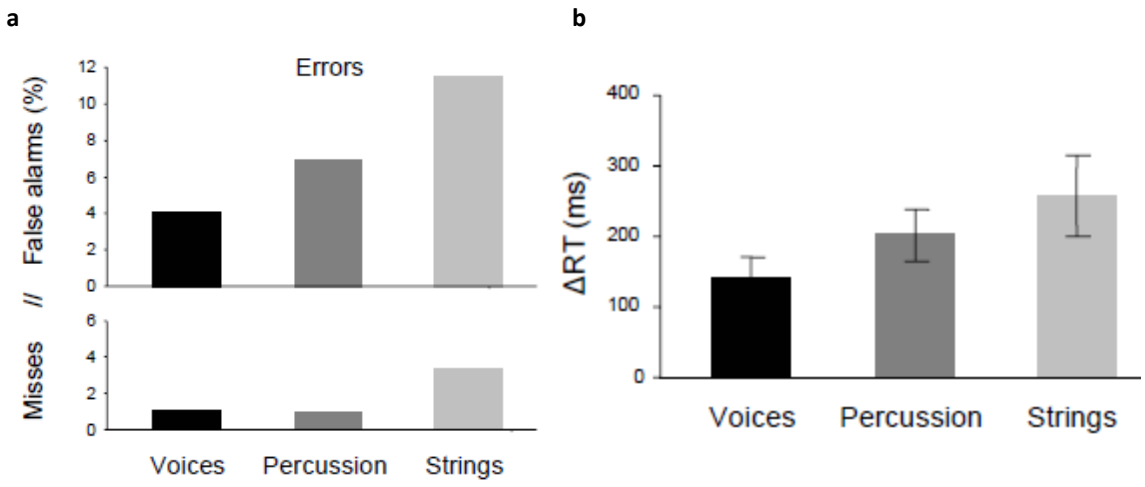

Figure 2. Psychoacoustical performance. a. Average false-alarm and miss rates for each target category. **b.** ΔRT (reaction times) calculated by subtracting RTs in the sound detection task ('go' task) from the RTs in the category recognition task ('go-no go' task) for each target category. Error bars show 95% confidence intervals centered on the mean. Voices were recognized faster and more accurately than other target categories (adapted from Agus et al., 2012).

Experiment 2

The second experiment (Agus et al., 2010) tested how short a sound can be and still support recognition. The sounds used in Experiment 1 were cut into short segments with durations of 2, 4, 8, 16, 32, 64, or 128 ms, selected randomly from the initial 100 ms of the sounds. Each segment was multiplied by a Hanning window with the corresponding duration. In each trial, listeners heard a short sound and had to indicate whether it was part of the target category or not. Target categories were presented in different blocks. Distractors were the same on all blocks. Feedback was provided.

Data were analyzed using the d' sensitivity index employed in signal detection theory (Wickens 2001). A high d' represents reliable recognition of the target category, whereas a d' of 0 indicates a performance not better than chance. Figure 3 depicts the average d' as a function of the sound duration. Overall, the voices were most accurately recognized, and the strings least accurately recognized ($p < 0.001$). Reducing the length

of the sound segments reduced detectability for all target categories. At 16 ms, all of the categories were recognized significantly above chance (i.e., d' > 1). At 8 ms, d' was still low for all categories, with d'>1 only for voices. At 4 ms, only voices had a d' significantly larger than 0 ($p < 0.005$). For all of the target categories, detection was at chance level at 2 ms.

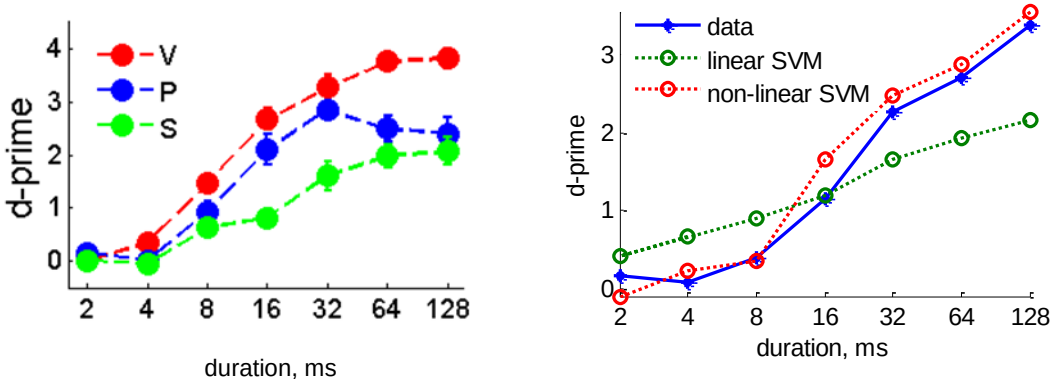

Figure 3. Recognition as a function of duration. a. Average performance (measured by d') across all subjects (n=9) is plotted for each target category (**red**: voice, **blue**: percussion, **green**: strings) as a function of the duration of the sound segments. Errorbars indicate standard errors. Adapted from Agus et al. (2010). **b.** Classifier performance (dotted lines) vs. behavioral data (continuous line), adapted from Suied et al. (2010). A linear SVM on the output of the auditory spectrogram (**green**) did not coincide with behavior. A non-linear SVM (**red**) on the auditory spectrogram fit the data much better. Training was done on 128 ms segments of sounds, testing on all sound durations.

Importantly, Agus et al. claimed that they failed to find any acoustical differences that correlated with the improved detection of the human voice. They used Support Vector Machines (SVM) on the output of the auditory spectrogram and found that linear SVMs did not perform similarly to behavior, being too good at short durations (<=8 ms) and not good enough at long durations (>=32 ms). Using non-linear SVMs resulted in classification that better fit the behavior (Fig. 3b, see Suied et al., 2010). The use of

nonlinear machines, however, leaves open the issue of the physical parameters underlying the classifier performance.

Predictive modeling of the sound set

The experimental sound set was used to test the application of the predictive framework for ASA in an experimental scenario. The rationale was to test the products of predictive modeling of the sounds against the psychoacoustic experimental results. The basic scheme was to treat the sounds used in the different experimental blocks as a combined sound set that constitutes a distinct relevant sound world to the subjects. Predictive models were then generated to fit the average statistics of the entire sound set and the success of the models in predicting the evolution of the sounds in time was compared between the sound categories.

The application of the Gaussian Information Bottleneck for predictive modeling of continuous signals was described in chapter 2: T, an optimal predictive representation of X, the past, is obtained by projecting the past on the eigenvectors of the normalized regression matrix, calculated from the sound's autocovariance matrix. The value of the future samples, Y, is then obtained by another linear transformation on T. The procedure described there was implemented to generate predictive models for the acoustic signals in the current sound set, with the following choices:

1. One common sound world:

   The sound ensemble selected for model construction consisted of all sounds used in the different experimental blocks. This way, the prediction from a specific sound segment was formed based on the regularities extracted from the entire sound set. The autocovariance used for model construction was therefore the average autocovariance of all 156 sounds in the sound set (13 timbres x 12 pitches).

2. The future (Y) consisted of one sample:

The goal of the predictive model was set to project one sample into the future (~0.11 ms at a 8.82 KHz sampling rate) based on the recent history of the signal at each step. The optimal representation T was therefore a time-varying, scalar quantity. According to the Gaussian IB solution, T is calculated by projecting the past samples on the left eigenvectors of $\Sigma xy(\Sigma yy^{-1})\Sigma yx(\Sigma xx^{-1})$ that have non-zero eigenvalues. In the case of projecting 1 sample into the future this matrix is of rank 1 and the single eigenvector with non zero eigenvalue is the vector $\Sigma xx^{-1}\Sigma xy$ (see chapter 2). The Gaussian IB solution is therefore identical in this case to the solution of the linear regression of the next sound sample on n preceding samples.

3. The length of the past (X)

   The number of past samples used to project one sample ahead ('model length') varied between 3 samples (~0.35 ms) and 320 samples (~36 ms)

Prediction evaluation

In order to evaluate the quality of prediction the Pearson correlation coefficient was applied between a sequence of predictions and the original waveform (see illustration in Fig. 4). Importantly, the predictions were made one sample a time based on the immediately preceding samples of the sound. As a consequence whole future segments used for prediction evaluation were not predicted from a fixed past. In order to simulate an on-going process of monitoring prediction quality, the correlations were computed on short, overlapping segments. The length of these segments - '*test length*' – ranged from 3 samples (~0.35 ms) to 320 samples (~36 ms). For each choice of model length and test length, 60 time points were randomly selected from the first 125 ms of the sound. This choice of time points was then used to extract segments from all sounds in the set. This procedure corresponded to the experimental procedure of the second psychoacoustic experiment and helped to avoid the irregularities during sound offset. Identical results were obtained with uniform sampling of the sounds instead of the random sampling.

Results

Generally, the task of predicting one sample ahead was relatively easy. This was true even when the model length was short, and prediction was only based on a few samples. Figure 4 illustrates the procedure with a 9-point (~1ms) model. Predictions were generated with this model for two sounds (voice /a/ singing G#4 and the piano playing D#4). The predicted traces resembled the original waveforms visually, and the correlations were high (see Fig. 4e, r=077±0.26, mean ± std for the entire sound set with this 1ms model).

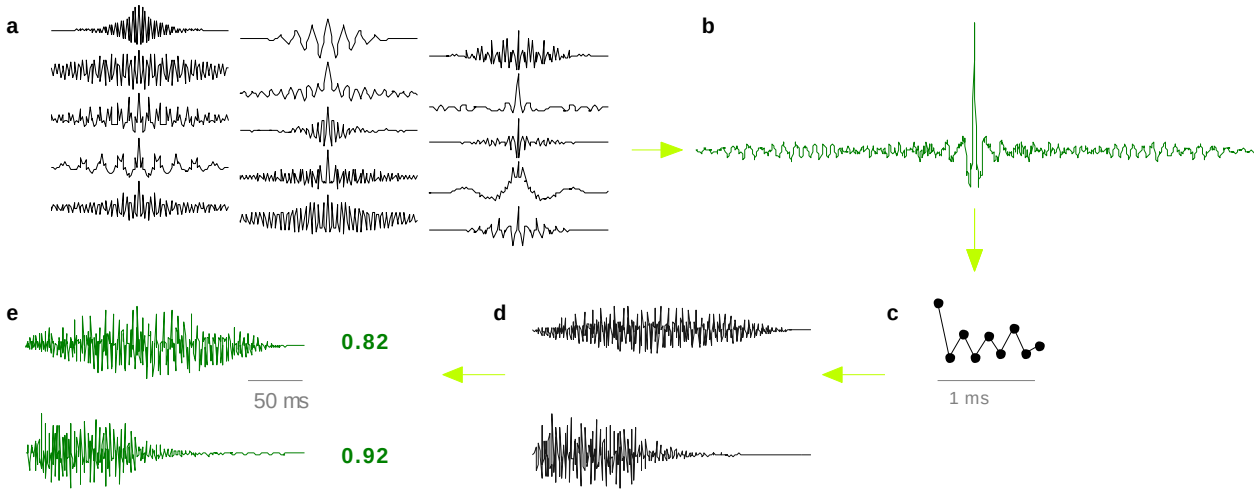

Figure 4. Implementing the Gaussian IB. Illustration of the elements in the implementation: The autocovariances of all individual sounds (**a,** represented here are 12 autocovariances of individual sounds selected randomly from the sound ensemble) were averaged across the entire sound set (**b**) and one predictive model was generated with model length= 9 samples, ~1 ms (**c**). The model was then convolved with individual sounds (**d**) to produce sequences of predictions (**e**). The correlations between predictions and entire sounds are shown in green (test length = 2205 samples, 250 ms). The sound waves in d and e were selected randomly, and are of voice /a/ (G#4) and piano (D#4).

The correlations shown in Fig. 4 were calculated for the entire sound duration, i.e. test length of 250 ms. However, it is more realistic to assume that the brain also evaluates

predictions over much shorter periods. For example, the reaction times for target identification (ΔRT) reported above were shorter than 250 ms. Therefore, I calculated correlations with much shorter sound segments, with test lengths ranging between ~0.35-35 ms (see above). Figure 5a illustrates the results obtained for one choice of parameters: a model length of 63 points (~7.1 ms) and a test length of 21 points (~2.4 ms). The average correlation per sound category was calculated using segments randomly drawn from each sound (see above) and across all timbres in each category. Although correlations were generally very high (for this choice of parameters the average correlation across all segments was 0.83 ± 0.17, mean ± std) there were significant differences between the different categories of sounds (3 way ANOVA for category X note X time point, with time points as a random factor. Main effect for category $F(3,9286)=122$, $p<<0.01$). Post-hoc comparisons showed that correlations were the lowest for the voice sounds, percussions ranked second and the string instruments had the highest correlations. This order matched the psychoacoustical results (compare to Fig. 2). The highest correlations were obtained for the sounds from the distractor group that consisted of the majority of sounds used for model construction (7/13 instruments, mostly winds, see methods). All these differences were highly significant in post-hoc tests.

This result was not unique to a specific choice of model and test durations. The average correlation between predictions and the actual sound samples was calculated for all the sounds in the ensemble using model and test durations ranging from ~0.3 to 35 ms. For all models longer than ~8 ms and test lengths longer than ~3 ms the correlations of the voice segments were lower on average than all the other instrument categories. The same result was obtained when prediction quality was quantified using the mean square error between the predicted trace and the actual sound segment instead of correlations.

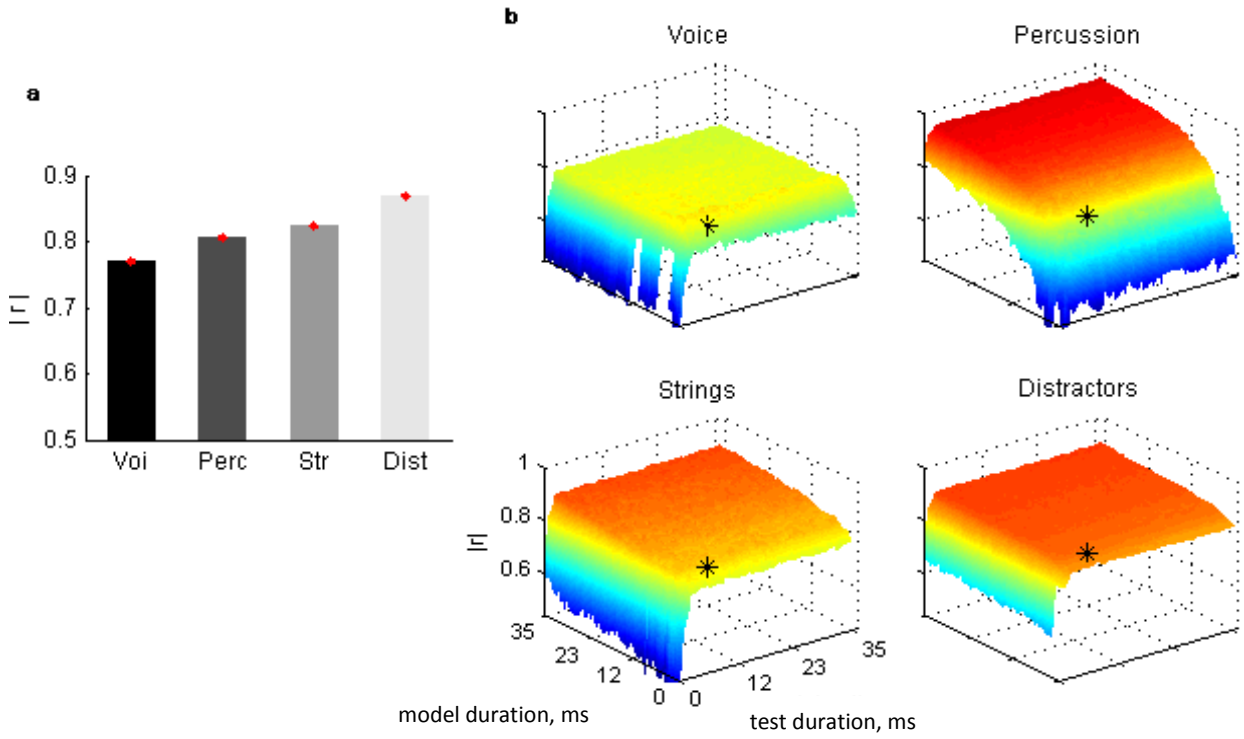

Figure 5. Quality of prediction by category. a. Correlations between predictions and waveforms by category, averaged across segments and instruments for a specific choice of parameters: model length of 63 points (~7.1 ms) and test length of 21 points (~2.4 ms). Error bars specify negligible sem, due to the large sample. **b.** Mean correlation between prediction and waveforms as a function of model duration (abscissa) and test duration (ordinate) for the entire parameter space sampled (model and test durations between ~0.3-35 ms). The z axis and the color axis denote the average correlation between sequences of predictions and the original waveforms (range 0.45 – 0.9). In the target categories, each point is the average of 1560 values (60 segments X 12 pitches X 2 timbres in each of the target categories). In the distractor group, each point is the average of 5460 values (60 segments X 12 pitches X 7 timbres). Asterisks mark the model from **a**.

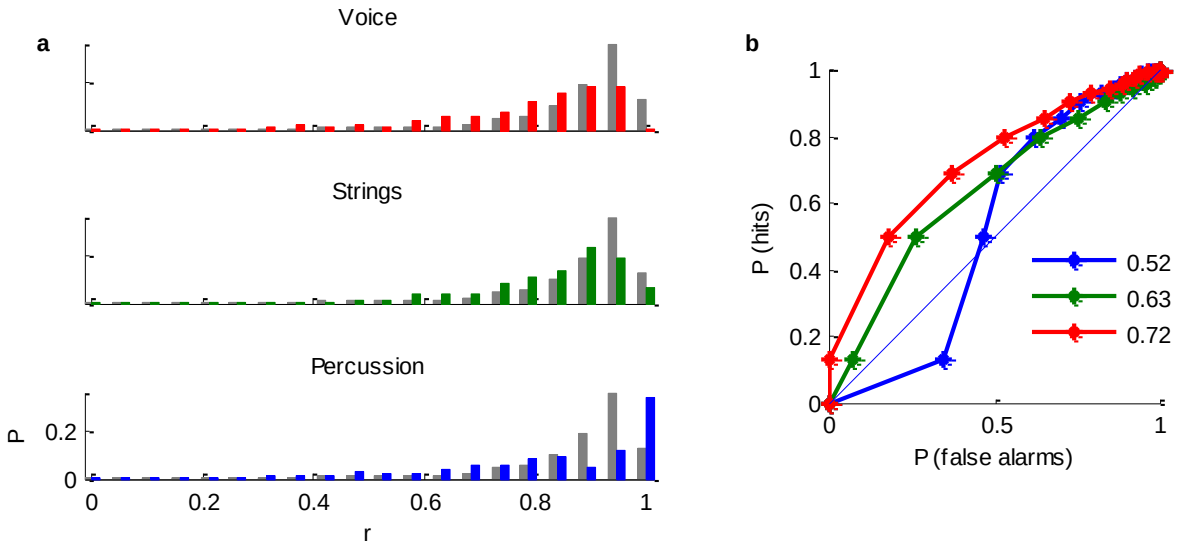

Figure 6. Target identification by correlation values. a. Distributions of correlations for the distractor group (**gray**) and for each of the target categories (**red** – voice, **green** – strings, **blue** – percussion) **b.** ROC curves generated from the distributions in **a** (same color code). Classification was done here by a threshold on the correlation value, which is suboptimal when likelihoods are not monotonic (see for example the percussion concave curve). Numbers indicate the area under the curve, an estimate of performance in a discrimination test using a classification criterion on the value.

Link to behavior - ROC analysis

Receiver operating characteristic (ROC) analysis was applied to test whether the differences in the distributions could support classification of the type performed in the psychoacoustic experiments. The distribution of correlations for the distractor group was compared to the distribution of each of the target categories to test whether the differences in prediction errors was consistent with psychoacoustic performance. Classification performance was quantified by the area under the ROC curve, estimating the probability of correct decision in a two-interval two-alternative forced choice test (see Figs. 6 and 9). Discrimination threshold was set at 70.7% corresponding to the threshold typically used in auditory psychophysics (see Chapter 2).

Figure 6 illustrates the ROC procedure for a model length of 63 points (~7.1 ms) and a test length of 21 points (~2.4 ms). These are the same parameters used in Fig. 5a. The classification criterion was imposed on the correlation values, and not on the likelihood (see Chapter 2 for more details). Discrimination was better for the voice category: 72%, compared to 63% for the strings and 52% for percussion. The ROC analysis was repeated with the classification of the correlation likelihood this time, as described in chapter 2. The performance for voice and strings categories did not change. However, performance for the percussion category reached a high of 74%, even better than the voice category. This is probably due to the pronounced bimodality of the distribution in this category that led to a nonmonotonous likelihood (see Chapter 2).

The analysis was repeated for the entire paramter space of model and test durations (Fig. 7). For the vast majority of parameters tested, the performance level for voices was above threshold level, regardless of choice of classification criterion: correlation value or likelihood. Performance for the strings, on the other hand, did not reach threshold for any choice of parameters, for either type of classification. For the percussion category, results were affected by the classification criterion. However, even the higher performance levels obtained using criteria on the likelihood only crossed the 70.0% discrimination threshold for short (<15ms) model durations.

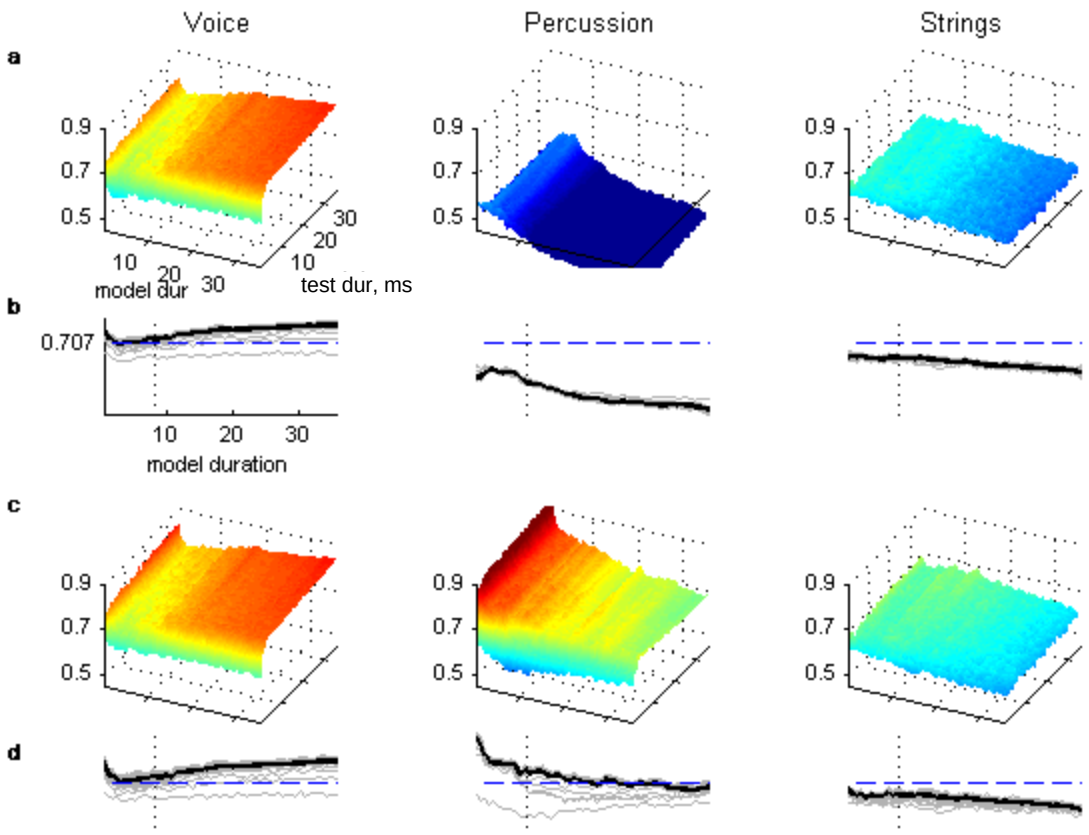

Figure 7. Performance in a classification task by target category. a. For each target category, the probability of discriminating between a target sound and a distractor was plotted as a function of model duration and test duration (in ms) for the tested parameter space (~0.35-35 ms). The classification criterion was imposed here on the correlation values, and not on the likelihood. The z axis and color axis denote the probability of correct discrimination (range 0.45 – 0.9). In this analysis the first 13 ms of the sounds were excluded to avoid the irregular spectral content of sound onset. Identical results were obtained when onsets were included. **b.** Since test duration had only slight effects on performance, the same results shown in **a** are plotted as a function of model duration. Results for each test duration (**gray lines**) are shown together with the average across all test durations (**black line**). Dashed blue lines mark the threshold at 0.707 and dotted lines mark model duration of 8 ms. **c** and **d** are the same as **a** and **b** respectively, when the discrimination was done with the criteria imposed on the likelihood. Classification performance was markedly better for the percussion category when using likelihoods.

The spectral interpretation

The correlation coefficient between a prediction and actual signal, used for evaluating the quality of prediction, had a simple interpretation in the spectral domain. As the prediction is the result of convolving a sound segment with the predictive model, the correlation in the temporal domain is approximately the correlation coefficient between the following two terms in the spectral domain: (1) The Fourier transform of the signal segment being predicted, and (2) The transfer function of the predictive model multiplied by the Fourier transform of the signal segment that was used for prediction generation (in the case of 1 – sample ahead prediction, this is a sound segment that starts 1 sample before the predicted one). This follows directly from applying the Parseval and convolution theorems:

$$r(pred_n, sig_{n+1})$$

$$\stackrel{\text{def}}{=} \frac{\sum_{i=1}^{t} p_i {s^{n+1}}_i}{\sqrt{\sum_{i=1}^{t}({p_i}^2)\sum_{i=1}^{t}\left({s^{n+1}}_i{}^2\right)}} \stackrel{Parseval}{=} \frac{\sum_{k=1}^{t} P_k \overline{{S^{n+1}}_k}}{\sqrt{\sum_{k=1}^{t}(|P_k|^2)\sum_{k=1}^{t}(|{S^{n+1}}_k|^2)}} \stackrel{convolution}{=}$$

$$\frac{\sum_{k=1}^{t} {S^n}_k H_k \overline{{S^{n+1}}_k}}{\sqrt{\sum_{k=1}^{t}(|{S^n}_k H_k|^2)\sum_{k=1}^{t}(|{S^{n+1}}_k|^2)}} \stackrel{\text{def}}{=} r(\mathcal{F}(sig_n)\cdot\mathcal{F}(h), \mathcal{F}(sig_{n+1}))$$

where p denotes the sequence of predictions, s_n, the sound segment used for generating the prediction, s_{n+1} the predicted sound segment of length t, and h the impulse response of the predictive model. The subscripts *i* and *k* indicate summation in the time and the frequency domains, respectively.

Total sound energy was equated by the experimenters. The difference in the quality of predictions thus suggested a difference between the spectral profile of the filter and the spectral profiles of the sounds in the different categories (Fig. 8).

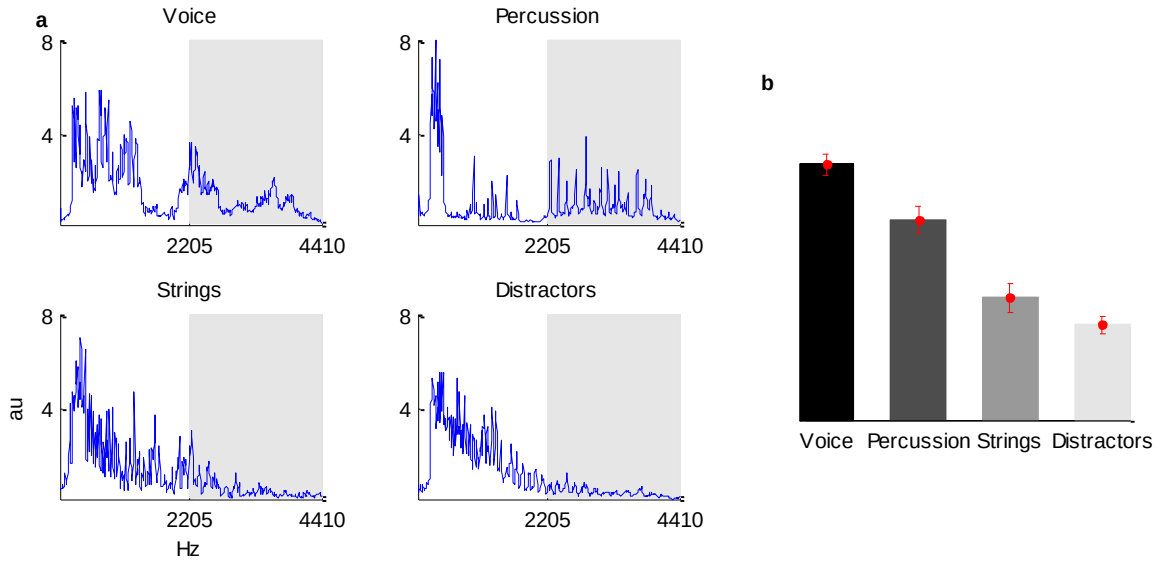

Figure 8. Spectral content by target category. a. The mean spectral content in arbitrary unite for each of the sound categories. Gray areas mark the 2.2-4.41 kHz range. **b.** Total spectral power between 2.2 and 4.41 kHz per category (mean ± sem marked in red). Compare order with that in Fig. 2b and Fig. 5a.

One obvious difference between the sound categories is marked by the shaded area: the spectral power between 2.2 and 4.41 kHz was significantly higher in the voice category than any other category, with percussion second, then strings and finally the distractor category. This order matched both the psychoacoustics and the prediction quality results (Fig. 2b and Fig. 5a). The extra energy in this spectral band found in the voice category was probably due to the higher formants of the sung vowels (the 2nd formant of the /i/ and the 3rd format of the /a/ both fall in this range).

The spectral profiles in Fig. 8 were calculated from the entire duration of the sounds (250 ms). The next step was to test whether the spectral differences in the same frequency range can support classification even when segments with a duration as short as 8 ms were used. The segment sampling procedure used for calculating the prediction error was repeated: sixty segments of 8 ms were randomly sampled from the first 125 ms of each of the sounds. In this analysis the first 13 ms of the sounds were excluded to avoid the irregular spectral content of sound onset. The exclusion of the onset only

affected results for the percussions, where the differences between the onset and on-going portions of the sounds were the largest. Using short sound segments from the onset portions of the percussions also affected psychoacoustic judgments in this experiment (see Agus et al. 2010).

The mean spectral power between 2.2 and 4.4 kHz was calculated for each segment, and the distributions estimated from the distractor group and from each of the target categories were compared. An ROC analysis was then carried out as above for these distributions (Fig. 9). Performance was substantially better for the voice category, which was the only category to pass the 70.7% discrimination threshold. Identical results were obtained when waveforms were first filtered using filters that mimicked the peripheral filtering of the auditory system (results not shown).

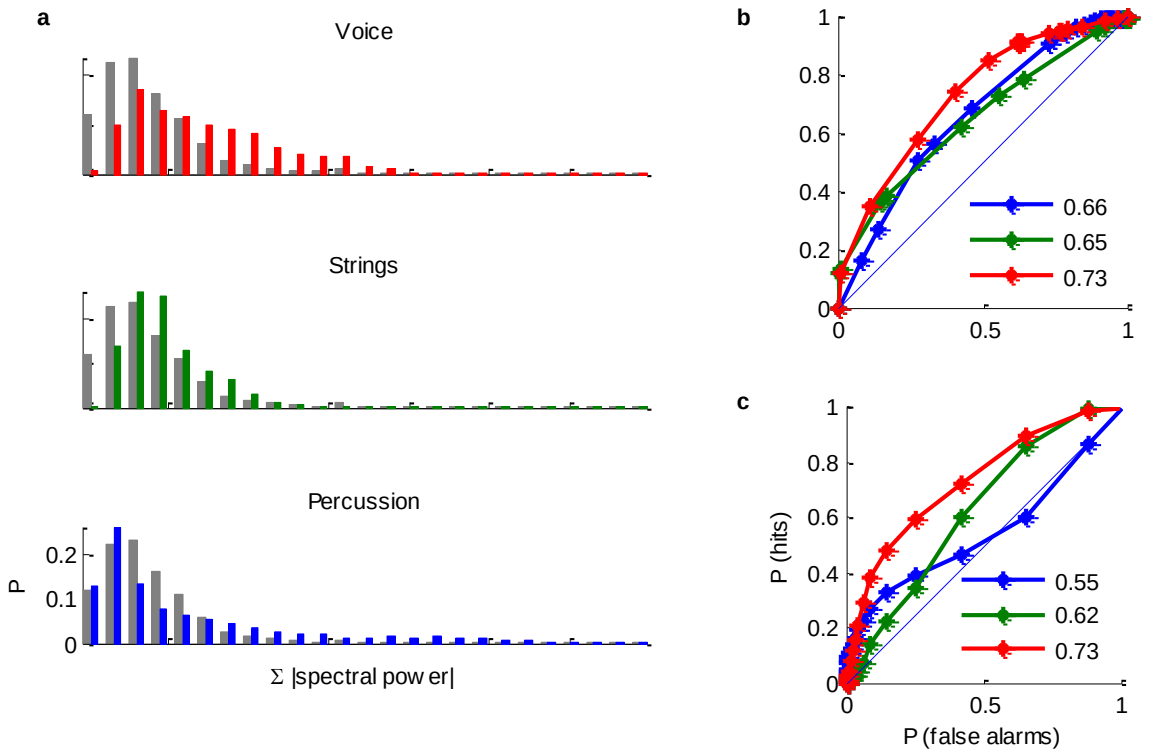

Figure 9. Target identification by spectral power. a. Distributions of the average spectral power between 2.2 and 4.41 kHz in 8 ms segments randomly drawn from the sounds. The distribution of the distractor group (gray) is shown with the distribution of each of the target categories. **Right:** ROC curves generated from the distributions in **a** (same color code). Classification criteria were imposed on the likelihood (**b**) or the spectral power value (**c**). The concave curve for the percussion category demonstrates the sub-optimality of criteria on the value. Numerical values indicate the area under the corresponding curves. Results for the percussion category were affected by the way the criterion was applied.

Conclusion

The results of implementing the predictive framework with an experimental sound set show that the framework is applicable to experimental scenarios, and that its outputs are relevant to explaining behavioral results. Performance of predictive models generated for the entire sound set mirrored the behavioral results from Agus et al. (2010, 2012). Prediction errors were highest on average for voice segments, with percussions next and strings last, with differences appearing already at short durations

like those tested behaviorally. The prediction error thus turned out to be a bridge between the abstract framework and the experimental observations. The spectral interpretation of the prediction error further suggests that the predictive framework can be implemented with spectral quantities that natural to the auditory system and are already available at the auditory periphery.

Is this finding trivial? It could be the case that the predictive framework worked because of the specific spectral differences between the voices and the other categories; i.e. the higher formants of the sung vowels. However it is likely that other spectral differences between the sounds would have been identified as well. The implementation of the predictive framework can thus be viewed as a principled method for highlighting differences between an average sound model and specific sound sets.

The assumptions underlying this implementation are obviously not met exactly by this sound set. For example, the sound waveforms rather than their frequency components were used, although the later would be more natural to the auditory system. The limited projection (one sample) into the future at each step is probably another choice that should be tested against other choices that might result in better fit with the data. Nevertheless the general abstract formulation was shown to be applicable in an experimental setup, and can account for experimental observations on a rather sophisticated behavior with real-world signals.

PREDICTION IN MUSIC

Evidence from extracellular recordings

** Amit Yaron, Itai Hershenhoren and Ayelet-Hashahar Shapira, from the lab of Israel Nelken, the Hebrew University of Jerusalem, conducted the experiments described below.*

Music as a stimulus - Ligeti's Musica Ricercata II

The organization of sounds in time and sound space is essential to music. Current music theories (e.g. Meyer, 1956, Huron, 2006) argue that rule-based expectations and their violations contribute to the effect of music, thus making music a good test case for the predictive modeling of sounds. Does the organization of sounds in time modulate brain responses, in addition to the effects of the short-term spectro-temporal content on the responses? To test this we recorded extracellular neuronal responses in rat auditory cortex during the presentation of a monophonic piano piece by Ligeti entitled *Musica Ricercata II* that has a simple structure, with obvious regularities as well as rule violations.

György Ligeti composed *Musica Ricercata,* a set of 11 short pieces for piano, between 1951 and 1953. The first piece, *Musica Ricercata I,* is confined to two notes from the chromatic scale (the second only appears as the final note). Each subsequent piece has exactly one more pitch class, so that the final piece of the cycle contains all 12 notes of the chromatic scale. Ligeti described his own work as follows: *"I started to experiment with simple structure of rhythm and sounds, in order to evolve a new music from nothing, so to speak. I regarded all the music I had known and loved up to then as*

something I couldn't use. I asked myself: what can I do with a single note: what can I do with the octave, or with interval, or two intervals, or a specific rhythmic situation."

Musica Ricercata II can thus be considered an experiment with a minimalistic pitch structure. It is composed of three chromatic neighbors, E♯, F♯, and G, that are either played solo (predominantly on a single octave, in one occasion on two octaves) or in unison on four octaves. The main theme is a plaintive alternation between E♯ and F♯ (see Fig. 1). Near the middle of the piece, after the theme has been repeated four times with small variations, the second theme - a G - is introduced with a vigorous, loud accelerando. The two motifs are then played loudly with the Gs either interfering with the main motif or playing in an unmetered soft tremolo along with it. A highly compressed version of the first theme follows, and then both motifs are used in the coda, gradually fading into silence. This tight minimal pitch structure allows for a relatively controlled experiment.

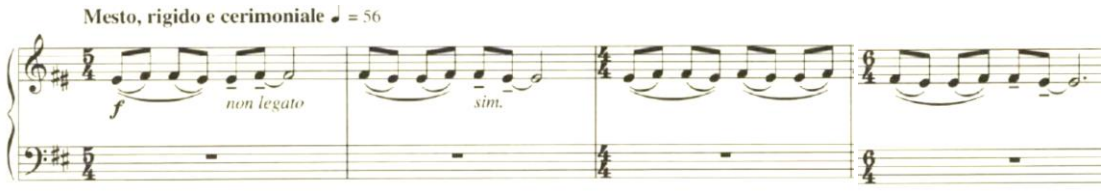

Figure 1. Ligeti's *Musica Ricercata II*, the first 4 bars. The principal motif at the beginning of the piece.

We used a piano recording of *Musica Ricercata II* that was tailored for the experiment. The piece was adapted by Nori Jacobi and played by pianist Rotem Luz on a Yamaha b1 SG2 upright piano with a silent mechanism allowing simultaneous recording of audio and MIDI. The frequency range of rat audition is ~500 Hz – 70 KHz, significantly higher than that the 20 Hz - 20 KHz auditory range of humans. To better fit rat audition, the original piece was transposed up by one octave and only the three upper octaves of the piece (octaves 5, 6, and 7) were used. The parts in the piece originally written to be played in unison on four octaves were played in this version on three octaves. During

the audio recording a MIDI file was produced with the exact timing and key velocities of the notes played by the pianist.

The experimental Setup

Preparation

Electrophysiological data were recorded in-vivo from the left auditory cortex of 6 adult female Sabra rats (weighing 210 gm-265 gm). The joint ethics committee (IACUC) for animal welfare of the Hebrew University and Hadassah Medical Center approved the study protocol. The Hebrew University is an AAALAC International accredited institute. Animals were initially anaesthetized with an intramuscular injection of ketamine (0.1 ml, Ketaset, 100mg/ml, Fort Dodge Animal Health, Fort Dodge, IA) and medetomidine (0.05 ml Domitor, Orion Pharma, Espoo, Finland). Additional smaller doses of ketamine were administered as needed to maintain anesthesia during surgery. Surgical level of anesthesia was verified by pedal-withdrawal reflex.

The trachea was cannulated and the animals were fixed to a custom-made head holder (Haidarliu, 1996), that left the scalp and ears free. The animals were ventilated (10-15 mm H_2O peak inlet pressure, 47/min, 15-30cc per stroke, 0.7-1.4 L/min) by a mixture of O_2 and halothane (Rhodia Organique Fine Ltd., Bristol, UK) using a small-animal ventilator (model AWS, Hallowell EMC, MA), and a halothane vaporizer (VIP 3000, Matrx, NY). Once an animal was ventilated, ketamine anesthesia was discontinued, and halothane concentration was set around 0.7% (as needed). Throughout the experiment, respiration quality was monitored by continuously measuring the CO_2 concentration in the tracheal cannula (Microcap, Oridion Medical Ltd., Jerusalem, Israel). The depth of anesthesia was judged by the lack of motion and resistance to the respirator, and levels of anesthesia and ventilation pressure were adjusted accordingly. Body temperature was monitored and maintained at 36-38°C using a rectal thermistor probe and a feedback-controlled heating pad (FHC Inc., ME).

Acoustic stimulation delivery

The experiment was conducted in a sound-proof chamber (IAC, Winchester, UK). Pure tones and broadband noise bursts were synthesized online using Matlab (The Mathworks, Inc., Natick, MA), transduced to voltage signals by a sound card (HDSP9632, RME, Germany), attenuated (PA5, TDT), and played through a sealed speaker (EC1, TDT) into the right ear canal of the rat. For pure tones, an attenuation level of 0 corresponded to about 100 dB SPL. Noise stimuli were synthesized at a spectrum level of -50 dB/sqrt(Hz) relative to pure tones at the same attenuation level.

Electrophysiological recordings

The left temporal portion of the skull was cleaned from skin, muscles, and connective tissue. A craniotomy was performed over the estimated location of the left auditory cortex: 2.5mm-6.5mm posterior to and 2mm-6mm ventral to the bregma (Swanson, 1992). A copper wire hook implanted in the neck muscles was used as the electrical reference. Recordings were made using glass-coated tungsten electrodes (Alpha-Omega Ltd., Nazareth-Illit, Israel). The electrodes were arranged into a fixed array and lowered into the cortex using a single microdrive (MP-225, Sutter Instrument Company, Novato, CA). The electrical signals were pre-amplified (×10), filtered between 3 Hz and 8 kHz to obtain both local field potential signals (LFP) and multiunit activity (MUA) and then amplified again, for a total gain of ×5000 (MCP, Alpha-Omega, Nazareth Illit, Israel), to yield the raw signals. The raw signals were sampled at 25 kHz and stored for offline analysis. During pure tone and noise presentations the analog signals were also sampled at 977 Hz after anti-aliasing filtering (RP2.1, TDT, Tucker-Davis Technologies, Alachua, FL), stored for LFP analysis, and used for online display. To detect MUA, large fast events in the filtered raw signals were marked as spikes with the threshold for spike detection set to seven times the median of the absolute deviations from the median (MAD) of the filtered voltage traces (corresponding to more than four SDs for Gaussian signals). The resulting spike trains were aligned to stimulus onset and averaged.

Experimental procedure

Recording sites were selected by their response to broad-band noise (BBN). The electrodes were driven into the cortex while presenting 200 ms BBN bursts (0-50 kHz) with an inter-stimulus time interval (ISI, onset to onset) of 500 ms and a level of 30 dB att. The LFP responses were averaged online, and the electrodes were positioned at a location and depth that showed large evoked LFP responses on as many electrodes as possible. Once selected, the BBN responses of the recording site were validated and recorded using a sequence of 280 BBN bursts with a duration of 200 ms, 10 ms linear onset and offset ramps, ISI of 500 ms, and seven different attenuation levels between 0 and 60 dB att with 10 dB steps which were presented pseudo-randomly so that each level was presented 40 times. The main data were collected if the noise threshold level was at least 30 dB; attenuation and noise-evoked potentials changed regularly with level. Otherwise, the electrodes were moved to a different location. Several quasi-random frequency sequences of 370 tone bursts (50 ms duration, 5 ms onset/offset linear ramps, 500 ms ISI) at 37 frequencies (6 tones/octave, frequency range was 1-64 kHz in animals 1-3 and 0.5-32 kHz in animals 4-6) at several attenuation levels from threshold and up to an attenuation of 20 dB, were used to measure the frequency response area (FRA) of the neuronal responses. Once the recording site was characterized in terms of its responses to pure tones, the piano recording of *Musica Ricercata II* by György Ligeti was played at 20 dB attenuation. The piece was played consecutively 10 times with a short break between repetitions. In some experiments, additional stimuli were played between successive repetitions of the piano recording.

Data Analysis

Responses to sounds in the music

The version of *Musica Ricercata II* presented to the rats had 208 main notes (termed 'main notes' below), when notes played in unison on three octaves were counted together:

1. The first half of the piece (~0-135 sec): 4 repetitions of the principal motif, 26 E♯ and F♯ notes each. This past includes 3 F♯ notes that break the template of the main motif and are played an octave higher (F♯6).
2. The middle part (~135-190 sec): the initial introduction of the G theme (17 Gs), a tremolo of 22 Gs, and the two motifs played together (26 E♯ and F♯ with 7 Gs interspersed between them). The 50 loud notes in this part (excluding the tremolo) are termed 'climax notes' below.
3. The last part (~190-250 sec): the principal motif (26 E♯ and F♯ notes each), and the coda ending with 6 notes, 2 at each pitch.

These 208 notes were included in the analysis. The piece also included additional tremolos on the Gs (G7/G6) which are played at the same time as the main notes, but were not included in the analysis. For each of the main notes, the timing and maximal sound level were extracted directly from the acoustic signal used in the experiment. For this purpose, the instantaneous sound level was estimated by lowpass filtering the squared waveform using a 20 ms Hanning window (corner frequency ~ 47Hz). The expected onset time from the MIDI recording was used to identify the peak envelope level for all notes. For each note, the exact timing and maximal sound level were then identified in a temporal window ±35 ms around the expected peak time. The time interval between peak times of successive notes was longer than 200 ms, with a typical interval around 700 ms (735 ± 113 ms, median ± MAD) corresponding to the duration of the eighth notes that dominate the piece.

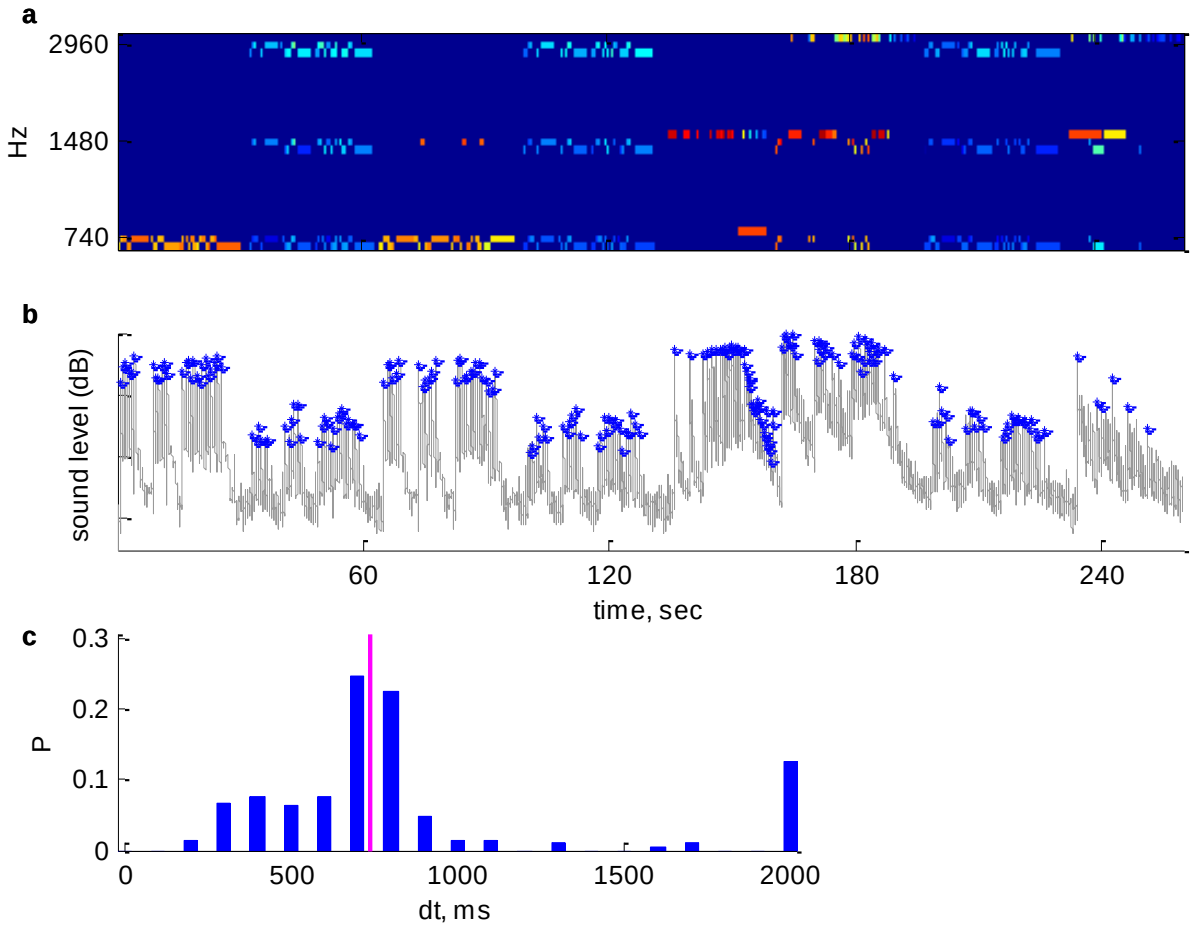

Figure 1. **a**. Schematic (MIDI) representation of Ligeti's *Musica Ricercata II* as used in the experiment. The color scale corresponds to MIDI key velocity (and therefore approximately to sound level). **b**. The sound level trace (gray) is presented together with the peak level of each of the main notes (blue dots). **c**. The distribution of intervals between successive notes for the main notes included in the analysis. The magenta line marks the median of the distribution (735 ms).

For each recording location, MUA and LFP recorded in response to the 10 repetitions of the musical piece were averaged. MUA responses were then smoothed using a 9ms moving average window to estimate the instantaneous spike rate. The responses to each note were quantified by the maximal response (absolute value of the maximal negativity in the case of the LFP) in a 135 ms window (-25 ms to +110 ms) around the peak time of the corresponding note. Recording sites were included in the final analysis

if they showed significant responses during the second part of the piece (starting at about 134 sec, with the first occurrence of the G until the end of the piece). To determine the significance, responses to the 104 main notes that appeared in this part of the piece were compared by paired t-tests ($p<0.05$) with maximas of the neural signal extracted the same way from the 135 ms just preceding each note (from -175 ms to -40 ms relative to the peak time of the note).

Frequency response areas (FRAs) and FRA based filtering

FRAs were constructed for each recording site for both the LFP and the MUA by summing responses in a 15 ms window centered on the maximal response. The exact time window was selected separately for the MUA and LFP signals in each animal, and was ~15-30 ms after tone onset for MUA recordings and ~20-35 ms after tone onset for LFP. FRAs were then smoothed with a 3X3 moving window (the product of two Hamming windows along the two axes). In case of an attenuation that was not surveyed in the recording session measuring the FRA, the FRA was interpolated linearly between it two neighboring attenuations.

In order to determine the expected response of a recording location based on its FRA, the acoustic signal had to be expressed in the same frequency and level terms of the FRA. The waveform was first filtered using second order bandpass Butterworth filters centered on the frequencies used for FRA evaluation (500 Hz - 22050Hz, the maximum possible for a 44.1 KHz sampling rate). The instantaneous sound level of the filtered signals was determined by squaring and lowpass filtering with a 20ms Hanning window (corner frequency ~47 Hz) and then downsampling to 1 KHz. In order to calibrate the sound levels of each note in the piece after FRA filtering with the sound levels during FRA estimation, the peak level of the first G was used as a reference and set to 10 dB att. This choice resulted in the highest correlation between the notes sound levels (expressed in dB attenuation) and the neuronal responses to the notes. Setting the sound level of the first G to 20 dB att resulted with very similar results. Because of the way the FRA was measured, the relevant level scale was between 20-80 dB att, in steps

of 10 dB att. The notes peak levels after FRA filtering, ranging from ~8 to ~110dB att, were quantized to 20-80 dB att by mapping values to the nearest discrete level in this range. As a result, the acoustic signal was expressed in the same terms of frequency and level as the FRAs, with each note corresponding to a line (level as function of frequency) in the FRA frequency-level plane. The FRA-based response to a note was then estimated by summing the FRA values along this line. When not all relevant tone frequencies were used to measure the FRA, the estimate was based only on the frequencies bands actually tested. This didn't affect the results because the missing frequencies usually didn't evoke any significant response.

Surprise measures

To test the general suggestion that sounds expectancy modulates the neuronal response, simple statistical modeling of the sound was implemented as proof of principle and tested against the neural responses to the music. Earlier findings (e.g. Ulanovsky et al., 2004) suggested that sounds that violate expectations are likely to elicit larger responses in auditory cortex (alternatively, responses to expected sounds are depressed). It follows that the log probability of a note should be negatively correlated with neural responses. I tested this hypothesis deriving two basic surprise measures from the musical piece.

An important form of regularity consists of the probability of single events. In this statistical model every note in the piece is expected according to the relative frequency of its occurrence in the melody. Thus, as a first approximation, the surprise value of each note was determined using its probability of occurrence among the main notes. There are different possible functions of the probability that can serve as surprise measure such that notes that are less expected generate a larger surprise. The definition used here is the negative log of the probability, so that: $surprise_i = -log_2(p_i)$.

A second measure of surprise was derived directly from the acoustic signal by predictive modeling of the sound, in the same manner described in Chapter 5. The acoustic signal

was downsampled to a sampling rate of 8.82 KHz. One common predictive model for the entire piece was generated that used 600 sample points (~70 ms) to predict the next sample point of the acoustic signal (see Chapters 2 and 5). For each major note in the piece, the Pearson correlation coefficient was calculated between the sequence of predictions and original waveform in temporal windows of ±500 ms around note peak time. A high correlation reflects a high level of predictability of the signal, so the correlation had to be transformed into a surprise measure by a monotonically-decreasing function. The negative log of the correlation $-log_2(r_i)$ was selected as the surprise measure for each note.

Results

Recordings were obtained in 80 recording locations from 6 rats. Significant LFP responses in response to the music were found in 59% (47/80) of the recording sites and significant MUA responses in 38% (30/80) of the recording sites (see Methods). Only these are further analyzed below.

General properties of the responses

The responses to individual notes were compared to the responses recorded in the same recording site to pure tones presented during FRA measurement (Fig. 2). The temporal structure of the responses was essentially identical, with typical response times that corresponded well to response latencies in rat primary auditory cortex: ~22 ms from tone onset to maximal spike rate and ~30-35 ms from tone onset to the negative peak of the LFP.

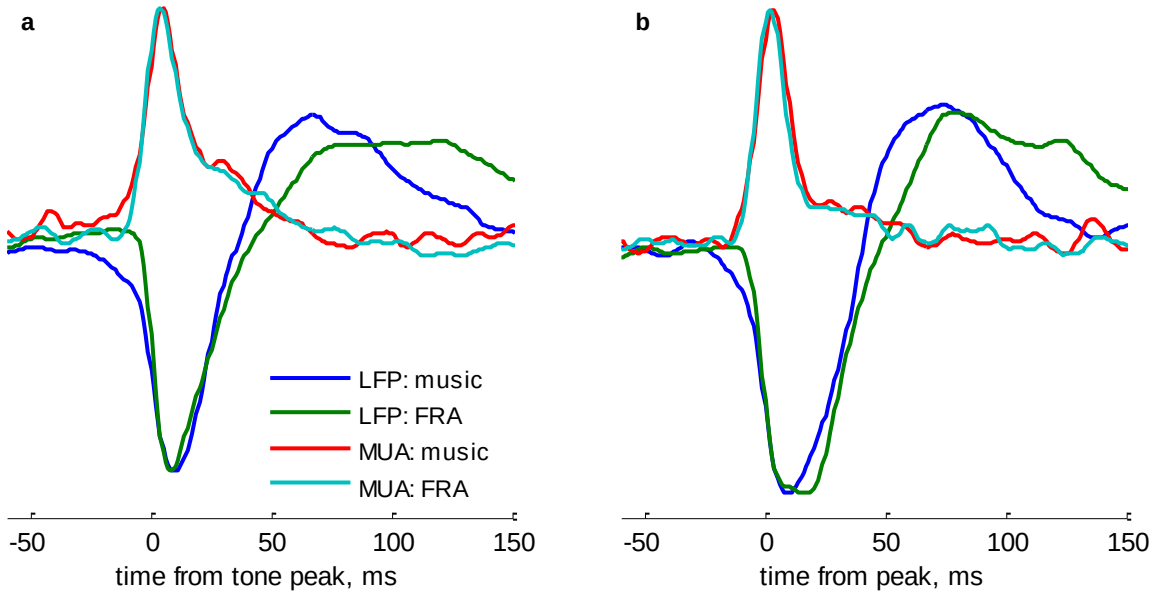

Figure 2. Response dynamics. A comparison between responses to pure tones (recorded during FRA measurement) and to piano notes (recorded during the presentation of the music piece) averaged across all responses in a single recording location (**a**) and for the entire set of responsive locations in one animal (**b**, LFP n=17, MUA n=6). Responses to the tones during FRA measurement were averaged across all tones at the highest sound level presented during FRA measurement (20 dB att, about 80 dB SPL). Responses to the notes were averaged across all *main notes* in the piece (N=208, see definitions above). Each average trace was normalized individually to have a maximal absolute response of one and corrected for baseline activity during the 200ms preceding tone onset.

Surprise affects neural responses

Figure 3 illustrates the response profile to the entire musical piece. Both in the single and in the average traces most responses were to the *climax notes* (see definitions above) during the middle part of the piece. These notes were considerably louder than all the notes played earlier in the piece (see Fig. 3b) and were played at higher tempo, in accordance with the score's instructions. The average level of these 50 notes (excluding the G tremolo) was -13 dB, ~9.5 dB louder than the average level of the 104 notes in the first part of the piece.

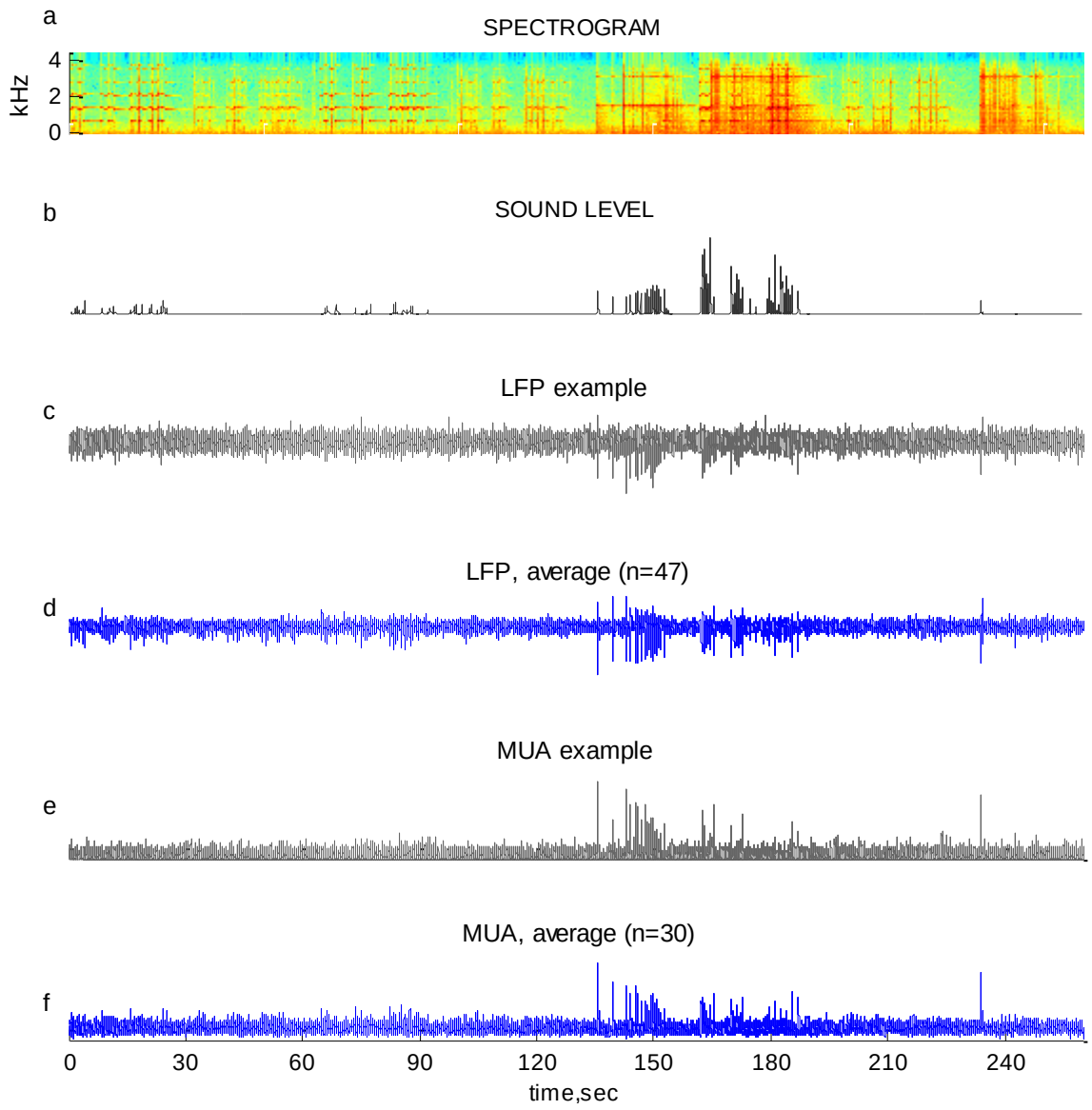

Figure 3. Responses to the music. a. The spectral content of the version of *Musica Ricercata II* presented to the rats. **b.** Instantaneous sound level (linear scale). **c-f.** Representative response traces demonstrate general consistency in the neural response to this piece and the prevalence of the responses to the climax notes in the loud middle part of the piece.

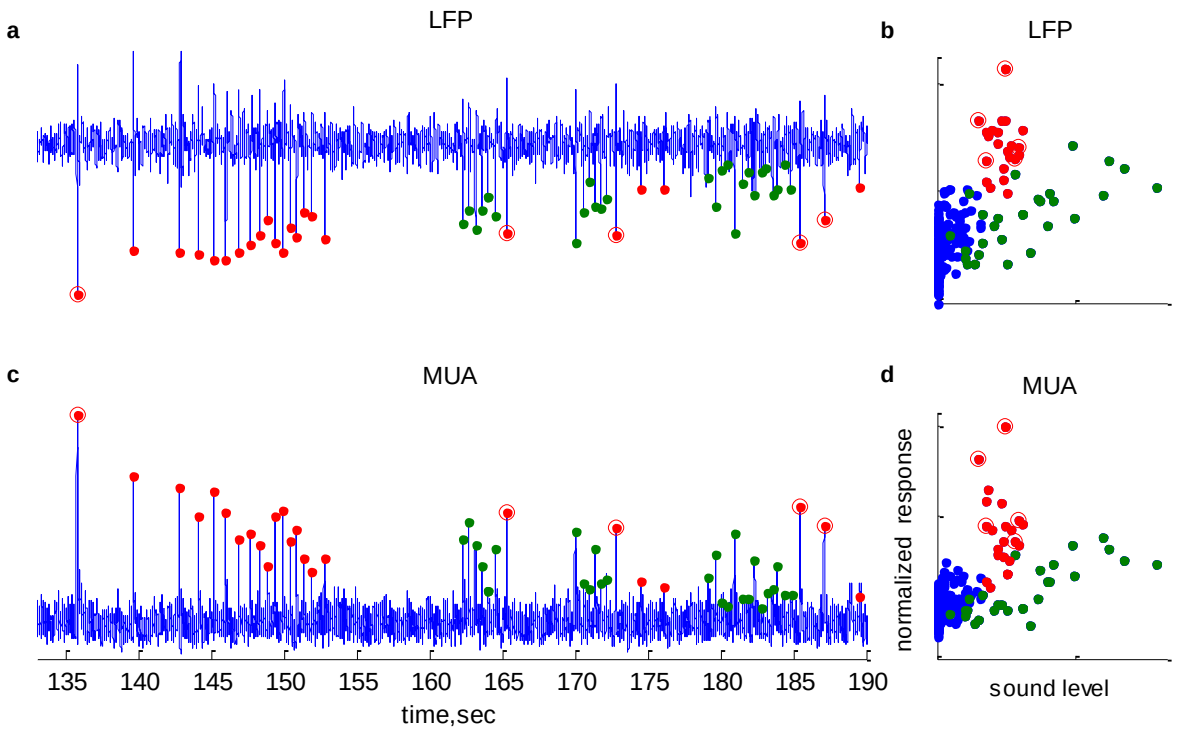

Figure 4. Neural responses - a detailed view. Left: Mean responses to the middle part of the piece (135-185 sec), magnified from panels d and f in Fig. 3 - LFP (**top**, n=47) and MUA (**bottom**, n=30). Dots mark the responses to the climax notes in this segment (n=50), colored according to the motif the notes belong to – Gs in red, E♯s and F♯s from the main motif in green. **Red empty circles** denote the 4 Gs that appear immediately after notes from the main motif (E♯ or F♯). **Right:** Responses to the main notes in the piece plotted as a function of note sound level (n=208, ISI>200 ms). Same color code as the left panels.

Each of the climax notes played in the middle part (n=50, 17 Gs played alone and 33 notes when the two motifs are played together, see above) evoked a distinct neural response. In Fig. 4, the average responses to all main notes are plotted against the peak sound level (right panels, N= 208 notes with ISI>200 ms). The overall correlation between sound level and responses for this set of notes was high – 0.62 and 0.53 for LFP and MUA respectively ($p<<0.01$). This mainly reflected the fact that low sound levels evoked weak or no response, whereas all responses were evoked by the louder sounds. However, when limiting the analysis to the climax notes that actually evoked the most

responses in the middle part of the piece, there was a highly nonlinear relation between sound level and response (Fig. 4, red and green dots).

To select only the louder sounds, a threshold was set at ½ standard deviation above the mean note level (~12 dB att), and calculations were repeated for the *above-threshold notes* (n=44/208 notes). The correlation between sound level and response to the notes above threshold (*above-threshold correlation*) was not significant (r=-0.08,-0.06 for LFP and MUA respectively, $p>0.05$, see for example the left panels in Fig. 9 below). The responses to the Gs included in this group of notes were larger than would have been expected based on the level of these notes alone. Moreover, the response to Gs that mark the interruption of the main motif (red circles in the panels of Fig. 4) were exceptionally large.

The presence of larger than expected responses to the Gs was observed in most individual recording sites, as illustrated in the representative traces in Figs 5 and 6 for LFP and MUA respectively. For MUA, the correlation between sound level and responses at the individual sites was 0.23 ± 0.23 (mean ± std) with the correlation reaching significance ($p<0.01$, t-test, see Chapter 2) in 19/30 units. However, when only the responses to above-threshold notes (½ std above mean level) were considered, the correlation of the responses with above-threshold sound levels vanished in all 30 units (-0.02 ± 0.12, mean ± std, $p>0.05$). The same pattern was present for the LFP responses, as measured by the absolute value of the maximal negativity. The correlation between level and LFP response at the individual sites was 0.36 ±0.21 (mean ± std), and was mostly significant ($p<0.01$ in 39/47 sites), whereas the above-threshold correlation was not significant ($r= -0.05 \pm 0.13$, $p>0.05$ in 47 recording sites).

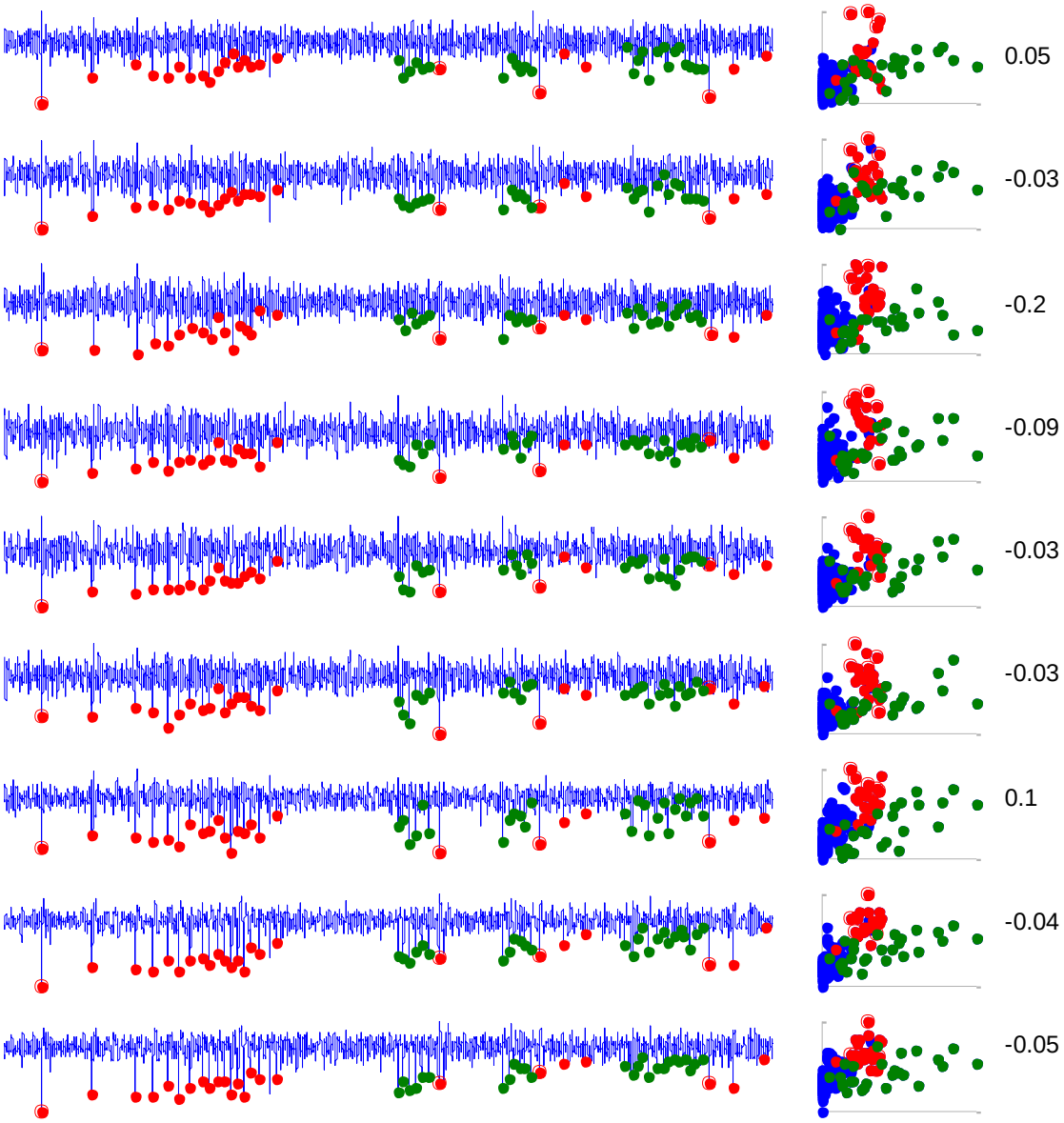

Figure 5. LFP responses – individual recording sites. Left: LFP response traces from individual sites to seconds 130-190 of the piece. **Right:** a scatter plot of LFP responses, measured as the absolute value of the maximal negativity, to all main notes (n=208) vs. note level. Each row is from one representative recording site. The above-threshold correlations, the correlations between sound level and responses restricted to above-threshold notes, are displayed on the right. Color conventions as in Fig. 4 above.

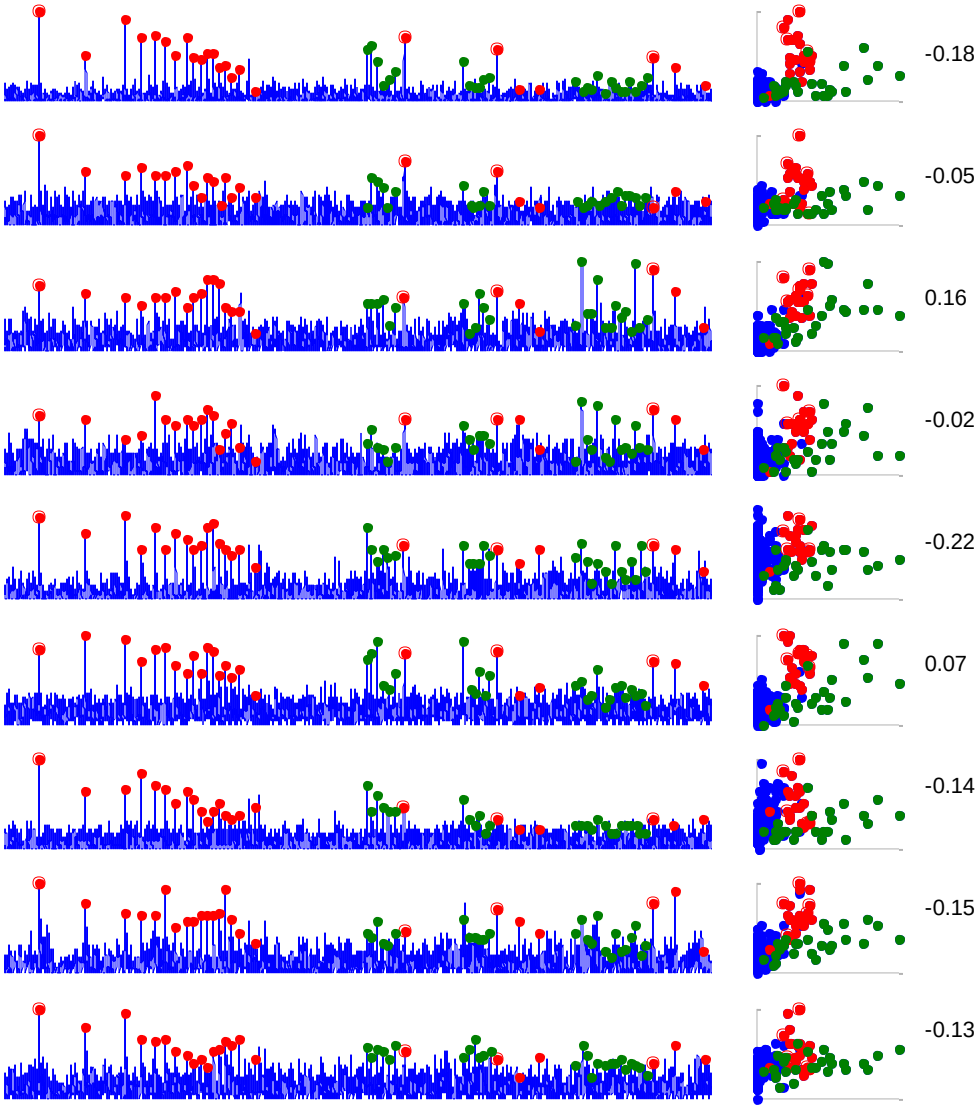

Figure 6. MUA responses – individual recording sites. MUA response traces from individual recording sites to seconds 130-190 of the piece. **Right:** a scatter plot of MUA responses to all main notes (n=208) vs. note level. Each row is from one representative recording site. The above-threshold correlations, the correlations between sound level and responses restricted to above-threshold notes, are displayed on the right. Color conventions as in Fig. 4 above.

Figures 5 and 6 illustrate these findings for representative recording sites. Each of the 50 climax notes in the middle part of the piece reliably elicited identifiable responses in these sites. More importantly, the responses to the Gs were consistently larger than those evoked by the E♯ and F♯ notes, in spite of the fact that some of the E♯ and F♯

notes were louder. In particular, the four Gs that interrupted the main motif (see Fig. 7) generally elicited larger responses than the neighboring notes that were part of the main motif, although the latter were ~1.4 dB louder than the Gs on average. In Fig. 7, the mean response to these four Gs is plotted against the mean response to the main motif notes, for each response from an individual recording site. The vast majority of the points lie above the diagonal.

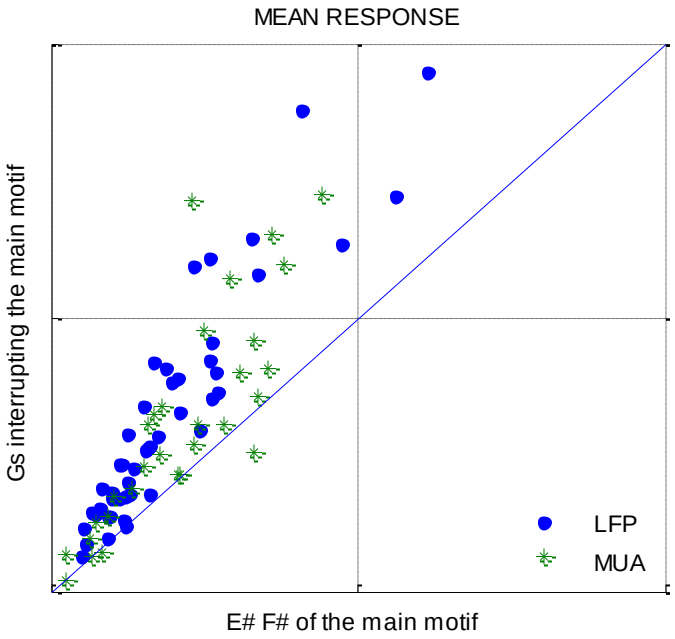

Figure 7. Larger responses to Gs that follow E♯ or F♯. The mean response to the four G notes that interrupted the principal motif (immediately followed an E♯ or an F♯) during the middle part of the piece (**ordinate**, n=4) versus the mean response to the motif notes played during the same part (**abscissa**, E♯s and F♯s, n=26). Each point represents an individual recording site (LFP n=47, MUA n=30). Values were normalized by the maximal response to aid visualization.

FRA-based response predictions

Thus far the overall sound level of the piano notes has been used to account for the neural responses. However, when presented with pure tones, neurons in A1 typically respond to the energy in a limited frequency range. The neural responses to isolated

pure tones measured at each recording site were used to construct FRAs that characterize the neural response in frequency-level space. Fras excitatory regions were typically v-shaped or doulepeaked, as demonstrated in Figure 8 presents by FRAs for three recording sites. As expected in A1, some of the FRAs exhibited non-monotonic behavior with level. There were obvious similarities between the FRAs derived from the LFP and FRAs derived from the MUA in the same recording site, with correlations of 0.63 ± 0.3 (mean ± std) between FRAs in sites where both responses were significant (n=27). This was also manifested in the similarity between the mean FRAs of the two response types, obtained by averaging across all the responsive sites (r = 0.97, p<<0.01. Figure 8, right panels). The range of best frequencies (BFs) of individual FRAs was 1.1 - 57 KHz and FRAs excitatory regions covered much of the frequency range of the music.

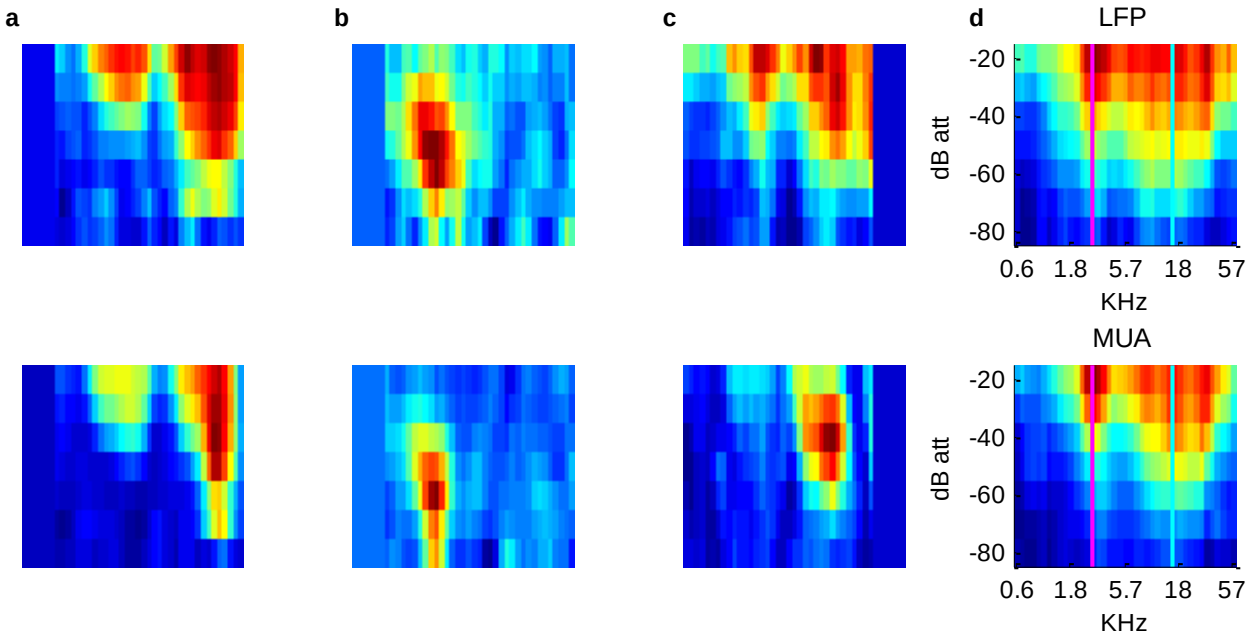

Figure 8. FRAs – response profiles in frequency-level space derived from LFP responses (**top**) and MUA (**bottom**). **a-c.** Three examples of FRAs from individual recording sites showing overall similarity (**a**, **b**) or some differences (**c**) between FRAs derived from LFP and MUA. **d.** FRAs averaged across all responsive sites. The general shape was similar between the two response types. The frequency and level that evoked the maximal mean response (2.8 KHz, **magenta line**), as well as the frequency with the lowest threshold (16 KHz, **cyan line**) were identical in the two response types.

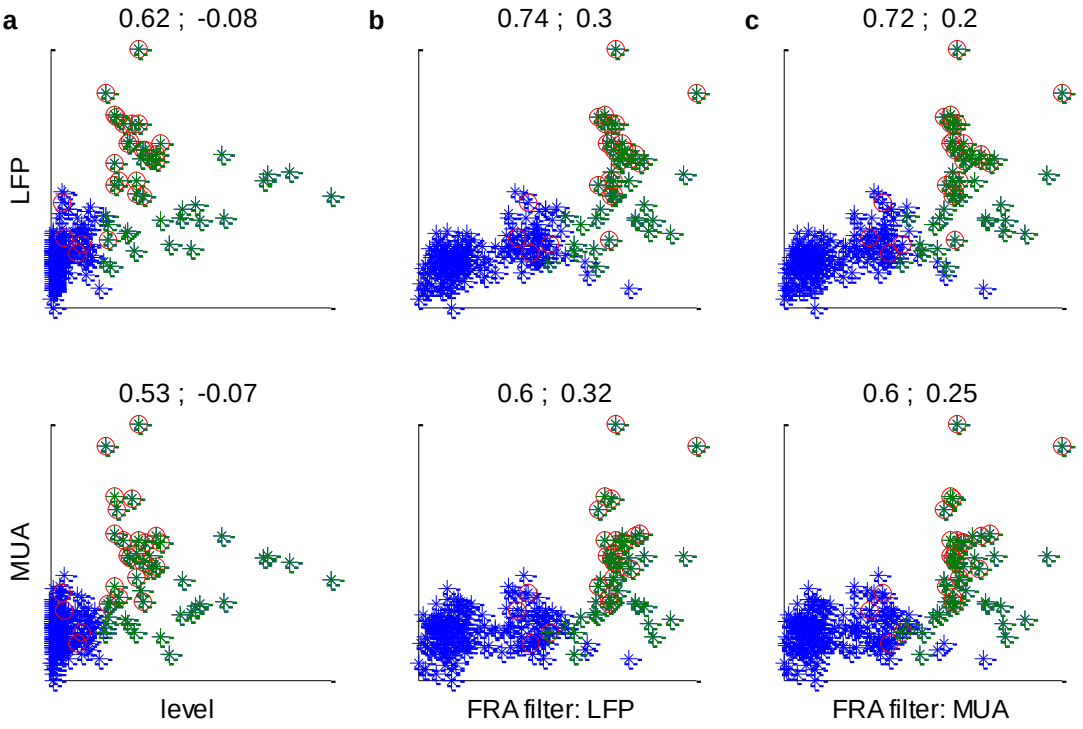

Figure 9. FRA-filtered sound levels. Averaged neuronal responses (**top** – LFP, **bottom** – MUA) are plotted against either the original note level (**a**) or the level after filtering by the average FRAs (**b** and **c**, LFP and MUA derived FRAs respectively). Responses for above-threshold notes (½ std above mean level) are in green. Red empty circles denote the main Gs in the piece. Correlation values are indicated above each panel for all the notes (**left**) and for the above-threshold notes (**right**).

FRAs from responsive recording sites were used to estimate the *FRA-filtered sound level* of each note - the total energy of the note relevant to the neural response at the recording site according to the FRA. The FRA-filtered sound levels indeed correlated better than the unfiltered notes sound levels with the actual neural responses. Figure 9 illustrates this for the average responses. To attain the average FRA-filtered sound levels, I used the FRAs averaged across all responsive sites to filter the sound. The overall correlation between responses and the average FRA-filtered sound level was 0.74 and 0.6 for LFP and MUA respectively. Moreover, when only above-threshold notes were considered, correlations reached ~0.3, although they were still not significant ($r=0.3$, $p=0.05$ for LFP, $r=0.25$, $p=0.09$ for MUA). MUA responses to above-threshold

notes were even slightly better accounted for by the average LFP FRA (r=0.32, p=0.03). Thus, after spectral filtering based on the FRAs measured at each site, the unexplained variance in the responses to above-threshold notes was somewhat reduced, although still very considerable (Fig. 9b, c).

The same trend exhibited in the averaged responses was found in the responses at individual recording sites (see Fig. 13 and Table 1 below). The predictions based on the FRAs correlated better with the neuronal responses, including the responses to above-threshold notes. On average, in comparison to unfiltered sound levels, both overall and above-threshold correlation values were higher with FRA-filtered sound levels, and there was a small increase in the number of sites where the correlation reached significance (see Table 1). Filtering the sound with FRAs measured from the LFP resulted with higher correlations with both LFP and MUA responses, although differences were small.

However, the nonlinearity of the responses to the above-threshold notes remained considerable. Above-threshold correlations obtained with FRA-filtered sound levels were negligible: 0.09 ± 0.24 and 0.18 ± 0.15 (mean ± std) for MUA and LFP respectively, and only ~10% of the responsive sites showed significant correlations ($p<0.05$). Thus, the explanatory power of the FRAs for the responses to above-threshold notes was still very low.

Models incorporating surprise

The first measure of surprise, the instantaneous surprise of each note, was defined using the negative log of the probability of occurrence of each note in the piece,$-log_2(p_i)$ (see above). This resulted in a straightforward prediction - that every time a certain pitch class appeared it would elicit the same response. Despite its limited range, this surprise measure by itself had considerable explanatory power. The correlation between the responses to all notes and the surprise measure was low - 0.33 and 0.29 for LFP and MUA respectively (see Fig. 10). These correlations are even lower

than the correlations of the responses with notes sound level alone. The above-threshold correlations, on the other hand, were high - 0.68 and 0.58 for LFP and MUA respectively. This is consistent with the notion that once a note was loud enough to elicit a response, the response was modulated by the surprise the note generated.

Next, I treated the surprise of each note as a modulator that facilitates/depresses the neuronal response this note would elicit in isolation. I combined the surprise associated with each note (as defined above) and the sound level, simply by taking their product: $-(log_2(p_i) - \alpha) * level_i$. In the results below, $\alpha = 1$ was used as it yielded a good fit with the neuronal data. Even this basic measure of surprise combined with sound level yielded a significant increase in explanatory power (Fig. 10c).

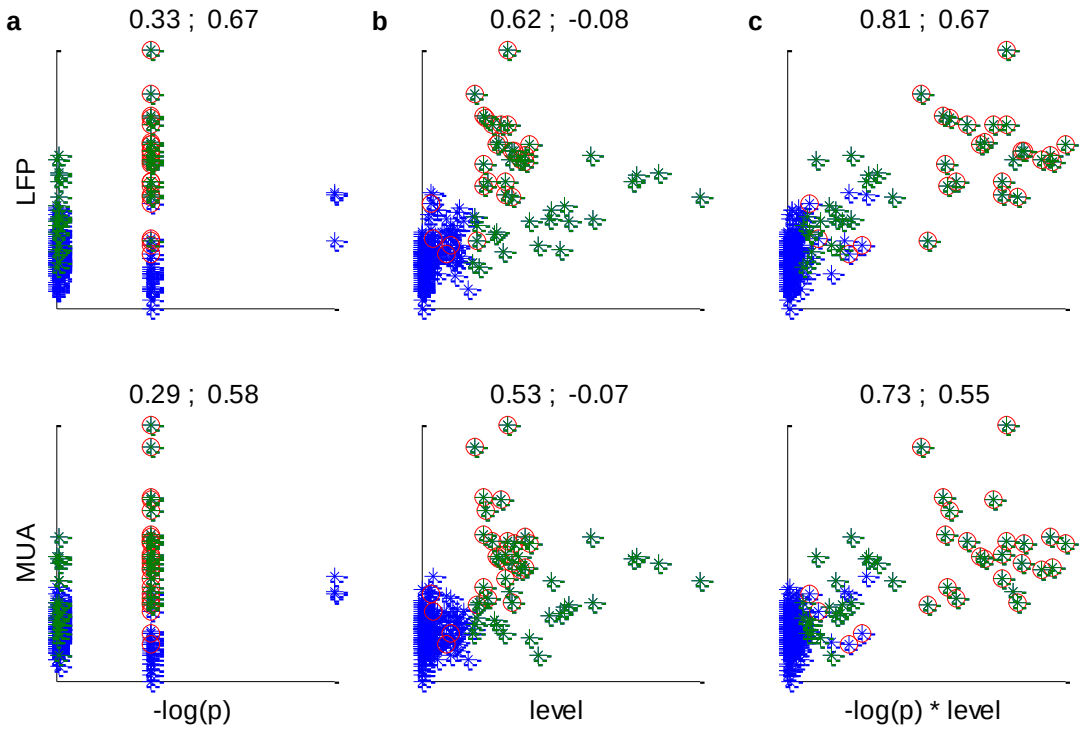

Figure 10. Combined surprise and sound level measure I. Average neuronal responses (**top** – LFP, **bottom** – MUA) were plotted against either the surprise alone, as measured by the negative log probability of each note (**a**), the sound level alone (**a**) or the surprise combined with the sound level of the notes (**c**). Color coding is the same as in Fig. 9. Correlation values are indicated above each panel for the all the notes (**left**) and for the above- threshold notes (**right**).

Combining surprise and sound level substantially increased the overall correlation, up to 0.81 and 0.73 for LFP and MUA respectively ($p<<0.01$). The above-threshold correlations were also high and significant and resembled those obtained with the surprise measure by itself: 0.67 for the LFP responses, and 0.55 for MUA responses ($p<0.001$ in both cases). This result was confirmed in the individual recording sites, as depicted in Fig. 13 and Table 1; the combined surprise and sound level measure resulted in significantly increased correlations that were large and significant in most responsive sites for both correlations for all responses (0.47 ± 0.21 / 0.3 ± 0.28 for LFP/MUA) and the above-threshold correlations (0.41 ± 0.16 / 0.22 ± 0.27 for LFP/MUA).

Next, I used an additional surprise measure based on the general procedure described in Chapter 5 for predictive modeling of acoustic signals. The second surprise measure was derived directly from the statistics of short segments of the acoustic signal, based on the notion that successful predictions reflect high predictability. Thus, a model for predicting one sample ahead based on 600 past samples (~70 ms) of the sound was generated for the entire piece, based on the covariance matrix of the sound. Then, sample-by-sample predictions were generated for short segments of 500 ms centered on the onset of each major note in the piece. For each segment, the Pearson correlation coefficient (r_i) was calculated between the predicted trace and the actual sound. Thus r_i can be considered as measuring how well the acoustic signal in the half second surrounding note peak time matched the expectation generated by the model based on the recent history of the sound (an additional ~70 ms to the 500 ms test segment). The surprise elicited by each note was then defined as the negative log of the correlation $-log_2(r_i)$.

This procedure yielded an instantaneous measure of surprise extracted directly from the physical acoustic signal that is formulated dynamically based on a relatively short memory trace. The covariance matrix of the signal used for constructing the model was averaged across the entire piece, and thus introduced some global statistical characteristics of the piece to the calculation. Figure 11 presents the surprise for each of

the major notes in the piece. The Gs of the second motif generated high surprise values in this realization. Interestingly, surprise values were also higher for the three high Fs that mark a change in the part of the piece, when only the first motif is played (see the piece description above).

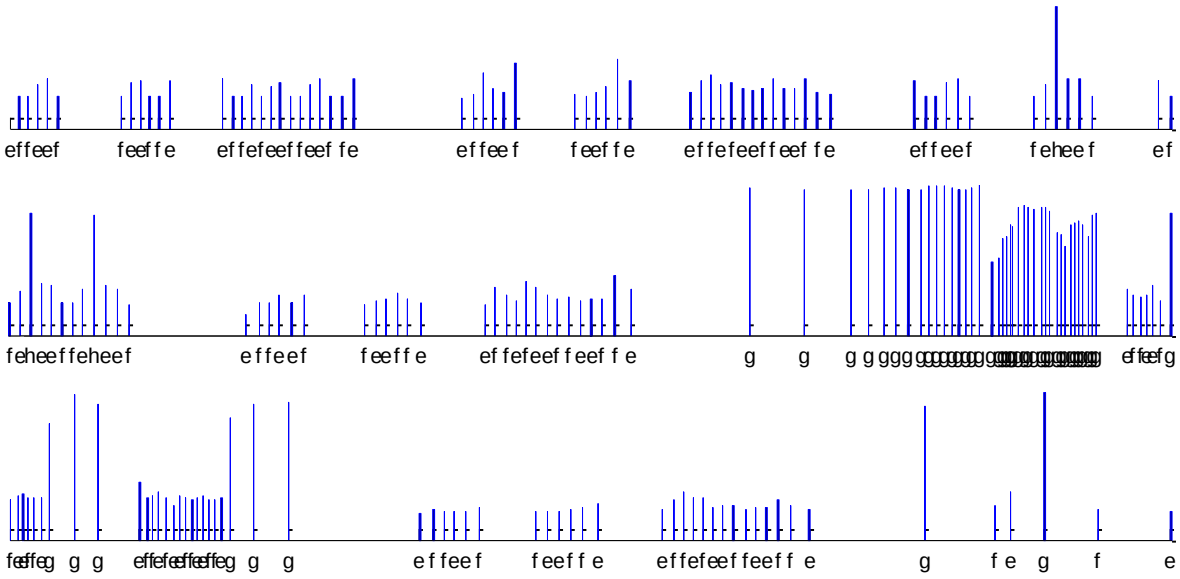

Figure 11. Instantaneous surprise measure II. Surprise was measured by $-log_2(r_i)$, where r_i is the correlation between the predicted and the actual signal 1 second around note peak level. The letters indicate the note played, with **e,f** : the E♯ and the F♯ notes of the main motif, played on one octave or three, **h** : the high F♯6 notes that interrupt the main motif, and **g**: the G notes of the second motif.

Notably, the 500 ms window used for calculating the correlation surrounded note peak time. The time window thus included the time the note was actually played and not only the onset dynamics, which is the most important determinant of the neural responses. Also, it is not clear to what extent the correlation values assigned to each individual note, were affected by neighboring notes, in particular by residues of the preceding note which may not be uniformly well predicted. Including the ~70 ms of the model, the total duration used for calculating the correlation included ~320 ms preceding note peak level. The median time interval between successive note peaks for the main notes included in the analysis was 735 ms (see above), with shorter intervals occurring in the

middle part (463 ± 117 ms, median ± MAD) in general accordance with the tempo instructions of the score. Therefore, additional analysis has to be performed to trace the precise sources of variation in this surprise measure.

The results found for the second surprise measure were almost identical to those found for the first surprise measure. Taken alone, this surprise measure already had explanatory power, especially for the responses to above-threshold notes. The correlations of the surprise measure by itself with all responses were 0.46 and 0.4 for LFP and MUA respectively (see Fig 12) - lower than the correlations with sound level alone. When only responses to above-threshold notes were included, the correlation with the surprise measure was high - 0.65 for LFP and LFP 0.58 for MUA, in agreement with the results obtained with the first surprise measure. Following the same rationale as above, the surprise measure was multiplied by the sound level, for a combined surprise and sound level measure: $-log_2(r_i) * level_i$. The result was a significant, large increase in the correlation with all responses: 0.82 and 0.73 for LFP and MUA respectively. The above-threshold responses were also high, and similar to those obtained for the surprise measure by itself: 0.65 for LFP and 0.55 for MUA. See Fig. 12 for the average population results, and Fig. 13/Table 1 for the results from the individual recording sites.

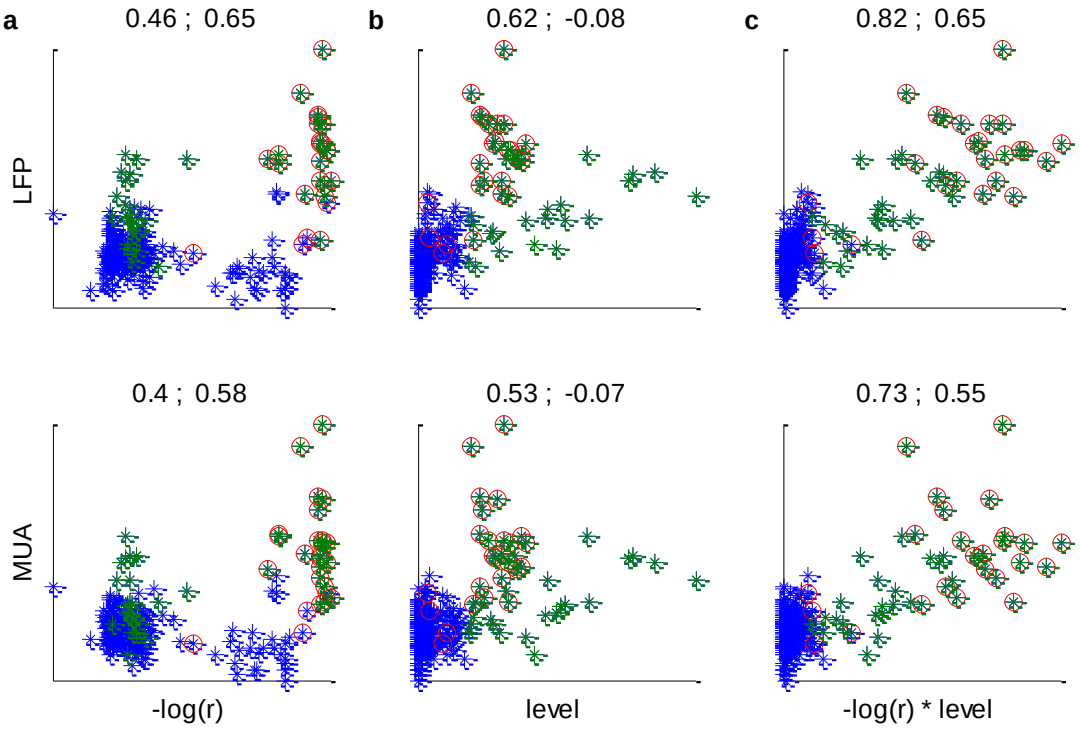

Figure 12. Combined surprise and level measure II. Average neuronal responses (**top** – LFP, **bottom** – MUA) were plotted against either the second surprise measure alone, as measured by the negative log probability of the correlation between prediction and actual signal (**a**), the sound level alone (**b**) or the surprise combined with the sound level of the notes (**c**). Color code is the same as in Fig. 9. Correlation values are indicated above each panel for all the notes (**left**) and for the above- threshold notes (**right**).

Interim summary and a caveat

Quite simple statistical modeling of the musical piece yielded an large increase in the explanatory power in relation to note level alone. Figure 13 and Table 1 summarize results for the individual recording sites. The results mirror the population results illustrated above; namely, both LFP and MUA responses from sindividual recording sites correlated significantly with the combined level-and-surprise measures. These was true even for the particularly non-linear set of responses to above-threshold notes: the combination of surprise measures and level resulted in significantly higher above-

threshold correlations. The average correlations from individual recording sites were in general lower than the correlations with the average responses across all sites. This is possibly due to the fact that the responses in individual sites were noisier, and to the sparsity of the responses, especially the MUA.

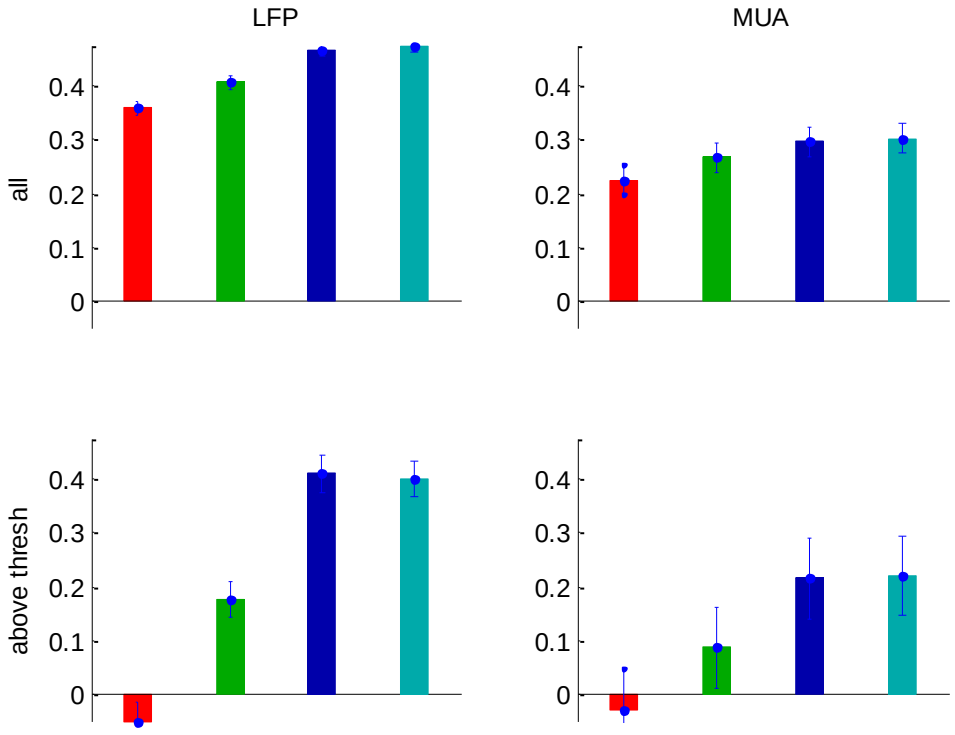

Figure 13. Correlations averaged across responses in individual recording sites– summary: mean correlation between neural responses to notes (**left**- LFP, **right** – MUA, **top** - to all notes, **bottom** – to above-threshold notes) and the different predictors tested – notes sound level (**red**); FRA predictions (**green**); combined surprise and sound level using the first surprise measure I ($-(log_2(p_i) - 1) * level$, **blue**) and combined surprise and sound level using the second surprise measure ($-log_2(r_i) * level$, **cyan**). Error bars mark the confidence interval around each mean as derived from post-hoc comparisons (Tukey Kramer test). Combining both surprise measures with sound level resulted in significantly higher correlations than sound level alone in all cases ($p<<0.01$).

		level	FRA-filtered level	Surprise and level (I)	Surprise and level(II)
LFP (47)	overall	0.36 ± 0.21 (39)	0.41 ± 0.22 (41)	0.47 ± 0.21 (43)	0.48 ± 0.23 (43)
	>threshold	-0.05 ± 0.13 (0)	0.18 ± 0.15 (4)	0.41 ± 0.16 (22)	0.40 ±0.17 (26)
MUA (30)	overall	0.23 ± 0.23 (20)	0.27 ±0.24 (25)	0.30 ± 0.28 (22)	0.30 ± 0.29 (22)
	>threshold	-0.03 ± 0.12 (0)	0.09 ±0.24 (5)	0.22 ± 0.22 (12)	0.22 ± 0.22 (12)

Table 1. Correlations - summary. Correlations between the responses to the notes in the piece averaged across all the individual responsive sites for the different predictors tested (mean ± std). **Gray** rows – overall correlation of the responses to all notes; **white** rows – correlation of the responses to above-threshold notes. In brackets – the number of sites with significant correlations ($p<0.05$).

The most notable gain in explanatory power was obtained for the above-threshold correlations when surprise measures were combined with sound level. Figure 14 illustrates the probable origin of the added explanatory power. The combined surprise and sound level measures of the structurally surprising notes in seconds 130-190 of the piece were increased in relation to notes of the main motif. This was true for the notes in the G motif, and in particular for the Gs that interrupt the E♯s and F♯*s* from the main motif (red circles in Fig. 16).

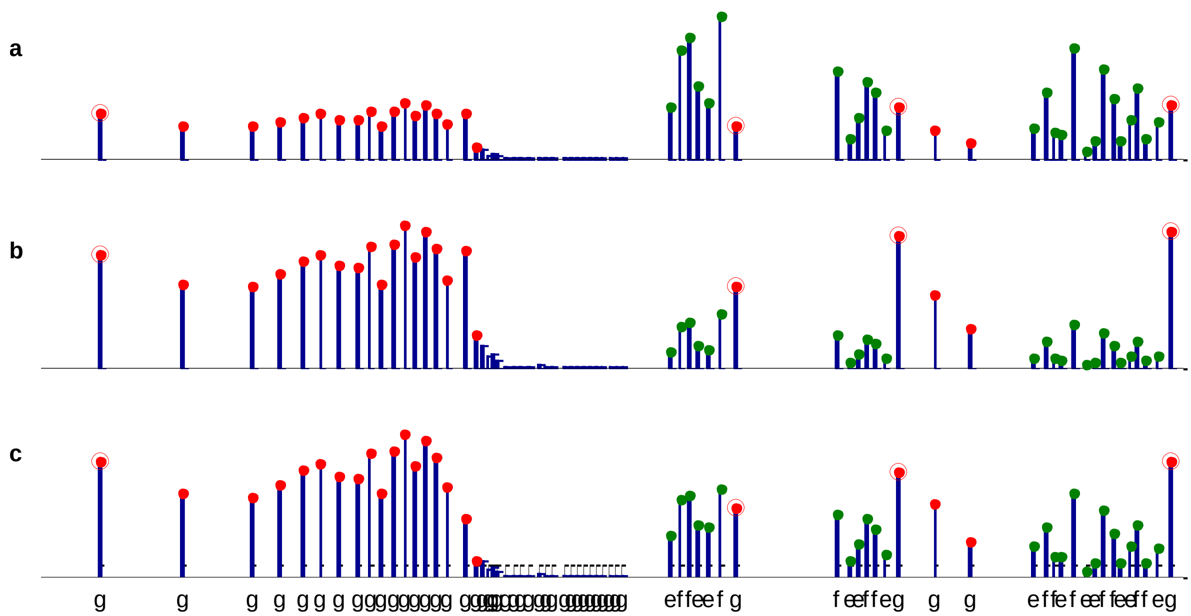

Figure 14. For all the 50 climax notes, the figure shows the note sound level (**a**) and the combined surprise and sound level measures (**b** and **c** for the first and second surprise measure respectively). Both of the combined measures of surprise and sound level were lower for the E♯ and F♯ notes that are part of the main motif (**green** dots) in relation to the G notes that make up the second motif (**red** dots). This is especially true for the Gs that interrupt the main motif (**red** circles).

It is possible that the responses to the notes of the main motif have been suppressed because of their predictability, either due to the prominence of their pitch class in the piece or to the establishment of the higher-order statistics of the principal motif. However, the findings are not inclusive as to the source of the larger responses.

Conclusion

This chapter explored the predictive framework under somewhat more natural conditions using extracellular responses of cortical neurons to music. The extracellular neural responses recorded in primary auditory cortex included local field potentials (LFP) and multiunit activity (MUA). Recordings were made while an adapted version of a musical piece, Ligeti's *Musica Ricercata II*, was presented. This piano piece comprises

only three pitch classes and has a very tight structure made up of a recurring theme constructed of two pitch classes. Because of its rather simple pitch structure, it is amenable to controlled experiments. Complex response patterns found both in the MUA and the LFP could not be satisfactorily explained by the spectral content of the sound but were roughly consistent with violations of expectations in the piece. Thus, the merit of this chapter is the realization of the predictive framework to a concrete complex auditory sound that is both very structured and meaningful. The results demonstrate the high explanatory potential of the notions developed above to neuronal responses as early as the rat primary auditory cortex. While additional controls are needed, the responses to this simple musical piece may constitute evidence of early processing of complex acoustic structure in the auditory system.

WHAT HAVE WE LEARNED?

We perceive sounds in terms of distinct objects. This we mostly do without paying attention or making any explicit effort. We hear a bird sing (a high pitched song), the neighbor talking (too soft to understand), cars passing by (coming nearer and then further away), fingers pounding on the keyboard (in rhythm). More often than not, these sound sources are concurrently active, and the waveforms they emit become entangled before entering the ear. Nevertheless, sounds of the sort encountered in everyday scenarios appear to us as a set of auditory objects, separate entities usually related to events in the environment and having a combination of perceptual features (e.g. pitch, timbre, loudness, duration, spatial location). This perceptual organization has but an indirect relationship to the vibrations of the tympanic membrane, especially when sounds are complex. The transformation from a pressure wave to auditory objects occurs in the auditory system through a long and complicated sequence of processes which are only partially understood. In this dissertation I suggested that it may be possible to bridge the gap between physical sounds and the perceptual auditory objects by applying a predictive framework for Auditory Scene Analysis (ASA).

The basic suggestion of this dissertation is to identify auditory objects with predictive models of auditory input. Predictive models can be generated directly from the statistics of the auditory data stream, tested against incoming input and adjusted according to the prediction success. By utilizing prediction errors, predictive models can be immediately evaluated, selected and adjusted without additional information. Under the predictive framework, the core process involved in recognizing an object in a sound is the extraction of regularities from past sensory data and the extrapolation of these regularities into the future. Forming a hypothesis regarding the organization of an

auditory scene into objects is to construct one or more predictive models for the sound. The test of a hypothesis is thus its ability to successfully extrapolate into the future and predict the incoming auditory input, not its agreement with the outside world. In this sense, the perceptual auditory object is inherently predictive. I propose that by the predictive modeling of auditory input the auditory system is able to identify auditory objects directly from the sound signal, even without direct access to information regarding the physical sources of the sound. Hence, brain processes that extract the regularities that govern the evolution of sounds in time underlie the recognition of objects in a sound signal. Specifically, the monitoring of prediction error – the mismatch between predictions and actual inputs – is suggested as key to the dynamic representation of a non-stationary auditory world, a trigger to the introduction of new auditory objects in the scene.

In this dissertation, I studied a number of different aspects of the prediction error in auditory processing of sounds in light of its potential role in signaling a new perceptual auditory object in a scene. This is in some contrast to the common computational approaches to ASA that attempt to reconstruct individual sources from a mixture (see Chapter 1). My main hypothesis is that prediction errors are continuously monitored by the brain and are explicitly represented in the responses of neurons in the auditory system. I specifically focused on the primary auditory cortex (A1), which has been suggested to play a central role in auditory scene analysis (e.g. Nelken and Bar-Yosef, 2008, Winkler et al., 2009), and has also been implicated in the encoding of regularity violations (e.g. Näätänen et al., 2001, Ulanovsky et al., 2003, Taaseh et al., 2011). Neurons in A1 are clearly sensitive to the short-term spectro-temporal content of sounds. My assumption is that their responses are also significantly modulated by the extent to which a current sound fulfills or violates the regularities detected in earlier acoustic context. The hypothesis is that A1 responses are facilitated when auditory statistical regularities are violated, and that enhanced responses signal the formation of a new auditory object. Although beyond the scope of the current work, the predictive

framework is in no way limited to A1 - it may well be that the predictive representations are hierarchically constructed in the different stations along the auditory pathway.

It should be noted that the suggested predictive formulation of auditory objects is related to prevailing notions regarding the role of prediction in perception. First, whereas sensory information is rarely complete, the common experience is that objects appear complete in perception. Perceptual sensory objects typically include information not actually transmitted from the sensory organs, but interpolated and extrapolated from it. This notion can be traced back to Helmholtz's view of perception as inference, to the Gestalt principles of perception, and to Gregory (1980) who suggested considering perceptions as scientific hypotheses (Gregory, 1980). It follows, that perception, like a scientific hypothesis, is validated by comparing its predictions to new information that was not used to generate the hypothesis. In vision, a likely source of validation data is information initially occluded and not projected to the retina - the fourth leg of a perceived table is rarely 'actually seen' but inferred. In case the extrapolation turns out to be wrong – a change of perspective reveals there are actually only three legs – perception is reassessed. In audition, while spectral data is also obviously inferred, a most useful extrapolation is in time – natural sounds have detectable underlying regular patterns that govern their evolution in time, that result from physics of the sound emission processes. Thus, the future emissions of a putative sound source are a likely validation data set for auditory based perceptions. In the present work 'predictive' has been used in this restricted temporal sense, to refer to the detection of dependencies within a time series of auditory samples and to the extrapolation of these dependencies to future time points.

Predictive accounts of sensory systems appeared both as a theoretical suggestion (Bar, 2007, Summerfield and Egner, 2009, Winkler et al., 2009, Clark, 2012) and in computational models (e.g. Creutzig et al., 2009, Friston and Kiebel, 2009a) in connection to the role of perception in guiding decisions and actions. The basic realization is that the functional role of sensory representations is to direct interactions

with the environment that lie exclusively in the future. Thus, a proactive or predictive aspect of sensory objects is beneficial for acting, particularly in an environment that is constantly changing. There is an evolutionary advantage in not only answering "what is going on out there right now?" but rather inferring what is coming next. The predictive framework for sensory perception thus has possible extensions that are beyond the scope of this work to the study of the perception-action cycle.

Context sensitivity of spectro-temporal models

In the first experiment described here (Chapter 3) I applied the traditional approach of using spectro-temporal receptive fields (STRFs) to model neuronal responses based on the short-term (<0.5 second) spectro-temporal content of the sound. STRFs are a linear approximation of the transformation from spectro-temporal content to neural responses. Under the assumption that neurons integrate signal energy linearly, STRFs can be generalized to predict responses to a new sound by convolving the STRF with a spectro-temporal representation of the sound (e.g. spectrograms, see Eggermont et al., 1983, Machens et al., 2004).

To test this, I contrasted STRFs calculated from responses of human cortical neurons to two acoustic contexts with very different statistics: 'real world' sounds, including speech and music, and random chords, strictly artificial sounds tailored to survey the spectral sensitivity of the neurons. First, STRFs were estimated from the responses of one set of sounds, correcting for first- and second-order correlations in the sound set (Theunissen et al., 2002). The Artificial STRFs, derived from the responses to the random sounds, resembled simple spectral filters, more selective than previously reported in the primary auditory cortex of any other mammal (except bats). The natural STRFs, calculated from the more natural sounds exhibited more specto-temporal structure. Moreover, whereas the STRFs did capture some of the variability of the neuronal response, their generalization across sound sets was limited. Each STRF predicted the responses to new sounds from the same ensemble used for its estimation than the STRF derived from the

other sound ensemble. Thus, artificial STRFs were worse in predicting responses to the natural sounds than natural STRFs and vice versa. Thus, the STRFs could not qualify as a general model for neuronal responses of these cortical units, regardless of acoustic context. The results showed that the short term spectro-temporal content did not entirely determine the responses when sounds consisted of elaborate auditory scenes instead of random, artificial sounds. This is consistent with earlier reports that the STRF of cortical neurons differs significantly when estimated by natural stimuli vs. artificial stimuli (Theunissen et al., 2000, Theunissen et al., 2001, David et al., 2009).

These findings were consistent with the notion that sounds of the kind encountered in real-world scenarios engage additional processing mechanisms in A1. The suggestion I made is that these cortical mechanisms are best understood under the framework of ASA. Further evidence discussed below supports the notion that in addition to the encoding of spectro-temporal features, neurons in A1 encode the appearance of an auditory object in the scene.

Change detection and object detection - ERPs

Chapter 4 used human MEG data to demonstrate the advantages of considering the composition of complex sounds in terms of objects when interpreting neural responses. I studied the neuronal representations of masked tones by comparing the responses to the maskers with the responses to the same maskers accompanied by tones. This comparison revealed a significant magnetic response locked to tone onset that appeared only in some of the tone and masker combinations used in the experiment. Importantly, the response was not correlated with the detectability of acoustic change. Rather the response was only elicited when a tone was heard as a new separate object in the presence of noise. I argued that this response, termed 'object potential', is indicative of a specific perceptual change in the auditory scene; namely, the appearance of a new object.

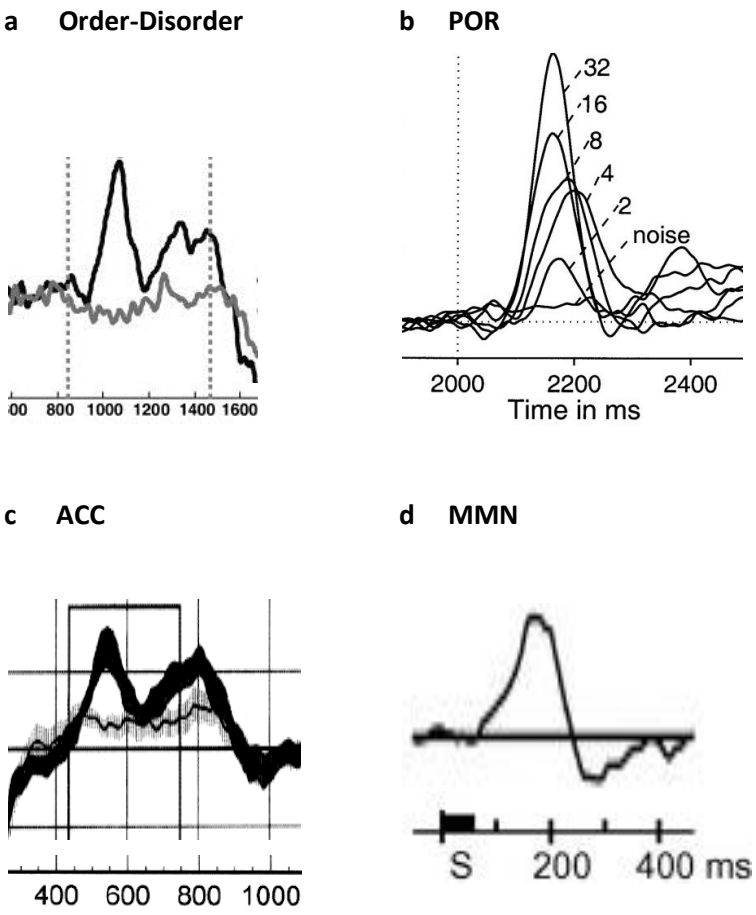

Figure 1. Different protocols, similar pattern. ERPS detected in response to auditory edges of unattended sounds in various procedures, all with a prominent peak at ~150 ms post change onset. For more details, see the original publications cited below. **a.** The transition between tone pips of random frequency to pips of constant frequency occurred at 840 ms. Adapted from Chait et al. (2012). **b.** The transition from noise to iterated ripples noise occurred at 2000 ms. Adapted from Krumbholz et al. (2003). **c.** The transition from a band of noise to tonal complex occurred between 390-400 ms. Adapted from Martin and Boothroyd (1999). **d**. The difference waves obtained by subtracting the responses to a standard stimulus (tone at 1KHz) from the responses to a deviant stimulus (tone at 1032 Hz). Adapted from Näätänen et al. (2007).

This interpretation links the object potential described in this work to earlier accounts relating auditory ERPs (event-related potentials) to change detection in sounds. In this literature, an abrupt change is usually introduced to an ongoing sound by varying a

specific feature of interest (e.g. pitch, timbre, vowel identity). This is done in a manner designed to imitate natural sounds, and in order to separate the response to the induced change from generic responses to stimulus energy-onset. For instance, Chait et al. (2007a, 2007b) identified a magnetic field that accompanied a transition between "disorder," modeled as a sequence of random frequency tone pips, and "order," modeled as a constant tone. The magnetic response had a temporal pattern similar to the temporal pattern of the object potential (Fig. 1a). Similarly, the Pitch Onset Response (POR; Krumbholz et al., 2003, Gutschalk et al., 2004, see Fig. 1b) in the term used to describe a magnetic response locked to transitions between no pitch (irregular click trains/white noise) and pitch (irregular click trains/iterated ripples noise) that were not accompanied by a change in energy or spectral content. Changes in frequency, complexity, intensity and speech syllables in ongoing sounds are typical stimuli reported to elicit the electric field termed Acoustic Change Complex (ACC), a cortical potential with components similar those that appear at the onset of a sound (the so called 'P1-N1-P2 complex', see Martin and Boothroyd, 1999, 2000, Fig. 1c). There seems to be a considerable overlap in appearance and timing among the various change related potentials (all peak in the 100-200 ms range, see Fig. 1). The differences in terminology may have more to do with the different experimental procedures than with underlying processing mechanisms or cortical generators. Notably, all the above responses were interpreted as marking a local edge in some low level auditory dimension. I hypothesize that the object potential was evoked by a change in a high level, perceptual dimension of the sound - the number of auditory objects in the scene. This raises the possibility that other reported responses were not evoked by the physical change per se, but rather by its effect on the scene organization.

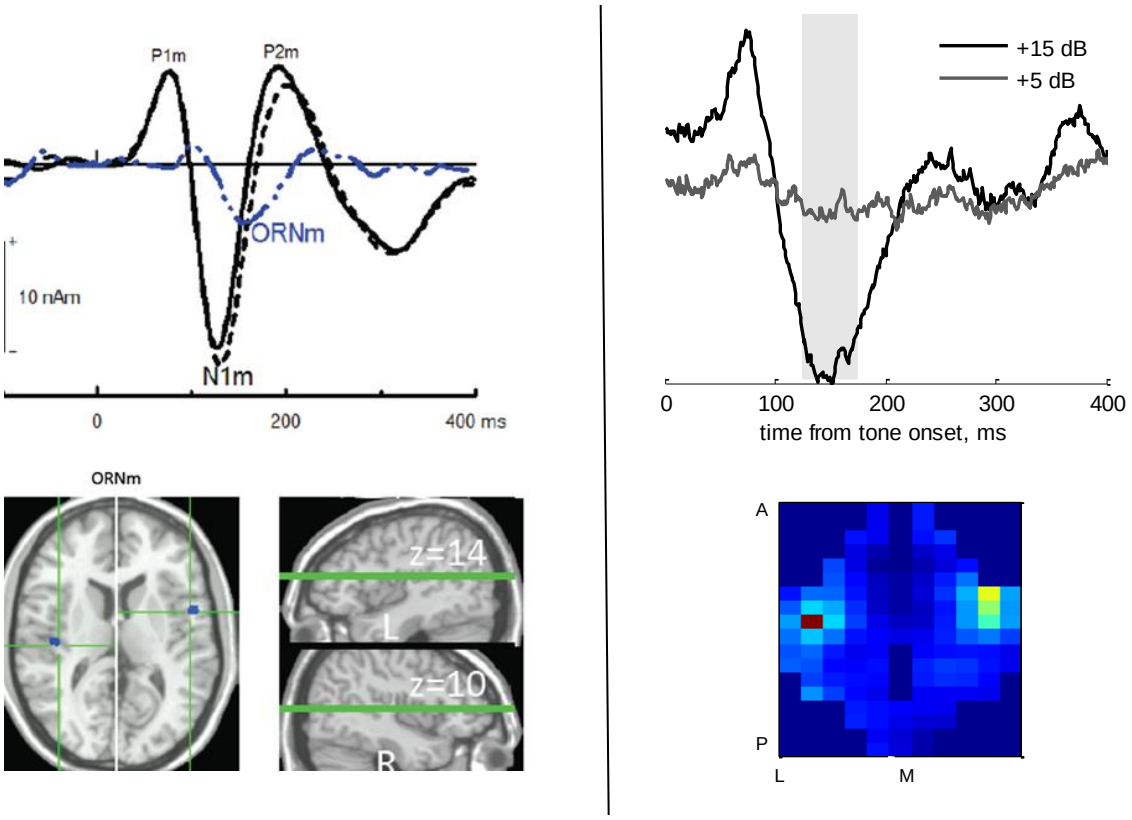

Figure 2. Left: The magnetic version of the ORN (ORNm) is the difference wave between responses elicited by a tuned stimulus (perceived as one sound) and a mistuned stimulus (typically perceived as two concurrent sounds). **Top:** The pattern in time, in relation to tone onset. **Bottom**: location of the peak maxima (green cross hairs) for the fitted ORNm sources (blue). Adapted from Snyder et al. (2012). **Right:** The pattern in time (**top**) and mean scalp distribution (**bottom**) of the object potential presented in this work (adapted from Chapter 4, Figs. 9a and 10). Scalp distribution was calculated by averaging all the panels in Figure 10 and smoothing with a 3X3 moving window (the product of two Hanning windows along the two axes).

Auditory ERPs are traditionally related to sequential object segregation - the segregation between auditory objects along the time axis. By contrast, the Object-Related Negativity (ORN) has been linked to simultaneous segregation - the segregation of auditory objects along the frequency axis. Alain et al. (2002) manipulated the tuning of harmonic complexes so that listeners perceived either one or two concurrent sounds, according to

the level of mistuning. The perception of the mistuned harmonic as a separate sound was associated with a response that had a first peak about 150 ms after sound onset. The amplitude of the ORN correlated with the likelihood of hearing two concurrent sound objects (Alain et al., 2001). McDonald and Alain (2005) further showed a correlation between the ORN and changes in grouping induced by a manipulation of the spatial location of the sounds. The ORN is thus very similar in its interpretation to the object potential. Moreover, the magnetic version of the ORN (ORNm) showed a temporal pattern and scalp distribution similar to those found for object potential (Snyder et al., 2012, see Fig. 2). Thus, with the caution required by the difference in procedures, I postulate that both the ORN and the object potential reflect the neural processing that accompanies the separation of concurrently presented sounds into distinct perceptual objects.

Deviance detection and ASA

The Mismatch Negativity (MMN) is a much studied pre-attentive ERP, emitted not only when a change occurs in an auditory signal, but when an acoustic regularity is broken (Näätänen et al., 2007). This potential, usually peaking at about 150-200 ms from change onset (Fig. 1d), is thought to reflect the discrepancy between expectations and actual input (e.g. Sams et al., 1985, Näätänen, 1992, Näätänen et al., 1993) or else the update of an internal representation of a regularity due to violation of the regularity (Winkler et al., 1996, Winkler, 2003). There is converging evidence that the auditory processes that generate MMN originate, in the first place, in the auditory cortex (see summary in Näätänen et al., 2001). MMN appears in the context of purely abstract regularities, including ones that only manifest at very long timescales (Winkler et al., 1996, Schröger et al., 2007, reviewed in Paavilainen, 2013). Electrical activity with patterns similar to that of the MMN are elicited by violations of musical expectations (see a review in Näätänen et al., 1994, Rohrmeier and Koelsch, 2012). Recent interpretations of the MMN link it directly to ASA (Winkler et al., 2009), and to the

adjustment of probabilistic models of the environment following a mismatch between the existing model and the actual sounds (Lieder et al., 2013).

For a deviance to be detected, some regular pattern in the sound must first be identified. What happens when more than one sound source is concurrently active? Typically, the regular patterns that identify one sound source overlap in time with signals emitted from the other active sound sources. It is generally assumed that regularity extraction cannot be performed upon a mixture. In this view, the auditory scene is first organized into separate entities, and only then are processes of regularity extraction applied to each object (usually termed stream in this context. e.g. Sussman, 2005). A recent study of MMN (Bendixen et al., 2012) challenges this view. Bendixen et al. (2012) found that MMN was elicited in response to violations of regularities in a sequence of non-adjacent sounds that was played together with other intervening sounds, but only when the intervening sequence itself also contained a regular pattern in time. Namely, when the intervening sounds varied randomly, the auditory system was unable to detect a deviance from an established regularity. Consistent evidence was obtained by Andreou et al. (2011) who tested the effect of rhythmic regularity on the segregation of sounds. In specific conditions, they showed that stream segregation is facilitated when a temporal regularity is introduced in the intervening sequence. The findings from both MMN studies suggest that regularities are available to the auditory system as a cue for identifying sound sources, and that regularity extraction may precede grouping. This is consistent with the notion advocated here that regularity extraction forms the basis of perceptual auditory objects.

Object detection in a pressure wave – the prediction error

Most studies of auditory object identification use controlled auditory stimuli, often consisting of only a few elements. As such, the possibilities for auditory source classification are limited and are typically manipulated explicitly in the experiment. This is generally not the case with complex auditory stimuli that approximate natural

acoustic environments. In order to study such acoustic scenarios without a-priori assuming their perceptual organization, general procedures are needed for identifying auditory objects in acoustic signals, that not limited to the specific scenario for which they were developed. The suggestion to identify auditory objects with predictive models of sounds is both a conceptual and a practical step forward. To test the utility of this approach, I used the Information Bottleneck (IB) method developed by N. Tishby (Tishby et al., 2000) to construct predictive models of sounds. The method employs information theory to extract relevant information from one variable with respect to another in a principled way. To evaluate the predictive framework I combined the general procedure based on first principles with concrete experimental scenarios. Using the IB method with minimal additional assumptions, predictive models were extracted directly from the pressure waves of complex sounds used in experiments. The prediction error - the mismatch between actual sound samples and the predictions of a predictive model – were then tested against the experimental results. Importantly, in the IB method, the prediction error is not a property of the sound sequence by itself, but a product of both the predictive model (the internal representation) and the actual sequence of sound samples. In this sense, the prediction error is a flexible measure that can change with task and listening context, even when the stimulus input sequence is unchanged.

Prediction error has been suggested as central to learning (Schultz and Dickinson, 2000), memory formation (Ranganath and Rainer, 2003), motor control (Shadmehr et al., 2010) and decision-making (Doya, 2007). The prevalent predictive coding theories of brain function today suggest the minimization of prediction error as a primary objective, with sensory perception mathematically formulated as a hierarchical generative model that dynamically adapts to minimize prediction error by a cascade of processing schemes (e.g. Rao and Ballard, 1999, Lee and Mumford, 2003, Friston and Kiebel, 2009b). Note that some of these studies specifically distinguish terms that were used interchangeably throughout this work. Lieder et al. (2013) for example, distinguishes between the encoding of surprise (a violation of prediction), prediction error (a violation, and also its direction) and model adjustment (index of an actual update of an internal model). Such

distinctions were not made here, as both the formulation and the data were not sufficiently detailed to support them.

Feasibility of the predictive approach in experimental scenarios

In Chapter 5 I applied the predictive formulation to sounds from a set of behavioral experiments designed to test the recognition of timbre, including musical instruments and human singing. The entire sound ensemble was treated as the acoustic context for predictive modeling. The prediction errors, derived from the predictions of general models generated for the entire ensemble, showed a sensitivity to the timbre categories defined in the experiments. The errors were consistently larger for voices than for musical instruments, matching the special behavioral status of this category. Furthermore, in accordance with the psychoacoustics, successful classification of sounds as voices or instruments was achieved based on prediction errors calculated with very short sound samples (~10 ms). Moreover, a spectral interpretation of the prediction measure assigned the source of the differences between categories to the higher formants of the vowel sounds. This important equivalence between temporal prediction and spectral analysis suggests that the predictive framework can be implemented with quantities available to the periphery of the auditory system. In this view, predictive modelling can function as a principled method for highlighting spectral differences between sound sets. Alternatively, the spectral representation of a sound can be considered as a way to estimate the average prediction error. More generally, the results highlight the feasibility of the predictive framework and its potential explanatory power - the prediction error corresponded in this implementation both to behavioral results and to the physical properties of the sounds, as well as pointed to a possible link between the two.

Encoding of prediction error in A1

Chapter 6 directly tested the proposal that prediction error is encoded in the primary auditory cortex, such that the neuronal response is a function of two aspects of a sound:

its spectro-temporal content and the degree to which it was expected, based on the recent acoustic content. I analyzed neuronal responses of rat cortical neurons to Ligeti's Musica Ricercata no. 2, a short piece that contains many relatively well-separated piano notes. Music in general is a good candidate for studying the effects of acoustic context on the representation of individual sounds. To a large extent, it is the organization that makes a collection of sounds into the music and that differentiates music from natural sound scenes (Nelken, 2011). For the sake of analysis, the well-separated piano notes in the musical piece can be regarded either as individual sounds or as part of the general context of the music. The basic assumption was that the context-based probability of a piano note to occur would modulate the neuronal response the note would have elicited in isolation. As could be expected, the first prominent effect on neuronal response was that of note level - the loud notes in the music generally elicited substantially larger responses in A1 than the soft notes. But sound level alone, calculated for each individual note regardless of the acoustic context, was not enough to account for the neuronal response. Next, a prediction error was associated with each of the main notes in the piece, which indicated how well the note fit the prediction of the predictive models. The predictive models were generated in two different ways from the statistics of the entire piece. The prediction error was in itself significantly and positively correlated with the neural response to the piano notes. Moreover, combining the level of the note with its predictability (by multiplication) explained a significantly large portion of the neural response, more than any of the measures separately.

This finding provides strong support for the claim that A1 neurons encode prediction error in addition to the encoding of the spectro-temporal features of the sound. By the identification of predictive models with objects, the predictive framework of ASA sets the encoding of prediction error in A1 in a wider functional context. What would have traditionally been discussed in terms of deviance detection can be interpreted as indication of object representation in the responses of single neurons in A1. The important consequence is that two lines of existing evidence regarding the encoding of sound in A1 are unified in one framework - that A1 responses are related to violations of

the statistical regularities of sounds, and that cortical responses are indicative of the appearance of a new auditory object in the auditory scene.

Future directions

The current work provided a connection between the predictive formulation of auditory objects, the algorithmic implementation of predictive modeling of sound signals, and experimental results. The more general issue is the correspondence between the vibrating objects of the world and perceptual auditory objects. The predictive framework offers insights into *what* information is being encoded by the auditory system (that which is relevant to the future of the auditory data stream), *how* it is encoded (in the form of predictive models and their associated prediction errors), *why* (to direct interactions with the environment that lie in the future), and even *where* the information is being encoded (perhaps ubiquitously in the auditory system, but predominantly starting in A1). Many questions remain to be further explored. First, the evidence I interpreted in favor of the predictive ASA consisted of effects expected from the processing of prediction errors (Chap. 5, 6) or from the identification of a new auditory object in a scene (Chap. 4). However, the experiments did not explicitly test the hypothesis that the core of ASA is the processing of prediction errors. The findings justify an effort to experimentally test for a more concrete link between auditory scene analysis and predictive modeling. The magnetic object potential, for example, could be used to test the correspondence between violations of regularities, conscious perception of objects, and neuronal responses in human subjects.

In the applications of predictive modeling described in this work, the probability distribution characterizing the acoustic environment, $p(past, future)$, was assumed to be known in advance. This implies a stationary acoustic environment and a learning stage that ends prior to the behavior being tested. The auditory system, however, is expected to interact continuously with natural environments that are non-stationary, and have a probability distribution $p(past, future)$ that changes over time. Furthermore, the notion

of a complete set of training data, assumed to be stored in memory and available prior to prediction, is inappropriate for the auditory system. An online learning scenario is more fitting, particularly as auditory input is essentially sequential by nature. Online learning algorithms are designed to be updated after each new input. An incoming sensory input can serve first as a test of the validity of the predictive model, and then as a training data sample to adjust the predictive model in the case of a mismatch. Thus, learning can be fast, can start from little available data, and be achieved with bounded memory usage. Prediction errors serve in an online scenario as signals that guide the establishment of the internal predictive representation. A concrete online algorithm for predictive modeling of non-stationary sound could also provide insights as to the brain processes that govern ASA in real-world sounds.

There are possible immediate practical implications for a better understanding of how the auditory system performs the transformation between pressure waves and perceptual objects. Current computer based systems are no match, in accuracy or speed, to the most basic abilities of the auditory system in recognizing auditory sources and in adjusting to changes in the environment. This is exemplified in the performance of auditory recognition systems, including state-of-the-art speech recognition systems and sonic-aware robots, particularly in conditions of noisy environments, intervening auditory sources and intrusions. The current use of artificial sound systems is limited due to difficulties in separating interferences from the target objects and recognizing the appearance of new objects in the scene. A related source of interest and research is the relative inadequacy of conventional hearing aids in confronting the same issues. In spite of restoration of lost sensitivity through amplification and dynamic range compression, hearing-impaired subjects still have difficulties separating mixtures of sounds in everyday scenarios. An understanding of the solutions applied by the auditory system to solve the mathematically ill-posed problem of perceptual organization might inform the construction of future sound classification systems that could enhance the performance of hearing aids.

The predictive framework was presented here in relation to auditory objects and auditory scene analysis. The defining feature of a perceptual object in the predictive ASA - the extrapolation of past sensory input into the future - is however not confined in any way to the auditory modality. First, the identification between predictive models and perceptual objects is general enough to be applied to the representation of visual objects as well as representations of multimodal objects. The temporal dimension of objects has not been as widely studied in vision as in audition. In most experiments on visual objects, subjects are presented with still pictures, and models of visual grouping are predominantly spatial and static. The temporal aspect of visual object formation is usually addressed in the context of either motion representation or in special cases of grouping without spatial cues (for example, dot stimuli and the so-called grouping 'by common fate'). However, in everyday visual scenes, perception is clearly far from static. In fact, a truly static visual object does not exist - without eye movements and the changes they entail in visual input, vision disappears. Parts of objects typically become visible at different times, as a perceiver and objects move and interact. Much important information can in fact only be derived by relating visual inputs from different times to each other. Specifically, the processing of temporal order is crucial to visual perception, as is the extrapolation of visual inputs into the future. Recent scientific approaches to visual object representation indeed include an explicit temporal dimension. For instance, Freyd (1987, 1993) demonstrated predictive aspects and heightened sensitivity to implicit dynamic information in the mental representation of static pictures. The notion of extrapolation into the future has been explicitly invoked in the interpretation of systematic errors in localizing visual objects (e.g. representational momentum and the related flash-lag effect, see review in Hubbard, 2005, 2013). Recent accounts of deviance detection in vision suggest a predictive aspect to visual MMN, based on similarities to the auditory MMN (e.g. Kimura, 2012, Winkler and Czigler, 2012). The predictive framework, accordingly adapted to vision, could provide a unified framework for the evidence regarding the non-static nature of visual objects and enrich the study of visual objects representation. Obviously, establishing formal similarities

between predictive models in the visual and auditory domains could facilitate the treatment of multisensory representations. In due course, the utilization of time as a principle dimension for object recognition and validation might not only prove beneficial for auditory research but could repay the conceptual debt of the auditory research community to the field of visual perception.

REFERENCES

Agus TR, Suied C, Thorpe SJ, Pressnitzer D (2010) Characteristics of human voice processing. In: Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, pp 509-512: IEEE.

Agus TR, Suied C, Thorpe SJ, Pressnitzer D (2012) Fast recognition of musical sounds based on timbre. The Journal of the Acoustical Society of America 131:4124.

Alain C, Arnott SR, Picton TW (2001) Bottom–up and top–down influences on auditory scene analysis: Evidence from event-related brain potentials. Journal of Experimental Psychology: Human Perception and Performance 27:1072.

Alain C, Schuler BM, McDonald KL (2002) Neural activity associated with distinguishing concurrent auditory objects. The Journal of the Acoustical Society of America 111:990.

Andreou L-V, Kashino M, Chait M (2011) The role of temporal regularity in auditory segregation. Hearing research 280:228-235.

Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Network: Computation in neural systems 3:213-251.

Attneave F (1954) Some informational aspects of visual perception. Psychological review 61:183.

Averbeck BB, Romanski LM (2006) Probabilistic encoding of vocalizations in macaque ventral lateral prefrontal cortex. The Journal of neuroscience 26:11023-11033.

Banai K, Ahissar M (2004) Poor frequency discrimination probes dyslexics with particularly impaired working memory. Audiology and Neurotology 9:328-340.

Banai K, Ahissar M (2006) Auditory processing deficits in dyslexia: task or stimulus related? Cerebral Cortex 16:1718-1728.

Bar-Yosef O, Rotman Y, Nelken I (2002) Responses of neurons in cat primary auditory cortex to bird chirps: effects of temporal and spectral context. The Journal of neuroscience 22:8619-8632.

Bar M (2007) The proactive brain: using analogies and associations to generate predictions. Trends in cognitive sciences 11:280-289.

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. Sensory communication 217-234.

Bartlett EL, Sadagopan S, Wang X (2011) Fine frequency tuning in monkey auditory cortex and thalamus. Journal of neurophysiology 106:849-859.

Bee MA, Klump GM (2004) Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. Journal of Neurophysiology 92:1088-1104.

Bee MA, Klump GM (2005) Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. Brain, Behavior and Evolution 66:197-214.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403:309-312.

Benasich AA, Tallal P (2002) Infant discrimination of rapid auditory cues predicts later language impairment. Behavioural Brain Research 136:31-49.

Bendixen A, Schröger E, Ritter W, Winkler I (2012) Regularity extraction from non-adjacent sounds. Frontiers in psychology 3.

Bendor D, Wang X (2005) The neuronal representation of pitch in primate auditory cortex. Nature 436:1161-1165.

Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. Neural Computation 13:2409-2463.

Bleeck S, Ives T, Patterson RD (2004) Aim-mat: the auditory image model in MATLAB. Acta Acustica United with Acustica 90:781-787.

Bregman AS (1990) Auditory scene analysis: The perceptual organization of sound.

Brown GJ (1992) Computational auditory scene analysis: A representational approach. In: Computer Science, vol. Ph.D.: University of Sheffield.

Campbell RA, Schulz AL, King AJ, Schnupp JW (2010) Brief sounds evoke prolonged responses in anesthetized ferret auditory cortex. Journal of neurophysiology 103:2783-2793.

Carlyon RP (2004) How the brain separates sounds. Trends in cognitive sciences 8:465-471.

Chait M, Poeppel D, de Cheveigné A, Simon JZ (2007a) Processing asymmetry of transitions between order and disorder in human auditory cortex. The Journal of neuroscience 27:5207-5214.

Chait M, Poeppel D, Simon JZ (2007b) Human auditory cortical processing of transitions between 'order'and 'disorder'. In: Hearing–From Sensory Processing to Perception, pp 323-331: Springer.

Chait M, Ruff CC, Griffiths TD, McAlpine D (2012) Cortical responses to changes in acoustic regularity are differentially modulated by attentional load. Neuroimage 59:1932-1941.

Chechik G, Anderson MJ, Bar-Yosef O, Young ED, Tishby N, Nelken I (2006) Reduction of information redundancy in the ascending auditory pathway. Neuron 51:359-368.

Chechik G, Globerson A, Tishby N, Weiss Y (2005) Information bottleneck for Gaussian variables. In: Journal of Machine Learning Research, pp 165-188.

Chechik G, Nelken I (2012) Auditory abstraction from spectro-temporal features to coding auditory entities. Proceedings of the National Academy of Sciences 109:18968-18973.

Ciocca V (2008) The auditory organization of complex sounds. Frontiers in bioscience : a journal and virtual library 13:148-169.

Clark A (2012) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 1-86.

Condon CD, Weinberger NM (1991) Habituation produces frequency-specific plasticity of receptive fields in the auditory cortex. Behavioral neuroscience 105:416.

Cooke M, Brown GJ (1993) Computational auditory scene analysis: Exploiting principles of perceived continuity. Speech Communication 13:391-399.

Creutzig F, Globerson A, Tishby N (2009) Past-future information bottleneck in dynamical systems. Physical Review E 79:041925.

Darwin CJ (1997) Auditory grouping. Trends in cognitive sciences 1:327-333.

David SV, Mesgarani N, Fritz JB, Shamma SA (2009) Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. The Journal of Neuroscience 29:3374-3386.

deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280:1439-1443.

Deouell LY, Heller AS, Malach R, D'Esposito M, Knight RT (2007) Cerebral responses to change in spatial location of unattended sounds. Neuron 55:985-996.

Dimitrov AG, Miller JP (2001) Neural coding and decoding: communication channels and quantization. Network: Computation in Neural Systems 12:441-472.

Doya K (2007) Bayesian brain: Probabilistic approaches to neural coding: MIT Press.

Duda RO, Lyon RF, Slaney M (1990) Correlograms and the separation of sounds. In Proceedings Asilomar Conference on Signals, Systems and Computers

Eggermont J, Johannesma P, Aertsen A (1983) Reverse-correlation methods in auditory research. Quarterly reviews of biophysics 16:341-414.

Ehret G, Merzenich MM (1988) Complex sound analysis (frequency resolution, filtering and spectral integration) by single units of the inferior colliculus of the cat. Brain Research Reviews 13:139-163.

Ehret G, Schreiner C (1997) Frequency resolution and spectral integration (critical band analysis) in single units of the cat primary auditory cortex. Journal of Comparative Physiology A 181:635-650.

Elhilali M, Shamma SA (2008) A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. The Journal of the Acoustical Society of America 124:3751.

Ellis DP (1996) Prediction-driven computational auditory scene analysis. Massachusetts Institute of Technology.

Evans E (1977) Frequency selectivity at high signal levels of single units in cochlear nerve and nucleus. Psychophysics and physiology of hearing 185-192.

Fawcett T (2006) An introduction to ROC analysis. Pattern recognition letters 27:861-874.

Fishman YI, Arezzo JC, Steinschneider M (2004) Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. The Journal of the Acoustical Society of America 116:1656.

Fishman YI, Reser DH, Arezzo JC, Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hearing research 151:167-187.

Fishman YI, Steinschneider M (2012) Searching for the mismatch negativity in primary auditory cortex of the awake monkey: deviance detection or stimulus specific adaptation? The Journal of Neuroscience 32:15747-15758.

Freyd JJ (1987) Dynamic mental representations. Psychological review 94:427.

Freyd JJ (1993) Five hunches about perceptual processes and dynamic representations. Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience 99-119.

Fried I, Wilson CL, Maidment NT, Engel Jr J, Behnke E, Fields TA, Macdonald KA, Morrow JW, Ackerson L (1999) Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients: technical note. Journal of neurosurgery 91:697-705.

Friedman N, Mosenzon O, Slonim N, Tishby N (2001) Multivariate information bottleneck. In: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, pp 152-161: Morgan Kaufmann Publishers Inc.

Friston K (2005) A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences 360:815-836.

Friston K, Kiebel S (2009a) Cortical circuits for perceptual inference. Neural Networks 22:1093-1104.

Friston K, Kiebel S (2009b) Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B: Biological Sciences 364:1211-1221.

Gaese BH, Ostwald J (2001) Anesthesia changes frequency tuning of neurons in the rat primary auditory cortex. Journal of neurophysiology 86:1062-1066.

Gifford III GW, Cohen YE (2005) Spatial and non-spatial auditory processing in the lateral intraparietal area. Experimental brain research 162:509-512.

Gordon N, Shackleton TM, Palmer AR, Nelken I (2008) Responses of neurons in the inferior colliculus to binaural disparities: insights from the use of Fisher information and mutual information. Journal of neuroscience methods 169:391-404.

Green DM, Swets JA (1966) Signal detection theory and psychophysics: Wiley New York.

Gregory RL (1980) Perceptions as hypotheses. Philosophical Transactions of the Royal Society of London B, Biological Sciences 290:181-197.

Griffiths TD, Warren JD (2004) What is an auditory object? Nature Reviews Neuroscience 5:887-892.

Gutschalk A, Patterson RD, Scherg M, Uppenkamp S, Rupp A (2004) Temporal dynamics of pitch in human auditory cortex. Neuroimage 22:755-766.

Haidarliu S (1996) An anatomically adapted, injury-free headholder for guinea pigs. Physiol Behav 60:111-114.

Hall JW, Haggard MP, Fernandes MA (1984) Detection in noise by spectro-temporal pattern analysis. The Journal of the Acoustical Society of America 76:50.

Heinz MG, Colburn HS, Carney LH (2001) Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. Neural Computation 13:2273-2316.

Hotelling H (1935) The most predictable criterion. Journal of educational Psychology 26:139.

Howard III MA, Volkov IO, Abbas PJ, Damasio H, Ollendieck MC, Granner MA (1996) A chronic microelectrode investigation of the tonotopic organization of human auditory cortex. Brain research 724:260-264.

Hsu A, Borst A, Theunissen FE (2004) Quantifying variability in neural responses and its application for the validation of model predictions. Network: Computation in Neural Systems 15:91-109.

Hubbard TL (2005) Representational momentum and related displacements in spatial memory: A review of the findings. Psychonomic Bulletin & Review 12:822-851.

Hubbard TL (2013) Do the flash-lag effect and representational momentum involve similar extrapolations? Frontiers in psychology 4.

Huron DB (2006) Sweet anticipation: Music and the psychology of expectation: The MIT Press.

Itatani N, Klump GM (2009) Auditory streaming of amplitude-modulated sounds in the songbird forebrain. Journal of neurophysiology 101:3212-3225.

Jacoby N, Jakoby H, Lieder N, Tishby N, Ahissar M (2012) Tapping Working Memory with Sensorimotor Synchronization. In: Israel Society for Neuroscience's 2012 Annual Conference Eilat, Israel.

Kanwal JS, Medvedev AV, Micheyl C (2003) Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment. Network: Computation in Neural Systems 14:413-435.

Kimura M (2012) Visual mismatch negativity and unintentional temporal-context-based prediction in vision. International Journal of Psychophysiology 83:144-155.

Krumbholz K, Patterson R, Seither-Preisler A, Lammertmann C, Lütkenhöner B (2003) Neuromagnetic evidence for a pitch processing center in Heschl's gyrus. Cerebral Cortex 13:765-772.

Las L, Shapira A-H, Nelken I (2008) Functional gradients of auditory sensitivity along the anterior ectosylvian sulcus of the cat. The Journal of Neuroscience 28:3657-3667.

Las L, Stern EA, Nelken I (2005) Representation of tone in fluctuating maskers in the ascending auditory system. The Journal of neuroscience 25:1503-1513.

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. JOSA A 20:1434-1448.

Levy DA, Granot R, Bentin S (2003) Neural sensitivity to human voices: ERP evidence of task and attentional influences. Psychophysiology 40:291-305.

Lieder F, Daunizeau J, Garrido MI, Friston KJ, Stephan KE (2013) Modelling Trial-by-Trial Changes in the Mismatch Negativity. PLoS computational biology 9:e1002911.

Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. The Journal of neuroscience 24:1089-1100.

Malone BJ, Scott BH, Semple MN (2002) Context-dependent adaptive coding of interaural phase disparity in the auditory cortex of awake macaques. The Journal of neuroscience 22:4625-4638.

Martin BA, Boothroyd A (1999) Cortical, auditory, event-related potentials in response to periodic and aperiodic stimuli with the same spectral envelope. Ear and Hearing 20:33-44.

Martin BA, Boothroyd A (2000) Cortical, auditory, evoked potentials in response to changes of spectrum and amplitude. The Journal of the Acoustical Society of America 107:2155.

Matthen M (2010) On the diversity of auditory objects. Review of Philosophy and Psychology 1:63-89.

McArthur G, Bishop DV (2005) Speech and non-speech processing in people with specific language impairment: A behavioural and electrophysiological study. Brain and Language 94:260-273.

McDonald KL, Alain C (2005) Contribution of harmonicity and location to auditory object formation in free field: evidence from event-related brain potentials. The Journal of the Acoustical Society of America 118:1593.

Mellinger DK (1991) Event formation and separation in musical sound. In: CCRMA, vol. Ph.D.: Stanford University.

Meyer LB (1956) Emotion and Meaning in Music: University of Chicago Press.

Micheyl C, Tian B, Carlyon RP, Rauschecker JP (2005) Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron 48:139-148.

Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. Journal of Neurophysiology 87:516-527.

Moore BC (1982) In: An introduction to the psychology of hearing, pp 74-114: Academic Press (London and New York).

Moore BC (1999) Neurobiology: Modulation minimizes masking. Nature 397:108-109.

Moore BC (2003) Temporal integration and context effects in hearing. Journal of Phonetics 31:563-574.

Moore BC, Gockel H (2002) Factors influencing sequential stream segregation. Acta Acustica United with Acustica 88:320-333.

Moshitch D, Las L, Ulanovsky N, Bar-Yosef O, Nelken I (2006) Responses of neurons in primary auditory cortex (A1) to pure tones in the halothane-anesthetized cat. Journal of neurophysiology 95:3756-3769.

Mukamel R, Gelbard H, Arieli A, Hasson U, Fried I, Malach R (2005) Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science 309:951-954.

Murray MM, Camen C, Andino SLG, Bovet P, Clarke S (2006) Rapid brain discrimination of sounds of objects. The Journal of neuroscience 26:1293-1302.

Näätänen R (1992) Attention and brain function: Psychology Press.

Näätänen R, Ilmoniemi RJ, Alho K (1994) Magnetoencephalography in studies of human cognitive brain function. Trends in neurosciences 17:389-395.

Näätänen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I (2001) 'Primitive intelligence'in the auditory cortex. Trends in neurosciences 24:283-288.

Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clinical Neurophysiology 118:2544-2590.

Näätänen R, Paavilainen P, Titinen H, Jiang D, Alho K (1993) Attention and mismatch negativity. Psychophysiology 30:436-450.

Nelken I (2002) In: Integrative functions in the mammalian auditory pathway, vol. 15 (Oertel, D. et al., eds), pp 358–416: Springer.

Nelken I (2011) Music and the auditory brain: where is the connection? Frontiers in human neuroscience 5.

Nelken I, Bar-Yosef O (2008) Neurons and objects: the case of auditory cortex. Frontiers in neuroscience 2:107.

Neuert V, Verhey JL, Winter IM (2004) Responses of dorsal cochlear nucleus neurons to signals in the presence of modulated maskers. The Journal of neuroscience 24:5789-5797.

Neyman J, Pearson ES (1992) On the problem of the most efficient tests of statistical hypotheses: Springer.

Nudds M (2010) What are auditory objects? Review of Philosophy and Psychology 1:105-122.

Paavilainen P (2013) The mismatch-negativity (MMN) component of the auditory event-related potential to violations of abstract regularities: A review. International Journal of Psychophysiology.

Pasnau R (1999) What is sound? The Philosophical Quarterly 49:309-324.

Patterson RD, Uppenkamp S, Johnsrude IS, Griffiths TD (2002) The processing of temporal pitch and melody information in auditory cortex. Neuron 36:767-776.

Phillips D, Semple M, Calford M, Kitzes L (1994) Level-dependent representation of stimulus frequency in cat primary auditory cortex. Experimental Brain Research 102:210-226.

Pressnitzer D, Meddis R, Delahaye R, Winter IM (2001) Physiological correlates of comodulation masking release in the mammalian ventral cochlear nucleus. The Journal of Neuroscience 21:6377-6386.

Qin L, Kitama T, Chimoto S, Sakayori S, Sato Y (2003) Time course of tonal frequency-response-area of primary auditory cortex neurons in alert cats. Neuroscience research 46:145-152.

Qin L, Wang JY, Sato Y (2008) Representations of cat meows and human vowels in the primary auditory cortex of awake cats. Journal of neurophysiology 99:2305-2319.

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435:1102-1107.

Ranganath C, Rainer G (2003) Neural mechanisms for detecting and remembering novel events. Nature Reviews Neuroscience 4:193-202.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience 2:79-87.

Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of “what” and “where” in auditory cortex. Proceedings of the National Academy of Sciences 97:11800-11806.

Read HL, Winer JA, Schreiner CE (2001) Modular organization of intrinsic connections associated with spectral tuning in cat auditory cortex. Proceedings of the National Academy of Sciences 98:8042-8047.

Recanzone GH, Guard DC, Phan ML (2000) Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. Journal of Neurophysiology 83:2315-2331.

Rohrmeier MA, Koelsch S (2012) Predictive information processing in music cognition. A critical review. International Journal of Psychophysiology 83:164-175.

Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nature neuroscience 2:1131-1136.

Rupp A, Las L, Nelken I (2007) Neuromagnetic representation of comodulation masking release in the human auditory cortex. In: Hearing–From Sensory Processing to Perception, pp 125-132: Springer.

Sams M, Hämäläinen M, Antervo A, Kaukoranta E, Reinikainen K, Hari R (1985) Cerebral neuromagnetic responses evoked by short auditory stimuli. Electroencephalography and clinical Neurophysiology 61:254-266.

Schnupp J, Nelken I, King A (2011a) Auditory neuroscience: The MIT Press.

Schnupp J, Nelken I, King A (2011b) Auditory Scene Analysis. In: Auditory neuroscience; pp 223 - 268: The MIT Press.

Schnupp JW, Mrsic-Flogel TD, King AJ (2001) Linear processing of spatial cues in primary auditory cortex. Nature 414:200-204.

Schröger E, Bendixen A, Trujillo-Barreto NJ, Roeber U (2007) Processing of abstract rule violations in audition. PLoS One 2:e1131.

Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. Annual review of neuroscience 23:473-500.

Schwarz DW, Tomlinson RW (1990) Spectral response patterns of auditory cortex neurons to harmonic complex tones in alert monkey (Macaca mulatta). Journal of neurophysiology 64:282-298.

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends in neurosciences 26:100-107.

Shadmehr R, Smith MA, Krakauer JW (2010) Error correction, sensory prediction, and adaptation in motor control. Annual review of neuroscience 33:89-108.

Slonim N, Tishby N (2000) Document clustering using word clusters via the information bottleneck method. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp 208-215: ACM.

Smaragdis P (2001) Redundancy reduction for computational audition, a unifying approach. Massachusetts Institute of Technology.

Snyder JS, Alain C (2007) Toward a neurophysiological theory of auditory stream segregation. Psychological bulletin 133:780.

Snyder JS, Gregg MK, Weintraub DM, Alain C (2012) Attention, awareness, and the perception of auditory scenes. Frontiers in psychology 3.

Suied C, Mesgarani N, Pressnitzer D, Slaney M (2010) Auditory Gisting. Retrieved Feb 21, 2014, from http://neuromorphs.net.

Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. Trends in cognitive sciences 13:403-409.

Sussman ES (2005) Integration and segregation in auditory scene analysis. The Journal of the Acoustical Society of America 117:1285.

Swanson LW (1992) Brain Maps: Structure of the Rat Brain. Amsterdam: Elsevier.

Taaseh N, Yaron A, Nelken I (2011) Stimulus-specific adaptation and deviance detection in the rat auditory cortex. PLoS One 6:e23369.

Takahashi H, Nakao M, Kaga K (2004) Cortical mapping of auditory-evoked offset responses in rats. Neuroreport 15:1565-1569.

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Computation in Neural Systems 12:289-316.

Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. The Journal of Neuroscience 20:2315-2331.

Tishby N, Pereira FC, Bialek W (2000) The information bottleneck method. arXiv preprint physics/0004057.

Tramo MJ, Shah GD, Braida LD (2002) Functional role of auditory cortex in frequency processing and pitch perception. Journal of Neurophysiology 87:122-139.

Ulanovsky N, Las L, Farkas D, Nelken I (2004) Multiple time scales of adaptation in auditory cortex neurons. The Journal of Neuroscience 24:10440-10453.

Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical neurons. Nature neuroscience 6:391-398.

Vercoe B, Cumming D (1988) Connection machine tracking of polyphonic audio. Proceedings of International Computer Music Conference 211-218.

Walker KM, Bizley JK, King AJ, Schnupp JW (2011) Multiplexed and robust representations of sound features in auditory cortex. The Journal of Neuroscience 31:14565-14576.

Wang X, Lu T, Snider RK, Liang L (2005) Sustained firing in auditory cortex evoked by preferred stimuli. Nature 435:341-346.

Winkler I (2003) Change detection in complex auditory environment: beyond the oddball paradigm. In: Detection of Change, pp 61-81: Springer.

Winkler I, Czigler I (2012) Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. International Journal of Psychophysiology 83:132-143.

Winkler I, Denham SL, Nelken I (2009) Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends in cognitive sciences 13:532-540.

Winkler I, Karmos G, Näätänen R (1996) Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. Brain research 742:239-252.

Woolley SM, Gill PR, Theunissen FE (2006) Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. The Journal of Neuroscience 26:2499-2512.

Yaron A, Hershenhoren I, Nelken I (2012) Sensitivity to complex statistical regularities in rat auditory cortex. Neuron 76:603-615.

Zatorre RJ, Bouffard M, Ahad P, Belin P (2002) Where is' where'in the human auditory cortex? Nature neuroscience 5:905-909.

Zhang J, Nakamoto KT, Kitzes LM (2005) Modulation of level response areas and stimulus selectivity of neurons in cat primary auditory cortex. Journal of neurophysiology 94:2263-2274.